Support Vector Regression-Based Monte Carlo Simulation of Flexible Water Clusters

Samik Bose, Suman Chakrabarty,* and Debashree Ghosh*

ABSTRACT: Molecular simulations based on classical force fields are computationally efficient but lack accuracy due to the empirical formulation of non-bonded interactions. Quantum mechanical (QM) methods, albeit accurate, have inhibitory computational costs for large molecules and clusters. Hence, to overcome the bottleneck, machine learning (ML)-based methods have been employed in the recent years. We had earlier reported a combined scheme of many-body expansion (MBE) and ML to predict the interaction energies of rigid water clusters. In this work, we proceed toward building a flexible water model using the ML-MBE scheme. This ML-MBE scheme has an error of <1% for interaction energy prediction in comparison to the parent QM method for flexible water decamers. Machine learning-based Monte Carlo simulations (MLMC) are performed with this water model, and the structural properties of these configurations are compared with those obtained from ab initio molecular dynamics (AIMD) and the TIP3P classical force field. The radial distribution functions, tetrahedral order parameters, and number of hydrogen bonds in AIMD and MLMC have a similar qualitative and quantitative trend, whereas the classical force fields show a significant deviation.

INTRODUCTION

Computational studies of biomolecules are generally carried out in aqueous solution.1–7 Complete solvation of the solute molecules often requires a large number of water molecules.5,6 The interaction energies of such significantly large systems are modeled by using empirical, non-polarizable force fields.7 Although these force fields produce a computationally affordable route to study larger systems, they are often criticized for their treatment of polarization.12,13 Non-bonded interactions in empirical, non-polarizable force fields are treated in a pairwise additive scheme where the point charges and van der Waals points are considered to be on the atom centers of a molecule, as opposed to a continuous electron density picture.14,15 These force fields generally exclude the effect of higher-order interactions like polarization and charge transfer.16,17

Polarization in water plays a key role in structural and dynamical properties of water across all the phases.18,19 For the last few decades, a lot of attention has been devoted toward generating force fields to accurately predict the various properties of water such as density, enthalpy of vaporization, radial distribution functions (RDF), etc.18,19 Also, there have been substantial efforts to improve the existing water models to achieve transferability across different thermodynamic conditions.21–23 In one of the earliest efforts, Stillinger and Rahman were able to simulate liquid water by molecular dynamics (MD).24 Among the non-polarizable force fields of water, which are widely used to date in MD simulation, the simple point-charge-based water model (SPC) was proposed in 1981 by Berendsen and co-workers. This model was later re-parameterized to include the effect of correction due to self-interaction.25 Jorgensen and co-workers presented a transferable intermolecular potential, positioned on three and four points (TIP3P and TIP4P).26 Again, these models were later modified to enhance accuracy by constructing force fields such as TIP5P, TIP4P-Ew, etc.21,22 Flexible non-polarizable water models such as SPC/flex were introduced to capture the vibrational properties of bulk water.23 These flexible models predict more compact water clusters in comparison to their parent rigid models and are in better agreement with ab initio predictions.23 However, with water being a highly polar system, incorporation of flexibility cannot fully compensate for ignoring the effects of polarizability.13

In order to incorporate the effect of polarization, Ren and Ponder proposed a force field, AMOEBA, where polarization is treated via self-consistent induced atomic dipoles.6,26 Later, Ren and co-workers modified the AMOEBA force field to produce comparable results at a lesser computational cost.28 Cummings and co-workers suggested a Gaussian charge-based polarizable model that predicts accurate dielectric and structural properties of water in clusters in supercritical conditions.29 Roux and co-workers presented a Drude
oscillator-based polarizable water model, which reproduces many properties of liquid water at ambient conditions. 36,31 Friesner and co-workers reported a five-site ab initio parameterized polarizable water model that performs well in room temperature producing RDFs and dipole moments comparable to experimental observations. 32 Bowman and co-workers modeled a flexible ab initio potential WHBB, which was reported to predict the various electronic and structural properties of water clusters accurately. 33 It was later used to develop potentials for hydrated ions as well. 34–36 Later, Kiss and Baranyai proposed a rigid polarizable model that performs reasonably well going from cluster to liquid water in terms of structural and dynamical properties. 37,38 Although, polarizable and ab initio potentials are superior in terms of accuracy compared to their non-polarizable counterparts, the computational expense often restricts the usage in large systems.

Data-driven approaches have also been used to generate polarizable force fields with the desire to achieve better accuracy in the prediction of structural and dynamical properties in a computationally efficient scheme. 39–42 This opens up a new avenue to explore the domain of polarizable water models. Artificial neural networks (ANNs) have been used extensively for building force fields. 43–52 Recently, regression-based ML approaches are also gaining popularity for achieving similar targets. 53–57 Paesani and co-workers developed a many-body water potential (MBpol model) based on the analytical form of the Thole type polarizable force field TTMF-4. 53–55 The short-range interactions are formulated as permutationally invariant polynomials. The MBpol model is able to predict a range of structural and dynamical properties of water in the condensed phase. 53–55 They have also successfully implemented the method toward developing a potential for ion–water interactions. 56,57 Gaussian approximation potentials (GAP) describe the atomic energy of a molecule as geometry-dependent analytical functions of the neighboring atoms using Gaussian process regression. This has been reported to estimate accurate interaction energies of water clusters. 60

Support vector regression (SVR) is an ML method that can be used to capture the complicated non-linear functional form of the potential energy. In our previous work, we have benchmarked interaction energies of rigid water clusters estimated by SVR against ab initio energies. 61 We have observed that the prediction of interaction energies by SVR was in accordance with its underlying QM method. This encouraged us to expand the method further and investigate the possibilities of predicting interaction energies of flexible water clusters (i.e., clusters with vibrational degrees of freedom). We have used only two-body and three-body interaction energy contributions in the many-body expansion of interaction energy in our previous rigid water model. Now, along with the two- and three-body terms, we also consider the one-body energy contributions (intramolecular vibrations) in the MBE of interaction energy to develop a flexible polarizable water model. It should be noted that our SVR-MBE scheme fits the total interaction energy of the system in the form of its constituents (one-, two-, and three-body energies) of many-body expansion. Thus, in our formulation, the analytical form of the non-bonded energies is omitted. The reasons for eliminating the analytical form of the non-bonded energy are twofold: (i) one is to explore the capacity of advanced ML methods to predict the many-body terms without any analytical form, which is successfully achieved with an RMSE of non-bonded energy less than 3% against the parent QM method. (ii) The total non-bonded energy (up to the three-body interaction in the MBE of interaction energy) estimated with interactions such as dispersion, polarization, exchange-correlation, or charge-transfer energies is intrinsically taken care of without an analytical form.

The major objectives of this work are (i) to build a flexible water model for water clusters combining the one-, two-, and three-body interaction energy contributions, (ii) to run a simulation to generate the configuration space of water cluster using the SVR-based force field where the energy will be predicted by our flexible water model, and (iii) to compare various features of this simulated configuration space with those of AIMD and classical molecular dynamics.

### THEORETICAL DETAILS

#### Many-Body Expansion

The total energy of a system with \( n \) interacting molecules is stated by using MBE as

\[
E_{\text{total}}^{n\text{-mer}} = \sum_{i}^{n} E_{\text{mono}}^{i} + \sum_{i > j}^{n} \Delta E_{\text{di}}^{ij} + \sum_{i > j > k}^{n} \Delta E_{\text{tri}}^{ijk}
\]

\[
+ \sum_{i > j > k > l} \Delta E_{\text{tetra}}^{ijkl} + \cdots + \Delta E_{\text{mer}}^{n\text{-mer}}
\]

(1)

where \( \Delta E^{\text{ii} \text{iii}} \) and \( \Delta E^{\text{Eii} \text{iii}} \) are the two- and three-body energy components toward the \( n \)-body energy, respectively. \( E_{\text{mono}}^{i} \) is the energy of the monomers, which can be formulated as

\[
E_{\text{mono}}^{i} = \frac{1}{2} E_{\text{mono}}^{\text{opt}} + \Delta E_{\text{mono}}^{i}
\]

(2)

where \( E_{\text{mono}}^{\text{opt}} \) is the energy of the global minima of the monomer molecule and \( \Delta E_{\text{mono}}^{i} \) is the one-body energy correction arising due to the intramolecular (vibrational) motions.

The two- and three-body energy contributions are defined as

\[
\Delta E_{\text{di}}^{ij} = E_{\text{ij}} - E_{i} - E_{j}
\]

(3)

\[
\Delta E_{\text{tri}}^{ijk} = E_{\text{ijk}} - E_{i} - E_{j} - E_{k} - \Delta E_{\text{ij}} - \Delta E_{\text{jk}} - \Delta E_{\text{ki}}
\]

(4)

From eqs 1 and 2, the interaction energy of an \( n \)-mer system can be written as

\[
E_{\text{int}}^{n\text{-mer}} = \sum_{i}^{n} \Delta E_{\text{mono}}^{i} + \sum_{i > j}^{n} \Delta E_{\text{di}}^{ij} + \sum_{i > j > k}^{n} \Delta E_{\text{tri}}^{ijk} + \cdots + \Delta E_{\text{mer}}^{n\text{-mer}}
\]

(5)

In general, the QM methods have a steep computational scaling. Thus, electronic structure calculations for large systems become prohibitively expensive after a certain system size. MBE provides a path to estimate interaction energies of larger clusters using the lower body interaction terms. In many cases, the significant contributors toward the total \( n \)-body interaction are often found to be two- and three-body interaction contributions along with the one-body energy corrections. In the case of water, it has been reported that the impact of neglecting four-body onward interactions results in a minimal error. 52,62 This fast convergence of the MBE of interaction energy using only the lower body interaction terms is leveraged in our model.

#### Supervised Machine Learning

Supervised machine learning techniques have emerged as a class of machine
learning methods where the algorithms are trained with a set of known input–output and used for predicting the unknown output. The objective of supervised learning methods is to find a mapping function between input and output from the unbiased and impartially represented dataset. SVR is an advanced regression-based supervised machine learning method that has been implemented successfully in our previous work.

Support vector-based methods were originally built as a classifier (support vector machine, SVM) and later reshaped for regression applications. SVR uses the kernel expansion of the feature space along with a dual formulation of regression. The details of SVR are provided in the literature of Smola and Schölkopf64 and Drucker et al.65

In the case of linear regression, a function \( f(x) \) is represented in the following equation

\[
f(x) = \langle w; x \rangle + b \tag{6}
\]

This is optimized in a way so that one can achieve maximum flatness of \( w \). \( \langle \cdot \rangle \) is the notation for the inner product, whereas \( b \) is the \( y \) intercept. The problem may be re-written as

\[
\text{minimize } \|w\|^2
\]

subject to

\[
y_i - \langle w; x_i \rangle - b \leq \epsilon \tag{7}
\]

where \((x_i, y_i)\) is the set of data that will be fitted. Here, \( \epsilon \) is the threshold of acceptable errors. However, while solving this problem we implicitly assume the existence of the function \( f(x) \) that will satisfyeq 6. This condition is not valid for all \((x_i, y_i)\). Therefore, one needs to introduce slack parameters, \( \xi^i \), which provides some amount of flexibility to solve the problem. One can then use Lagrange multipliers to solve this constrained minimization problem. The Lagrange function including the slack parameters can be expressed as

\[
L = \frac{1}{2} \|w\|^2 + C \sum_i (\xi_i^+ + \xi_i^-) - \sum_i (\eta_i^+ + \eta_i^- \xi_i^-) \\
- \sum_i \alpha_i (\epsilon + \xi_i^+ - \chi - \langle w; x_i \rangle + b) \\
- \sum_i \alpha_i^- (\epsilon + \xi_i^- + \chi - \langle w; x_i \rangle - b) \tag{8}
\]

where \( \eta, \eta^+, \alpha, \alpha^- \) are Lagrange multipliers. The derivatives of \( L \) with respect to \( w, b, \xi_i^+, \xi_i^- \) will be used to estimate the Lagrange multipliers.

\( w \) and \( f(x) \) in eq 6 can be reformed as

\[
w = \sum_i (a_i - a_i^-)x_i
\]

\[
f(x) = \sum_i (a_i - a_i^+)\langle x_i; x \rangle + b \tag{9}
\]

Hence, the problem is reduced to finding the flattest function \( f(x) \), not in the input space, rather in the feature space.

The \( \epsilon \)-SVR is inconvenient when one is not aware of the error bar of the prediction. A variant of this scheme, \( \nu \)-SVR is often used in such cases. The parameter \( \nu \) is given by \( 0 \leq \nu \leq 1 \). It can be seen as an upper limit of the margin errors and a lower limit of the support vectors.66 Here, we have used the \( \nu \)-SVR algorithm to train the one-body energy dataset.

**Input Descriptor.** Training of the one-, two-, and three-body energy terms is the key for predicting interaction energies of large clusters using the combination SVR and MBE. The input descriptors are a set of structural information (interatomic distance reciprocals). The choice of \( 1/r \), rather than \( r \), is justified as the interaction energies are polynomial functions of \( 1/r \). These interatomic distance reciprocals are reported to be a suitable function for training the ML algorithms to predict energy.67 Furthermore, it is important to incorporate the spatial symmetry of water, i.e., the two H-atoms of one water are symmetric and when they are interchanged, the interaction energy remains the same. Thus, the spatial symmetry must be integrated in the input descriptor set. Behler and Parrinello used symmetry functions to include the effects of spatial symmetry.46 In our work, the spatial symmetry is accounted for in the pre-processing step by sorting the interatomic distance reciprocals separately. It should be noted here that the reciprocal mass-weighted distances were also tested as input descriptors but they do not improve the accuracy of the prediction (reported in the Supporting Information). Hence, we have worked with reciprocal distances set as inputs in our model.

Here, in Figure 1, we represent the input descriptor sets for one-body energy and two-body energy. In the case of a one-body energy dataset, there are reciprocals of two interatomic O–H distances and one H–H distance. The two O–H distance reciprocals are sorted in a descending order. For two- and three-body energies, we include the reciprocal distances that deal with the intermolecular degrees of freedom since the intramolecular effects are already taken care of in the one-body energy. For the two-body training set, there are nine such distances as shown in Figure 1. It consists of reciprocals of one O–O distance, four O–H distances, and four H–H distances.

The reciprocals of O–H and H–H distances are sorted separately. Similarly, for the three-body energy, there exists a total of 27 distance reciprocals. Among them, 3 are for O–O

![Figure 1. Input descriptors of water in dimers and monomers.](https://dx.doi.org/10.1021/acsomega.9b02968)
distance reciprocals, 12 are for O–H distance reciprocals, and 12 are for H–H distance reciprocals. These sets are then sorted separately to form the input descriptors.

It should be noted here that we have utilized the dataset of two- and three-body energies as well as the optimized SVR parameters from previous work.60 Henceforth, in this article, the one, two- and three-body energy datasets are referred to as DS1b, DS2b, and DS3b respectively. The details of the training sets and test sets are discussed in the following section.

■ COMPUTATIONAL DETAILS

Ab Initio Calculation and Interaction Energies. The configuration space of a flexible water dimer is generated from an ab initio NVT simulation at 300 K in the B97X-D/6-31G* level of theory using a Langevin thermostat. The time step of the simulation is 0.48 fs, and the simulation is run for a total of 150 ps. The first 50 ps is discarded, and equally spaced 2000 dimer snapshots are chosen from the final 100 ps. Each dimer provides 10 monomer configurations, thus accounting to a total of 20,000 monomers to build the one-body dataset (DS1b). Each set corresponds to one configuration. The one-body energy term is computed using a B97X-D/6-31+G* level of theory. The decamer interaction energies are estimated in the same level of theory with a basis set superposition error (BSSE) correction. The details of classical molecular dynamics simulation are provided in the Supporting Information.

Training and Testing: Cross Validation. The training and testing of DS1b and the SVR parameter optimization are carried out with kernlab and CVST packages of R.68,69 The number of randomly selected training data points are varied from 2000 to 8000. The training and test set are termed as TNb and TSb in the article. A fivefold cross validation across DS1b is carried out to benchmark the method. Cross validation is a convention for selecting the training and testing datasets in a way that the root mean square error (RMSE) in the output becomes invariant to the datasets. For an n-fold cross validation process, the dataset is divided into n subsets from which n − 1 subsets are taken for training and one for testing. Also, the training and testing datasets are rotated n-fold times to reduce the effect of subset selection in the error.

Monte Carlo Simulation. We have performed a machine learning-based Monte Carlo simulation (MLMC) with a water dimer cluster to generate the configuration space and compare the structural properties against ab initio and classical dynamics. Randomly chosen subsets of DS1b, DS2b, and DS3b are used as the training datasets for the SVR method. Here, the starting point is a random dimer configuration whose interaction energy is predicted by an MBE-SVR scheme. Previously optimized SVR parameters are used as input parameters for the simulation. The simulation is run at 300 K with 500,000 steps. Here, it should be noted that the training procedure of the machine is done only once in the whole MLMC simulation. It has been reported that the training is the computationally expensive part of the machine learning algorithm in comparison to prediction. Thus, by limiting the expensive step, one gains a significant advantage in computational wall time in comparison to ab initio methods. Each step of the MC simulation provides a random displacement toward all the atoms of the system and produces a new configuration. The interaction energy of the new configuration is estimated by the combined MBE-SVR method. Figure 2 is the schematic representation of the MLMC method with a detailed stepwise description. The accepted configurations are labeled as MLMC datasets. We have presented a detailed description of the MLMC simulation in the Supporting Information (Section S3).

RESULTS AND DISCUSSION

Optimization of SVR Parameters. We have used ν-SVR to fit the interaction energies of the flexible water molecules. The kernel used for SVR is a radial basis function (RBF)-based kernel, which was reported to be significantly more efficient than the polynomial function-based kernels.61 Therefore, the two parameters that need to be optimized for the SVR fitting are ν (which denotes the error insensitive loss function) and σ (which denotes the Gaussian width of the kernel). The two- and three-body datasets and the optimized parameters are reused from the rigid water model. The parameters, σ and ν, are optimized for DS1b.

The root mean square error (RMSE) in a one-body energy component of the test dataset (1000 data points) is plotted against seven different values of σ in Figure 3a. Three different values of ν are considered as well to optimize the kernel. It is observed that the RMSE is minimal in the range of 10 ≤ σ ≤ 100. The dependence of ν on the RMSE of one-body energy is minimal in the optimal range of sigma (10 ≤ σ ≤ 100). Therefore, to choose one particular combination of σ and ν, we compare the computational costs in the optimal range of σ and ν. From Figure 3b, it is evident that σ = 100 and ν = 0.1 is the optimized set of parameters for DS1b.

Energy of Flexible Water Molecules. An RBF-based SVR algorithm with σ = 100 and ν = 0.1 is used to train and test DS1b. Different combinations of training and testing data points are chosen maintaining the fivefold cross validation. The results are provided in Table 1. The RMSE in a one-body energy contribution is 0.12 kcal/mol when the complete DS1b is considered as the test set.

Table 1 indicates the effectiveness of SVR to estimate for a larger test set when the training set is comparatively small.

Figure 4 shows the distribution of the decamer interaction energies of 500 configurations obtained by using three different methods: (i) parent ab initio method (ωB97X-D/6-31+G*) in green, (ii) ML with only two- and three-body terms in blue (denoted as ML-MBE23), i.e., configurations without flexibility, and (iii) ML with one-body, two-body, and three-body terms in red (denoted as ML-MBE123), i.e., flexible configurations.

The ab initio interaction energies and the ML-trained interaction energies of the flexible water model are in excellent agreement (a most probable energy of 88 kcal/mol), while the
ML-trained non-flexible water interaction energies are quite different (a most probable energy of 76 kcal/mol). Therefore, the percentage errors of interaction energies for flexible and non-flexible ML models are 0.91 and 13.6%, respectively, as compared to ab initio interaction energies of flexible water molecules.

**Structural Analysis of Water Molecules.** We have performed an MLMC simulation of the water decamer cluster to understand the structural properties of the generated configuration space. The structural properties of the AIMD simulation produced configurations are considered to be the benchmark for the water decamer. We have compared the properties of MLMC and AIMD with those of a classical simulation of a water decamer by the TIP3P model as well.

**Radial Distribution Function.** The RDF \( g(r) \) gives the probability of finding a particle at a distance \( r \) from another particle. Figures 5 and 6 show the O–O and O–H RDFs with the three different sets of configurations: (i) AIMD, (ii) MLMC, and (iii) TIP3P classical MD. The first peak is the sharpest one and indicates the first solvation shell. We notice that the AIMD and MLMC peak maxima as well as the intensities are in excellent agreement with each other. However, the configurations of the classical TIP3P simulation have a significantly different structural form. In the case of the O–O RDF, the peak maxima are shifted toward the right in TIP3P indicating a more loosely bound cluster in TIP3P than in AIMD or MLMC. Interestingly, Kiss and Baranyai have reported a similar performance of classical MD when they compared their polarizable water model with classical water
The more compact configurations by AIMD as well as MLMC are a result of accurate treatment of all possible short-range interactions such as polarization, dispersion, charge transfer, and exchange-correlation between water molecules. Here, it is important to note that all the non-bonded interactions are organically inherited in the SVR-MBE model as the many-body energy terms are obtained from the ab initio method. The water decamer cluster, being a considerably small one, will have significant short-range non-bonded effects, which is not exactly incorporated in the TIP3P water model. It should be noted that the RDF does not converge to 1 like bulk water due to the finite size of the cluster. The intermolecular O–H RDFs have two peaks with AIMD and MLMC configurations showing similar nature, while TIP3P water clusters follow a slightly different trend. The two peaks in the intermolecular O–H RDF arise due to the presence of two different O–H pairs in a set of two water molecules.

Tetrahedral Order Parameter and Hydrogen Bond Numbers. The tetrahedral order parameter provides a description of the structure and orientation of molecules in liquids. It is a useful parameter to compare the configurations obtained from AIMD, MLMC, and classical MD. Errington and Debenedetti formulate the tetrahedral order parameter in the following way:\(^\text{(11)}\)

\[
Q_i = 1 - \frac{3}{8} \sum_{j=1}^{3} \sum_{k=j+1}^{4} \left( \cos \theta_{jk} + \frac{1}{3} \right)^2
\]

In the case of a perfect arrangement of water, the four nearest molecules around a central molecule orient themselves in such a way that \(Q = 1\). In the case of completely random and uniform distribution of solvents, \(Q = 0\).

Figure 7 shows the comparison of the tetrahedral order parameters calculated with AIMD, MLMC, and TIP3P generated configurations. The distributions corresponding to AIMD and MLMC configurations are similar to each other but significantly different from the distribution obtained from TIP3P generated configurations. The relatively higher values of \(Q\) in AIMD and MLMC configuration spaces indicate more structured water clusters in comparison to the TIP3P water model. This signature is confirmed from the distributions of H-bonds in AIMD, MLMC, and TIP3P generated configuration spaces, as shown in Figure 8. The intermolecular O–H–O distance and angle cutoff are considered as 3.2 Å and 150°, respectively.

Moreover, we have also compared the intramolecular O–H bond lengths, H–O–H angles, and the interaction energy distributions of water clusters in AIMD and MLMC generated configuration spaces and have found good agreement between them. The figures corresponding to these observations are provided in the Supporting Information.

Effect of Flexibility: Comparison with Flexible Water Models. The difference in the structural parameters obtained with different water models could have two major underlying causes: polarizability and flexibility of water. To estimate the relative effects of these two factors, we also compare the RDFs, tetrahedral order parameters, and number of hydrogen bonds with the flexible TIP3P model. From Figure 9, we notice that the introduction of flexibility in the TIP3P model corrects the peak position of the O–O RDF with respect to the AIMD. However, the intensity of the first peak of the flexible TIP3P model is significantly different from AIMD.

Figure 10 shows the comparison of the distributions of tetrahedral order parameters obtained from AIMD, MLMC, TIP3P, and flexible-TIP3P model generated configurations. Surprisingly, the most intense peak of the flexible TIP3P distribution has a shift in the opposite direction with respect to AIMD. However, on a closer inspection, one can see that (i) there is a single peak in the flexible TIP3P as opposed to two peaks in AIMD and (ii) the peak position of the flexible TIP3P water is in the middle of the two peaks obtained in AIMD. Comparison of the number of H-bonds, as shown in Figure 11, indicates that flexibility has minimal effects in ameliorating the errors in the TIP3P model when compared against AIMD.

From the observations on all three structural parameters, we can conclude that inclusion of all the major components of non-bonded interactions, such as electrostatics, polarization, and dispersion, from the ab initio energies enables us to estimate accurate structural information.

CONCLUSIONS

In this work, we have built a scheme for a Monte Carlo simulation of flexible water clusters using the combination of \(\nu\)-SVR and MBE. To that effect, we have formulated and validated a \(\nu\)-SVR-MBE scheme for flexible water molecules. In comparison to the parent QM estimates, we see that the flexible ML-MBE model has an average error of 0.91% in the...
prediction of interaction energy of water decamers. In the case of structural properties such as O−O and O−H RDFs or tetrahedral order parameters, we see that ab initio and MLMC generated conformations are in excellent agreement, whereas both flexible and non-flexible empirical classical water models have significant deviations. Our results clearly indicate a compact and rigid configuration space in MLMC and AIMD as compared to classical TIP3P dynamics. The average number of H-bonds per conformation are 14 and 13 in ab initio and MLMC, respectively, whereas the flexible and non-flexible variants of TIP3P configurations have 10 and 8 H-bonds, respectively. This is in good agreement with the reports of Kiss and Baranyai37 where they have suggested that incorporating the effect of polarization reduces the O−O distances and makes the water cluster compact. Since the many-body energies of our SVR-MBE scheme are estimated by the ab initio method, our model consists of major components of the non-bonded interactions such as electrostatics, polarization, dispersion, etc. It can be concluded that this ML-based water model is capable of reproducing the interaction energies and structural properties of water clusters with a comparable level of accuracy against its parent QM method. This is a stepping stone for building water models and force fields for molecular simulation in larger length scales, and we are working on that. It should be recognized that the major drawback of eliminating the analytical form of the potential is the difficulty to obtain the gradients for MD simulation. We are working on that aspect to ensure a condensed-phase simulation of water with our model in the future via either MC or MD. Also, in the future, we will be trying to generate an SVR-MBE-based model for ion–water interactions as well.

**ASSOCIATED CONTENT**

**Supporting Information**
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.9b02968.

Details of classical MD simulation, training and testing datasets, MLMC simulation, the comparison of the interaction energies, intramolecular bond lengths, bond angle from ab initio and MLMC generated configurations, the comparison of RMSE in the estimation of two body energies by mass-weighted reciprocal distances and reciprocal distances, comparison of RDFs and number of H-bonds with the MB-Pol MD, MLMC, and AIMD generated configuration spaces, and a section of the two-body energy dataset (PDF)

Dimer dataset at 300 K (ZIP)

**AUTHOR INFORMATION**

**Corresponding Authors**

Suman Chakrabarty — Department of Chemical, Biological & Macromolecular Sciences, S. N. Bose National Centre for Basic Sciences, Kolkata 700106, West Bengal, India; orcid.org/0000-0002-9461-0015; Email: sumanc@bose.res.in

Debashree Ghosh — School of Chemical Sciences, Indian Association for the Cultivation of Science, Kolkata 700032, West Bengal, India; orcid.org/0000-0003-0726-7878; Phone: +91-033-24734971 (Ext 1103); Email: pcdg@iacs.res.in

**Author**

Samik Bose — School of Chemical Sciences, Indian Association for the Cultivation of Science, Kolkata 700032, West Bengal, India; orcid.org/0000-0002-0273-9162

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.9b02968

**Notes**
The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**
The authors thank the Indian Association for the Cultivation of Science (IACS) for providing computational facilities. S.B.
acknowledges IACS for providing fellowship. D.G. thanks SERB, India, for project no. ECR/2018/002903.

■ REFERENCES

(1) De Beer, S.; Vermeulen, N. P.; Oostenbrink, C. The role of water molecules in computational drug design. Curr. Top. Med. Chem. 2010, 10, 55−66.
(2) Feig, M.; Brooks, C. L. III Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr. Opin. Struct. Biol. 2004, 14, 217−224.
(3) Sagui, C.; Darden, T. A. Molecular dynamics simulations of biomolecules: long-range electrostatic effects. Annu. Rev. Biophys. Biomol. Struct. 1999, 28, 155−179.
(4) Canuto, S. Solvation effects on molecules and biomolecules: computational methods and applications; Springer Science & Business Media: 2010, Vol. 6.
(5) Kumawat, A.; Chakrabarty, S. Hidden electrostatic basis of dynamic alloster in a PDZ domain. Proc. Natl. Acad. Sci. 2017, 114, ES825−ES834.
(6) Bose, S.; Chakrabarty, S.; Ghosh, D. Electrostatic Origin of the Red Solvatochromic Shift of DFHBDI in RNA Spinach. J. Phys. Chem. B 2017, 121, 4790−4798.
(7) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Intermolecular forces; Springer: 1981; pp 331−342, DOI: 10.1007/978-94-015-7658-1_21.
(8) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys. 1983, 79, 926.
(9) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J. Am. Chem. Soc. 1988, 110, 1657−1666.
(10) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. J. Comput. Chem. 2004, 25, 1157−1174.
(11) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, L.; et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J. Comput. Chem. 2010, 31, 671−690.
(12) Warschel, A.; Kato, M.; Pislakow, A. V. Polarizable force fields: history, test cases, and prospects. J. Chem. Theory Comput. 2007, 3, 2034−2045.
(13) Halgren, T. A.; Damm, W. Polarizable force fields. Curr. Opin. Struct. Biol. 2001, 11, 236−242.
(14) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. Second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 1995, 117, 5179−5197.
(15) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; et al. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. J. Comput. Chem. 2003, 24, 1999.
(16) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. The effective fragment potential method: A QM-based MM approach to modeling environmental effects in chemistry. J. Phys. Chem. A 2001, 105, 293−307.
(17) Gordon, M. S.; Slupchenko, L.; Li, H.; Jensen, J. H. Chapter 10 The effective fragment potential: a general method for predicting intermolecular interactions. Annu. Rev. Comput. Chem. 2007, 3, 177−193.
(18) Matsumoto, M.; Saito, S.; Ohmine, I. Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing. Nature 2002, 416, 409.
(19) Sprik, M.; Hutter, J.; Parrinello, M. Ab initio molecular dynamics simulation of liquid water: Comparison of three gradient-corrected density functionals. J. Chem. Phys. 1996, 105, 1142−1152.
(20) Cisneros, G. A.; Wikfeldt, K. T.; Ojämäe, L.; Lu, J.; Xu, Y.; Torabifard, H.; Bäröck, A. P.; Csányi, G.; Molinero, V.; Paesani, F. Modeling molecular interactions in water: From pairwise to many-body potential energy functions. Chem. Rev. 2016, 116, 7501−7528.
(21) Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J. Chem. Phys. 2000, 112, 8910−8922.
(22) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J. Chem. Phys. 2004, 120, 9665−9678.
(23) Wu, Y.; Tepper, H. L.; Voth, G. A. Flexible simple point-charge water model with improved liquid-state properties. J. Chem. Phys. 2006, 124, No. 024503.
(24) Rahman, A.; Stillinger, F. H. Molecular dynamics study of liquid water. J. Chem. Phys. 1971, 55, 3336−3359.
(25) Caldwell, J.; Dang, L. X.; Kollman, P. A. Implementation of nonadditive intermolecular potentials by use of molecular dynamics: development of a water-water potential and water-ion cluster interactions. J. Am. Chem. Soc. 1990, 112, 9144−9147.
(26) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. J. Chem. Phys. B 2003, 107, 5933−5947.
(27) Wang, L.-F.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. Systematic improvement of a classical molecular model of water. J. Phys. Chem. B 2011, 117, 9956−9972.
(28) Qi, R.; Wang, L.-P.; Wang, Q.; Pande, V. S.; Ren, P. United polarizable multipole water model for molecular mechanics simulation. J. Chem. Phys. 2015, 143, No. 014504.
(29) Parcaud, P.; Predota, M.; Chialvo, A. A.; Cummings, P. T. From dimer to condensed phases at extreme conditions: Accurate predictions of the properties of water by a Gaussian charge polarizable model. J. Chem. Phys. 2005, 122, 244511.
(30) Yu, W.; Lopes, P. E. M.; Roux, B.; MacKerell, A. D., Jr. Six-site polarizable model of water based on the classical Drude oscillator. J. Chem. Phys. 2013, 138, No. 034508.
(31) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. A polarizable model of water for molecular dynamics simulations of biomolecules. Chem. Phys. Lett. 2006, 418, 245−249.
(32) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. Combined fluctuating charge and polarizable dipole models: Application to a five-site water potential function. J. Chem. Phys. 2001, 115, 2237−2251.
(33) Wang, Y.; Huang, X.; 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, No. 094509.
(34) Wang, Y.; Bowman, J. M.; Kamarchik, E. Five ab initio potential energy and dipole moment surfaces for hydrated NaCl and NaF. I. Two-body interactions. J. Chem. Phys. 2016, 144, 114311.
(35) Kamarchik, E.; Toffoli, D.; Christiansen, O.; Bowman, J. M. Ab initio potential energy and dipole moment surfaces of the F(H2O) complex. Spectrochim. Acta, Part A 2014, 119, 59−62.
(36) Kamarchik, E.; Bowman, J. M. Coupling of low-and high-frequency vibrational modes: Broadening in the infrared spectrum of F(H2O); J. Phys. Chem. Lett. 2013, 4, 2964−2969.
(37) Kiss, P. T.; Baranyai, A. A systematic development of a polarizable potential of water. J. Chem. Phys. 2013, 138, 204507.
(38) Kiss, P. T.; Baranyai, A. Anomalous properties of water predicted by the B3K model. J. Chem. Phys. 2014, 140, 154505.
(39) Hunger, J.; Huttner, G. Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. J. Comput. Chem. 1999, 20, 455−471.
(40) Cho, K.-H.; No, K. T.; Scheraga, H. A. A polarizable force field for water using an artificial neural network. J. Mol. Struct. 2002, 641, 77–91.
(41) Darley, M. G.; Handley, C. M.; Popelier, P. L. A. Beyond point charges: dynamic polarization from neural net predicted multipole moments. J. Chem. Theory Comput. 2008, 4, 1435–1448.
(42) Li, Y.; Li, H.; Pickard, F. C.; IV, Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. J. Chem. Theory Comput. 2017, 13, 4492–4503.
(43) Handley, C. M.; Hawe, G. I.; Kell, D. B.; Popelier, P. L. A. Optimal construction of a fast and accurate polarizable water potential based on multipole moments trained by machine learning. Phys. Chem. Chem. Phys. 2009, 11, 6365–6376.
(44) Balabin, R. M.; Lamakina, E. I. Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. J. Chem. Phys. 2009, 131, No. 074104.
(45) Handley, C. M.; Popelier, P. L. A. Potential energy surfaces fitted by artificial neural networks. J. Phys. Chem. A 2010, 114, 3371–3383.
(46) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett. 2007, 98, 146401.
(47) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. Phys. Chem. Chem. Phys. 2011, 13, 17930–17955.
(48) Gatsegger, M.; Marquetand, P. High-dimensional neural network potentials for organic reactions and an improved training algorithm. J. Chem. Theory Comput. 2015, 11, 2187–2198.
(49) Morawietz, T.; Behler, J. A density-functional theory-based neural network potential for water clusters including van der waals corrections. J. Phys. Chem. A 2013, 117, 7356–7366.
(50) Yao, K.; Parkhill, J. Kinetic energy of hydrocarbons as a function of electron density and convolutional neural networks. J. Chem. Theory Comput. 2016, 12, 1139–1147.
(51) Manzhos, S.; Carrington, T., Jr. Using neural networks to represent potential surfaces as sums of products. J. Chem. Phys. 2006, 125, 194105.
(52) Janet, J. P.; Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. Chem. Sci. 2017, 8, 5137–5152.
(53) Babin, V.; Leforestier, C.; Paesani, F. Development of a first principles water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient. J. Chem. Theory Comput. 2013, 9, 5395–5403.
(54) Babin, V.; Medders, G. R.; Paesani, F. Development of a first principles water potential with flexible monomers. II: Trimer potential energy surface, third virial coefficient, and small clusters. J. Chem. Theory Comput. 2014, 10, 1599–1607.
(55) Medders, G. R.; Babin, V.; Paesani, F. Development of a first-principles water potential with flexible monomers. III. Liquid phase properties. J. Chem. Theory Comput. 2014, 10, 2906–2910.
(56) Fracchia, F.; Del Frate, G.; Mancini, G.; Rocchia, W.; Barone, V. Force Field Parametrization of Metal Ions from Statistical Learning Techniques. J. Chem. Theory Comput. 2018, 255.
(57) Reddy, S. K.; Straight, S. C.; Bajaj, P.; Huy Pham, C.; Riera, M.; Moberg, D. R.; Morales, M. A.; Knight, C.; Götz, A. W.; Paesani, F. On the accuracy of the MBpol many-body potential for water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice. J. Chem. Phys. 2016, 145, 194504.
(58) Riera, M.; Mardrossian, N.; Bajaj, P.; Götz, A. W.; Paesani, F. Toward chemical accuracy in the description of ion–water interactions through many-body representations. Alkali-water dimer potential energy surfaces. J. Chem. Phys. 2017, 147, 161715.
(59) Arismendi-Arrieta, D. J.; Riera, M.; Bajaj, P.; Prosmiti, R.; Paesani, F. i-TTM model for ab initio-based ion–water interaction potentials. I. Halide–water potential energy functions. J. Phys. Chem. B 2016, 120, 1822–1832.