Machine Learning Force Field Parameters from Ab Initio Data

Journal:	*Journal of Chemical Theory and Computation*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Li, Ying; Argonne National Laboratory, Li, Hui; University of Chicago, Department of Biochemistry and Molecular Biology Pickard, Franck; NHLBI, NIH, Laboratory Of Computational Biology Narayanan, Badri; Argonne National Laboratories, Center for Nanoscale Materials Sen, Fatih G.; Argonne National Laboratory - Center fro Nanoscale Materials, Chan, Maria; Argonne National Laboratory, Center for Nanoscale Materials Sankaranarayanan , Subramanian ; Argonne National Laboratory Brooks, Bernard; NHLBI, NIH, Laboratory Of Computational Biology Roux, Benoît; University of Chicago, Department of Biochemistry and Molecular Biology

Machine Learning Force Field Parameters from *Ab Initio* Data

Ying Li,[a] Hui Li,[b] Frank C Pickard IV,[c] Badri Narayanan,[d] Fatih Sen,[d] Maria K. Y. Chan,[d,e] Subramanian Sankaranarayanan,[d,e] Bernard R. Brooks,[c] and Benoît Roux[b,e]

[a] Leadership Computing Facility, Argonne National Laboratory, IL 60439, USA
[b] Department of Biochemistry and Molecular Biophysics, University of Chicago, IL 60637, USA
[c] Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[d] Center for Nanoscale Materials, Argonne National Laboratory, IL 60439, USA
[e] Computational Institute, University of Chicago, IL 60637, USA

ABSTRACT

Machine learning (ML) techniques with the genetic algorithm (GA) have been applied to explore a polarizable force field parameters using only *ab initio* data from quantum mechanics (QM) calculations of molecular clusters at the MP2/6-31G(d,p), DFMP2(fc)/jul-cc-pVDZ, and DFMP2(fc)/jul-cc-pVTZ levels to predict experimental condensed phase properties (*i.e.*, density and heat of vaporization). The performance of this ML/GA approach is demonstrated on 4,943 dimers electrostatic potentials and 1,250 clusters interaction energies for methanol. Excellent agreement between the training dataset from QM calculations and the optimized force field model can be achieved. Better results are achieved by introducing an offset factor during the machine learning process to compensate for the discrepancy of the QM calculated energy and the energy reproduced by optimized force field, where the offset factor maintain the local "shape" of the QM energy surface. Throughout the machine learning process, experimental observables were not involved in the objective function, but were only used for model validation. The best model, optimized from the QM data at the DFMP2(fc)/jul-cc-pVTZ level, appears to perform even better than the original AMOEBA force field (amoeba09.prm), which was optimized empirically to match liquid properties. The present effort shows the possibility of using machine learning techniques to develop descriptive polarizable force field using only QM data. The ML/GA strategy to optimize force field parameters described here could easily be extended to other molecular systems.

SECTION: Molecular Mechanics, Quantum Chemistry, Machine Learning and Force Field

INTRODUCTION

Classical simulations based on detailed atomic models and molecular mechanical (MM) potential functions are increasingly playing an important role in physics, chemistry, biochemistry, and materials science. To obtain meaningful results, the accuracy of the potential function, or force field, underlying such molecular dynamics (MD) simulations is of critical importance.[1-8] The simple force fields that are widely used for biomolecular simulations, in

particular, are empirically optimized to reproduce a number of calculated quantum mechanics (QM) and experimental properties.[4] While such an empirical approach can provide atomic models of useful accuracy, they can do so only if the simulated system of interest is built from the set of previously parameterized chemical functionalities. This has a critical impact on the predictive power of MD simulations in general. For instance, in the absence of any prior knowledge from experiment, it is unclear if the properties of a neat liquid of a compound comprising previously untreated atom types could be reliably predicted. A force field that requires an empirical calibration against experimental data are more acutely felt, where a departure from a pre-defined set of chemical functionalities is often required.[8]

The primary objective of the present study is to test the ability of machine learning (ML) techniques with the genetic algorithm (GA) to parameterize a polarizable force field by relying exclusively on *ab initio* QM data. A central question that we are trying to answer with the present effort is whether the resulting model can accurately predict neat liquid properties of a compound over a range of temperatures from only QM data, without any *a priori* information from experiment. The general idea that one could develop molecular potential functions from first principles QM data rests on ideas from early efforts by Clementi and co-workers in the 1970's.[9] They constructed an additive force field model by fitting the potential energies for the water dimer in various geometrical configurations calculated using the QM configuration-interaction method. However, the resulting MCY model did not accurately simulate liquid water. In retrospect, a serious limitation was the total neglect of many-body (MB) effects. To test our ML/GA strategy, it is important to choose a functional form that incorporates a correct physical representation of many-body effects and that provides an accurate description of physicochemical properties of the condensed phase. Force fields that explicitly account for induced electronic polarization offer a step in the right direction.[10-16] While non-polarizable additive force fields used in biomolecular simulations, which represent polarization in an average manner with effective fixed partial charges, can achieve a reasonably accurate representation of the condensed phase, they do so as the result of numerous empirical compromises.[4, 17] Consequently, it is possible that the functional form of non-polarizable additive force fields might be too limited to achieve the desired physical accuracy over a range of thermodynamic states.

Different functional forms of polarizable force fields have been proposed for simulating chemical and biomolecular systems. Those polarizable force fields include models based on classical Drude oscillators,[16, 18-19] the CHARMM charge equilibration (CHEQ) model based on fluctuating partial charges,[15] the AMBER ff02 based on inducible dipoles,[20] and the Atomic Multipole Optimized Energetics for Biomolecular Applications (AMOEBA) force field.[21-22] While most models remain at the level of atomic point charges, the AMOEBA force field further incorporates static field contributions from atomic dipoles and quadrupoles. More recently, there has been considerable progress in the design of potential energy functions to simulate water by building upon rigorous representations of the many-body expansion of the interaction energies.[23-25] Schmidt and collaborators have advanced a strategy on the basis of the symmetry-adapted perturbation theory (SAPT) methodology,[26-27] yielding models with relatively good properties for dense liquid systems. Similarly, the latest SAPT models may need to incorporate Axilrod−Teller three-body dispersion and exchange to yield accurate models of dense systems.[28]

Similarly, the MB-pol model developed by Paesani and co-workers has demonstrated that it is possible to achieve high accuracy across all phases of water through a quantitative and physically correct representation of both short- and long-range many-body contributions.[25] Of particular importance, the MB-pol model incorporates many-body van der Waals interactions. However, it is not yet clear whether such a rigorous treatment of many-body effects could be routinely implemented with more complicated molecules. The present effort, we have chosen to work with the AMOEBA functional form because of its relative mathematical simplicity and flexibility.

We have chosen the methanol molecule to explore the feasibility of our ML/GA parameterization strategy. This choice is motivated by the observation that methanol represents a molecule of sufficient complexity comprising both a hydroxyl polar group, able to form hydrogen bonds, as well as a bulky non-polar alkyl group.[29] The competing nature of these interactions makes methanol an excellent test case to explore the feasibility of our ML/GA strategy to generate force field parameters that relies exclusively on QM data.

METHODS

The complete functional form of the AMOEBA force field has been described in detail elsewhere.[21, 30] The present focus is mainly on the non-bonded electrostatic and van der Waals (vdW) interaction parameters. Because the internal covalent interaction (bond, angle, dihedral and torsional parameters) are not expected to greatly affect the properties of the condensed liquid phase, they remained unchanged from the original AMOEBA force field (amoeba09.prm).[22] To compute the electrostatic energy (E_{elec}) between methanol molecules, the AMOEBA force field requires 44 independent parameters. Those parameters are monopole (q), dipole (μ_x, μ_y, μ_z), quadrupole expressed as a traceless symmetric matrix with five independent elements (Q_{xx}, Q_{yx}, Q_{yy}, Q_{zx}, Q_{zy}), atomic polarizability (α) for the four methanol atom types (O, C, H_C, and H_O). A unique Thole damping factor (a) was used for all atoms, following the AMOEBA convention. The AMOEBA force field represents the vdW interactions as a buffered 14-7 potential.[31] To compute the vdW energy (E_{vdW}) between methanol molecules, the functional form requires 10 parameters. Eight of those parameters are the R_{min}, and ε_{min} for the four atoms types of methanol. Finally, two additional parameters (the so-called reduction factor λ) are required to scale the position of the two hydrogen atoms interaction site along their corresponding covalent bond.[14]

Force field parameterization, *i.e.* the determination of the optimal parameters associated with a complex functional form, is a challenging optimization problem in a space of high dimensionality. Local optimization approaches generally carry out this task by attempting to minimize an objective function using a simple gradient-based algorithm (*e.g.*, steepest descent, conjugate gradient method),[32-33] and are only effective if the starting point is sufficiently close to a satisfactory solution. However, even with pre-existing knowledge about the objective function and the parameter space, a number of problems can arise. For example, the convergence to optimal parameters is not guaranteed when the objective function is rugged, non-differentiable, or when the initial value deviates significantly from the global minimum. Ultimately, the range of parameter values that must be explored grows dramatically as the complexity of the force field functional form increases. Here, to overcome the challenge encountered with force field parameterization using ML techniques, we have used the genetic algorithm (GA),

which is an evolutionary algorithm that mimics the process of natural selection.[34] Previous successful applications of the genetic algorithm in similar contexts include the determination of the parameters for the ReaxFF reactive force field,[2-3, 5] and various force fields including Morse+QEq charge transfer ionic potential (CTIP),[6] and a new hybrid bond-order potential (HyBOP) for materials system.[2-3, 5-6, 35-36]

In the present work, electrostatic parameters were determined in a first stage, and vdW parameters were optimized in a second stage (*i.e.*, the electrostatic parameters were kept unchanged while the vdW parameters are optimized on the second stage). The optimization of the electrostatic parameters of methanol was carried out in two steps. In a first step, the value of the atomic multipoles was obtained from the QM electrostatic potential on the Connolly surface of a single isolated methanol molecule. The QM electrostatic potential of a methanol monomer, ϕ^{QM}, was calculated at the MP2 level of theory with various basis sets including Pople-style[37] and correlation consistent[38] basis sets. The results are given in Table 1. A Distributed Multipole Analysis (DMA)[39-41] was carried out using the Gaussian Distributed Multipole Analysis (GDMA 2.2) program[39, 42-43] for the electronic density results from different MP2 basis sets. The multipole parameters of each atomic site were identified using the Tinker[21, 44-48] package (the Poledit and Potential program). In a second step, all 44 electrostatic parameters were refined from the QM electrostatic potential calculated from an extensive training dataset of 4,943 methanol dimers using the genetic algorithm. This second step differs from the standard protocol used to determine the electrostatic parameters for the AMOEBA polarizable force field from QM calculations,[22, 30] which relies on the unperturbed monomer. By considering dimers in this situation, the mutual induced polarization of the two molecules is explicitly taken into account, allowing a non-perturbative determination of the atomic polarizabilities. The methanol dimer configurations were sampled through placing two methanol molecules, where one methanol was sampled over shell radius (1- 4 Å) on another methanol's Connolly surface. In total, 4,943 methanol dimer configurations were sampled. These 4,943 configurations of methanol dimers were relaxed through MP2 geometry optimization via Gaussian09 program,[49] while fixing the position of the carbon atoms. The optimization of electrostatic parameters seeks to minimize the objective function Δ as the averaged root mean square deviation (ARMSD) between the QM result and the parameterized force field calculation, shown in equation (1):

$$\begin{aligned}\Delta &= \frac{1}{N_{\mathrm{cluster}}}\sum_{j=1}^{N_{\mathrm{cluster}}}\sqrt{\chi^2(p_1,p_2,\ldots p_m)} \\ &= \frac{1}{N_{\mathrm{cluster}}}\sum_{j=1}^{N_{\mathrm{cluster}}}\left(\sqrt{\frac{1}{n^{\mathrm{grid}}}\sum_{k=1}^{n^{grid}}\left(\phi_k^{\mathrm{QM}}-\phi_k^{\mathrm{FF}\,(p_1,p_2,\ldots p_m)}\right)^2}\right)_j\end{aligned} \qquad (1)$$

where $(p_1, p_2, \ldots p_m)$ represent the electrostatic parameters of the force field and $N_{cluster}$ = 4,943. It is noteworthy that the static monopole, dipoles and quadrpoles remained essentially unchanged during this second step, which primarily affected the optimization of the atomic polarizabilies.

In a second stage, the parameters for the vdW interactions were optimized to best-match the interaction between a central methanol molecule and its nearest neighbors via supermolecular QM computations for a large number of clusters extracted from classical MD simulations of liquid methanol using the original amoeba09.prm.[22]

The electrostatic parameters optimized in the previous step are kept unchanged at this stage. Second-order Møller–Plesset perturbation theory[50] (MP2) was employed as the QM method to calculate the interaction energy, with basis sets superposition error (BSSE) counterpoise correction.[51] This level of theory and basis sets has been widely verified by previous studies to predict energetics and structural properties.[52] One set of QM calculations was carried out at the MP2/6-31G(d,p) with BSSE level using the Gaussian09 program.[49] Two additional sets density-fitted MP2 and frozen core approximation QM calculations were carried out at the DFMP2(fc)/jul-cc-pVDZ and DFMP2(fc)/jul-cc-pVTZ level with BSSE counterpoise corrections using the Psi4 quantum chemistry package.[53] A total of 1,250 clusters were included in the training set (999 clusters of 9 molecules, 157 clusters of 11 molecules, and 94 clusters of 13 molecules). The MD simulation system used to generate the snapshots from which the clusters were extracted comprises 344 methanol molecules. The MD trajectory was generated under periodic boundary conditions (PBC) in the *NpT* ensemble at constant pressure and temperature using the original AMOEBA force field (amoeba09.prm).[22] To acquire a wide range of clusters configurations, *NpT* ensemble simulations were performed on the system at various temperatures and pressures, from T = 193.15 K, p = 1.0 atm, T = 293.15 K, p = 1.2 atm, to T = 393.15 K, p = 6.293 atm, with final stable density of the system as ρ = 0.933, 0.794 and 0.627 g/ml, respectively. These MD simulations were generated with an integration time step of 1 fs using the GPU dynamics program in the Tinker-OpenMM suite (http://biomol.bme.utexas.edu/tinker-openmm). Table 3 provides further details about the clusters extracted from the different MD simulation systems. The optimization of the vdW parameters seeks to minimize the objective function F given by,

$$F(p_1, p_2, ..., p_m) = \sum_j \left(\Delta E_j^{\mathrm{QM}} - \delta - \Delta E_j^{\mathrm{MM}}(p_1, p_2, ..., p_m) \right)^2 \qquad (2)$$

where δ is an offset factor and ΔE_j^{QM} and ΔE_j^{MM} are the interaction energies between the central molecule and the surrounding molecules for the *j-th* cluster in the training set calculated from the QM and MM force field, respectively. The vdW parameters in the force field were optimized using the interaction energy from 1,250 methanol clusters by employing the developed ML/GA framework.[6-7, 35]

A flowchart of the ML/GA optimization protocol is depicted in Figure 1. Briefly, the optimization starts with a set of parent parameters, which is defined as the population. For parameters optimization of the force field, the parents may have different ranges of settings. Some of the individuals in this population present a better fit, which in the context of parameters optimization means lower value in the objective function, *e.g.*, Δ and F in Eqs (1) and (2). Fit parents survive and are allowed to mate, which is accomplished by crossing patterns with other fit individuals. During crossover, random mutations in the genes are also allowed to a certain degree to avoid a stagnant gene pool and a better sampling of the parameters space. The offspring individuals form the next generation of parents and this process continues until some pre-defined criteria are met. For the ML/GA optimization of electrostatic parameters, the process was initiated by generating a population of sets of parameters randomly, such that their values lie within physically allowable limits, which are ±1.5 times of value from Tinker Poledit and Potential programs for methanol monomer at MP2/6-311G(d,p) level of QM calculations. The

searching range of the 44 independent electrostatic parameters over which the ML/GA optimization was is given in Table 2.

For the ML/GA optimization of the vdW parameters, the process was initiated by generating a population a population of N_p = 120 parameters sets randomly, such that their values lie within physically allowable limits. The searching range of the 11 independent vdW parameters over which the ML/GA optimization was performed is given in Table 4. For each set of parameters, we compute the interaction energies for all structures in the training dataset using Tinker and evaluate the objective function F given by Eq. (2). Sets of parameters were then ranked in ascending order. After the ranking, non-linear roulette wheel selection [54-55] was performed to select the best 60% members, *i.e.*, the ones with lowest values of F, which were then subjected to genetic operations: mutation and crossover with crossover-rate 3%. These mutations introduce sufficient diversity into the population, and the non-linear selection scheme helps to avoid premature convergence of the ML/GA run. After the genetic operations, both the parent and offspring sets of parameters are ranked by their value of F. The best N_p parameters sets are then chosen to constitute the next generation. Such an optimization routine ensures that only satisfactory parameters sets survive after each generation; upon repeating this workflow for sufficient generations and sampling viable regions in the parameters space, we performed three separate ML/GA runs starting with different random populations. From each of the converged ML/GA run, we chose the final parameters set corresponding the lowest value of F. This code is freely available for download at https://github.com/AmYingLi/GA4AMOEBA.

RESULTS AND DISCUSSION

We successfully generated electrostatic and vdW parameters for the methanol molecule consistent with the functional form of the AMOEBA force field by exclusively taking QM data of molecular clusters as training datasets and using machine learning techniques. For the sake of clarity, we emphasize that the intended objective of the present effort is not to offer a re-calibration of the original AMOEBA force field for methanol that could be transferable to other systems. Our central goal is to test whether it is practically feasible to get parameters for a force field of a physically reasonable functional form without any prior experimental data, and to be sufficiently accurate for liquid simulations. For this reason, we refer to these results as "optimized force field models", and reserve the acronym AMOEBA only for the original force field: amoeba09.prm.

In the following, we present the results for the electrostatic parameters and the vdW parameters from the genetic algorithm optimization by calibrating the electrostatic potential and the interaction energy for methanol molecules. Finally, a number of additional classical MD simulations were carried out using the set of optimized parameters to calculate the density (ρ) and heat of vaporization (ΔH_{vap}) of liquid methanol to verify the accuracy of the ML/GA strategy to generate force field parameters. The condensed liquid phase system consists of 344 methanol molecules simulated with periodic boundary conditions (PBC) in the NpT ensemble at a constant pressure of 1 atm as temperature varies from -5°C to 60°C. The density and heat of vaporization were averaged from three different simulations up to 2.5 ns initiated with different random velocities.

One important methodological question for a machine learning strategy is whether the number of clusters included in the training set is sufficiently large to yield a statistically meaningful sampling for parameters determination. As a control, to ascertain the validity of the ML/GA strategy, we first attempted to recover the vdW parameters using interaction energy generated directly from the original amoeba09.prm force field (the original electrostatic parameters from amoeba09.prm were kept unchanged for this test). If the ML/GA strategy is functioning properly, then the generated parameters should be very close to the original vdW parameters amoeba09.prm. In practice, it was found that a minimum dataset of 600 clusters of 9 methanol molecules was necessary to accurately recover the amoeba09.prm parameters, where 300 clusters were extracted from MD trajectories generated at atmospheric pressure and corresponding density, and 300 clusters were extracted from MD trajectories generated at high pressure. The resulting liquid properties from MD simulations using the ML/GA generated vdW parameters were essentially identical to those of the original amoeba09.prm. This test confirmed that our ML/GA protocol was effective, as long as the training set was sufficiently large (more than several hundred configurations).

The electrostatic parameters were determined by training the atomic multipoles (q, $\vec{\mu}$, $\vec{\vec{Q}}$), atomic polarizabilities (α) and Thole's factor (a) from the QM electrostatic potential on the Connolly surface of a methanol molecule using the Gaussian Distributed Multipole Analysis (GDMA 2.2) tool[39, 42] and the Tinker package.[44, 46-47] We then refined the atomic polarizabilities and Thole's factor with a set of methanol dimers using the ML/GA. The ML/GA generated electrostatic parameters from the MP2/6-311G(d,p) level are given in Table 5. Because the AMOEBA force field adopts a universal Thole's damping factor, during the refinement step, the only varying parameter for each atom type is the polarizability. From table 5, we can see what are the change of the polarizability (α) and Thole's factor (a) through learning the electrostatic potential from a monomer to 4,943 dimers (denoted as →). As shown in Table 6, the electrostatic potential calculated from this ML/GA generated force field model closely matches the QM at the MP2/6-311G(d,p) level for 4,943 methanol dimers. In fact, the deviations are smaller than those given by the original parameterization of the AMOEBA force field (amoeba09.prm).[22] The new optimized force fields were cross-validated by comparing the MM and QM electrostatic potential for 502 clusters of 9 methanol molecules, which were not used in the training process. The results, given in Table 6, show that the deviations between the force field with ML/GA generated parameters and the QM calculations are smaller than those obtained with amoeba09.prm.[22] It is important to note that such comparisons with AMOEBA are meaningful because essentially the same QM methods [MP2/6-311G(d,p) and MP2/aug-cc-pVTZ] were originally used to determine the amoeba09.prm electrostatic parameters.[22] Therefore, the discrepancies of the resulting parameters is likely due to sampling differences.

The current procedure used to determine the electrostatic parameters, which considers the mutual induced polarization of the two molecules in methanol dimers, differs from the standard AMOEBA protocol.[22, 30] Nonetheless, all sets of electrostatic parameters remain fairly similar to one another. This overall consistency reflects the fact that the electrostatic features of a small molecule like methanol are accurately constrained by the QM calculations. In contrast, the determination of the vdW parameters without any experimental data is much more

challenging. The vdW parameters in the AMOEBA force field (amoeba09.prm) were empirically adjusted to yield accurate liquid properties. A key question is whether a strategy exclusively based on QM calculations has the ability to yield force field models of equivalent accuracy.

The vdW parameters were optimized to best match the interaction between a central methanol molecule and its nearest neighbors via non-perturbative supermolecular QM computations for a large number of clusters representative of the liquid phase. This approach is different than perturbative approaches that seek to estimate intermolecular interactions from the wave function of an isolated monomer such as symmetry-adapted perturbation theory (SAPT).[26-28, 56-57] In principle, SAPT could be used as an alternative route to generate the target data for the ML/GA optimization. A total of 1,250 methanol clusters of 9, 11 and 13 molecules were considered in the supermolecular QM cluster computations. Typical cluster configurations from the training set are depicted in Figure 2. The interactions were characterized by considering three different levels of QM calculations: MP2/6-31G(d,p), DFMP2(fc)/jul-cc-pVDZ, and DFMP2(fc)/jul-cc-pVTZ. The distribution of the interaction energy for 1,250 methanol clusters calculated at the MP2/6-31G(d,p) is shown in Figure 2(d). The distribution converging to a Gaussian-like shape suggests that the number of the configurations is adequate to provide a sufficient sampling of these systems. The ML/GA generated vdW parameters for all force field models are given in Table 7.

First we discuss the results from force field models optimized without the offset factor δ in Eq. (2). This means that the parameters are generated to match the absolute value of the QM interaction energies for the training set. The optimized force field models generated from the different QM calculations, MP2/6-31G(d,p), DFMP2(fc)/jul-cc-pVDZ, and DFMP2(fc)/jul-cc-pVTZ, converge reasonably well, with correlation coefficients of 0.974, 0.977 and 0.979, respectively. This is confirmed by observing Figure 3 (a), (b), and (c), where the interaction energies from QM and the force field models are strongly correlated. However, the ultimate test is to verify the ability of the optimized force field model to accurately predict condensed phase properties of liquid methanol. Table 8 shows the result of density and heat of vaporization from polarizable force fields optimized from QM data at the MP2/6-31G(d,p), DFMP2(fc)/jul-cc-pVDZ and DFMP2(fc)/jul-cc-pVTZ levels. Results from the original AMOEBA force field (amoeba09.prm)[22] are included for comparison. The performance of the force field models optimized to match the absolute value of the QM interaction energies (offset factor δ set to zero) is disappointing. The liquid density from the model based on MP2/6-31G(d,p) is 0.401 g/ml, and the density from the model based on DFMP2(fc)/jul-cc-pVDZ is 0.568 g/ml, which are in poor agreement with experiment (0.786 g/ml). The heat of vaporization from these two models is also too small. The model based on DFMP2(fc)/jul-cc-pVTZ performs slightly better, with a density of 0.686 g/ml and a heat of vaporization of 8.58 kcal/mol.

To address this shortcoming, an offset factor δ was introduced in Eq. (2). The purpose of the offset factor is to free the parameter optimization from the absolute magnitude of the QM interaction energies.[58-60] It should be noted that the offset factor δ is not directly used in the force field functional form, but only acting as a hyper-parameter for the GA. In essence, when δ is included in Eq. (2), the force field parameters are generated to reproduce the local "shape" of the energy surface of the cluster as a function of the atomic coordinates, while releasing the requirement to match the absolute magnitude of the QM interaction energies. This strategy has some similarities with force-

matching methodologies, where a force field is optimized on the basis of the local first derivative of the energy function,[61] though here the emphasis is put on relative interactions energies rather than on trying to match local forces. With a non-zero offset factor ($\delta \neq 0$), the force field parameters generated from ML/GA based on the different QM training datasets, MP2/6-31G(d,p), DFMP2(fc)/jul-cc-pVDZ, and DFMP2(fc)/jul-cc-pVTZ, converges similarly well, with correlation coefficients of 0.955, 0.971, and 0.975, respectively. The effect of the offset factor between the interaction energies from QM and the force field models is noticeable in Figure 3 (d), (e), and (f). However, as observed in Table 8, these force field models clearly produced liquid properties that are closer to experiment. While the inaccuracy of force field optimized without the offset δ becomes less important with better QM data, the trend in Table 8 suggests that meaningful information about the overall shape of the energy surface and the relative interaction energies is already contained within lower level QM calculations. The best parameter set, obtained from the QM data at the DFMP2(fc)/jul-cc-pVTZ level, appears to perform even better than the original AMOEBA force field (amoeba09.prm),[22] which was optimized empirically to match liquid properties. This force field model is also able to reproduce the radial distribution functions of liquid methanol determined from neutron scattering,[62] as shown in Figure 4.

Figures 3g, 3h, and 3i compare the interaction energies from QM and the original AMOEBA force field (amoeba09.prm),[22] which signifies that different basis sets of the MP2 method give out different accuracy. The original AMOEBA force field (amoeba09.prm)[22] is directly compared with the QM data in Figure 3 (g, h and i). This comparison leads to an interesting observation; there appears to be a shift between AMOEBA and the QM calculations that closely reflects the offset factor δ used in the force field models optimization. While the vdW parameters of the AMOEBA force field (amoeba09.prm)[22] field are empirically adjusted to yield accurate liquid properties, the shifts observed in Figures 3g, 3h, and 3i essentially mirror the offset factor δ used in the force field models optimization very consistently. This observation reinforces the conclusion that seeking to match the absolute interaction energies based on MP2/6-31G(d,p) and DFMP2(fc)/jul-cc-pVDZ levels leads to inaccurate force field models.

A comparison of the interaction energy from the QM calculations and from the force field with the optimized parameters is shown in Figure 3 for the 1,250 clusters. Considering Figure 3, one can see that the ML/GA generated hyper-parameters, offset factor δ, decreases with increasingly accurate QM method, equal to 4.4 kcal/mol for MP2/6-31G(d,p), 2.5 kcal/mol for DFMP2(fc)/jul-cc-pVDZ, and 1.2 kcal/mol for DFMP2(fc)/jul-cc-pVTZ). This is also consistent with the observation that DFMP2(fc)/jul-cc-pVTZ is the QM data leading to the best force field model optimized to match the absolute value of the QM interaction energies (offset factor δ set to zero). By introducing a hyper-parameter in GA, an offset factor δ in Eq. (2), we relieve the parameter optimization from the constraint of matching the absolute value of the interaction energy. ML in the presence of the offset factor yields force field models that try to reproduce the "shape" of the potential energy surface in the multi-dimensional space of the cluster coordinates. The relative accuracy of the different parameters generated by ML/GA for the AMOEBA force field model in Table 8 suggests that the relative "shape" of the potential energy surface is more important than matching the absolute value of the interaction energies. Ultimately, discrepancies (the offset factor δ) of interaction

energy calculated from MP2 and the ML/GA generated parameters for AMOEBA force field model are likely due to the incompleteness of the basis sets for the particular level of QM (MP2) theory and the insufficiency sampling of configurations in parameterization of force field.[52, 58-60] This might be the reason that higher level QM data is able to elucidate some details that are difficult to capture at the lower level. The obvious lesson from this is that the QM level must be sufficiently high.

As a final test of the accuracy of the polarizable force field optimized from QM data, additional sets of MD simulations were generated to examine the density and heat of vaporization of liquid methanol as a function of temperature. For this test, we used the vdW parameters optimized from the QM data at the DFMP2(fc)/jul-cc-pVDZ and DFMP2(fc)/jul-cc-pVTZ levels. Figure 5 shows ρ and ΔH_{vap} as temperature varies from -5°C to 60°C calculated from *NpT* MD simulations. We observe that the polarizable force field optimized from QM data accurately predicts the density and heat of vaporization, in excellent accord with experiments, over the range of temperatures.

Finally, it is important to emphasize that consideration of many-body dispersion effects, even with pair-wise long-range vdW interactions, appears to be crucial to generate accurate force field models. For instance, attempts to use ML of the vdW parameters based on the training dataset from the interaction energy calculated at the MP2/6-31G(d,p) level for a large set of methanol dimers spanning a wide range of carbon-carbon (2 to 10 Å) distance failed to yield an accurate ML/GA generated parameters for AMOEBA force field. While the force field model with ML/GA generated parameters succeeds in accurately reproducing the dimer QM training data, the resulting force field model fails to provide an accurate representation of the bulk liquid system. MD simulations (at 25°C and 1 atmosphere) showed that such a force field model, with machine learning of vdW parameters only from dimers, severely underestimates the liquid density (0.572 g/ml) and heat of vaporization (7.32 kcal/mol) compared to experiment. We also used CCSD(T)[63-65] to calculate the interaction energies of methanol dimers. While a good correlation between the QM target data and the interaction energy from the force field model was obtained (R= 0.998), the model did not yield accurate condensed matter properties when used in MD simulations. This shows that using training data from sufficiently large cluster for the parameterization is important. The lesson here is that machine learning of the vdW parameters using QM data on clusters larger than simple dimers was essential to incorporate many-body dispersion effects into the force field model. This observation is consistent with previous results from McDaniels and Schmidt, who showed that Axilrod-Teller three-body dispersion and exchange terms needed to be incorporated in their SAPT force field of methanol to obtain a satisfactory model of the liquid phase.[28] In force field optimization, it is often tempting to put more emphasis on training data obtained from the highest possible level of QM. The challenge is that, to remain computationally tractable, one is typically limited to smaller systems of a few molecules that are not able to capture all the pertinent information to represent the features of a condensed phase system. This example shows that the training data must include QM calculations on sufficiently large cluster sizes. Even though the functional form for the vdW interactions comprises only pair-wise terms, it appears that many-body dispersion effects are "effectively" incorporated during the optimization of the vdW parameters based on the cluster QM data. While this may appear surprising, the situation is analogous to the fixed

charges in an additive non-polarizable force field, which are empirically optimized to account for polarization in an effective manner.

CONCLUSIONS

The present study demonstrates the feasibility of predicting the condensed phase properties (*i.e.*, density and heat of vaporization) of a substance by optimizing the parameters of a polarizable force field using QM data exclusively—without any *a priori* information from experiment. Experimental data was only used for validating the final force field models. This idea was tested on methanol, a small yet challenging molecule containing both polar and nonpolar moieties. A genetic algorithm was utilized to overcome the challenges of parameter optimization in a high dimensional space. For the electrostatic component of the force field, we optimized all the multipoles, polarizabilities, and Thole damping factors of methanol. The ML/GA based on dimers was able to nearly reproduce the canonical AMOEBA electrostatic parameters, the optimized parameters displaying smaller deviations relative to the QM electrostatic potential than the AMOEBA original parameters.

The vdW parameters were optimized by attempting to match the interaction energy between a central molecule and the surrounding molecules for a large number of methanol clusters. These clusters were extracted from snapshots of condensed phase MD simulations to be as representative as possible of the liquid phase. Following this protocol, excellent agreement between the optimized force field model and the QM calculations could be achieved. However, only the QM calculation from fairly high level provides sufficiently accurate vdW parameters for MD simulation to get proper condensed phase properties of methanol. Consistently better results were obtained by a using a parameterization strategy allowing for an offset factor δ between the force field and the QM target data. Through this strategy, which bears some similarities with force-matching methodologies,[61] the force field is optimized to reproduce the local "shape" of the QM energy surface without matching its absolute value. Without the offset factor δ, the absolute magnitude of the QM interactions is incorporated into the MM force field via the optimized vdW parameters, yielding models that appear to be less accurate in simulations of dense liquids.

One of the main difficulties of parameterizing a force field without any experimental input stems from the reliability of the QM calculations. Here, we consider three QM methods: MP2/6-31G(d,p), DFMP2(fc)/jul-cc-pVDZ and DFMP2(fc)/jul-cc-pVTZ. It is well known that MP2 with different basis sets can lead to different accuracy. Furthermore, there appears to be an additional issue with the overestimation of interaction energy from MP2 calculations, when using a small basis set.[52, 58-60] The present results show that QM calculations relying on more extensive basis sets lead to a more accurate determination of the interaction energies. In principle, sampling of larger molecular clusters would also be expected to increase accuracy. Nevertheless, the current strategy was able to leverage moderate computation resource to tune the force field parameters for reasonable agreement between the MD simulated data and experimental values.

ACKNOWLEGDMENTS

We thank Nichols A. Romero and Stephen Gray for their help. Y.L. was supported by the Margaret Butler Postdoctoral Fellowship at Argonne Leadership Computing Facility. We gratefully acknowledge the computing resources provided on Blues, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. Use of the Center for Nanoscale Materials, an Office of Science user facility, was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. This work was completed in part with resources provided by the University of Chicago Research Computing Center. B.R. was supported by grant R01-GM072558 from the National Institutes of Health. FCP and BRB acknowledge support from the intramural research program of the National Heart, Lung and Blood Institute of the National Institutes of Health; and utilized high-performance computational capabilities of the LoBoS clusters at the National Institutes of Health (http://www.lobos.nih.gov).

REFERENCES

1. van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A., ReaxFF: A reactive force field for hydrocarbons. *J Phys Chem A* **2001,** *105* (41), 9396-9409.
2. Islam, M. M.; Kolesov, G.; Verstraelen, T.; Kaxiras, E.; van Duin, A. C., eReaxFF: A Pseudoclassical Treatment of Explicit Electrons within Reactive Force Field Simulations. *J Chem Theory Comput* **2016,** *12* (8), 3463-72.
3. Pahari, P.; Chaturvedi, S., Determination of best-fit potential parameters for a reactive force field using a genetic algorithm. *J Mol Model* **2012,** *18* (3), 1049-1061.
4. Mackerell, A. D., Jr., Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem* **2004,** *25* (13), 1584-604.
5. Larsson, H. R.; van Duin, A. C. T.; Hartke, B., Global optimization of parameters in the reactive force field ReaxFF for SiOH. *J Comput Chem* **2013,** *34* (25), 2178-2189.
6. Sen, F. G.; Kinaci, A.; Narayanan, B.; Gray, S. K.; Davis, M. J.; Sankaranarayanan, S. K. R. S.; Chan, M. K. Y., Towards accurate prediction of catalytic activity in IrO2 nanoclusters via first principles-based variable charge force field. *J Mater Chem A* **2015,** *3* (37), 18970-18982.
7. Narayanan, B.; Sasikumar, K.; Mei, Z. G.; Kinaci, A.; Sen, F. G.; Davis, M. J.; Gray, S. K.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S., Development of a Modified Embedded Atom Force Field for Zirconium Nitride Using Multi-Objective Evolutionary Optimization. *J Phys Chem C* **2016,** *120* (31), 17475-17483.
8. Massobrio, C.; Du, J.; Bernasconi, M.; Salmon, P. S., *Molecular Dynamics Simulations of Disordered Materials: From Network Glasses to Phase-Change Memory Alloys*. 2015; Vol. 215, p 1-529.
9. Matsuoka, O.; Clementi, E.; Yoshimine, M., CI Study of water dimer potential surface. *J. Chem. Phys.* **1976,** *64*, 1351-1361.
10. Belle, D. V.; Couplet, I.; Prevost, M.; Wodak, S. J., Calculations of electrostatic properties in proteins. *J. Mol. Biol.* **1987,** *198*, 721-735.
11. Cieplak, P.; Kollman, P. A.; Lybrand, T., A new water potential including polarization: Application to gas-phase, liquid and crystal properies of water. *J. Chem. Phys.* **1990,** *92*, 6755-6760.
12. Rick, S. W.; Stuart, S. J.; Berne, B. J., Dynamical fluctuating charge force field: Application to liquid water. *J Chem Phys* **1994,** *101*, 6141-6156.
13. Lopes, P. E. M.; Roux, B.; Mackerell, A. D., Jr., Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability. Theory and applications. *Theor Chem Acc* **2009,** *124* (1-2), 11-28.
14. Ponder, J. W.; Wu, C. J.; Ren, P. Y.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T., Current Status of the AMOEBA Polarizable Force Field. *J Phys Chem B* **2010,** *114* (8), 2549-2564.
15. Bauer, B. A.; Patel, S., Recent applications and developments of charge equilibration force fields for modeling dynamical charges in classical molecular dynamics simulations. *Theor Chem Acc* **2012,** *131* (3).
16. Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D., Jr., An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem Rev* **2016,** *116* (9), 4983-5013.

17. Mackerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **1998,** *102* (18), 3586-3616.
18. Lamoureux, G.; MacKerell, A. D.; Roux, B., A simple polarizable model of water based on classical Drude oscillators. *J Chem Phys* **2003,** *119* (10), 5185-5197.
19. Lamoureux, G.; Roux, B., Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J Chem Phys* **2003,** *119* (6), 3025-3039.
20. Wang, Z. X.; Zhang, W.; Wu, C.; Lei, H. X.; Cieplak, P.; Duan, Y., Strike a balance: Optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *J Comput Chem* **2006,** *27* (6), 781-790.
21. Ren, P. Y.; Ponder, J. W., Polarizable atomic multipole water model for molecular mechanics simulation. *J Phys Chem B* **2003,** *107* (24), 5933-5947.
22. Ren, P. Y.; Wu, C. J.; Ponder, J. W., Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J Chem Theory Comput* **2011,** *7* (10), 3143-3161.
23. Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A., Predictions of the properties of water from first principles. *Science* **2007,** *315* (5816), 1249-1252.
24. Wang, Y. M.; Huang, X. C.; Shepler, B. C.; Braams, B. J.; Bowman, J. M., Flexible, ab initio potential, and dipole moment surfaces for water. I. Tests and applications for clusters up to the 22-mer. *J Chem Phys* **2011,** *134* (9).
25. Paesani, F., Getting the Right Answers for the Right Reasons: Toward Predictive Molecular Simulations of Water with Many-Body Potential Energy Functions. *Acc Chem Res* **2016,** *49* (9), 1844-51.
26. Schmidt, J. R.; Yu, K.; McDaniel, J. G., Transferable Next-Generation Force Fields from Simple Liquids to Complex Materials. *Acc Chem Res* **2015,** *48* (3), 548-556.
27. McDaniel, J. G.; Schmidt, J. R., Next-Generation Force Fields from Symmetry-Adapted Perturbation Theory. *Annu Rev Phys Chem* **2016,** *67*, 467-488.
28. McDaniel, J. G.; Schmidt, J. R., First-Principles Many-Body Force Fields from the Gas Phase to Liquid: A "Universal" Approach. *J Phys Chem B* **2014,** *118* (28), 8042-8053.
29. Kobayashi, T.; Shishido, R.; Mizuse, K.; Fujii, A.; Kuo, J. L., Structures of hydrogen bond networks formed by a few tens of methanol molecules in the gas phase: size-selective infrared spectroscopy of neutral and protonated methanol clusters. *Phys Chem Chem Phys* **2013,** *15* (24), 9523-9530.
30. Wu, J. C.; Chattree, G.; Ren, P. Y., Automation of AMOEBA polarizable force field parameterization for small molecules. *Theor Chem Acc* **2012,** *131* (3).
31. Halgren, T. A., Representation of Vanderwaals (Vdw) Interactions in Molecular Mechanics Force-Fields - Potential Form, Combination Rules, and Vdw Parameters. *J Am Chem Soc* **1992,** *114* (20), 7827-7843.
32. Powell, M. J. D., Nonconvex Minimization Calculations and the Conjugate-Gradient Method. *Lect Notes Math* **1984,** *1066*, 122-141.
33. Fliege, J.; Svaiter, B. F., Steepest descent methods for multicriteria optimization. *Math Method Oper Res* **2000,** *51* (3), 479-494.
34. McCall, J., Genetic algorithms for modelling and optimisation. *J Comput Appl Math* **2005,** *184* (1), 205-222.
35. Narayanan, B.; Kinaci, A.; Sen, F. G.; Davis, M. J.; Gray, S. K.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S., Describing the Diverse Geometries of Gold from Nanoclusters to Bulk—A First-Principles-Based Hybrid Bond-Order Potential. *The Journal of Physical Chemistry C* **2016,** *120* (25), 13787-13800.
36. Cherukara, M. J.; Narayanan, B.; Kinaci, A.; Sasikumar, K.; Gray, S. K.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S., Ab Initio-Based Bond Order Potential to Investigate Low Thermal Conductivity of Stanene Nanostructures. *J Phys Chem Lett* **2016,** *7* (19), 3752-3759.
37. Binkley, J. S.; Pople, J. A.; Hehre, W. J., Self-Consistent Molecular-Orbital Methods .21. Small Split-Valence Basis-Sets for 1st-Row Elements. *J Am Chem Soc* **1980,** *102* (3), 939-947.
38. Dunning, T. H., Gaussian-Basis Sets for Use in Correlated Molecular Calculations .1. The Atoms Boron through Neon and Hydrogen. *Journal of Chemical Physics* **1989,** *90* (2), 1007-1023.
39. Stone, A. J., Distributed Multipole Analysis, or How to Describe a Molecular Charge-Distribution. *Chem Phys Lett* **1981,** *83* (2), 233-239.
40. Stone, A. J.; Alderton, M., Distributed Multipole Analysis - Methods and Applications. *Mol Phys* **1985,** *56* (5), 1047-1064.
41. Cisneros, G. A., Application of Gaussian Electrostatic Model (GEM) Distributed Multipoles in the AMOEBA Force Field. *J Chem Theory Comput* **2012,** *8* (12), 5072-5080.
42. Stone, A. J.; Alderton, M., Distributed multipole analysis - Methods and applications (Reprinted from Molecular Physics, vol 56, pg 1047-1064, 1985). *Mol Phys* **2002,** *100* (1), 221-233.
43. Stone, A. J., Distributed multipole analysis: Stability for large basis sets. *J Chem Theory Comput* **2005,** *1* (6), 1128-1132.

44. Ponder, J. W.; Richards, F. M., An Efficient Newton-Like Method for Molecular Mechanics Energy Minimization of Large Molecules. *J Comput Chem* **1987,** *8* (7), 1016-1024.
45. Kundrot, C. E.; Ponder, J. W.; Richards, F. M., Algorithms for Calculating Excluded Volume and Its Derivatives as a Function of Molecular-Conformation and Their Use in Energy Minimization. *J Comput Chem* **1991,** *12* (3), 402-409.
46. Dudek, M. J.; Ponder, J. W., Accurate Modeling of the Intramolecular Electrostatic Energy of Proteins. *J Comput Chem* **1995,** *16* (7), 791-816.
47. Kong, Y.; Ponder, J. W., Calculation of the reaction field due to off-center point multipoles. *J Chem Phys* **1997,** *107* (2), 481-492.
48. Pappu, R. V.; Hart, R. K.; Ponder, J. W., Analysis and application of potential energy smoothing and search methods for global optimization. *J Phys Chem B* **1998,** *102* (48), 9725-9742.
49. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Gaussian, Inc.: Wallingford, CT, USA, 2009.
50. Headgordon, M.; Pople, J. A.; Frisch, M. J., Mp2 Energy Evaluation by Direct Methods. *Chem Phys Lett* **1988,** *153* (6), 503-506.
51. Boys, S. F.; Bernardi, F., The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors (Reprinted from Molecular Physics, vol 19, pg 553-566, 1970). *Mol Phys* **2002,** *100* (1), 65-73.
52. Li, H.; Ngo, V.; Da Siva, M. C.; Salahub, D. R.; Callahan, K.; Roux, B.; Noskov, S. Y., Representation of Ion-Protein Interactions Using the Drude Polarizable Force-Field. *J Phys Chem B* **2015,** *119* (29), 9401-9416.
53. Turney, J. M. S., A. C.; Parrish, R. M.; Hohenstein, E. G.; Evangelista, F. A.; Fermann, J. T.; Mintz, B. J.; Burns, L. A.; Wilke, J. J.; Abrams, M. L.; Russ, N. J.; Leininger, M. L.; Janssen, C. L.; Seidl, E. T.; Allen, W. D.; Schaefer, H. F., III; King, R. A.; Valeev, E. F.; Sherrill, C. D.; Crawford, T. D., *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012,** *2* (4), 556-565.
54. Lipowski, A.; Lipowska, D., Roulette-wheel selection via stochastic acceptance. *Physica A* **2012,** *391* (6), 2193-2196.
55. Razali, N. M.; Geraghty, J. In *Genetic Algorithm Performance with Different Selection Strategies in Solving TSP*, Proceedings of the World Congress on Engineering, 2011.
56. Parker, T. M.; Burns, L. A.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D., Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *J Chem Phys* **2014,** *140* (9), 094106.
57. McDaniel, J. G.; Schmidt, J. R., Physically-Motivated Force Fields from Symmetry-Adapted Perturbation Theory. *J Phys Chem A* **2013,** *117* (10), 2053-2066.
58. Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P., Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys Chem Chem Phys* **2006,** *8* (17), 1985-1993.
59. Cybulski, S. M.; Lytle, M. L., The origin of deficiency of the supermolecule second-order Moller-Plesset approach for evaluating interaction energies. *J Chem Phys* **2007,** *127* (14).
60. Tkatchenko, A.; DiStasio, R. A.; Head-Gordon, M.; Scheffler, M., Dispersion-corrected Moller-Plesset second-order perturbation theory (vol 131, 094106, 2009). *J Chem Phys* **2009,** *131* (12).
61. Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A., Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching. *J Chem Phys* **2004,** *120* (23), 10896-10913.
62. Yamaguchi, T.; Hidaka, K.; Soper, A. K., The structure of liquid methanol revisited: a neutron diffraction experiment at −80 °C and +25 °C. *Mol. Phys.* **1999,** *96* (8), 1159- 1168.
63. Woon, D. E., Accurate Modeling of Intermolecular Forces - a Systematic Moller-Plesset Study of the Argon Dimer Using Correlation Consistent Basis-Sets. *Chem Phys Lett* **1993,** *204* (1-2), 29-35.
64. Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P., Scaled MP3 Non-Covalent Interaction Energies Agree Closely with Accurate CCSD(T) Benchmark Data. *ChemPhysChem* **2009,** *10* (1), 282-289.
65. Řezáč, J.; Hobza, P., Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *J Chem Theory Comput* **2013,** *9* (5), 2151-2155.
66. Yamaguchi, T.; Hidaka, K.; Soper, A. K., The structure of liquid methanol revisited: A neutron diffraction experiment at -80 degrees C and +25 degrees C. *Mol Phys* **1999,** *96* (8), 1159-1168.

Table 1. The electrostatic potential energy (E_{ESP}) for methanol monomer from MP2 calculations using different basis sets and from the corresponding fitted multipole parameters, with the corresponding root mean square deviation and relative error.

Basis	ESP from MP2 (kcal/mol)	ESP from fitted multipole (kcal/mol)	RMSD (kcal/mol)	Relative Error (%)
6-31G(d, p)	4.878	4.865	0.165	0.267
6-31+G*	5.675	5.649	0.172	0.458
6-31G*	5.059	5.046	0.170	0.257
6-311G(d, p)	5.947	5.935	0.172	0.202
6-311G(2df, 2pd)	4.462	4.447	0.154	0.336
6-311G*	5.187	5.172	0.156	0.289
6-311+G*	5.648	5.626	0.167	0.390
6-311++G**	5.307	5.282	0.165	0.471
6-311+G**	5.309	5.285	0.166	0.452
Aug-CC-pvDz	4.758	4.729	0.177	0.609
Aug-CC-pvTz	4.748	4.723	0.165	0.527
Aug-CC-pvQz	4.745	4.720	0.165	0.527

Table 2. Range of the 44 independent electrostatic parameters over which the ML/GA was performed.

Parameters		Atom types			
		O	H (-O)	C	H (-C)
E_{ESP}	Monopole (q)	-1.0 ~ 0	0 ~ 0.4	0 ~ 0.3	0 ~ 0.06
	Dipole (μ_x, μ_y, μ_z)	$(0 \sim 0.5, 0.0, 0 \sim 0.4)$	$(-0.14 \sim 0, 0.0, -0.4 \sim 0)$	$(-0.4 \sim 0, 0.0, 0 \sim 0.9)$	$(-0.1 \sim 0, 0.0, -0.1 \sim 0)$
	Quadrupole $\begin{bmatrix} Q_{xx} & * & * \\ Q_{yx} & Q_{yy} & * \\ Q_{zx} & Q_{zy} & * \end{bmatrix}$	$\begin{bmatrix} 0 \sim 0.6 & * & * \\ 0.0 & -0.8 \sim 0 & * \\ 0 \sim 0.06 & 0.0 & * \end{bmatrix}$	$\begin{bmatrix} 0 \sim 0.28 & * & * \\ 0.0 & 0 \sim 0.06 & * \\ -0.2 \sim 0 & 0.0 & * \end{bmatrix}$	$\begin{bmatrix} 0 \sim 0.06 & * & * \\ 0.0 & -0.11 \sim 0 & * \\ -0.6 \sim 0 & 0.0 & * \end{bmatrix}$	$\begin{bmatrix} 0 \sim 0.15 & * & * \\ 0.0 & -0.17 \sim 0 & * \\ -0.08 \sim 0 & 0.0 & * \end{bmatrix}$
	Polarizability (α)	0.5 ~ 1.0	0.2 ~ 0.7	1.0 ~ 1.6	0.2 ~ 0.7
	Thole's factor (a)	0.3 ~ 0.5	0.3 ~ 0.5	0.3 ~ 0.5	0.3 ~ 0.5

Table 3. Number distribution of the extracted clusters (9 molecules, 11 molecules and 13 molecules) from different liquid MD systems.

MD simulations	9 methanol	11 methanol	13 methanol
ρ = 0.933 g/ml, T = 193.15 K, p = 1.000 atm	372	157	94
ρ = 0.794 g/ml, T = 293.15 K, p = 1.200 atm	502	N/A	N/A
ρ = 0.627 g/ml, T = 393.15 K, p = 6.293 atm	125	N/A	N/A

Table 4. Range of the 11 independent vdW parameter over which the ML/GA optimization were performed.

Parameters		Atom types			
		O	H (-O)	C	H (-C)
E_{vdW}	Minimum energy depth (ε_{min})	3.5 ~ 4.0	2.5 ~ 3.0	2.2 ~ 2.8	3.5 ~ 4.0
	Minimum energy distance (R_{min})	0.12 ~ 0.15	0.001 ~ 0.0015	0.4 ~ 0.6	0.001 ~ 0.009
	H reduction factor (λ)	N/A	0.9 ~ 0.95	N/A	0.9 ~ 0.95
	Energy offset (δ)	0.05 ~ 8.0			

Table 5. ML/GA generated the electrostatic parameters.

Parameters		Atom types			
		O	H (-O)	C	H (-C)
E_{ESP}	Monopole (q)	-0.45883	0.21357	0.15811	0.02905
	Dipole (μ_x,μ_y,μ_z)	(0.23922, 0.0, 0.16079)	(-0.06982, 0.0, -0.20347)	(-0.19328, 0.0, 0.45462)	(-0.04887, 0.0, -0.04808)
	Quadrupole $\begin{bmatrix} Q_{xx} & * & * \\ Q_{yx} & Q_{yy} & * \\ Q_{zx} & Q_{zy} & * \end{bmatrix}$	$\begin{bmatrix} 0.31570 & 0.0 & 0.03229 \\ 0.0 & -0.40859 & 0.0 \\ 0.03229 & 0.0 & 0.09289 \end{bmatrix}$	$\begin{bmatrix} 0.14001 & 0.0 & -0.10133 \\ 0.0 & 0.03002 & 0.0 \\ -0.10133 & 0.0 & -0.17003 \end{bmatrix}$	$\begin{bmatrix} 0.02941 & 0.0 & -0.30849 \\ 0.0 & -0.05306 & 0.0 \\ -0.30849 & 0.0 & 0.02365 \end{bmatrix}$	$\begin{bmatrix} 0.07423 & 0.0 & -0.04155 \\ 0.0 & -0.08433 & 0.0 \\ -0.04155 & 0.0 & 0.01010 \end{bmatrix}$
	Polarizability (α)	0.986→0.852	0.573→0.438	1.240→1.351	0.573→0.438
	Thole's factor (a)	0.3908→0.390	0.3908→0.390	0.3908→0.390	0.3908→0.390

Table 6. Root mean square deviation (RMSD) of electrostatic potential between different force field models and the QM calculations at the MP2/6-311G(d,p) level. The force fields include the new optimized electrostatic parameters and the original AMOEBA force field (amoeba09.prm). The methanol clusters include 4943 dimers and 502 nonamers (9 molecules cluster).

RMSD of E_{elec} (kcal/mol)	AMOEBA	Optimized force field model
4,943 dimers	0.723	0.305
502 nonamers	0.718	0.495

Table 7. ML/GA generated van der Waals parameters.

vdW parameters		Amoeba09	MP2/6-31G(d,p)		MP2/jul-cc-pVDZ		MP2/jul-cc-pVTZ		CCSD(T)*
Atom Type			$\delta = 0$	$\delta \neq 0$	$\delta = 0$	$\delta \neq 0$	$\delta = 0$	$\delta \neq 0$	$\delta = 0$
O	ε_{min}	3.4050	3.7132	3.4745	3.6377	3.4687	3.5617	3.4372	3.6165
	R_{min}	0.1100	0.1274	0.1484	0.1505	0.1483	0.2037	0.1484	0.1026
H_O	ε_{min}	2.6550	3.1559	2.1463	2.4495	2.3580	2.9640	2.2861	2.0481
	R_{min}	0.0135	0.0026	0.0100	0.0045	0.0100	0.0026	0.0103	0.0100
	λ	0.910	0.9131	0.9049	0.9198	0.9001	0.9399	0.9014	0.9090
C	ε_{min}	3.7600	2.8536	3.9092	3.1137	3.8046	3.4887	3.8005	3.8744
	R_{min}	0.1010	0.2803	0.1018	0.1373	0.1185	0.1920	0.1198	0.1051
H_C	ε_{min}	2.8700	3.8178	2.9984	3.3133	2.9583	3.3784	2.7496	2.0066
	R_{min}	0.0240	0.0048	0.0200	0.0118	0.0215	0.0068	0.0212	0.0203
	λ	0.910	0.9120	0.9333	0.9077	0.9085	0.9312	0.9490	0.9357

ε_{min} is the minimum energy depth; R_{min} is the minimum energy distance; λ is the hydrogen reduction factor
*Parameters optimized only from QM data on methanol dimer

Table 8. Density (ρ) and heat of vaporization (ΔH_{vap}) of liquid methanol calculated from MD simulations (T = 25 °C and p = 1 atm) using different optimized polarizable force fields. The optimized vdW parameters were determined from according to the Eq (1) without δ and with δ offset. For comparison, the results from the AMOEBA force field (amoeba09.prm) are given. The MD results were averaged from 3 simulations of 2.5 ns.

Property	Exp.	MD						
		MP2/6-31G(d,p)		DFMP2(fc)/jul-cc-pVDZ		DFMP2(fc)/jul-cc-pVTZ		AMOEBA
		$\delta = 0$	$\delta \neq 0$	$\delta = 0$	$\delta \neq 0$	$\delta = 0$	$\delta \neq 0$	
ρ (g/ml)	0.786	0.401	0.759	0.568	0.763	0.686	0.781	0.774
ΔH_{vap} (kcal/mol)	8.95	6.54	9.44	7.23	9.43	8.58	8.98	9.17

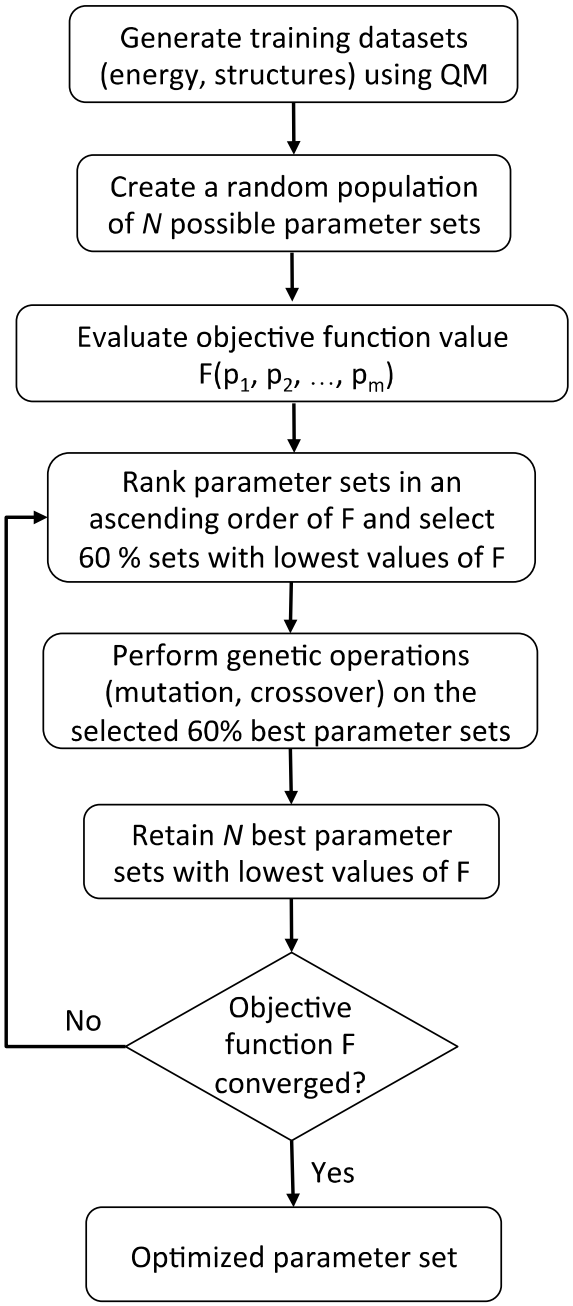

Figure 1. Flowchart of the machine learning with genetic algorithm (ML/GA) strategy describing the sequence of steps employed in this work.

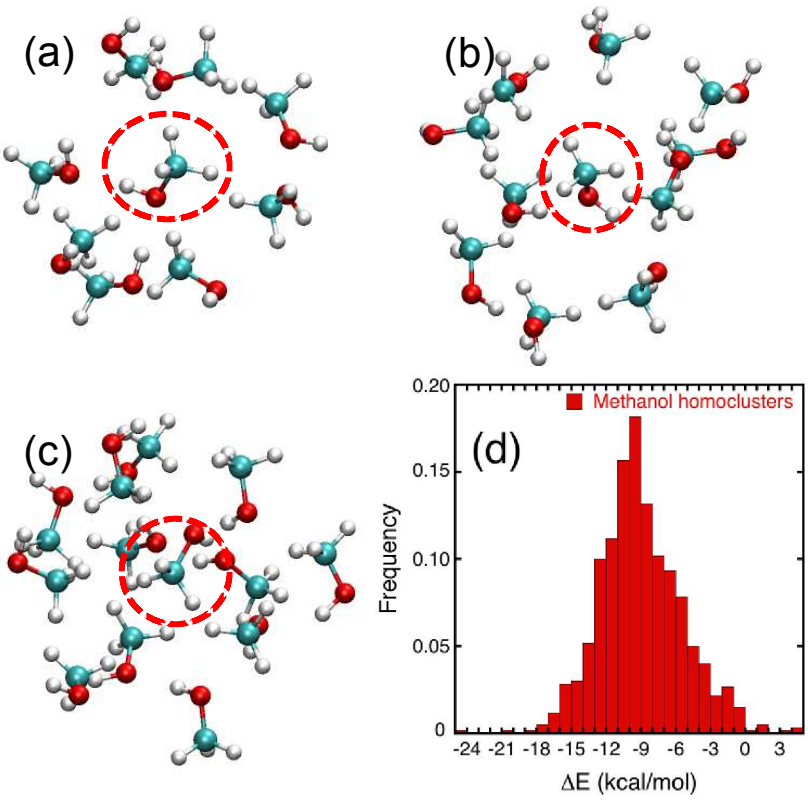

Figure 2. Illustration of the methanol cluster configurations used for the force field parameterization. Typical configurations of the (a) 9 molecules, (b) 11 molecules, and (c) 13 molecules of methanol are shown (the H atom is displayed in white, C atom in cyan and O atom in red). The interaction energy between the central molecule (circled by a dashed line) and the surrounding molecules is used in the optimization of the vdw parameters. (d) Distribution of interaction energies from MP2/6-31G(d,p) of 1,250 methanol clusters (999 clusters of 9 molecules, 157 clusters of 11 molecules, and 94 clusters of 13 molecules).

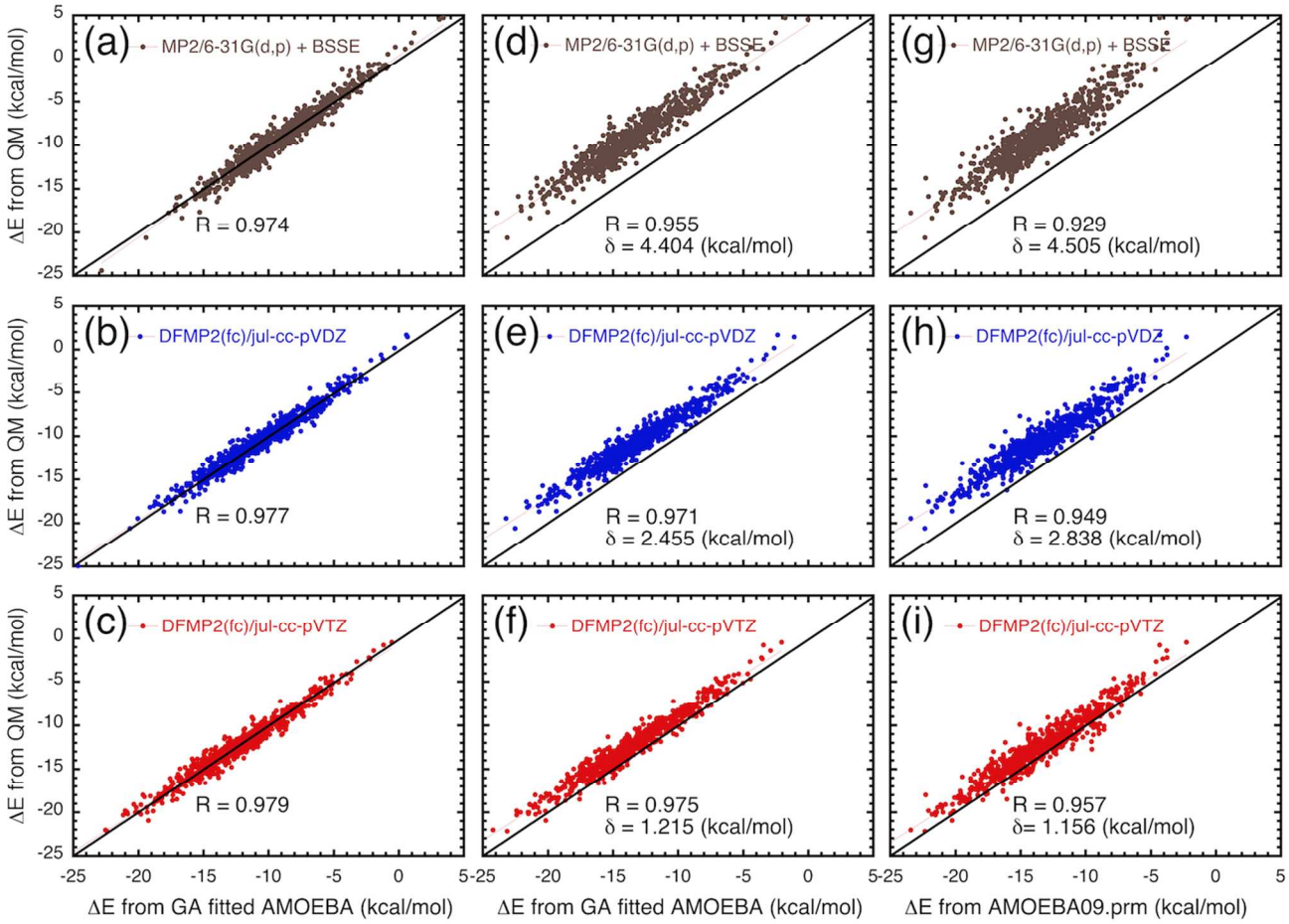

Figure 3. Comparison of the interaction energy (ΔE) for 1,250 methanol clusters, including 999 clusters of 9 molecules, 157 clusters of 11 molecules, and 94 clusters of 13 molecules, computed from (a) MP2/6-31G(d,p) (b) DFMP2(fc)/jul-cc-pVDZ, and (c) DFMP2(fc)/jul-cc-pVTZ and the optimized force field model fitted without the offset parameter δ in Eq. (2). (d), (e) and (f) are the optimized force field model fitted with the offset parameter δ in Eq. (2).

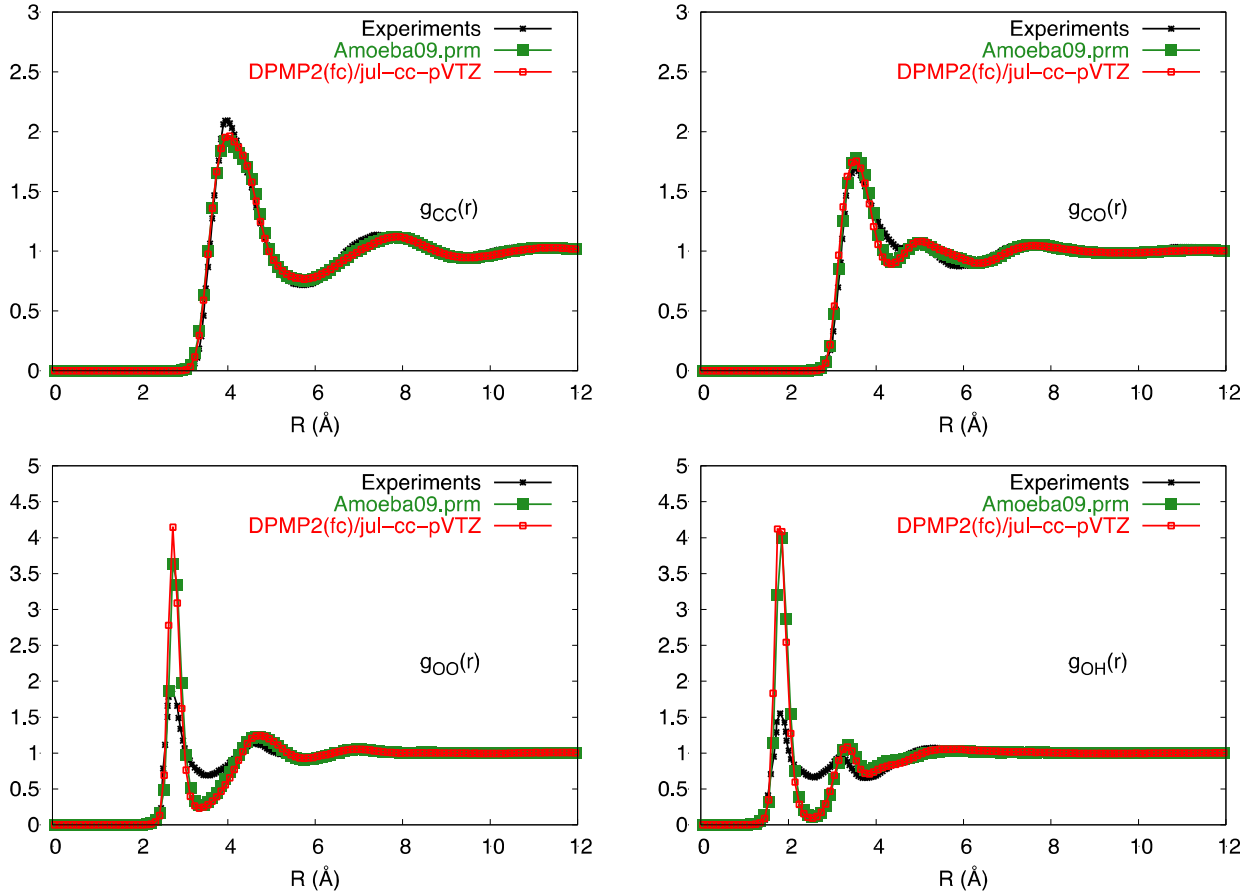

Figure 4. Liquid structure of methanol. The radial distribution function of methanol determined from neutron scattering[62] compares favorably with the present force field model.[66] The O-H hydrogen bonding peak is, however, higher than the experiment. This may reflect the neglect of quantum effects of the nuclei in the simulations,[25] as well as inaccuracies in the experimentally determined pair correlation functions.

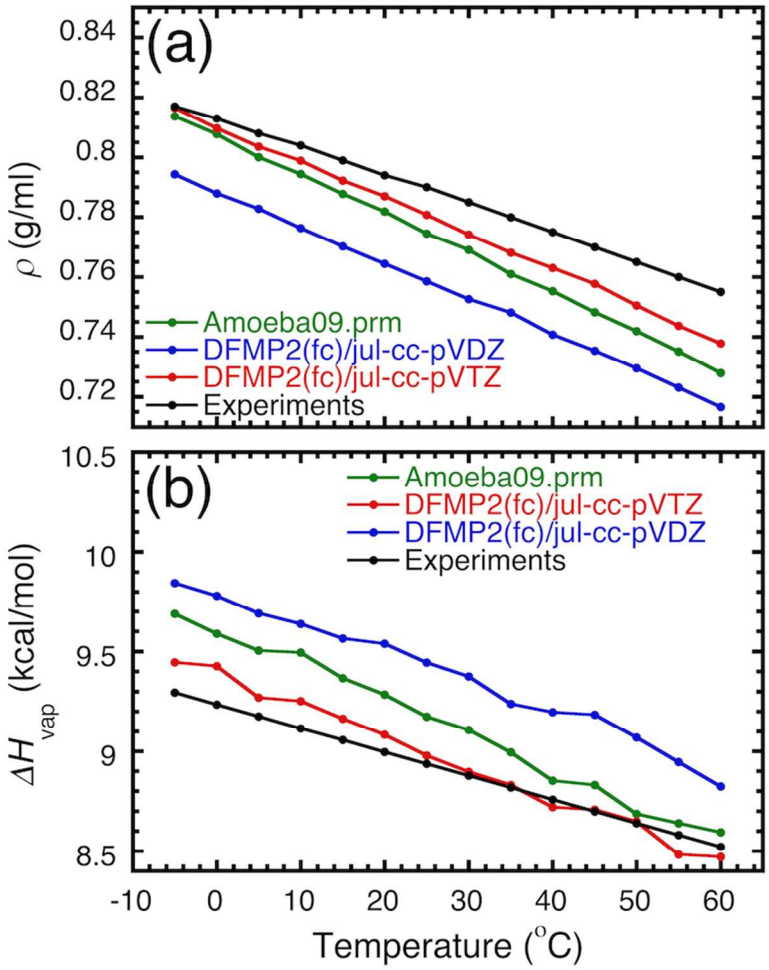

Figure 5. Density (ρ) and heat of vaporization (ΔH_{vap}) of methanol as a function of temperature from MD simulations using the optimized polizable force field optimized from the DFMP2(fc)/jul-cc-pVDZ and DFMP2(fc)/jul-cc-pVTZ data. For comparison the results from experiments and the original AMOEBA force field (amoeba09.prm) are also shown.