Hybrid Ensembles for Improved Force Matching

Lee-Ping Wang and Troy Van Voorhis
Department of Chemistry, Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139 USA

(Dated: 29 October 2010)

Force matching is a method for parameterizing empirical potentials in which the empirical parameters are fitted to a reference potential energy surface (PES). Typically, training data is sampled from a canonical ensemble generated with either the empirical potential or the reference PES. In this Communication, we show that sampling from either ensemble risks excluding critical regions of configuration space, leading to fitted potentials that deviate significantly from the reference PES. We present a hybrid ensemble which combines the Boltzmann probabilities of both potential surfaces into the fitting procedure, and we demonstrate that this technique improves the quality and stability of empirical potentials.

Molecular mechanics (MM) simulation is a powerful method for investigating the dynamical behavior of complex atomistic systems, but its utility depends critically on the accuracy of the empirical potential being used. A common approach is to fit the parameters in the empirical potential to a high-accuracy reference potential energy surface (PES), which can be obtained from quantum mechanical (QM) calculations. This approach, known as force matching[1–6] or potential fitting[7–10], aims to find the optimal parameters k that minimize an objective function χ^2 of the difference between a set of properties $\mathbf{Q}$ computed using the reference PES, and the analogous properties $\mathbf{M}$ computed using the empirical potential:

$$\chi^2 \equiv \int_{\mathbb{R}^{3N}} P(\mathbf{r})|\mathbf{X}(\mathbf{r},k)|^2\, d\mathbf{r}. \tag{1}$$

Here, the norm squared of the difference vector $\mathbf{X}(\mathbf{r},k) \equiv \mathbf{M}(\mathbf{r},k) - \mathbf{Q}(\mathbf{r},k)$ is integrated over the $3N$-dimensional configuration space. $\mathbf{X}$ may contain energies, atomistic forces, and/or other computable quantities that one wishes to match. The integral is typically performed by quadrature, or sampling techniques such as the Metropolis algorithm; in the latter case, $P(\mathbf{r})$ is a probability density corresponding to some ensemble.

The ensemble being sampled may be generated with the reference PES; we will call this the QM ensemble and denote the corresponding objective function using χ^2_{QM}. Accurate sampling of the QM ensemble is generally desirable but costly. Alternatively, training configurations may be rapidly sampled from the empirical (MM) ensemble; we will denote the corresponding objective function using χ^2_{MM}. Ischtwan and Collins,[11] and more recently Akin-Ojo and Wang[6] have proposed a scheme in which the MM ensemble is resampled after parameterization, with the process repeated in *generations* until self-consistency. Both types of force matching utilizing χ^2_{QM} and χ^2_{MM} have been applied widely to parameterize empirical potentials for gas-phase molecules[11–13], condensed phase systems[2,6,14], and solids.[1,3,15–17] The force matching method is not restricted to using MM for the empirical potential or QM as the reference PES; in multiscale coarse graining methods[18], the reference PES is an atomistic MM potential and is used to parameterize an empirical potential for coarse-grained particles.

In this Communication, we first present a simple heuristic example where force matching using either χ^2_{MM} or χ^2_{QM} does not optimally reproduce the QM PES. Instead, configurations from both QM and MM ensembles are important for force matching, and using either ensemble alone is insufficient. As a solution, we propose a hybrid-ensemble approach that combines the probability densities of both ensembles and generates a MM potential that optimally reproduces the QM PES in the heuristic example. Finally, we demonstrate the effects of hybrid-ensemble force matching when it is applied to parameterize MM potentials for a helium dimer and a water hexamer.

As an example, take both the QM PES and the MM potential to be hard repulsive walls as illustrated in Fig. 1. The MM potential has one adjustable parameter σ_{MM} which sets the location of the hard wall, and the optimal match is obtained when $\sigma_{\mathrm{MM}} = \sigma_{\mathrm{QM}}$. We make the simplifying assumption that the wall height is $\gg kT$, such that only the configurations where $r \geq \sigma$ are thermally accessible, and the repulsive region is excluded from the integral in Equation 1.

Consider first the case where the QM PES has a hard wall at $\sigma_{\mathrm{QM}} = 50$, and the MM potential underestimates this value at $\sigma_{\mathrm{MM}} = 30$; we will denote this potential using $\mathrm{MM}_<$ (Fig. 1a). The "error region", with finite $\mathbf{X}(r, \sigma_{\mathrm{MM}})$, is given by $r \in \{30, 50\}$; the QM-accessible region with finite $P_{\mathrm{QM}}(r)$ is given by $r > 50$, and the MM-accessible region with finite $P_{\mathrm{MM}}(r)$ is given by $r > 30$. The error region overlaps with the MM-accessible region and contributes to χ^2_{MM}; thus, χ^2_{MM} is minimized by increasing σ_{MM} until the optimal value, $\sigma_{\mathrm{MM}} = \sigma_{\mathrm{QM}}$, is obtained. However, the error region does *not* overlap with the QM-accessible region and does not contribute to χ^2_{QM}; in fact, any $\sigma_{\mathrm{MM}} < 50$ minimizes χ^2_{QM}! Thus, minimizing χ^2_{QM} risks severely underestimating σ_{MM}.

Now, consider the case where the initial guess is an overestimate ($\sigma_{\mathrm{MM}} = 70$), denoted by $\mathrm{MM}_>$ (Fig. 1b). The error region is now $r \in \{50, 70\}$; the QM-accessible region is the same ($r > 50$), and the MM-accessible region

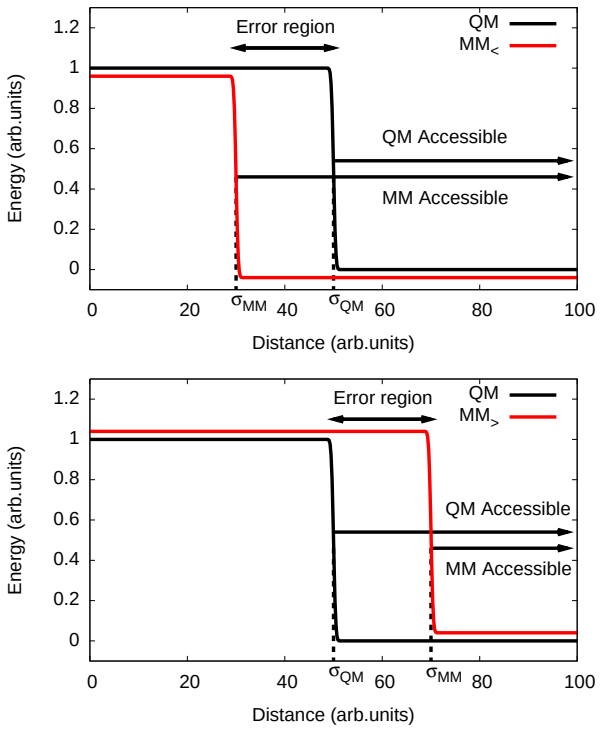

FIG. 1. (**a**) QM and $MM_<$ potential energy curves. (**b**) QM and $MM_>$ potential energy curves. The error region and thermally accessible regions are labeled using arrows.

is $r > 70$. The error region now overlaps with the QM-accessible region but not the MM-accessible region. The fitting situation is reversed; there is a contribution to $\chi^2_{\rm QM}$ but not to $\chi^2_{\rm MM}$. We now expect minimizing $\chi^2_{\rm QM}$ to effectively optimize $\sigma_{\rm MM}$, while any $\sigma_{\rm MM} > 50$ will minimize $\chi^2_{\rm MM}$. Thus, in a realistic situation where the relationship between the QM PES and the MM potential is unknown, we would expect force matching using $\chi^2_{\rm QM}$ and $\chi^2_{\rm MM}$ to respectively underestimate and overestimate σ; clearly, both choices are inadequate. The problem is further worsened when approximate functional forms are used for the MM potential.

The key observation is that the optimization fails when the chosen ensemble excludes configurations that are included in the other ensemble. Evidently, it is essential for configurations from both ensembles to contribute to χ^2 in order to produce an accurate MM potential. A simple treatment would be to construct χ^2 using a hybrid ensemble which combines the QM and MM probabilities:

$$\chi^2_{\rm H} \equiv \int_{\mathbb{R}^{3N}} (\omega P_{\rm QM}(\mathbf{r}) + (1-\omega) P_{\rm MM}(\mathbf{r}))|\mathbf{X}(\mathbf{r},k)|^2 \, d\mathbf{r}. \quad (2)$$

Here, ω is an adjustable mixing parameter which mixes the QM and MM probabilities. In the hard-wall example, the entire error region always has a finite probability density in $\chi^2_{\rm H}$, and minimizing $\chi^2_{\rm H}$ would effectively optimize $\sigma_{\rm MM}$ starting from any initial value. Note that after obtaining a perfect fit (i.e. the QM PES and MM potential are exactly the same), the hybrid ensemble reduces to the standard QM or MM ensemble.

Our main results in this Communication are based upon force-matching using $\chi^2_{\rm H}$. We now demonstrate the application of hybrid-ensemble force matching to two example systems; the helium dimer and water hexamer. The QM PES, in all cases, is computed at the Hartree-Fock (HF)/3-21G level of theory using Q-Chem.[19] MM simulations were performed using GROMACS.[20] Force matching was performed using the ForTune force matching program developed within our group, which minimizes χ^2 using a Newton-Raphson algorithm in similar fashion to previously published software packages.[21,22] We include both energy and force contributions in χ^2; each component is inverse variance-weighted, and the force components are attenuated by $\frac{1}{3N}$ such that energy and force differences contribute approximately equally. In all optimizations, a penalty proportional to $|\Delta k|^2$ is added to prevent large fluctuations in the parameters when the change in χ^2 is small.

The examples presented in this Communcation are designed to highlight the effects of using the hybrid ensemble for force-matching to the QM PES and not intended to reproduce experimental properties. Thus, we will discuss the quality of the parameterization by direct comparison of the QM PES and MM potentials in the former case, and by comparing radial distribution functions from QM and MM molecular dynamics in the latter.

Our first example, the parameterization of an MM potential for a helium dimer, is a direct computational realization of the hard-wall example above. In HF theory, the interaction between helium atoms is well characterized by a van der Waals (vdW) repulsive interaction. Our MM potential uses the Lennard-Jones (LJ) functional form:

$$E_{\rm LJ}(r_{ab}) = 4\epsilon \left(-\left(\frac{\sigma_{ab}}{r_{ab}}\right)^6 + \left(\frac{\sigma_{ab}}{r_{ab}}\right)^{12} \right) \quad (3)$$

In all cases, the LJ well depth ϵ is fixed at 0.01 kJ/mol, and only the X-intercept σ is optimized. Training configurations were obtained from a uniform grid of 800 He-He internuclear separations ranging from 1.5 Å to 4.0 Å. $\chi^2_{\rm QM}$ and $\chi^2_{\rm MM}$ were computed by performing a weighted sum over snapshots, with the appropriate Boltzmann probabilities: $P_{{\rm QM},s} = e^{-\beta E_{\rm QM}(\mathbf{r}_s)}$ and $P_{{\rm MM},s} = e^{-\beta E_{\rm MM}(\mathbf{r}_s)}$. The hybrid $\chi^2_{\rm H}$ uses an average probability with ω set to 0.5: $P_{{\rm H},s} = \frac{1}{2}(P_{{\rm QM},s} + P_{{\rm MM},s})$.

Two initial MM potentials, $MM_<$ and $MM_>$, were chosen such that they respectively underestimate and overestimate the position of the repulsive wall. For each initial potential, we obtained three optimized MM potentials using the three objective functions above; each optimized MM potential was obtained from three generations of the iterative force matching scheme described earlier. Fig. 2 shows the QM PES, initial MM potential, and the optimized MM potentials obtained using each of the three

objective functions. Here, $\mathrm{MM_{MM}}$ stands for "optimized MM potential using χ^2_{MM}". The values of σ and χ^2 for the three optimizations are given in Tables S1 and S2.[26]

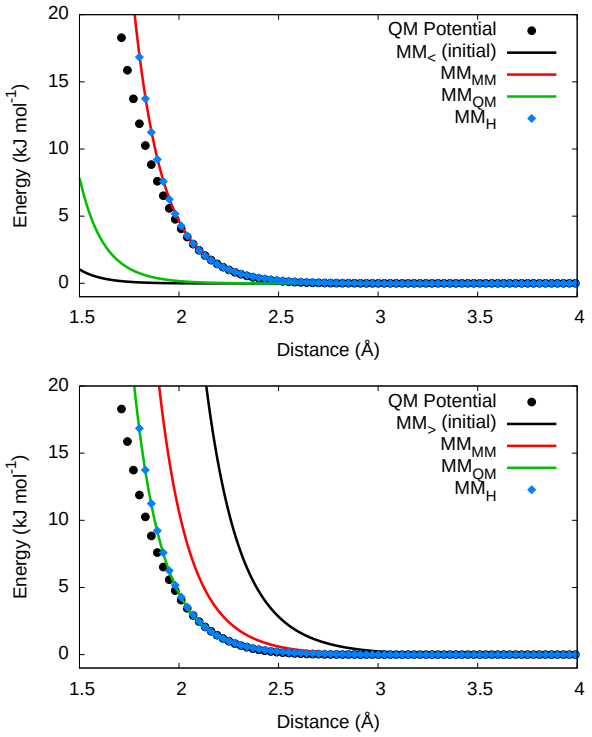

FIG. 2. Optimized MM potentials from force matching, starting from the initial guess (black line) and fitting to the quantum PES (black dots). Initial guesses are $\mathrm{MM_<}$ (**a**) and $\mathrm{MM_>}$ (**b**); optimized MM potentials using χ^2_{MM} (red), χ^2_{QM} (green), and χ^2_{H} (blue) are shown.

From Fig. 2, the role reversal of QM vs. MM weights can be clearly seen. Starting from $\mathrm{MM_<}$, using χ^2_{MM} reproduces the quantum result while using χ^2_{QM} underestimates the solution; the opposite effect is observed when we start from $\mathrm{MM_>}$. In both cases, using χ^2_{H} produces the correct solution, because the regions that contribute to either χ^2_{MM} and χ^2_{QM} are both counted.

It should be noted that in this example, all methods will eventually lead to the correct result; the convergence will just be very slow, taking many generations. This is because the repulsive interaction still contributes a very small amount to χ^2, and that all regions are perfectly sampled in this ideal case. When more parameters and degrees of freedom are involved, changing the sampling method can lead to different self-consistent MM potentials, as we describe in the following example.

In our second example, our system is a cluster of six water molecules; the intermolecular interactions in small water clusters is highly nontrivial[23] and MM potential development for water remains a highly active field. We chose the flexible SPC/E model[24] as our initial MM potential; the MM potential contains 7 empirical parameters. In each generation of force matching, the MM potential is used to generate 3.0 ns of dynamics in the canonical ensemble (T = 300K), from which 3,000 snapshots are sampled at 1ps time intervals and added to the training data. We utilize WHAM[25] to include training data from past generations. A shallow harmonic potential with force constant 0.01 kJ $\mathrm{mol^{-1}}$ $\mathrm{nm^{-2}}$ was applied to prevent divergent trajectories.

The MM dynamics directly samples the MM ensemble, so in computing χ^2_{MM} each snapshot has an equal probability. In computing χ^2_{H}, each snapshot's probability contains a constant from the MM ensemble, plus a non-Boltzmann factor corresponding to that snapshot's probability in the QM ensemble:

$$\chi^2_{\mathrm{H}} = \sum_s^S (\omega \frac{e^{-\beta(E_{\mathrm{QM}}(\mathbf{r}_s) - E_{\mathrm{MM}}(\mathbf{r}_s))}}{S_{\mathrm{QM}}} + (1-\omega)\frac{1}{S}) |\mathbf{X}(\mathbf{r}_s, k)|^2. \tag{4}$$

Here, S is the total number of snapshots, and S_{QM} is the sum of all non-Boltzmann factors.

Three objective functions were used with the following ensembles: χ^2_{80} ($\omega = 0.8$), χ^2_{50} ($\omega = 0.5$), and χ^2_0 ($\omega = 0.0$, pure MM). Due to large fluctuations in the non-Boltzmann factors, χ^2_{100} ($\omega = 1.0$, pure QM) was not usable in force matching and χ^2_{80} was used as a substitute. Force matching was performed for 19 generations, and the final MM potentials ($\mathrm{MM_{80}}$, $\mathrm{MM_{50}}$, $\mathrm{MM_0}$) were used to obtain radial distribution functions (RDFs) for comparison with an 1.8 ns AIMD reference trajectory, generated using the same temperature and harmonic potential. All three cases produced improved agreement with the AIMD RDFs compared to the initial SPC/E parameters; however, none of the MM potentials reproduced the exact shape of the AIMD RDFs, possibly due to the limitations of the functional form.

Fig. 3 shows the RDFs generated using the three optimized MM potentials compared to the AIMD RDF. $\mathrm{MM_{80}}$ underestimates the contact distances and peak positions for all three atom pairs; notably, this occurred starting from an overestimated initial guess. This may have been caused by fitting to components of χ^2 from other intermolecular degrees of freedom (i.e. torques) which couple to this one. $\mathrm{MM_0}$ overestimates the distances, while $\mathrm{MM_{50}}$ provides the best estimate with peak positions agreeing with the AIMD result to within 0.1 Å. Thus, the hybrid-ensemble force matching leads to the most accurate MM potential in this more complex case and corresponds well with the earlier examples.

We also performed force matching by sampling from the QM ensemble directly. 50,000 snapshots were sampled at 1ps intervals from the AIMD trajectories, and force matching was used to obtain the optimized potential $\mathrm{MM_{100}}$. We find that $\mathrm{MM_{100}}$ underestimates the approach distance, but not as severely as $\mathrm{MM_{80}}$; this result is surprising as we expect $\mathrm{MM_{80}}$ to interpolate between $\mathrm{MM_0}$ and $\mathrm{MM_{100}}$. This may be due to differences in the sampling trajectories; the training data for $\mathrm{MM_{100}}$ came from direct AIMD, while the data for $\mathrm{MM_{80}}$ came from an MM trajectory and required non-

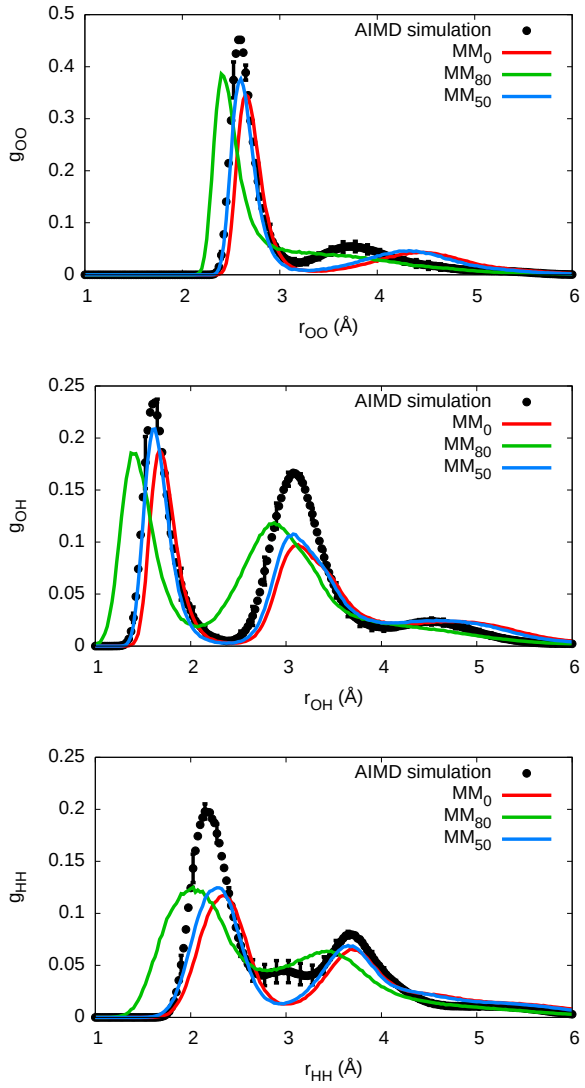

FIG. 3. Radial distribution functions of water for (**a**) O—O distances, (**b**) O—H distances, and (**c**) H—H distances. Thirty AIMD RDFs of length 60 ps were combined to produce the reference curve.

Boltzmann weights. This indicates that more complete sampling may be achieved by combining trajectories from both AIMD and MM dynamics.

In summary, using a single canonical ensemble to obtain training data for force matching can lead to substantial errors in the optimized MM potential. In particular, there is a risk of obtaining a poor fit if χ^2 neglects regions which make large contributions to $|\mathbf{X}(\mathbf{r},k)|^2$ and have finite probability in *either* the QM or MM ensemble. As a solution, we proposed constructing χ^2 from a hybrid ensemble which combines the probabilities from both ensembles. We demonstrated the advantages of the method by applying force matching to parameterize MM potentials for a helium dimer and a water cluster. In both cases the anomalies associated with single-ensemble objective functions χ^2_{QM} and χ^2_{MM} were shown, and the hybrid χ^2_{H} provided the correct behavior. Our finding makes intuitive sense: if training data is sampled from the MM ensemble, the MM-accessible configurations will naturally be included in the fit. Meanwhile, QM-accessible configurations that are MM-inaccessible should still be penalized for their absence.

The hybrid ensemble may be applied to improve the parameterization of atomic charges and vdW-type interactions (e.g. LJ, n-6 or Buckingham), which traditionally are the most difficult to parameterize yet are essential for determining the dynamical and bulk properties of any multimolecular system.

This work was funded by ENI S.p.A. as part of the Solar Frontiers Research Program.

[1] F. Ercolessi and J. B. Adams, *Europhys. Lett.* **26**, 583 (1994).
[2] T. G. A. Youngs, M. G. D. Popolo, and J. Kohanoff, *J. Phys. Chem. B* **110**, 5697 (2006).
[3] X. Y. Liu, J. B. Adams, F. Ercolessi, and J. A. Moriarty, *Model. Simul. Mater. Sc.* **4**, 293 (1996).
[4] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, *J. Chem. Phys.* **120**, 10896 (2004).
[5] P. Maurer, A. Laio, H. W. Hugosson, M. C. Colombo, and U. Rothlisberger, *J. Chem. Theory Comput.* **3**, 628 (2007).
[6] O. Akin-Ojo, Y. Song, and F. Wang, *J. Chem. Phys.* **129**, 064108 (2008).
[7] A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, R. Narulkar, S. Bukkapatnum, and R. Komanduri, *J. Chem. Phys.* **130**, 134101 (2009).
[8] H. Sun, *J. Phys. Chem. B* **102**, 7338 (1998).
[9] P. Brommer and F. Gahler, *Philos. Mag.* **86**, 753 (2006).
[10] G. Toth, *J. Phys.-Condens. Mat.* **19**, 335222 (2007).
[11] J. Ischtwan and M. A. Collins, *J. Chem. Phys.* **100**, 8080 (1994).
[12] L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, and R. Komanduri, *J. Chem. Phys.* **122**, 084104 (2005).
[13] M. Malshe, R. Narulkar, L. M. Raff, M. Hagan, S. Bukkapatnam, and R. Komanduri, *J. Chem. Phys.* **129**, 044111 (2008).
[14] S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 6573 (2005).
[15] P. Tangney and S. Scandolo, *J. Chem. Phys.* **117**, 8898 (2002).
[16] G. Grochola, S. P. Russo, and I. K. Snook, *J. Chem. Phys.* **123**, 204719 (2005).
[17] S. Paramore, L. W. Cheng, and B. J. Berne, *J. Chem. Theory Comput.* **4**, 1698 (2008).
[18] W. G. Noid, J. W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, *J. Chem. Phys.* **128**, 244114 (2008).
[19] Y. S. *et al*, *Phys. Chem. Chem. Phys.* **8**, 3172 (2006).
[20] D. V. der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *J. Comp. Chem.* **26**, 1701 (2005).
[21] P. Brommer and F. Gahler, *Model. Simul. Mater. Sc.* **15**, 295 (2007).
[22] B. Waldher, J. Kuta, S. Chen, N. Henson, and A. E. Clark, *J. Comput. Chem.* **31**, 2307 (2010).
[23] C. J. Tsai and K. D. Jordan, *Chem. Phys. Lett.* **213**, 181 (1993).
[24] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *J. Phys. Chem.* **91**, 6269 (1987).
[25] A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988).
[26] See Supplementary Material Document No. _____ for force field parameters and additional RDF data. For information on Supplementary Material, see http://www.aip.org/pubservs/epaps.html. (Supporting Information)