

MIT Open Access Articles

Hybrid Ensembles for Improved Force Matching

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Wang, Lee-Ping, and Troy Van Voorhis. "Communication: Hybrid Ensembles for Improved Force Matching." *The Journal of Chemical Physics* 133.23 (2010): 231101.

As Published: <http://dx.doi.org/10.1063/1.3519043>

Publisher: American Institute of Physics

Persistent URL: <http://hdl.handle.net/1721.1/69604>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike 3.0



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: 24 April 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,¹ 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 chemistry (QM) calculations. This approach, known as force matching²⁻⁹ or potential fitting¹⁰⁻¹³, 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:

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

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 energy and 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 canonical ensemble.

The canonical 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_{QM}^2 . Accurate sampling of the QM ensemble is generally desirable, but the algorithms involved generally have a high computational cost. Alternatively, training configurations may be rapidly sampled from the empirical (MM) ensemble; we will denote the corresponding objective function using χ_{MM}^2 . Ischtwan and Collins,¹⁴ and more recently Akin-Ojo and Wang^{9,15} 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_{QM}^2 and χ_{MM}^2 have been applied widely to parameterize empirical potentials for gas-phase molecules^{14,16,17}, condensed phase systems^{3,7,9,15,18}, and solids.^{2,4,5,19-23} The force matching method is not restricted to using MM for the empirical potential or QM as the reference PES; in

multiscale coarse graining methods²⁴, 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_{MM}^2 or χ_{QM}^2 does not optimally reproduce the QM PES. We claim that 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, consider where both the QM PES and the MM potential are hard repulsive walls as illustrated in Fig. 1. The MM potential has one adjustable parameter σ_{MM} which describes the location of the hard wall, and the optimal match is obtained when $\sigma_{\text{MM}} = \sigma_{\text{QM}}$. We make the simplifying assumption that the wall height is $\gg kT$, such that only the configurations where $r \geq \sigma$ are thermally accessible. In a given ensemble, the repulsive region does not contribute to that particular χ^2 .

Consider first the case where the QM PES has a hard wall at $\sigma_{\text{QM}} = 50$, and the MM potential underestimates this value at $\sigma_{\text{MM}} = 30$; we will denote this potential using $\text{MM}_<$ (Fig. 1, top pane). The “error region”, with finite $\mathbf{X}(r, \sigma_{\text{MM}})$, is given by $r \in \{30, 50\}$; the QM-accessible region with finite $P_{\text{QM}}(r)$ is given by $r > 50$, and the MM-accessible region with finite $P_{\text{MM}}(r)$ is given by $r > 30$. Clearly, the error region overlaps with the MM-accessible region, so there is a finite contribution to χ_{MM}^2 . It follows that minimization of χ_{MM}^2 would result in increasing σ_{MM} until the optimal value, $\sigma_{\text{MM}} = \sigma_{\text{QM}}$, is obtained. However, the error region does not overlap with the QM-accessible region, and there is no contribution to χ_{QM}^2 ; consequently, minimization of χ_{QM}^2 would still underestimate σ_{MM} .

Now, consider the case where the initial guess is an overestimate ($\sigma_{\text{MM}} = 70$), denoted by $\text{MM}_>$ (Fig. 1, bottom pane). The error region is now $r \in \{50, 70\}$; the

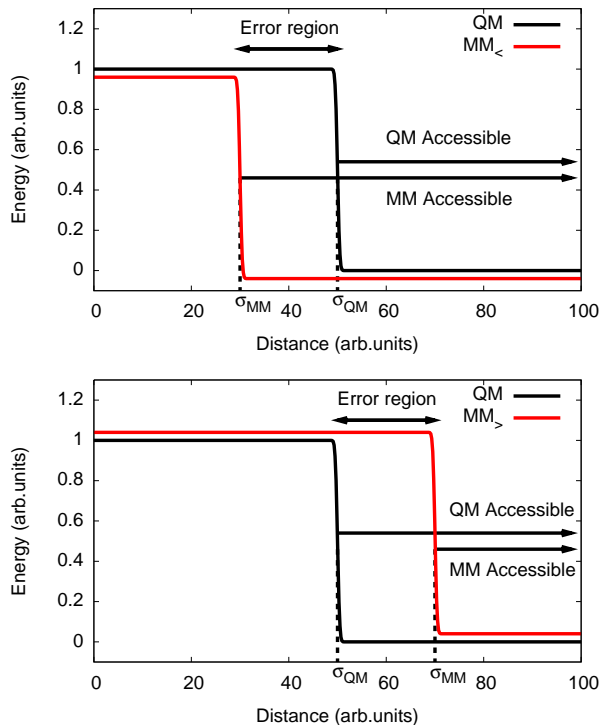


FIG. 1. Top: QM and $MM_{<}$ potential energy curves. Bottom: QM and $MM_{>}$ potential energy curves. The error region and thermally accessible regions are labeled using arrows.

QM-accessible region is the same ($r > 50$), and the MM-accessible region 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 finite contribution to χ_{QM}^2 , but no contribution to χ_{MM}^2 . We now expect minimizing χ_{QM}^2 to effectively optimize σ_{MM} , while minimizing χ_{MM}^2 will result in an overestimate. In a realistic situation where the relationship between the QM PES and the MM potential is unknown, we would thus expect force matching using χ_{QM}^2 and χ_{MM}^2 to respectively underestimate and overestimate σ ; clearly, both choices are inadequate.

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:

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

Here, ω is a mixing parameter which mixes the QM and MM probabilities; its value is to be determined. In the hard-wall example, the entire error region always has a finite probability density in χ_H^2 , and minimizing χ_H^2 would effectively optimize σ_{MM} starting from any initial

value.

Our main results in this Communication are based upon force-matching using χ_H^2 . We now demonstrate the benefits of hybrid-ensemble force matching by applying it 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.²⁵ MM simulations were performed using GROMACS.²⁶ Force matching was performed using the ForTune force matching program developed within our group, which optimizes MM potential parameters using a Newton-Raphson minimization of χ^2 . 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 MM potentials presented in this Communication are solely designed to fit the chosen QM PES, and not intended to describe 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:

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

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. χ_{QM}^2 and χ_{MM}^2 were computed by performing a weighted sum over snapshots, with the appropriate Boltzmann probabilities: $P_{QM,s} = e^{-\beta E_{QM}(\mathbf{r}_s)}$ and $P_{MM,s} = e^{-\beta E_{MM}(\mathbf{r}_s)}$. The hybrid χ_H^2 uses an average probability with ω set to 0.5: $P_{H,s} = \frac{1}{2}(P_{QM,s} + P_{MM,s})$.

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 quantum potential surface, initial MM potential, and the optimized MM potentials obtained using each of the three objective functions. Here, MM_{MM} stands for “optimized MM potential using χ_{MM}^2 ”. The values of σ and χ^2 for the three optimizations are given

in Tables S1 and S2 (Supporting Information).

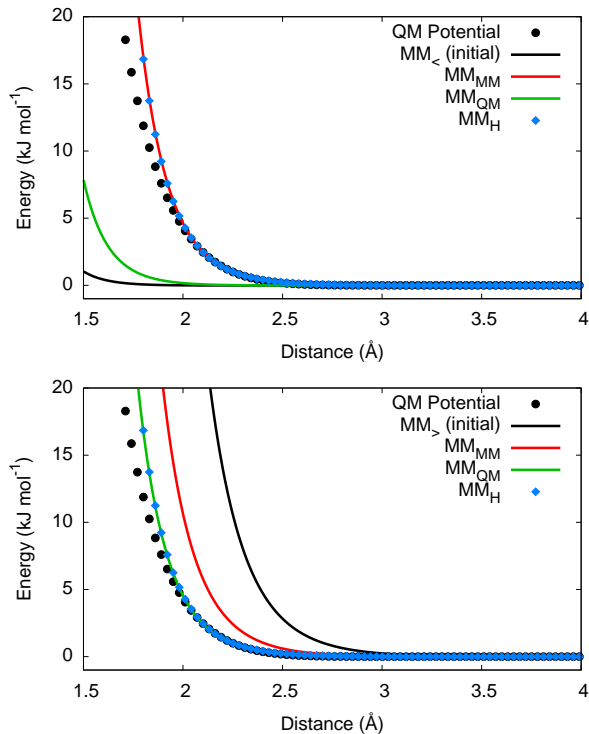


FIG. 2. Optimized MM potentials from force matching, starting from the initial guess (black line) and fitting to the quantum PES (black dots). Initial guesses are $MM_{<}$ (top) and $MM_{>}$ (bottom); optimized MM potentials using χ_{MM}^2 (red), χ_{QM}^2 (green), and χ_H^2 (blue) are shown.

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

It should be noted that in this simple example with a single parameter, all methods will eventually converge to the correct result given a large enough number of generations; the convergence will just be very slow. This is because the repulsive interaction is not an ideal hard wall, and the repulsive region still contributes a small amount to χ^2 . 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 initial MM potential is the flexible SPC/E model²⁷ consisting of intermolecular electrostatic interactions, a Lennard-Jones interaction between oxygen atoms, and flexible bond and angle interactions; there are 7 parameters in total. In each generation of force matching, the MM potential is used to generate 3.0 ns of dynamics in the MM canonical ensemble (T

= 300K), from which 3,000 snapshots are sampled at 1ps time intervals and added to the training data. We utilize WHAM^{28–30} to include training data from past generations. A shallow harmonic potential with force constant 0.01 kJnm^{-2} was applied to prevent trajectory divergence without significantly affecting the dynamics.

The MM dynamics directly samples the MM ensemble, so in computing χ_{MM}^2 each snapshot has an equal probability. In computing χ_H^2 , 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:

$$\chi_H^2 = \sum_s \left(\omega \frac{e^{-\beta(E_{QM}(\mathbf{r}_s) - E_{MM}(\mathbf{r}_s))}}{S_{QM}} + (1 - \omega) \frac{1}{S} \right) |\mathbf{X}(\mathbf{r}_s, k)|^2. \quad (4)$$

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_{80}^2 ($\omega = 0.8$), χ_{50}^2 ($\omega = 0.5$), and χ_0^2 ($\omega = 0.0$, pure MM). Due to large fluctuations in the non-Boltzmann factors, χ_{100}^2 ($\omega = 1.0$, pure QM) could not be constructed with this sampling technique and χ_{80}^2 was used as a substitute. Force matching was performed for 19 generations, and the final MM potentials (MM_{80} , MM_{50} , MM_0) were used to obtain radial distribution functions (RDFs) for comparison with an AIMD reference.

All three cases produced improved agreement with the AIMD RDF compared to the initial SPC/E parameters (Fig. S2); 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 an AIMD reference. MM_{80} underestimates the distance of nearest approach and peak positions for all three atom pairs; notably, this occurred starting from an overestimated initial guess (see Fig. S2). 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. MM_0 overestimates the distances, while MM_{50} provides the best estimate; the MM_{50} peak positions agree with the AIMD result to within 0.1 \AA .

We also performed force matching by sampling from the QM ensemble directly. 50,000 snapshots were sampled at 1ps time intervals from the AIMD trajectories, and force matching was performed until convergence to obtain the optimized potential MM_{100} . We find that MM_{100} underestimates the approach distance (Fig. S3), but not as severely as MM_{80} ; this result is counterintuitive in that we expect MM_{80} to interpolate between MM_0 and MM_{100} . We conjecture that this may be due to differences in the sampling technique; the training data for MM_{100} came from an AIMD trajectory which directly sampled the QM ensemble, while the training data for MM_{80} came from an MM trajectory. This indicates that the hybrid-ensemble force matching may benefit from im-

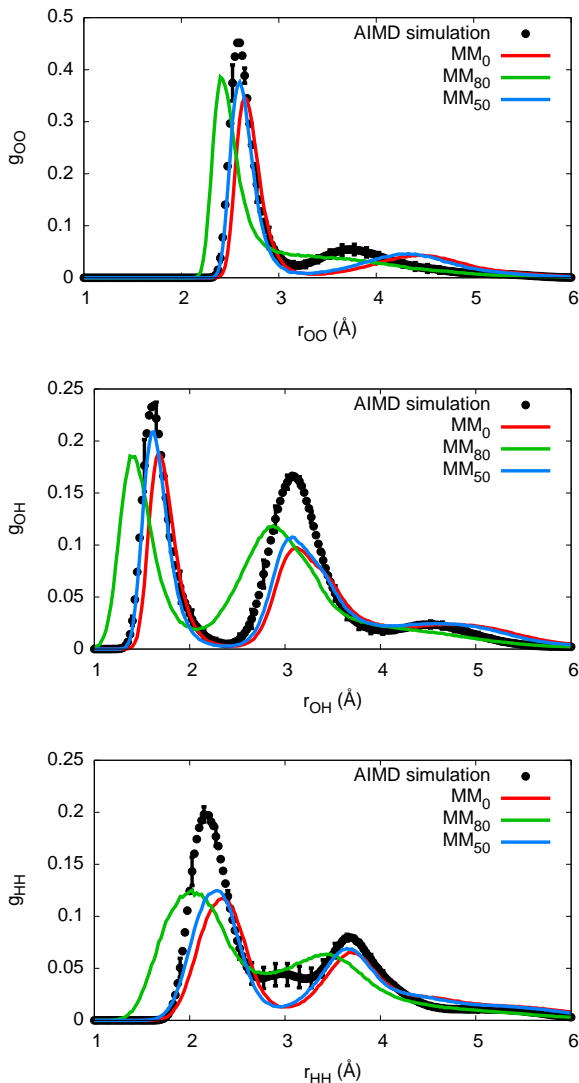


FIG. 3. Radial distribution functions of water. Thirty AIMD RDFs of length 60 ps were combined to produce the reference curve.

proved sampling techniques.

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_{QM}^2 and χ_{MM}^2 were shown, and the hybrid χ_{H}^2 provided the correct behavior. Our finding makes intu-

itive 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.

We plan to refine this method by investigating the optimality of the 50% choice in χ_{H}^2 , and perhaps to improve our sampling technique by combining data from MM molecular dynamics and AIMD simulations. In terms of applications, the hybrid ensemble may be used 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.

- ¹K. Machida, *Principles of Molecular Mechanics* (John Wiley and Sons, Inc., 1999).
- ²F. Ercolessi and J. B. Adams, *Europhys. Lett.* **26**, 583 (1994).
- ³T. G. A. Youngs, M. G. D. Popolo, and J. Kohanoff, *J. Phys. Chem. B* **110**, 5697 (2006).
- ⁴X. Y. Liu, J. B. Adams, F. Ercolessi, and J. A. Moriarty, *Model. Simul. Mater. Sc.* **4**, 293 (1996).
- ⁵A. Landa, P. Wynblatt, D. J. Siegel, J. B. Adams, O. N. Mryasov, and X. Y. Liu, *Acta Mater.* **48**, 1753 (2000).
- ⁶J. W. Chu and G. A. Voth, *Biophys. J.* **90**, 1572 (2006).
- ⁷S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, *J. Chem. Phys.* **120**, 10896 (2004).
- ⁸P. Maurer, A. Laio, H. W. Hugosson, M. C. Colombo, and U. Rothlisberger, *J. Chem. Theory Comput.* **3**, 628 (2007).
- ⁹O. Akin-Ojo and F. Wang, *J. Phys. Chem. B* **113**, 1237 (2009).
- ¹⁰A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, R. Narulkar, S. Bukkapatnum, and R. Komanduri, *J. Chem. Phys.* **130**, 134101 (2009).
- ¹¹H. Sun, *J. Phys. Chem. B* **102**, 7338 (1998).
- ¹²P. Brommer and F. Gahler, *Philos. Mag.* **86**, 753 (2006).
- ¹³G. Toth, *J. Phys.-Condens. Mat.* **19**, 335222 (2007).
- ¹⁴J. Ischtwan and M. A. Collins, *J. Chem. Phys.* **100**, 8080 (1994).
- ¹⁵O. Akin-Ojo, Y. Song, and F. Wang, *J. Chem. Phys.* **129**, 064108 (2008).
- ¹⁶L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, and R. Komanduri, *J. Chem. Phys.* **122**, 084104 (2005).
- ¹⁷M. Malshe, R. Narulkar, L. M. Raff, M. Hagan, S. Bukkapatnum, and R. Komanduri, *J. Chem. Phys.* **129**, 044111 (2008).
- ¹⁸S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 6573 (2005).
- ¹⁹X. Y. Liu, P. P. Ohotnicky, J. B. Adams, C. L. Rohrer, and R. W. Hyland, *Surf. Science* **373**, 357 (1997).
- ²⁰X. Y. Liu and J. B. Adams, *Acta Mater.* **46**, 3467 (1998).
- ²¹P. Tangney and S. Scandolo, *J. Chem. Phys.* **117**, 8898 (2002).
- ²²G. Grochola, S. P. Russo, and I. K. Snook, *J. Chem. Phys.* **123**, 204719 (2005).
- ²³S. Paramore, L. W. Cheng, and B. J. Berne, *J. Chem. Theory Comput.* **4**, 1698 (2008).
- ²⁴S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).
- ²⁵Y. S. *et al.*, *Phys. Chem. Chem. Phys.* **8**, 3172 (2006).
- ²⁶D. V. der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *J. Comp. Chem.* **26**, 1701 (2005).
- ²⁷H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, *J. Phys. Chem.* **91**, 6269 (1987).
- ²⁸A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988).
- ²⁹S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comp. Chem.* **13**, 1011 (1992).
- ³⁰S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *J. Comp. Chem.* **16**, 1339 (1995).

I. SUPPORTING INFORMATION

The supporting information contains parameters for the helium dimer and water cluster examples, as well as RDFs for the water cluster obtained using the SPC/E and AIMD-fitted MM potentials.

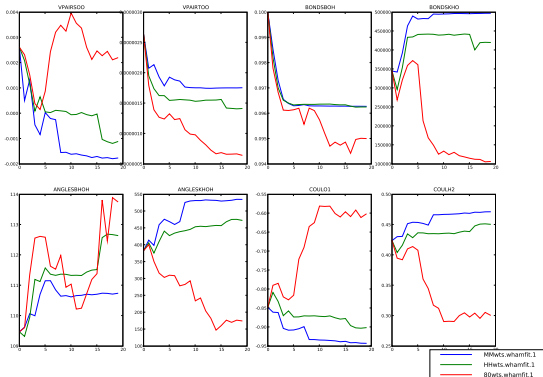


FIG. S1. Evolution of parameters during force matching for MM_0 (blue), MM_{50} (green), and MM_{80} (red).

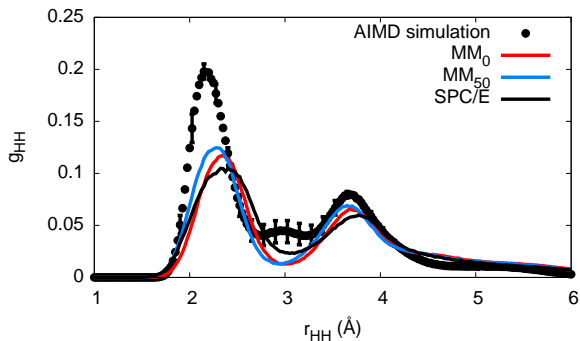


FIG. S2. Comparison of g_{HH} using MM_0 , MM_{50} , and the SPC/E initial guess.

TABLE S1. Force matching for the helium dimer, starting from $MM_{<}$.

Gen.	σ			χ^2		
	MM	QM	50	MM	QM	50
0	0.200	0.200	0.200	297.440	50.448	237.750
1	0.283	0.206	0.272	17.977	50.298	30.823
2	0.298	0.213	0.296	0.613	49.975	1.267
3	0.299	0.222	0.299	0.346	49.230	0.385

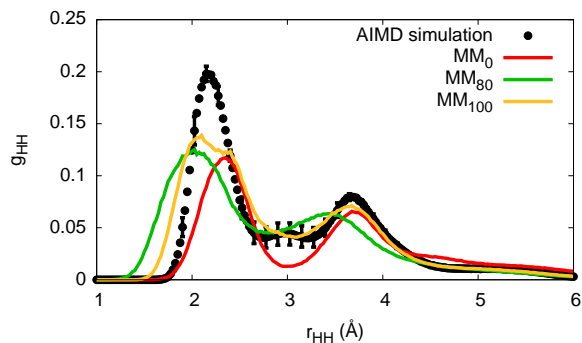


FIG. S3. Comparison of g_{HH} using MM_0 , MM_{80} , and MM_{100} ; the latter is fitted to an AIMD trajectory.

TABLE S2. Force matching for the helium dimer, starting from $MM_{>}$.

Gen.	σ			χ^2		
	MM	QM	50	MM	QM	50
0	0.360	0.360	0.360	15.476	4106.800	3795.100
1	0.344	0.304	0.304	3.815	2.841	2.873
2	0.334	0.299	0.299	1.498	0.410	0.386
3	0.326	0.299	0.299	0.699	0.409	0.385