Active learning of many-body configuration space: Application to the Cs^+–water MB-nrg potential energy function as a case study

Yaoguang Zhai,[1, a)] Alessandro Caruso,[2, b)] Sicun Gao,[1, c)] and Francesco Paesani[2, 3, 4, d)]

[1)]*Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093, United States*

[2)]*Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States*

[3)]*Materials Science and Engineering, University of California San Diego, La Jolla, California 92093, United States*

[4)]*San Diego Supercomputer Center, University of California San Diego, La Jolla, California 92093, United States*

The efficient selection of representative configurations that are used in high-level electronic structure calculations needed for the development of many-body molecular models poses a challenge to current data-driven approaches to molecular simulations. Here, we introduce an active learning (AL) framework for generating training sets corresponding to individual many-body contributions to the energy of a N-body system, which are required for the development of MB-nrg potential energy functions (PEFs). Our AL framework is based on uncertainty and error estimation, and uses Gaussian process regression (GPR) to identify the most relevant configurations that are needed for an accurate representation of the energy landscape of the molecular system under exam. Taking the Cs+–water system as a case study, we demonstrate that the application of our AL framework results in significantly smaller training sets than previously used in the development of the original MB-nrg PEF, without loss of accuracy. Considering the computational cost associated with high-level electronic structure calculations for training set configurations, our AL framework is particularly well-suited to the development of many-body PEFs, with chemical and spectroscopic accuracy, for molecular simulations from the gas to condensed phase.

[a)]Electronic mail: yazhai@ucsd.edu; These authors contributed equally

[b)]Electronic mail: acaruso@ucsd.edu; These authors contributed equally

[c)]Electronic mail: sig049@ucsd.edu

[d)]Electronic mail: fpaesani@ucsd.edu

I. INTRODUCTION

Computer simulations provide fundamental insights into the properties and behavior of molecular systems.[1–3] Since both accuracy and predictive ability of a molecular model are primarily limited by the computational cost associated with the model itself, developing cost-effective simulation approaches is key to studying increasingly more complex systems. It has recently become possible to perform molecular dynamics (MD) simulations of aqueous systems, from the gas to the condensed phase, retaining high accuracy in the description of the underlying molecular interactions.[4] This is achieved by employing many-body potential energy functions (PEFs) derived from high-level electronic structure data that are carried out on selected molecular configurations representative of the corresponding global many-body potential energy surfaces (PESs).[5–13] An optimal approach to the development of many-body PEFs would require identifying a minimal pool of configurations that can guarantee an accurate description of the system under exam and, at the same time, computation time is not lost on calculations on redundant configurations describing similar regions of the many-body PES.

Efficient sampling of the configuration space is challenging due to the high dimensionality of the associated molecular configurations. In principle, a regular grid search would provide a homogeneous representation of all regions of the many-body PES. This approach, however, becomes unfeasible as the number of degrees of freedom increases. To reduce the size of the configuration space, it is common practice in the development of many-body PEFs to apply biases on the relative translations and rotations of the individual molecular species constituting the system under exam.[9,10,12,13] Although of practical use, this approach can lead to redundant training sets containing several molecular configurations representing similar regions of the target many-body PES. While algorithms designed to remove geometrically similar configurations exist, it is not guaranteed that screening based on structural similarity is sufficient for identifying only configurations necessary for a faithful description of the target many-body PES.

The success of machine learning (ML) in many areas of molecular sciences (e.g., see Refs. 14–29) makes it a promising tool for efficiently screening large pools of molecular configurations for the development of many-body PEFs. Most common ML approaches rely on supervised learning, which, however, requires large set of known labeled data to train a model capable to accurately predict the labels of previously unseen data.[30–32] Active learning (AL) provides a potential solution to the need for constructing beforehand large training sets by interactively generating training

configurations at runtime. AL schemes are thus particularly appealing when using large training sets is prohibitively expensive either because of the high cost associated with determining the data labels or because of the high computational cost of the training stage.

In this study, we investigate the application of AL to generating representative training sets of molecular configurations necessary for the development of many-body PEFs, with a specific focus on two-body (2B) and three-body (3B) contributions to the Cs^+–water interaction energies. Our AL framework consists of a finite pool of molecular configurations (i.e., $Cs^+(H_2O)$ dimers for the 2B pool and $Cs^+(H_2O)_2$ trimers for the 3B pool) whose energies are unknown, a training set with configurations selected from the pool, a predictive model (predictor) thirsting for the training set, and a learner that actively selects configurations from the pool. We assume that the size of the pool is beyond awareness of the learner and only a subset of the configurations (referred to as candidates) in the pool are available to the learner at each iteration. Through the application of our AL approach, we demonstrate that the size of the original pool of configurations used to develop the Cs^+–water MB-nrg PEF can be greatly reduced without compromising the accuracy with which the new MB-nrg PEFs describe Cs^+–water interactions, from small clusters to aqueous solutions.

II. METHODS

A. MB-nrg potential energy functions

The total energy of a system containing N (atomic or molecular) monomers ("bodies"), can be rigorously expressed through the many-body expansion (MBE) of the energy,[33]

$$V_N = \sum_{i}^{N} V_i^{1B} + \sum_{i<j}^{N} V_{ij}^{2B} + \sum_{i<j<k}^{N} V_{ijk}^{3B} + \ldots + V^{NB} \tag{1}$$

where the V_i^{1B} corresponds to the energy required to distort monomer i from its equilibrium geometry. Therefore, $V^{1B}(i) = 0$ for atomic monomers, and $V^{1B}(i) = E(i) - E_{eq}(i)$ for molecular monomers, where $E(i)$ and $E_{eq}(i)$ are the energies of monomer i in distorted and equilibrium geometries, respectively. All higher n-body (nB) interaction terms (V^{nB}) in Eq. II A are defined recursively through

$$V^{nB}(1,\ldots,n) = E_n(1,\ldots,n) - \sum_i V^{1B}(i) - \sum_{i<j} V^{2B}(i,j) - \ldots - \sum_{i<j<\cdots<n-1} V^{(n-1)B}(i,j,\ldots,(n-1)) \quad (2)$$

Within the MB-nrg framework, the water–water interactions are described by the MB-pol PEF,[9–11] which has been shown to correctly reproduce the properties of water[34,35] from small clusters in the gas phase,[36–48] to bulk water,[49–52] the air/water interface,[53–56] and ice.[57–59] The interactions between Cs^+ ions and water molecules are described through the MBE of Eq. II A. Specifically, the Cs^+–water MB-nrg PEF includes explicit 2B Cs^+–H_2O and 3B Cs^+–$(H_2O)_2$ terms, with all higher-order interactions being implicitly taken into account through a classical many-body polarization term.[13,60] The 2B term includes three contributions,

$$V^{2B} = V^{2B}_{short} + V^{2B}_{TTM} + V^{2B}_{disp} \quad (3)$$

where V^{2B}_{disp} is the 2B dispersion energy, and V^{2B}_{TTM} is the 2B classical polarization contribution described by a Thole-type model.[61] V^{2B}_{short} in Eq. 3 describes 2B short-range contributions represented by a 5th-degree permutationally invariant polynomial (PIP) in variables that are functions of the distances between the Cs^+ ion and each of the six sites of the MB-pol water molecule.[13]

Similarly, the 3B term of the Cs^+–water MB-nrg PEF includes two contributions,

$$V^{3B} = V^{3B}_{short} + V^{3B}_{TTM} \quad (4)$$

where V^{3B}_{TTM} is the 3B classical polarization contribution described by the same Thole-type model as in V^{2B}_{TTM}, and V^{3B}_{short} describes 3B short-range contributions that are represented by a 4th-degree PIP in variables that are functions of the same distances as in V^{2B}_{short}.[60] The coefficients of both 2B and 3B PIPs were optimized using Tikhonov regression (also known as ridge regression)[62] to reproduce reference interaction energies obtained from high-level electronic structure calculations.

B. Interaction energies, fitting procedure, and MD simulations

The 2B and 3B reference energies were taken from Refs. 13 and 60 where MOLPRO (version 2015.1) was used to carry out electronic structure calculations at the coupled cluster level of theory using single, double and perturbative triple excitations, i.e., CCSD(T), the “gold standard" for chemical accuracy.[63] In Ref. 13, the 2B CCSD(T) energies were calculated in the complete

basis set (CBS) limit that was achieved through a two-point extrapolation[64,65] between the values obtained with the correlation-consistent polarized valence triple zeta (aug-cc-pVTZ for H,O, and cc-pwCVTZ for Cs^+) and quadruple zeta (aug-cc-pVQZ for H,O, and cc-pwCVQZ for Cs^+) basis sets.[66–69] In Ref. 60, the 3B CCSD(T) energies were calculated using the aug-cc-pVTZ basis set for the O and H atoms, and the cc-pwCVTZ basis set for Cs^+, and were corrected for the basis set superposition error using the counterpoise method.[70] In both 2B and 3B energy calculations, the ECP46MDF pseudopotential was used for the core electrons of Cs^+.[71]

The original 2B training set consisted of $Cs^+(H_2O)$ dimer configurations generated on a uniform spherical grid, with the Cs^+–O distance in the 1.6 - 8 Å range.[13] For the present study, dimer configurations with interaction energies larger than 100 kcal/mol were removed since they were found to be not necessary for representing $Cs^+(H_2O)$ configurations sampled in MD simulations at ambient conditions. The 2B pool was then further reduced to 13525 dimer configurations after randomly removing 1547 configurations for the 2B test set.

Due to the larger number of degrees of freedom, the original 3B training set was generated in Ref. 60 by extracting $Cs^+(H_2O)_2$ trimer configurations from MD simulations of a single Cs^+ ion in liquid water at 298.15 K. For the present study, the original 3B set of Ref. 60 was reduced to a 3B pool of 34441 configurations after randomly removing 4480 configurations for the 3B test set.

The MD simulations presented in Section III D were carried out in the isobaric-isothermal (NPT) ensemble for a box containing a single Cs^+ ion and 277 H_2O molecules. The equations of motion were propagated using the velocity-Verlet algorithm with a timestep δt of 0.2 fs. The temperature of 298.15 K was controlled by Nosé-Hoover chains of 4 thermostats attached to each degree of freedom while the pressure of 1.0 atm was controlled following the algorithm described in Ref. 72. All MD simulations were carried out using an in-house software based on DL_POLY 2.0.[73]

C. Active learning

An AL framework based on uncertainty and error estimation was used to generate optimal 2B and 3B training sets with the goal of reducing the number of dimers and trimers necessary to develop Cs^+–water MB-nrg PEF, without compromising accuracy. The major difficulty faced by the active learner in generating optimal 2B and 3B training sets is represented by the need to determine the relevance of candidate dimer and trimer configurations before knowing the associated

2B and 3B energies.[74] It is apparent that the more accurate the active learner is, the more precise its assessment of a molecular configuration is. In addition, for efficiency purposes, the energy estimation made by the learner should be computationally inexpensive compared to the energy determination performed by the predictor.

In this context, Gaussian process regression (GPR) provides a general approach to assessing the relevance of a candidate configuration by accurately estimating the associated energy.[75] GPR implies a correlation between the unknown energies of the candidate configurations and the energies determined for configurations that are already in the training sets. The correlation is expressed by the covariance matrix between known and unknown values of the energies, with the elements of the covariance matrix being calculated by a kernel function. GPR assumes that both known and unknown energies are distributed according to a multidimensional Gaussian distribution and then uses the covariance matrix to predict the conditional probability distribution of the unknown energies given the known energies. The ability of GPR to interpolate between known energy values makes it a good model for local uncertainty prediction. It should be noted that a similar approach is exploited by Gaussian Approximation Potential (GAP) models that have been developed to represent interatomic interactions.[76]

Our AL framework, shown in Fig. 1, consists of a pool of an unknown number of molecular configurations, corresponding to $Cs^+(H_2O)$ dimers for the 2B pool and $Cs^+(H_2O)_2$ trimers for the 3B pool, a predictor, and a learner that, based on feedback from the predictor, selects configura-

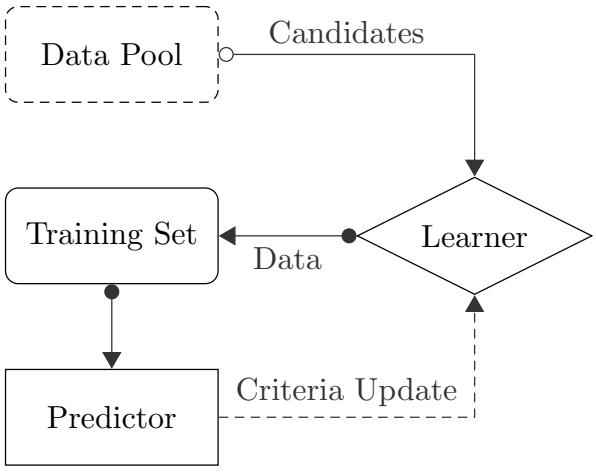

FIG. 1. Schematic representation of the AL framework introduced in this study.

tions from the pool and adds them to the training set. The complete AL protocol is summarized below:

- At each iteration t, the pool S sends a subset of configurations with unknown energies ($C_t = \{x_j\}_t \subseteq S$) to the learner as training set candidates.
- Depending on the iteration index t, a training set T_t is formed:
  - At $t = 0$, all configurations in C_0 are added to the training set T_0 and their actual energies are determined.
  - For $t > 0$:
    * The training set T_{t-1} from the previous iteration is divided into clusters $\{\tau_{t-1,k}\}$ containing a fixed number of molecular configurations, independent of the training set size.
    * A cluster label k_j is predicted for each candidate configuration x_j in C_t (i.e., each candidate configuration x_j is assigned to one of the clusters $\{\tau_{t-1,k}\}$).
    * The uncertainty ΔE_j on the energy of the candidate configuration x_j is estimated as the GPR variance calculated for the entire cluster τ_k, $k = k_j$.
    * The error Err_j on the energy of the candidate configuration x_j is defined as the average error associated with the energies predicted by the model for all the configurations in the cluster τ_k, $k = k_j$.
    * A selection probability $P_t(x_j)$, proportional to the weighted sum of the energy uncertainty and the energy error, is assigned to each candidate configuration x_j in C_t,
    $$P_t(x_j) \propto [w_{\Delta E} * \Delta E_j + w_{Err} * Err_j] \tag{5}$$
    * A subset of configurations $\{\hat{x}_i\}_t \subseteq C_t$ is selected and, after determining the associated actual energies ε_i, added to the training set, $T_t = \{(\hat{x}_i, \varepsilon_i)\}_t \cup T_{t-1}$.
- The model M is trained on the training set T_t.
- The errors associated with the energies predicted by the model for all configurations in the training set T_t are updated
- The cycle is stopped when the gradient of the test error becomes lower than a predefined value.

The division into clusters $\{\tau_{t-1,k}\}$ of equal size reduces the computational cost associated with GPR, which typically scales as $O(n^3)$.[75] Since a radial basis function (RBF) kernel, which is based on the L2 distance, is used to determine the similarity between two configurations, it follows that configurations close to the candidate configuration play a central role in the GPR process. The use of the RBF kernel function allows interpolation with GPR only between configurations that are in the same cluster as the candidate, which, in turn, helps reduce the computational cost without losing predictive accuracy. As shown in Eq. 5, the learner selects configurations based on the weighted sum of uncertainty and model error. This procedure ensures a balanced exploration of the configuration space, exploiting the decision-making process. At each iteration, a small subset (5%) of candidates is selected and added to the training set to improve the reliability of the learner.

It should be noted that a similar method based on Pearson correlation has recently been proposed for the generation of training sets for interatomic potentials.[77] This method exploits the correlation between atomic features to build a probability distribution that is then used to select new candidates but does not consider the feedback provided by the fitting model which allows our AL framework to adapt continually to changes in the training set after each iteration.

In this study, we used the KMeans module available in the Scikit-learn Python package, *version 0.21.3*, to cluster both the $Cs^+(H_2O)$ dimers and the $Cs^+(H_2O)_2$ trimers in the corresponding 2B and 3B training sets and the cluster size was fixed at 50 configurations. For GPR we used the class *GaussianProcessRegressor* and the *RBF* kernel available in the same Python package.

Both GPR and KMeans require a vector representation of the 2B and 3B structures in the high-dimensional configuration space. For this purpose, we used the many-body tensor representation (MBTR) of atomic environments.[78] MBTR defines a structural descriptor that is easily computable and well suited to calculate the kernels for both GPR and KMeans. The MBTR descriptor is constructed by storing the terms of the Coulomb matrix[15] associated with each pair of the N_e chemical elements constituting the molecular system of interest into an $N_e \times N_e \times d$ tensor, where d is the largest number of unique pairs of the same two chemical elements. The MBTR descriptor thus takes the form

$$f_k(x,\boldsymbol{z}) = \sum_{\boldsymbol{i}}^{N_a} w_k(\boldsymbol{i})\mathscr{D}(x, g_k(\boldsymbol{i})) \prod_{j=1}^{k} C_{z_j,Z_{i_j}}, \tag{6}$$

where $\boldsymbol{z} \in \mathbf{N}^k$ are atomic numbers, $\boldsymbol{i} = (i_1,\ldots,i_k) \in \{1,\ldots,N_a\}$ are index tuples, k runs over the number of atoms, $\mathscr{D}$ is a broadening function, C is the element correlation matrix, and g_k is a function that assigns a scalar to the k atoms based on a k-body physical feature. The MBTR

descriptor is then discretized and rearranged in the form of a vector.

We used the Python package *qmmlpack* for the vector representation of the 2B and 3B configurations in their respective training sets. The broadening function $\mathscr{D}$ was chosen to be the normal distribution with $k = 2,3$. The inverse of the distance, r^{-1}, and the angle, θ, were used as g_k for $k = 2,3$, respectively. The number of bins and the width of the normal distribution were tuned to guarantee the efficiency of the MBTR calculations, without compromising accuracy.

III. RESULTS

The results of our AL framework are presented in the following three subsections. First, we discuss the learning curves for the 2B and 3B energies, and comparisons are made between our AL framework and a generic approach based on a random selection (RS) of molecular configurations. Second, we introduce sketch-maps[79] of different 2B and 3B training sets generated through our AL framework and discuss the corresponding distributions of 2B and 3B energies. Third, we analyze the interaction and many-body energies of small water clusters as well as the Cs^+-oxygen radial distribution functions (RDFs) of liquid water calculated using different 2B and 3B training sets generated through our AL framework.

A. Learning curves of 2B and 3B energies

Figs. 2 and 3 show the learning curves for the 2B Cs^+–H_2O and 3B Cs^+–$(H_2O)_2$ energies, respectively, calculated for the training (left panels) and test (right panels) sets as a function of the training set size. Learning curves obtained using both our AL framework (blue) and RS approach (magenta) are shown.

The training root-mean-square errors (RMSEs) associated with the RS approach increase as a function of the training set size for both 2B and 3B energies while the corresponding AL curves display steeper increases for smaller training sets, reach a maximum, and then decrease. The test RMSEs show different trends, with the curves obtained with our AL framework displaying a significantly faster decrease as a function of the training set size. Since our AL framework specifically targets configurations with higher uncertainties and neighborhood training errors, these configurations are selected more frequently by the learner and added to the training set. It follows that the configurations that are left in the pool after each iteration are associated with progressively

smaller uncertainties and training errors. This implies that, when added to the training sets in subsequent iterations, these configurations necessarily lead to a decrease of the training RMSEs and only negligible variations in the test RMSEs as shown in in Figs. 2 and 3.

Based on the analysis of both training and test RMSEs obtained with our AL framework, the optimal numbers of configurations in the 2B and 3B training sets for the Cs^+–water MB-nrg PEFs were determined to be 5000 $Cs^+(H_2O)$ dimers and 10000 $Cs^+(H_2O)_2$ trimers, respectively.

B. Sketch-maps

Sketch-maps have been shown to be useful tools for representing high-dimensional configuration spaces with lower-dimensional projections that are easily interpretable in terms of well-defined structural features.[79–81]

To provide structural insights into the composition of the 2B and 3B training sets, with varying sizes, obtained with our AL framework, MBTR was used to generate the sketch-maps shown in Figs. 4 and 5, respectively. Panel a) of Fig. 4 is a representation of the entire 2B pool projected onto a 2-dimensional space. Each point on the map corresponds to a $Cs^+(H_2O)$ dimer configuration and the associated color indicates the corresponding CCSD(T) reference 2B energy. Since the 2B pool was generated on a grid by varying the Cs^+-O distance and distorting the water molecule, these features are reflected in the resulting sketch-map where points cluster together, in an orderly

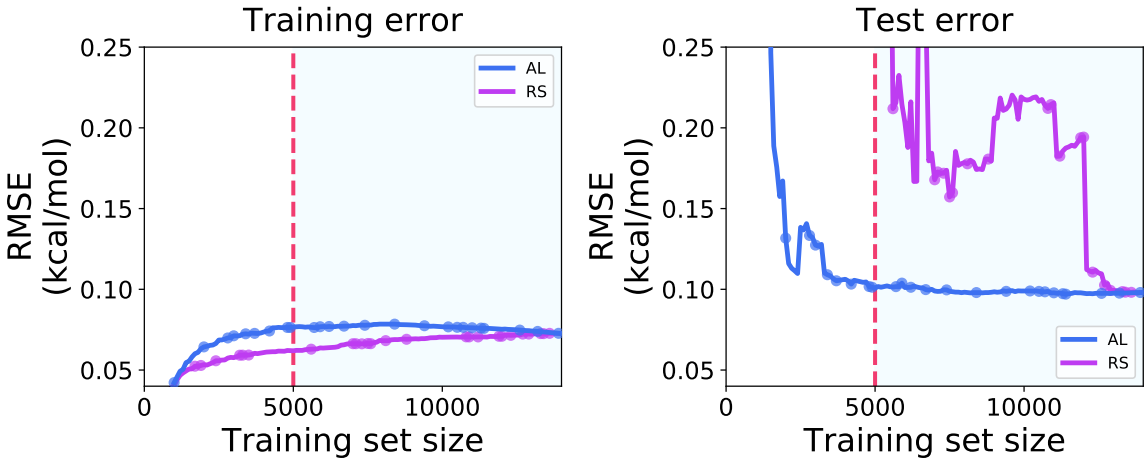

FIG. 2. RMSEs (in kcal/mol) associated with the 2B training (left) and test (right) sets displayed as a function of the training set size. Blue and magenta curves correspond to AL and RS learning curves, respectively. The dashed line indicates the optimal training set size as determined in this study.

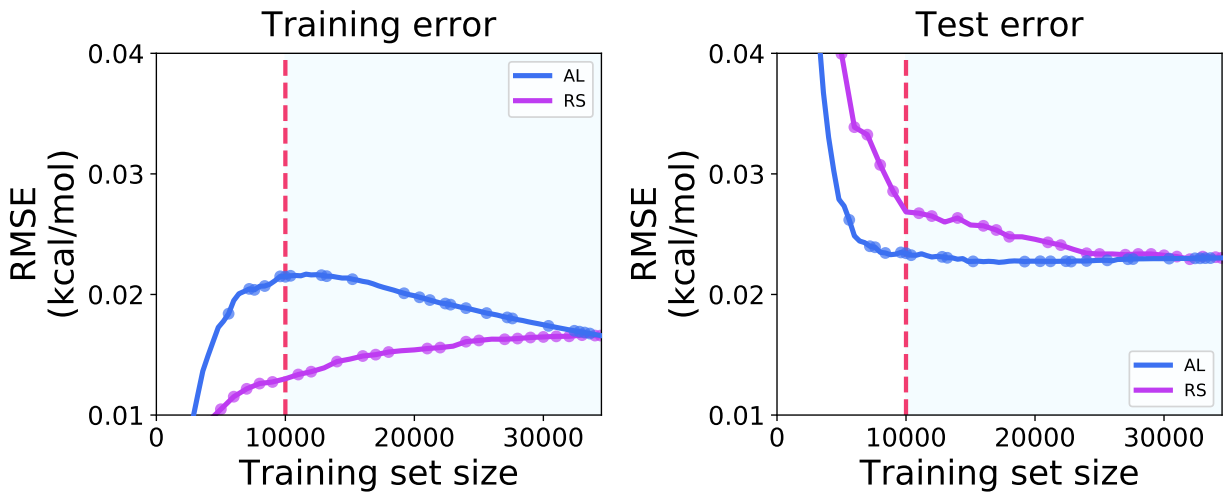

FIG. 3. RMSEs (in kcal/mol) associated with the 3B training (left) and test (right) sets displayed as a function of the training set size. Blue and magenta curves correspond to AL and RS learning curves, respectively. The dashed line indicates the optimal training set size as determined in this study.

fashion.

Panel b) of Fig. 4 shows a sketch-map of the energy differences between the reference 2B energies and the corresponding values predicted by the MB-nrg PEF trained on the full 2B pool (13525 configurations). This comparison shows that the the MB-nrg PEF provides an accurate description of the overall 2B energy landscape, with deviations larger than 0.5 kcal/mol only found for $Cs^+(H_2O)$ dimers with associated binding energies larger than 40 kcal/mol, and deviations on the order of 0.04 kcal/mol for $Cs^+(H_2O)$ dimer configurations with lower binding energies (less than 40 kcal/mol). It should be noted that dimer configurations with larger binding energies are unlikely to be visited in MD simulations at ambient conditions and are included in the 2B training sets to guarantee that the PIPs representing short-range interactions within the the MB-nrg PEF are well-behaved at short Cs^+–water distances. Panels c-f) show sketch-maps of the differences between 2B energies predicted by the MB-nrg PEF trained on the full 2B pool and the corresponding values predicted by MB-nrg PEFs trained on progressively smaller training sets containing 10000, 8000, 6000, 4000 configurations generated using our AL framework. As expected, systematically reducing the training set size introduces progressively larger errors, with training sets with fewer than 4000 dimer configurations leading to overfitting. This analysis shows that our AL framework allows for significantly reducing the original 2B Cs^+–H_2O training set without compromising the overall accuracy of the resulting MB-nrg PEF. In this context, it should be noted that the areas

of the sketch-maps in panels c-f) that display larger deviations from the original MB-nrg PEF of Ref. 13, as the training set size decreases, correspond to dimer configurations for which the original MB-nrg PEF also shows larger deviations from the CCSD(T) reference data (panel b).

Similar conclusions can be drawn from the analysis of the sketch-maps of the 3B training sets shown in Fig. 5. Since the original 3B pool was generated by extracting $Cs^+(H_2O)_2$ trimers from MD simulations of a single Cs^+ ion in liquid water, the resulting sketch-map (panel a) displays a more uniform distribution in the 2-dimensional space compared to the corresponding sketch-map obtained for the 2B pool. Depending on the associated CCSD(T) reference 3B energies, trimer configurations broadly cluster in two areas, with the "dividing surface" being between -5.0 and -3.0 kcal/mol. Also in this case, the original MB-nrg PEF closely reproduces the CCSD(T) reference 3B energies over the entire configuration space of the 3B pool. As for the 2B energies, progressively smaller training sets of 20000, 15000, 10000, 5000 configurations, generated using

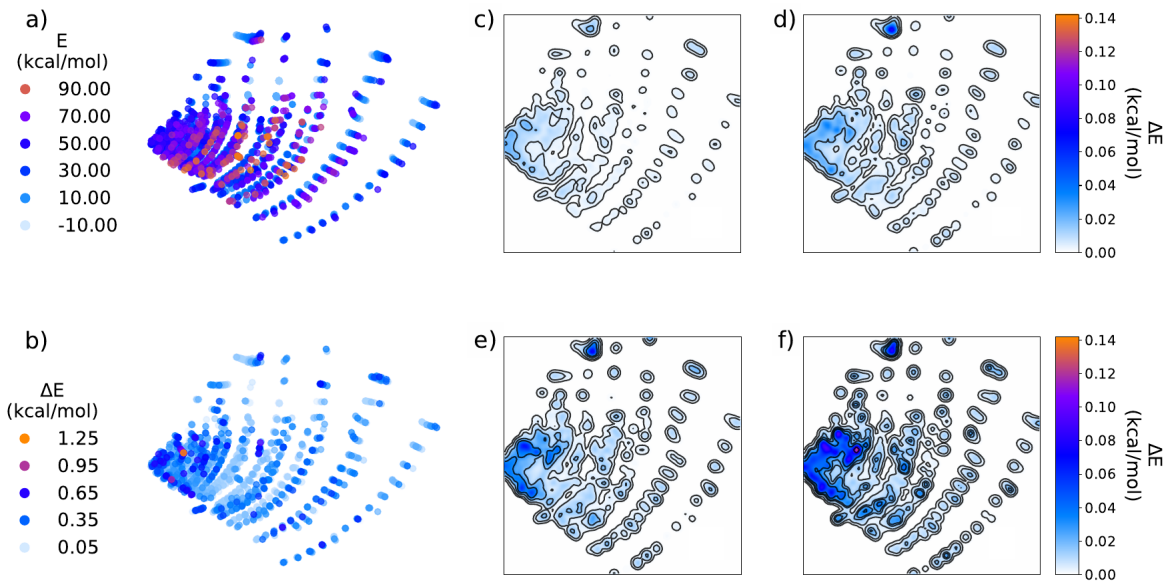

FIG. 4. Sketch-maps of the 2B configurations. The map in in a) represents the reference CCSD(T) energies while the map in b) represents the difference, ΔE, between the reference CCSD(T) energies and the energies predicted by the MB-nrf PEF trained on the full pool of 2B configurations. The maps in c) to f) represent the difference, ΔE, between the energies predicted by the MB-nrg PEF trained on of the full training set and the corresponding values predicted by MB-nrg PEFs trained on the reduced-size training sets of 10000, 8000, 6000, and 4000 configurations generated using the AL framework, respectively).

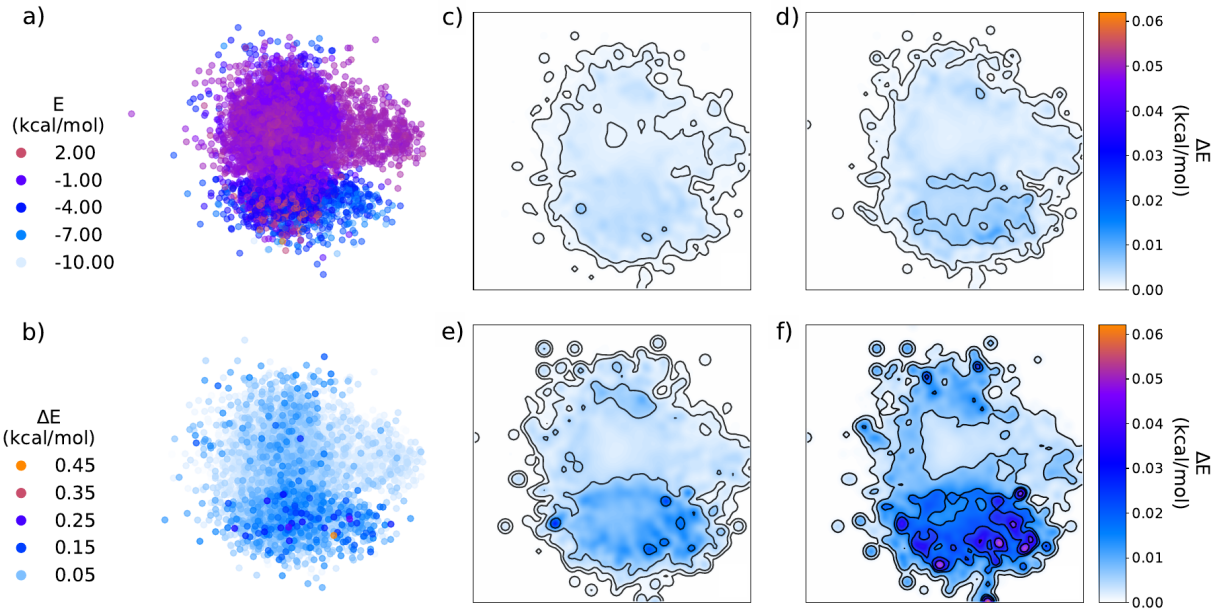

FIG. 5. Sketch-maps of the 3B configurations. The map in in a) represents the reference CCSD(T) energies while the map in b) represents the difference, ΔE, between the reference CCSD(T) energies and the energies predicted by the MB-nrf PEF trained on the full pool of 2B configurations. The maps in c) to f) represent the difference, ΔE, between the energies predicted by the MB-nrg PEF trained on of the full training set and the corresponding values predicted by MB-nrg PEFs trained on the reduced-size training sets of 20000, 15000, 10000, and 5000 configurations generated using the AL framework, respectively).

our AL framework and analyzed through the sketch-maps shown in panels c-f), lead to progressively larger deviations from the original MB-nrg PEF. It should be noted thart, on average, the deviations remain smaller than 0.06 kcal/mol even for the smallest training set (5000 trimer configurations).

C. Clusters analysis

To assess the relative accuracy of the various training sets generated using our AL framework and determine how the associated differences in the representation of 2B and 3B energies affect the ability of the resulting MB-nrg PEFs to reproduce the properties of water from the gas to the condensed phase, deviations from the reference 2B and 3B energies of low-lying isomers of the $Cs^+(H_2O)_{n=1-3}$ clusters are analyzed in Fig. 6. This analysis is carried out for several MB-nrg PEFs generated from fits to various combinations of the 2B and 3B training sets shown in Figs. 4

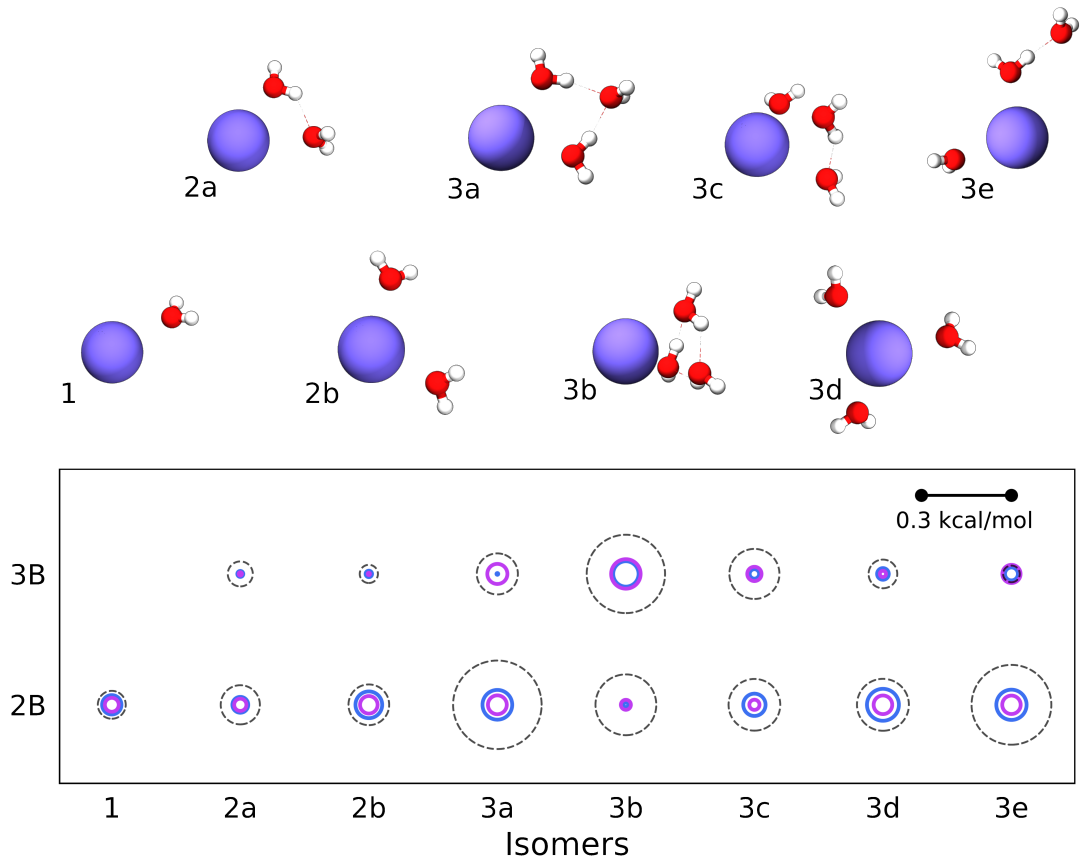

FIG. 6. Schematic representation of the errors associated with the 2B and 3B energies of low-lying isomers of $Cs^+(H_2O)_{n=1-3}$ clusters. The dashed black circles represent the difference between the reference CCSD(T) energies and the corresponding values obtained with the MB-nrg PEF trained on the full 2B and 3B pools. The other solid circles represent the differences between the energies predicted by the MB-nrg PEF trained on the full 2B and 3B pools and the corresponding values predicted by MB-nrg PEFs trained on 4000 2B configurations and 5000 3B configurations, with blue and magenta corresponding to the AL and RS training sets, respectively.

and 5. In particular, we consider the performance of three MB-nrg PEFs that use the smallest 2B training set (4000 dimer configurations) and the full 3B training set, the full 2B training set and the smallest 3B training set (5000 configurations), and the smallest 2B and 3B training sets, respectively. Also shown for comparison are the deviations obtained with the corresponding combinations of the same training sets generated from random selection. In all cases, the differences between the 2B and 3B energies predicted by the different MB-nrg PEFs are comparable for all clusters, and often smaller than the corresponding differences between the original MB-nrg PEF

fitted to the full 2B and 3B training sets and the CCSD(T) reference data. This analysis thus indicates that the reduction of the training set sizes does not affect the ability of the resulting MB-nrg PEFs to correctly represent 2B and 3B energies in small water clusters. It should be noted that this is true for both families of MB-nrg PEFs derived from training sets generated through AL and RS. This similarity can be attributed to the intrinsic low dimensionality of the $Cs^+(H_2O)$ dimers and $Cs^+(H_2O)_2$ trimers that make up the corresponding 2B and 3B training sets, which allowed for extensive sampling of the relevant configurations for the development of the original training sets in Refs. 13 and 60 . However, while no appreciable differences exist in the performance of the two sets of MB-nrg PEFs, AL clearly provides a more efficient approach to the selection of the training set sizes as demonstrated by the significantly higher variability associated with the learning curves obtained with the RS approach. The efficiency of the AL framework thus becomes particular important when, differently from the present case of the Cs^+–water MB-nrg PEF, no prior information on training sets is provided. This aspect of our AL framework will be the subject of a forthcoming study.

D. Radial distribution functions

To investigate the effects of training set reduction on modeling the properties of bulk solutions, the Cs^+–O RDFs calculated using different MB-nrg PEFs obtained from fits to different combinations of 2B and 3B training sets generated using AL (left panels) and RS (right panels) are analyzed in Figs. 7 and 8. The effects of the 2B training set is first assessed in Fig. 7 by analyzing the performance of five MB-nrg PEFs generated by fitting the 2B term to 2B training sets of various sizes (full, 10000, 8000, 6000, and 4000 dimer configurations) while fitting the 3B term to the full 3B training set for training the 3B term (34441 trimer configurations). The resulting RDFs calculated from MD simulations with the resulting MB-nrg PEFs generated from both AL and RS training sets are shown in the top left and right panels of Fig.7, respectively. As discussed in more detail in Ref. 60, the Cs^+–O RDF displays a narrow peak, corresponding to the first hydration shell, at 3.16 Å, and a broader peak, corresponding to the second hydration shell, at $\sim$6 Å. No appreciable differences are found between the RDFs obtained using MB-nrg PEFs with progressively smaller 2B training sets. This is further evidenced by the curves shown in the bottom panels of Fig. 7 representing the differences between the RDFs calculated with each of the MB-nrg PEFs trained on reduced 2B training sets and the RDF calculated with the MB-nrg PEF trained on the

full 2B training set.

Similarly, the effects of the 3B training set size reduction are investigated in Fig. 8 through the analysis of five MB-nrg PEFs generated by fitting the 3B term to 3B training sets of various sizes (full, 20000, 15000, 10000, and 5000 trimer configurations) while fitting to the 2B term to the full 2B training set. In this case, reducing the 3B training set size to less than 10000 trimer configurations results in small differences in the Cs^+–water RDF for distances larger than 5.0 Å, which lead to a shift of the second hydration shell to slightly larger distances. However, as shown

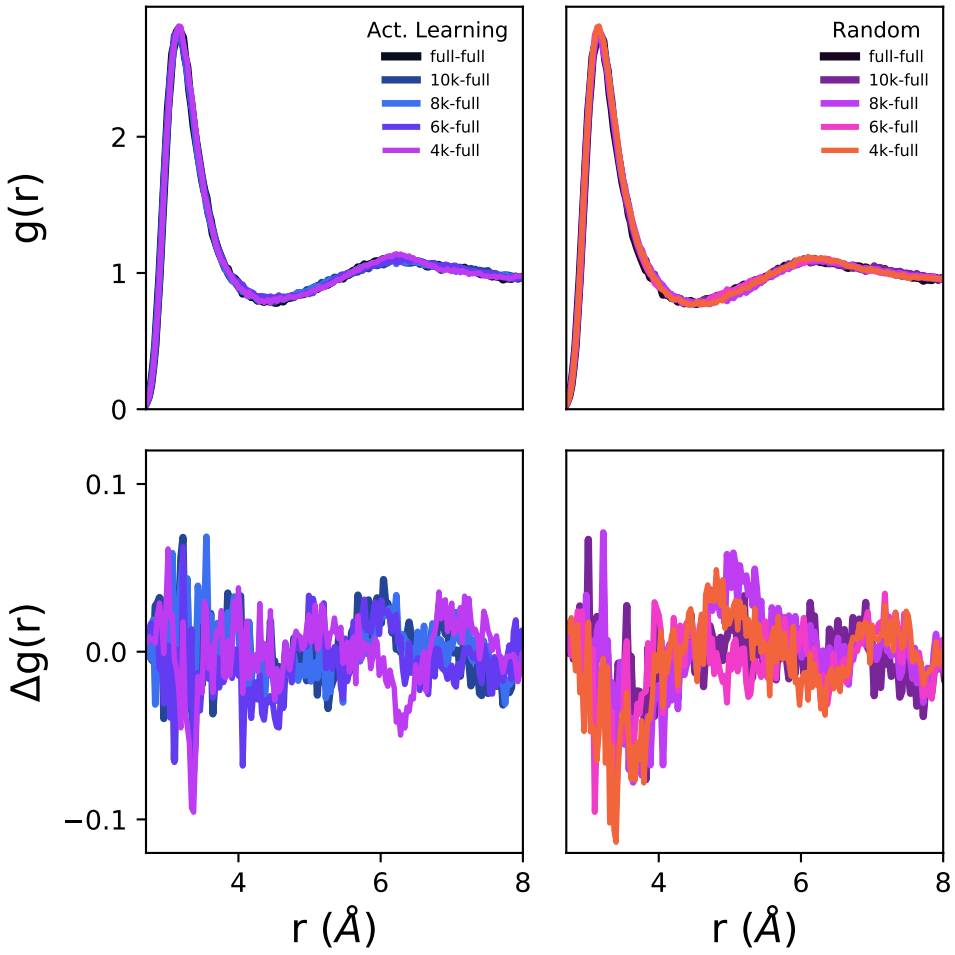

FIG. 7. Top panels: Cs^+-O RDFs calculated from MD simulations with MB-nrg PEFs trained on progressively smaller 2B training sets (in the range of 10000-4000 dimer configurations) generated through AL (left) and RS (right), and the full 3B pool. Bottom panels: Differences between the RDF calculated with the MB-nrg PEF trained on the full 2B and 3B pool and the corresponding RDFs calculated with MB-nrg PEFs trained on the reduced-size AL (left) and RS (right) 2B training sets, and the full 3B pool.

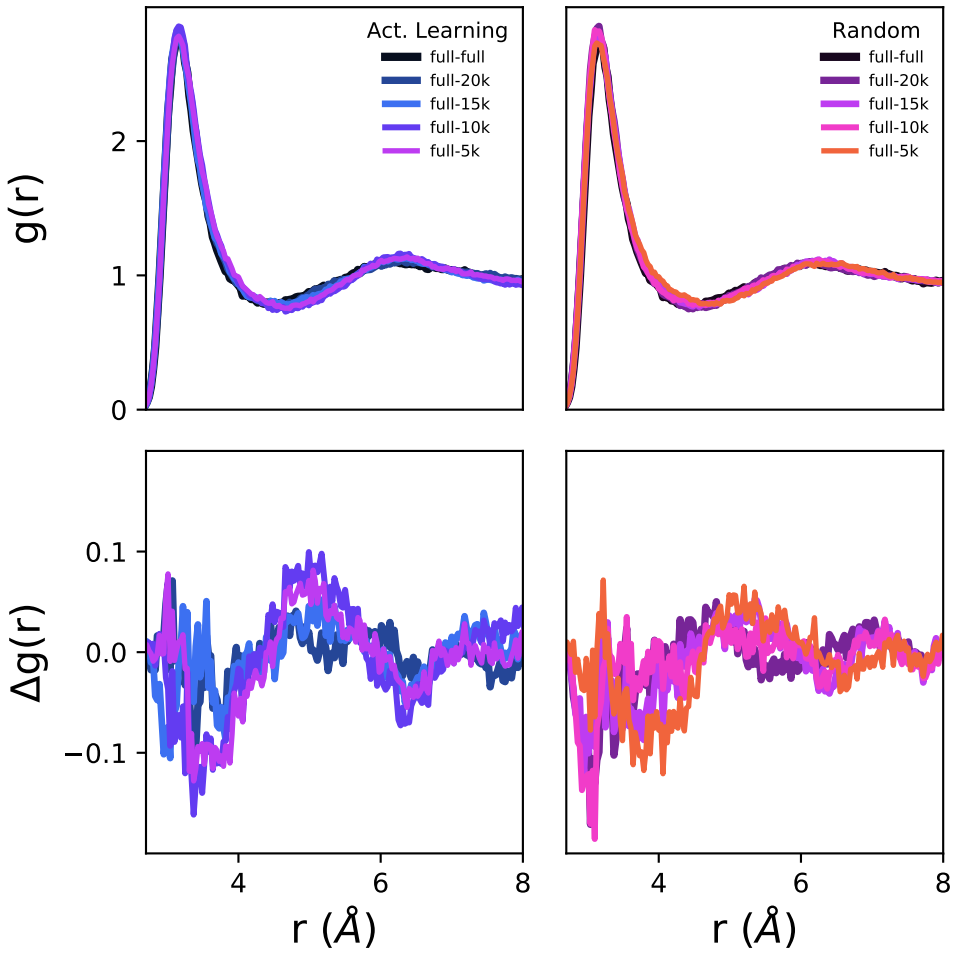

FIG. 8. Top panels: Cs^+-O RDFs calculated from MD simulations with MB-nrg PEFs trained on progressively smaller 3B training sets (in the range of 20000-5000 trimer configurations) generated through AL (left) and RS (right), and the full 2B training set. Bottom panels: Differences between the RDF calculated with the MB-nrg PEF trained on the full 2B and 3B pool and the corresponding RDFs calculated with MB-nrg PEFs trained on the reduced-size AL (left) and RS (right) 3B training sets, and the full 2B pool.

in the bottom panels of Fig. 8, these differences are barely noticeable and do not lead to any qualitative change in the hydration structure of Cs^+ in liquid water.

Overall, the analysis of both cluster and bulk properties demonstrates that the application of our AL framework to the original pools of 2B and 3B configurations of Refs. 50 and 60, respectively, leads to significantly smaller training sets, without loss of accuracy, which, in turn, largely reduces the cost associated with the development of CCSD(T)-level MB-nrg PEFs.

IV. CONCLUSIONS

In this study, we introduced an AL framework for generating representative training sets needed for the development of MB-nrg PEFs.[12,13] Our AL framework is based on uncertainty and error estimation, and consists of a pool of an unknown number of molecular configurations, a predictor, and a learner that, based on feedback from the predictor, selects configurations from the pool and adds them to the training set. The selection process relies on Gaussian process regression and clustering of the configurations in the training set, which allows for efficiently identifying the most relevant configurations needed to accurately represent the target many-body PES.

The application of our AL framework to the development of a Cs^+–water MB-nrg PEF chosen as a case study led to significantly smaller training sets than those found necessary for the development of the original MB-nrg PEF. Analyses of the interaction and many-body energies calculated for small $Cs^+(H_2O)_n$ cluster as well as the Cs^+-oxygen RDF calculated from MD simulations of a single Cs^+ ion in water demonstrate that the new MB-nrg PEFs derived from the reduced-size training sets generated through AL display the same accuracy of the original MB-nrg PEF derived from the full 2B and 3B pools.[13,60]

Given the computational cost associated with reference CCSD(T) calculations of individual many-body energies, our AL framework is particularly well-suited to the development of many-body PEFs, with chemical and spectroscopic accuracy, which can then be used in MD simulations of the target molecular system, from the gas to the condensed phase.

V. ACKNOWLEDGEMENT

We thank Andreas Götz for helpful discussions on the selection of optimal training sets, Matthias Rupp for guidance on using the MBTR descriptor, and Giulio Imbalzano and Michele Ceriotti for suggestions on generating of the 2B and 3B sketch-maps. This research was supported by the National Science Foundation through grant no. ACI-1642336. All calculations and simulations used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation through grant no. ACI-1053575, under allocation TG-CHE110009, and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC).

REFERENCES

[1]M. Karplus, "Development of multiscale models for complex chemical systems: From H + H_2 to biomolecules (Nobel lecture)," Angew. Chem. Int. Ed. **53**, 9992–10005 (2014).

[2]A. Warshel, "Multiscale modeling of biological functions: From enzymes to molecular machines (Nobel lecture)," Angew. Chem. Int. Ed. **53**, 10020–10031 (2014).

[3]M. Levitt, "Birth and future of multiscale modeling for macromolecular systems (Nobel lecture)," Angew. Chem. Int. Ed. **53**, 10006–10018 (2014).

[4]G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartok, G. Csanyi, V. Molinero, and F. Paesani, "Modeling molecular interactions in water: From pairwise to many-body potential energy functions," Chem. Rev. **116**, 7501–7528 (2016).

[5]R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. Van der Avoird, "Predictions of the properties of water from first principles," Science **315**, 1249–1252 (2007).

[6]Y. Wang, B. C. Shepler, B. J. Braams, and J. M. Bowman, "Full-dimensional, ab initio potential energy and dipole moment surfaces for water," J. Chem. Phys. **131**, 054511 (2009).

[7]Y. Wang, X. Huang, B. C. Shepler, B. J. Braams, and J. M. Bowman, "Flexible, ab initio potential, and dipole moment surfaces for water. I. tests and applications for clusters up to the 22-mer," J. Chem. Phys. **134**, 094509 (2011).

[8]V. Babin, G. R. Medders, and F. Paesani, "Toward a universal water model: First principles simulations from the dimer to the liquid phase," J. Phys. Chem. Lett. **3**, 3765–3769 (2012).

[9]V. Babin, C. Leforestier, and F. Paesani, "Development of a "first principles" water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient," J. Chem. Theory. Comp. **9**, 5395–5403 (2013).

[10]V. Babin, G. R. Medders, and F. Paesani, "Development of a "first principles" water potential with flexible monomers. II: Trimer potential energy surface, third virial coefficient, and small clusters," J. Chem. Theory. Comp. **10**, 1599–1607 (2014).

[11]G. R. Medders, V. Babin, and F. Paesani, "Development of a "first-principles" water potential with flexible monomers. III. Liquid phase properties," J. Chem. Theory. Comp. **10**, 2906–2910 (2014).

[12]P. Bajaj, A. W. Götz, and F. Paesani, "Toward chemical accuracy in the description of ion–water interactions through many-body representations. I. Halide–water dimer potential energy surfaces," J. Chem. Theory. Comp. **12**, 2698–2705 (2016).

[13]M. Riera, N. Mardirossian, P. Bajaj, A. W. Götz, and F. Paesani, "Toward chemical accuracy in the description of ion–water interactions through many-body representations. alkali-water dimer potential energy surfaces," J. Chem. Phys. **147**, 161715 (2017).

[14]J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).

[15]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).

[16]G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New J. Phys. **15**, 095003 (2013).

[17]K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," J. Phys. Chem. Lett. **6**, 2326 (2015).

[18]E. Y. Lee, B. M. Fulan, G. C. Wong, and A. L. Ferguson, "Mapping membrane activity in undiscovered peptide sequence space using machine learning," Proc. Natl. Acad. Sci. U.S.A. **113**, 13588–13593 (2016).

[19]J. P. Janet and H. J. Kulik, "Predicting electronic structure properties of transition metal complexes with neural networks," Chem. Sci. **8**, 5137–5152 (2017).

[20]J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost," Chem Sci. **8**, 3192–3203 (2017).

[21]Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: a benchmark for molecular machine learning," Chem. Sci. **9**, 513–530 (2018).

[22]L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, "Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).

[23]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[24]Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: A benchmark for molecular machine learning," Chem. Sci. **9**, 513–530 (2018).

[25]A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, "Transferable machine-learning model of the electron density," ACS Cent. Sci. **5**, 57–64 (2018).

[26]L. Tallorin, J. Wang, W. E. Kim, S. Sahu, N. M. Kosa, P. Yang, M. Thompson, M. K. Gilson, P. I. Frazier, M. D. Burkart, *et al.*, "Discovering de novo peptide substrates for enzymes using machine learning," Nat. Commun. **9**, 1–10 (2018).

[27]M. H. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic ai," Nature **555**, 604–610 (2018).

[28]F. Noé, S. Olsson, J. Köhler, and H. Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," Science **365**, eaaw1147 (2019).

[29]J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, and C. Clementi, "Machine learning of coarse-grained molecular dynamics force fields," ACS Cent. Sci. **5**, 755–767 (2019).

[30]R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06 (ACM, New York, NY, USA, 2006) pp. 161–168.

[31]S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies* (IOS Press, Amsterdam, The Netherlands, The Netherlands, 2007) pp. 3–24.

[32]H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, AAAI'91 (AAAI Press, 1991) pp. 547–552.

[33]D. Hankins, J. W. Moskowitz, and F. H. Stillinger, "Water molecule interactions," J. Chem. Phys. **53**, 4544–4554 (1970).

[34]F. Paesani, "Getting the right answers for the right reasons: Toward predictive molecular simulations of water with many-body potential energy functions," Acc. Chem. Res. **49**, 1844–1851 (2016).

[35]S. K. Reddy, S. C. Straight, P. Bajaj, C. Huy Pham, M. Riera, D. R. Moberg, M. A. Morales, C. Knight, A. W. Götz, and F. Paesani, "On the accuracy of the mb-pol 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. **145**, 194504 (2016).

[36]J. O. Richardson, C. Pérez, S. Lobsiger, A. A. Reid, B. Temelso, G. C. Shields, Z. Kisiel, D. J. Wales, B. H. Pate, and S. C. Althorpe, "Concerted hydrogen-bond breaking by quantum

tunneling in the water hexamer prism," Science **351**, 1310–1313 (2016).

[37]W. T. Cole, J. D. Farrell, D. J. Wales, and R. J. Saykally, "Structure and torsional dynamics of the water octamer from thz laser spectroscopy near 215 μm," Science **352**, 1194–1197 (2016).

[38]J. D. Mallory and V. A. Mandelshtam, "Diffusion monte carlo studies of mb-pol (h2o) 2- 6 and (d2o) 2- 6 clusters: Structures and binding energies," J. Chem. Phys. **145**, 064308 (2016).

[39]P. Videla, P. Rossky, and D. Laria, "Communication: Isotopic effects on tunneling motions in the water trimer." J. Chem. Phys. **144**, 061101–061101 (2016).

[40]S. E. Brown, A. W. Götz, X. Cheng, R. P. Steele, V. A. Mandelshtam, and F. Paesani, "Monitoring water clusters "melt" through vibrational spectroscopy," Journal of the American Chemical Society **139**, 7082–7088 (2017).

[41]C. L. Vaillant and M. T. Cvitaš, "Rotation-tunneling spectrum of the water dimer from instanton theory," Phys. Chem. Chem. Phys. **20**, 26809–26813 (2018).

[42]C. L. Vaillant, D. J. Wales, and S. C. Althorpe, "Tunneling splittings from path-integral molecular dynamics using a langevin thermostat," J. Chem. Phys. **148**, 234102 (2018).

[43]M. Schmidt and P.-N. Roy, "Path integral molecular dynamic simulation of flexible molecular systems in their ground state: Application to the water dimer," J. Chem Phys. **148**, 124116 (2018).

[44]K. P. Bishop and P.-N. Roy, "Quantum mechanical free energy profiles with post-quantization restraints: Binding free energy of the water dimer over a broad range of temperatures," J. Chem. Phys. **148**, 102303 (2018).

[45]P. E. Videla, P. J. Rossky, and D. Laria, "Isotopic equilibria in aqueous clusters at low temperatures: Insights from the mb-pol many-body potential," J. Chem. Phys. **148**, 084303 (2018).

[46]N. R. Samala and N. Agmon, "Temperature dependence of intramolecular vibrational bands in small water clusters," J. Phys. Chem. B **123**, 9428–9442 (2019).

[47]N. R. Samala and N. Agmon, "Thermally induced hydrogen-bond rearrangements in small water clusters and the persistent water tetramer," ACS Omega (2019).

[48]M. T. Cvitaš and J. O. Richardson, "Quantum tunnelling pathways of the water pentamer," Phys. Chem. Chem. Phys. (2020).

[49]G. R. Medders and F. Paesani, "Infrared and raman spectroscopy of liquid water through "first-principles" many-body molecular dynamics," J. Chem. Theory Comput. **11**, 1145–1154 (2015).

[50]S. K. Reddy, D. R. Moberg, S. C. Straight, and F. Paesani, "Temperature-dependent vibrational spectra and structure of liquid water from classical and quantum simulations with the mb-pol

potential energy function," J. Chem. Phys. **147**, 244504 (2017).

[51]Z. Sun, L. Zheng, M. Chen, M. L. Klein, F. Paesani, and X. Wu, "Electron-hole theory of the effect of quantum nuclei on the x-ray absorption spectra of liquid water," Phys. Rev. Lett. **121**, 137401 (2018).

[52]K. M. Hunter, F. A. Shakib, and F. Paesani, "Disentangling coupling effects in the infrared spectra of liquid water," J. Phys. Chem. B **122**, 10754–10761 (2018).

[53]G. R. Medders and F. Paesani, "Dissecting the molecular structure of the air/water interface from quantum simulations of the sum-frequency generation spectrum," J. Am. Chem. Soc. **138**, 3912–3919 (2016).

[54]D. R. Moberg, S. C. Straight, and F. Paesani, "Temperature dependence of the air/water interface revealed by polarization sensitive sum-frequency generation spectroscopy," J. Phys. Chem. B **122**, 4356–4365 (2018).

[55]S. Sengupta, D. R. Moberg, F. Paesani, and E. Tyrode, "Neat water–vapor interface: Proton continuum and the nonresonant background," J. Phys. Chem. Lett. **9**, 6744–6749 (2018).

[56]S. Sun, F. Tang, S. Imoto, D. R. Moberg, T. Ohto, F. Paesani, M. Bonn, E. H. Backus, and Y. Nagata, "Orientational distribution of free oh groups of interfacial water is exponential," Phys. Rev. Lett. **121**, 246101 (2018).

[57]C. H. Pham, S. K. Reddy, K. Chen, C. Knight, and F. Paesani, "Many-body interactions in ice," J. Chem. Theory. Comp. **13**, 1778–1784 (2017).

[58]D. R. Moberg, S. C. Straight, C. Knight, and F. Paesani, "Molecular origin of the vibrational structure of ice Ih," J. Phys. Chem. Lett. **8**, 2579–2583 (2017).

[59]D. R. Moberg, P. J. Sharp, and F. Paesani, "Molecular-level interpretation of vibrational spectra of ordered ice phases," J. Phys. Chem. B **122**, 10572–10581 (2018).

[60]D. Zhuang, M. Riera, G. K. Schenter, J. L. Fulton, and F. Paesani, "Many-body effects determine the local hydration structure of Cs^+ in solution," J. Phys. Chem. Lett. **10**, 406–412 (2019).

[61]C. J. Burnham, D. J. Anick, P. K. Mankoo, and G. F. Reiter, "The vibrational proton potential in bulk liquid water and ice," J. Chem. Phys. **128**, 154519 (2008).

[62]A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," Soviet Math. **4**, 1035–1038 (1963).

[63]J. Rezac and P. Hobza, "Benchmark calculations of interaction energies in noncovalent complexes and their applications," Chem. Rev. **116**, 5038–5071 (2016).

[64]J. G. Hill, K. A. Peterson, G. Knizia, and H.-J. Werner, "Extrapolating MP2 and CCSD explicitly correlated correlation energies to the complete basis set limit with first and second row correlation consistent basis sets," J. Chem. Phys. **131**, 194105 (2009).

[65]U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, "Interaction energies of large clusters from many-body expansion," J. Chem. Phys. **135**, 224102 (2011).

[66]T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. I. the atoms boron through neon and hydrogen," J. Chem. Phys. **90**, 1007–1023 (1989).

[67]R. A. Kendall, T. H. Dunning, and R. J. Harrison, "Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions," J. Chem. Phys. **96**, 6796–6806 (1992).

[68]D. E. Woon and T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. V. core-valence basis sets for boron through neon," J. Chem. Phys. **103**, 4572–4585 (1995).

[69]J. G. Hill and K. A. Peterson, "Gaussian basis sets for use in correlated molecular calculations. XI. pseudopotential-based and all-electron relativistic basis sets for alkali metal (K–Fr) and alkaline earth (Ca–Ra) elements," J. Chem. Phys. **147**, 244106 (2017).

[70]S. F. Boys and B. F., "The calculation of small molecular interactions by the differences of separate total energies. some procedures with reduced errors," Mol. Phys. **19**, 553–566 (1970).

[71]I. S. Lim, P. Schwerdtfeger, B. Metz, and H. Stoll, "All-electron and relativistic pseudopotential studies for the group 1 element polarizabilities from K to element 119," J. Chem. Phys. **122**, 104103 (2005).

[72]G. J. Martyna, A. Hughes, and M. E. Tuckerman, "Molecular dynamics algorithms for path integrals at constant pressure," J. Chem. Phys. **110**, 3275–3290 (1999).

[73]W. Smith and T. Forester, "Dl_poly_2. 0: A general-purpose parallel molecular dynamics simulation package," J. Mol. Graph. **14**, 136–141 (1996).

[74]B. Settles, "Active learning literature survey," Tech. Rep. (University of Wisconsin-Madison Department of Computer Sciences, 2009).

[75]C. E. Rasmussen and C. K. I. Williams, *Gaussian Process for Machine Learning* (MIT Press, 2006).

[76]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).

[77]G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," J. Chem. Phys **148**, 241730 (2018).

[78]H. Huo and M. Rupp, "Unified representation for machine learning of molecules and crystals," arXiv:1704.06439 (2017).

[79]M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," Proc. Natl. Acad. Sci. U.S.A. **108**, 13023–13028 (2011).

[80]M. Ceriotti, G. A. Tribello, and M. Parrinello, "Demonstrating the transferability and the descriptive power of sketch-map," J. Chem. Theory. Comp. **9**, 1521–1532 (2013).

[81]T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, "Comparison of permutationally invariant polynomials, neural networks, and gaussian approximation potentials in representing water interactions through many-body expansions," J. Chem. Phys. **148**, 241725 (2018).