The Journal of Chemical Physics

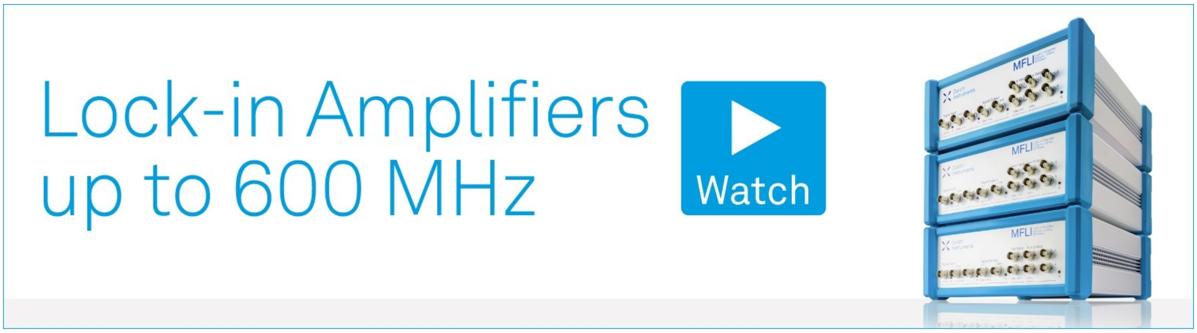
A look inside the black box: Using graph-theoretical descriptors to interpret a Continuous-Filter Convolutional Neural Network (CF-CNN) trained on the global and local minimum energy structures of neutral water clusters

Cite as: J. Chem. Phys. **153**, 024302 (2020); https://doi.org/10.1063/5.0009933
Submitted: 07 April 2020 . Accepted: 24 June 2020 . Published Online: 13 July 2020

Jenna A. Bilbrey, Joseph P. Heindel, Malachi Schram, Pradipta Bandyopadhyay, Sotiris S. Xantheas, and Sutanay Choudhury

COLLECTIONS

Paper published as part of the special topic on Machine Learning Meets Chemical Physics
Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.



AIP Publishing

A look inside the black box: Using graph-theoretical descriptors to interpret a Continuous-Filter Convolutional Neural Network (CF-CNN) trained on the global and local minimum energy structures of neutral water clusters

Cite as: J. Chem. Phys. 153, 024302 (2020); doi: 10.1063/5.0009933
Submitted: 7 April 2020 • Accepted: 24 June 2020 •
Published Online: 13 July 2020

View Online Export Citation CrossMark

Jenna A. Bilbrey,[1] Joseph P. Heindel,[2] Malachi Schram,[3] Pradipta Bandyopadhyay,[4] Sotiris S. Xantheas,[2,3,a)] and Sutanay Choudhury[3,b)]

AFFILIATIONS

[1] Computing and Analytics Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, P.O. Box 999, Richland, Washington 99352, USA

[2] Department of Chemistry, University of Washington, Seattle, Washington 98195, USA

[3] Advanced Computing, Mathematics and Data Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, P.O. Box 999, Richland, Washington 99352, USA

[4] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.

[a)] **Author to whom correspondence should be addressed:** sotiris.xantheas@pnnl.gov. **Tel.:** +1-509-375-3684
[b)] **Electronic mail:** sutanay.choudhury@pnnl.gov

ABSTRACT

We describe a method for the post-hoc interpretation of a neural network (NN) trained on the global and local minima of neutral water clusters. We use the structures recently reported in a newly published database containing over 5×10^6 unique water cluster networks $(H_2O)_N$ of size $N = 3$–30. The structural properties were first characterized using chemical descriptors derived from graph theory, identifying important trends in topology, connectivity, and polygon structure of the networks associated with the various minima. The code to generate the molecular graphs and compute the descriptors is available at https://github.com/exalearn/molecular-graph-descriptors, and the graphs are available alongside the original database at https://sites.uw.edu/wdbase/. A Continuous-Filter Convolutional Neural Network (CF-CNN) was trained on a subset of 500 000 networks to predict the potential energy, yielding a mean absolute error of 0.002 ± 0.002 kcal/mol per water molecule. Clusters of sizes not included in the training set exhibited errors of the same magnitude, indicating that the CF-CNN protocol accurately predicts energies of networks for both smaller and larger sizes than those used during training. The graph-theoretical descriptors were further employed to interpret the predictive power of the CF-CNN. Topological measures, such as the Wiener index, the average shortest path length, and the similarity index, suggested that all networks from the test set were within the range of values as the ones from the training set. The graph analysis suggests that larger errors appear when the mean degree and the number of polygons in the cluster lie further from the mean of the training set. This indicates that the structural space, and not just the chemical space, is an important factor to consider when designing training sets, as predictive errors can result when the structural composition is sufficiently different from the bulk of those in the training set. To this end, the developed descriptors are quite effective in explaining the results of the CF-CNN (a.k.a. the "black box") model.

I. INTRODUCTION

The use of artificial intelligence (AI) for scientific applications has rapidly increased over the past decade. Neural networks (NN), in particular, have proven useful for advances in domain areas such as computer-assisted drug discovery,[1] inverse materials design,[2] computer-aided synthesis planning,[3] and most notably as surrogate models for high-level computational methods.[4–6] Such surrogate models allow for the generation of interaction potentials that demonstrate the accuracy of higher-level computational methods while being orders of magnitude less expensive. However, the training of such a network requires a large amount of data generated at the desired level of accuracy, and the more general the network, the larger the coverage of chemical space required.

Neural networks are often viewed as "black boxes"—producing the answer they were trained to give without any indication of how they arrived there. Recently, efforts to interpret the learning process of NNs have been reported.[7–10] NN interpretability is often loosely defined as any method that allows a glimpse inside the box. In this study, we suggest a post-hoc interpretation of the performance of a trained NN through the analysis of the network output based on graph-theoretical descriptors of the molecular system (see Fig. 1). By correlating descriptors based on the properties of the fitted data with the network errors, we are able to elucidate the regions of configurational space for which the trained NN has predictive power. We focus the training of such a NN using a large database of neutral water cluster global and local minima, recently published by some of the authors in this study,[11] and show that structural space, in addition to chemical space, must be considered when training a NN.

Our drive for using water as an example is fueled by its importance in sustaining life on Earth through reactions in aqueous media that have earned water the moniker of "universal solvent." Therefore, an understanding of the properties of water represents a necessary first step toward the modeling of chemical and biological processes in aqueous environments. Clusters of water molecules allow for a quantitative probe of the nature and magnitude of intermolecular interactions within a water network.[12,13] Over the years, several important insights have been gained through the application of high-level *ab initio* methods, but the computational cost of these methods often prohibits the study of large systems over long time scales while maintaining a level of accuracy that produces meaningful macroscopic properties compared to experiment. A number of flexible, polarizable classical potentials for water have been developed[14–18] based on the results of high-level electronic structure calculations of clusters. These interaction potentials have been shown to accurately reproduce several of the macroscopic properties of water.[19–22] At the same time, a number of machine learning (ML) techniques have been developed for the interpolation of potential energy surfaces (PESs).[23–27]

Recently, Morawietz and co-workers demonstrated the use of neural networks to obtain high-quality potentials.[28–30] Radial and angular symmetry functions were developed as input to their neural networks to describe the environment of each atom through consideration of the position of all atoms in the system.[4,31,32] Notably, these symmetry functions satisfy the required invariance with respect to rotation, translation, and ordering, and they are continuous and differentiable. To create the potential, a NN was trained for each element in the system. The energy contribution from each atom was predicted, and the collection was summed to obtain the total system energy. The forces were calculated analytically as the negative gradients of the energy due to the well-defined functional form of the neural network potential. Recently, this method has been

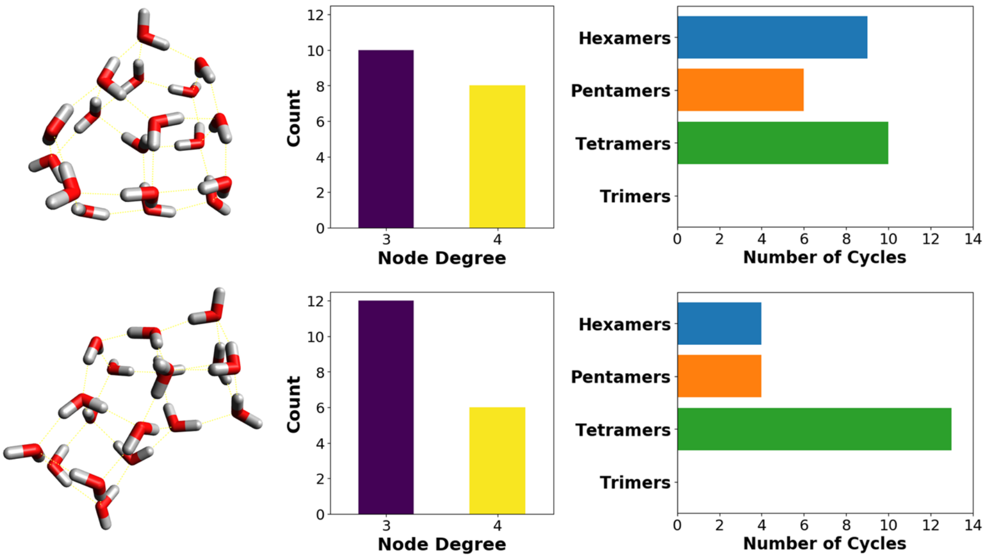


FIG. 1. Examples of two of the descriptors (node degree and number of geometric cycles) used to assess post-hoc interpretability of a CF-CNN trained to predict the potential energy of water clusters. The examples shown here are for two isomers of the $(H_2O)_{20}$ cluster contained in the database recently published by some of us.[11]

incorporated in the open-source LAMMPS molecular dynamics software program.[33,34]

The Behler–Parrinello method has been applied to both neutral water clusters and protonated water clusters. A water cluster potential was developed by training on ~40 000 dispersion-corrected density functional theory (DFT) reference computations, which resulted in errors on the order of 2 meV/H_2O.[28] More recently, the Behler–Parrinello method has been used to construct neural network PESs for protonated water clusters based on both DFT[29] and coupled-cluster[30,35] reference data. The most recent of these potentials uses DFT *ab initio* molecular dynamics simulations to generate configurations, which are then refined at the CCSD(T)-F12/VTZ level of theory. The resulting potential provides binding energies with an accuracy close to 0.1 kcal/mol. Based on these successes, we have chosen in this study to benchmark the accuracy of the CF-CNN against the Behler–Parrinello RuNNer model.[36]

Schütt *et al.* built upon Behler's atom-centered symmetry function approach, but with a key difference, namely, that the network learns the ideal atom representation.[37–39] Atom-based filters are formed from the vectors between atoms and their neighbors, a procedure providing a unique internal representation that incorporates symmetry and ordering invariances. The filters are learned during training of the network, and each atom is represented by an array of filters. The learned filters then act as features, from which convolutional layers learn a representation of pair-wise interactions between atoms in the cluster to predict the contribution of each atom to the desired property. Altogether, this architecture is described as a Continuous-Filter Convolutional Neural Network (CF-CNN). Multiple interaction blocks can be stacked, and because the filter generator is contained within the interaction block, the learned features will be different for each block. The atom-wise contributions are then summed to give the value of the desired molecular-level property. The network employs an assignable cutoff value and function to each atom to properly model the decay of the interaction energy between atoms at long distances. Overall, this architecture, known as SchNet, is able to pick up subtle changes in structure to produce continuous energy surfaces.

Here, we use the codebase from SchNetPack[39] to train a CF-CNN that can predict the potential energy of water clusters using a training set of 500 000 unique water cluster networks—to our knowledge, the largest molecular training set applied to neural networks to date. The CF-CNN is able to produce energies with a mean absolute error of 0.002 ± 0.002 kcal/mol per water molecule.

Note that our study focuses more on the interpretation of the accuracy of the trained NN through the development of graph-theoretical descriptors rather than just the simple training of the NN. Given this goal, the paper is organized as follows: In Sec. II, we discuss our approach providing details of the generation of the previously published water cluster network database, the definition of the graph-theoretical descriptors used to analyze the structural patterns of the various networks, and the details and setup of the CF-CNN. In Sec. III, we present the results of the graph-theoretical characterization of the water cluster network database, the optimization of the CF-CNN training, and the interpretation of these results using the graph-theoretical descriptors. Final conclusions are drawn in Sec. IV.

II. APPROACH

A. Details of the database of neutral water cluster networks

We recently published a database of networks corresponding to the global and local minima for neutral water clusters of sizes $N = 3$–30 obtained using the flexible, polarizable Thole-Type Model (TTM2.1-F, version 2.1) interaction potential for water in conjunction with the Monte Carlo Temperature Basin Paving (MCTBP) sampling method.[11] The database contains over 5×10^6 local minima that lie within 5 kcal/mol from the putative minimum for each cluster size.

The TTM2.1-F interaction potential is a many-body, flexible, polarizable potential with parameters derived from high-level *ab initio* calculations.[14,15] MCTBP is a global optimization method that aims to improve the convergence rate of global optimization techniques such as basin hopping,[40] with the algorithmic details presented elsewhere.[41,42] In short, all possible energies for the system of interest are split into finely spaced bins. Each of these bins are given a parameter, conceptually referred to as the temperature. This temperature controls the acceptance criteria for Monte Carlo moves. That is, for a single step from a bin with energy E_{old} and inverse temperature $\beta_{\text{old}} = 1/kT_{\text{old}}$ to a bin with energy E_{new}, the acceptance condition for this move is

$$\min\left(1, \exp(-\beta_{\text{old}}(E_{\text{new}} - E_{\text{old}}))\right). \tag{1}$$

This sampling method tends to decrease the number of times one revisits the same structure by increasing the temperature associated with a particular bin each time that bin is visited.

B. Graph-theoretical descriptors

NNs depend on exposure to numerous examples of the desired subset of configuration space to predict the behavior of new candidates in that space. As the configuration space in question grows, the number of structures in the training set must also increase. The configuration space examined here is defined by hydrogen-bonded networks present in water clusters, which exhibit rich structural diversity. To gain an understanding of the configuration space of these water clusters, we characterize the full database of over 5×10^6 structures using graph-theoretical descriptors. The natural invariances present in chemistry (translational, rotational, and atom ordering) thus correspond to invariances under isomorphism of the corresponding molecular graph.[43] Therefore, a chemical system is represented by a set of isomorphic graphs, and conversely, each isomorphic set represents a single chemical system. An exception is stereoisomers, which have isomorphic graphs because 3-dimensional information is lost upon conversion to a graph-based representation.

When generating the molecular graphs from the Cartesian coordinates provided in the database, we determine the connectivity using the definition of a hydrogen bond given by Kumar *et al.*[44] in which the hydrogen bond is parameterized by a distance r and an angle ψ, where r is the distance between a hydrogen atom and its neighboring oxygen and ψ is the angle between the O–H vector and the vector normal to the plane formed by the molecule receiving the hydrogen bond. This definition comports with the

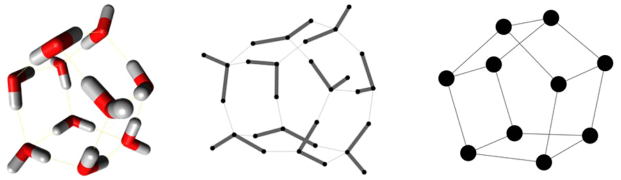


FIG. 2. Structure (left), all-atoms graph (middle), and projected graph (right) for the lowest-energy water cluster of size $N = 10$.

idea that electron density is donated from an O–H bond into a π^* orbital of the receiving water molecule. Each cluster is transformed into an all-atoms graph and a projected graph describing the oxygen atom frame (see Fig. 2). Specifically, in the all-atoms graph, each atom is represented by a node and the edges represent the covalent and hydrogen bonds present in the cluster, while in the projected graph, each water molecule is a node and the edges represent the hydrogen-bond network. The collection of graphs is available alongside the original database of Cartesian coordinates and relative energies at https://sites.uw.edu/wdbase/, whereas the code to generate the graphs and compute the descriptors is available at https://github.com/exalearn/molecular-graph-descriptors.

Both of these graph representations have been used successfully in the past for various applications. For instance, projected graphs can be employed for the complete enumeration of all digraphs corresponding to a particular oxygen frame. This analysis has been performed for all 30 026 hydrogen-bond networks of the pentagonal dodecahedron structure of $(H_2O)_{20}$.[45–48] Notably, this quantified the energy difference of the highest- and lowest-energy arrangements of hydrogen-bond networks that obey the Bernal–Fowler rules to be on the order of 35 kcal/mol.[45,46] This result is important for locating low-energy structures of water clusters of a particular size. That is, because the number of digraphs corresponding to a particular projected graph increases exponentially with the number of water molecules in the network, searches for low-energy structures must include the examination of oxygen frames that could correspond to low-energy structures as well as the particular arrangement of hydrogen bonds. This arrangement could be composed of thousands or even millions of unique digraphs, one of which gives the lowest-energy structure within the specific oxygen frame. Therefore, as N increases past 20, it becomes extremely difficult to confidently identify the global minimum structure. Due to the density of this manifold of states, it is not clear whether the true global minimum is of utmost importance as many structures will be thermally populated even at low temperatures.

Graph-based representations allow for the fast quantification of physical metrics, such as the number of atoms or water molecules (by counting the number of nodes), the number of hydrogen bonds (by counting the number of edges in the projected graph), and the number of dangling hydrogen atoms (by counting the number of nodes with less than four neighbors in the all-atoms graph). If a water molecule is not actually bound to the cluster, there will be a node with zero neighbors in the projected graph, making this representation a convenient way to determine when computations fail to produce a fully connected hydrogen-bond network.

Shared characteristics between two chemical systems can be quantified as a single value by computing the similarity of the corresponding graphs. In this work, we use the eigenvalue method to compute the similarity of two graphs.[49] This metric relies on the eigenvalue (λ) of the Laplacian of each graph, defined as the diagonal matrix of the degrees minus the adjacency matrix of the graph. The similarity s of graphs 1 and 2 is then computed as

$$s = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2, \tag{2}$$

where k represents the top k eigenvalues that contain 90% of the energy (note that this is the graph energy and not the energy of the molecular system). This metric is unbounded, $[0, \infty)$, where isomorphic graphs will show $s = 0$, with s increasing to infinity as the graphs become more dissimilar.

To compare networks within a specific cluster size, we compute s against the lowest-energy cluster of that size given in the database. Among the all-atoms graphs, each cluster will have a different value of s, although some may be very close. However, among the projected graphs, clusters with the same oxygen framework will have identical values of s. In this way, unique oxygen families within a cluster size can be identified.

Translating the Cartesian coordinates of the clusters into graphs also allows us to compute standard graph-based metrics. Two useful topological metrics are (i) the average shortest path length and (ii) the Wiener index. The average shortest path length is defined as the average number of steps along the shortest path between each pair of nodes, while the Wiener index is defined as the sum of the shortest path lengths between all non-hydrogen atoms. These two metrics are similar but have a key difference: the Wiener index will always grow with system size, while the average shortest path length will not. In conjunction, the two metrics provide information about the connectivity of the cluster. In this work, both metrics are computed for the projected graphs.

We also calculate the degree of each node, which is simply the number of edges connected to that node, on the projected graph to give an indication of the connectivity of the hydrogen-bond network. In the system under study, fully connected nodes have a degree of 4, although in some cases, a degree of 5 is observed, as it has been shown that a water molecule can accept up to three hydrogen bonds while still donating two.[11] A holistic view of the connectivity of the full cluster can be obtained by averaging the degree of all nodes in the graph, while the regularity of the cluster can be examined by calculating the variance in the degrees of all nodes in the graph.

Finally, we compute geometric shapes (polygons) present in the water clusters through the use of their projected graphs. Here, we compute the number of 3–6-membered rings in each cluster. This is accomplished through a depth-first search of the number of rings associated with each node. Fused rings are discounted by only considering non-chordal graphs in which the degree of each node in the ring subgraph is 2. For example, a fused 5-membered ring is considered to be composed of one trimer and one tetramer; if one of the water molecules contributing to the fused bond is rotated and the bond breaks, the ring is then considered to be a pentamer (see Fig. 3 for a visual explanation).

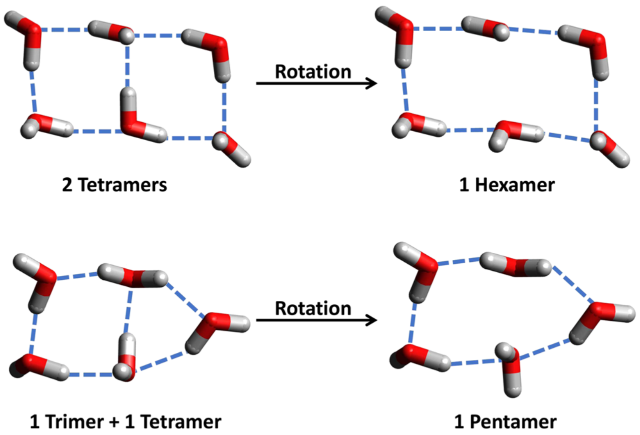


FIG. 3. Visual explanation of the definition of rings in a cluster. Fused rings (shown by the molecules on the left-hand side) are counted as the smallest component rings. A simple rotation of a molecule resulting in the breaking of a hydrogen bond can alter the ring topology. In this work, we counted the number of 3–6-membered rings in the projected graph of each cluster.

C. Continuous-filter convolutional neural networks

In this section, we briefly describe the CF-CNN used in SchNet.[38] The computation graph in this architecture has two phases. The first phase learns an atom-level representation of the energy function, followed by an aggregation phase in the sum pooling layer that sums all atom-level energies to predict the energy of the overall structure. The entire computation process is visually described in Fig. 4.

The key component in this computation is known as the interaction block, which is described in detail below. The interaction block is accompanied by a number of layers that perform a series of standard transformations.

1. Embedding layer

This layer maps each atom and its associated features into a fixed-size vector representation. The vectors are randomly initialized and updated through the training process.

2. Atom-wise layer

These are fully connected layers that recombine the features of an atom from an intermediate layer by considering the contribution of a feature to every other feature. This is implemented via matrix multiplication followed by an additive bias stated as $X_t = W_t X_{t-1} + b_t$, where X_t represents the intermediate features after layer t and W_t represents the learned weight matrix.

3. Activation and pooling layers

The activation layer performs a non-linear transformation via a shifted softplus function (Fig. 4). The shifting is shown to improve the convergence of the model. The sum-pooling layer enables down-sampling of the features through an additive process, aiming to reduce the sensitivity of the output to small perturbations in the input features.

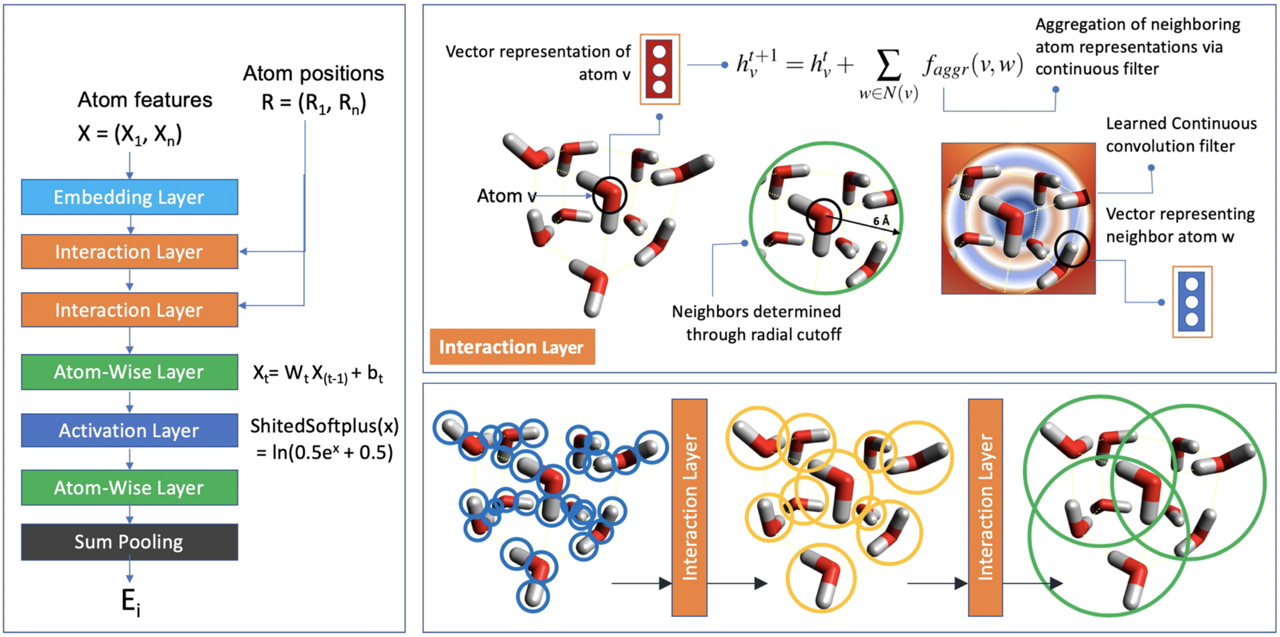


FIG. 4. Illustration of the network architecture and atomic neighborhoods examined by the CF-CNN[38] during learning. From these neighborhoods arises a representation of the geometry of the underlying system.

4. Interaction block

The vector representation for each atom is expressed as an iterative computation similar to those employed by message-passing neural networks. Each atom is assigned a hidden state representation referred to as h_v^t (for atom v and iteration t),

$$h_v^{t+1} = h_v^t + \sum_{w \in N(v)} f_{aggr}(v, w), \quad (3)$$

where f_{aggr} is a learned differentiable function, referred to as the continuous-convolution filter or the filter-generation network component.[38] Given an atom v, we first identify its neighbors within a cutoff [denoted $N(v)$]. The filter-generation network function is implemented as $f_{aggr}(v, w) = h_w \circledast W(\mathbf{r}_v - \mathbf{r}_w)$.

The continuous-filter operation is designed to be rotationally invariant. The rotational invariance is obtained by expanding the inter-atomic distance (d_{vw}) through N_f radial basis functions $\psi_k(\|\mathbf{r}_v - \mathbf{r}_w\|) = \exp(-\gamma\|d_{vw} - \mu_k\|^2)$, where μ_k is the mean of each Gaussian basis function located at a fixed interval ranging from 0 Å to a cutoff value (r_{cutoff} Å).

Finally, the interaction block is repeated a number of times (specified by the hyperparameter N_{iter}), which allows the representation of an atom to be updated by propagating the influence of other atoms that are more than r_{cutoff} away. Figure 4 illustrates how this propagation expands in an atomistic system through each repetition of the interaction block. For a description of the specific configurations of each layer in the CF-CNN architecture, refer the reader to Ref. 38.

D. CF-CNN setup

SchNet implementations are provided in both TensorFlow[38] and PyTorch[39] frameworks; however, the PyTorch implementation, distinguished by the name SchNetPack, includes additional tools for the prediction of PESs and other quantum-chemical properties and is the current, most-up-to-date implementation. We use SchNetPack to train all CF-CNNs in this work.

To optimize the model, we examined training hyperparameters such as the number of interaction blocks, the number of atom-wise features, and the variance in the network itself, along with several data-sampling strategies and the variance in the sampling. In these examinations, each network was trained on ∼100 000 water cluster structures taken from the published database for cluster sizes N = 11–29 using 90 000 of the clusters to learn the weights and the remainder to validate the learned weights during each epoch; for the exact counts of each-size cluster in the training sets, see Table S1. The trained networks were tested on a set of 10 500 clusters not included in the training set from cluster sizes N = 10–30 (500 clusters per size). Depending on the hyperparameters, each network required 11–17 h to train when distributed over four NVIDIA V100 GPUs. Two interaction blocks were used during training; although Schütt *et al.* found three to be the optimal number of interaction blocks in prior analyses,[37,38] we found two to be sufficient (see Table I and Fig. S1). A nearest-neighbor cutoff of 6 Å was applied, as this corresponds to the outer boundary of the second solvation shell in liquid water based on the O–O radial distribution function. Beyond this distance, water molecules are considered to essentially be de-correlated, indicating that their interactions are negligible. The batch size was set to 50, and the maximum number of epochs was set to 7500, although all training trials converged well before reaching this cutoff. Unless otherwise stated, a random seed of 19 was used to obtain reproducible initial weights for the network. After optimizing the training hyperparameters and sampling strategy, we trained a CF-CNN on 500 000 water clusters of size N = 11–29 and again tested on 10 500 unseen clusters of size N = 10–30.

III. RESULTS AND DISCUSSION

A. Graph-theoretical characterization of the full water cluster database

The similarity of each graph, as described in Sec. II B, of a certain cluster size N was computed against the lowest-energy cluster of

TABLE I. Error analysis of CF-CNNs during optimization of the number of interaction blocks, number of atom-wise features, and strategy for sampling from the database. Each network was trained on ∼100 000 water clusters of size N = 11–29. From the full training set, 90 000 clusters were used to learn the network weights, and the remainder were used to validate the weights during training; the test set consisted of 10 500 clusters of size N = 10–30. The mean absolute error (MAE) and root mean squared error (RMSE) of the validation and test sets for the final model are given.

Interaction blocks	Atom-wise features	Sampling strategy	Training loss	Validation loss	Validation MAE	Validation RMSE	Test MAE	Test RMSE
1	100	Even	0.0281	0.0290	0.1282	0.1704	0.1246	0.1678
2	100	Even	0.0061	0.0091	0.0690	0.0953	0.0682	0.0941
3	100	Even	0.0035	0.0082	0.0650	0.0908	0.0648	0.0905
2	50	Even	0.0172	0.0189	0.1013	0.1373	0.0994	0.1351
2	100	Even	0.0061	0.0091	0.0690	0.0953	0.0682	0.0941
2	200	Even	0.0020	0.0096	0.0705	0.0982	0.0692	0.0967
2	500	Even	0.0002	0.0201	0.1014	0.1418	0.1018	0.1424
2	100	Even	0.0061	0.0091	0.0690	0.0953	0.0682	0.0941
2	100	Linear	0.0079	0.0112	0.0792	0.1060	0.0695	0.0955
2	100	Exponential	0.0088	0.0132	0.0869	0.1151	0.0708	0.1404

that size present in the database. By examining the range of similarity values of the all-atoms graphs and the projected graphs of each cluster, we obtain a sense of the structural diversity present in the dataset. As discussed earlier, the set of all-atom graphs that correspond to a particular projected graph (or oxygen frame) can be highly dissimilar in relative energy. Correspondingly, the all-atoms graphs show a wide range of similarity values even with an oxygen frame family. Meanwhile, the similarity values of the projected graphs show variations in the oxygen frame families present in the database. Figure 5 shows the similarity plotted against the projected similarity of clusters of each cluster size N. The $N = 16$ group appears to have the largest amount of structural diversity according to the range of similarity and projected similarity values. Notably, although the $N = 25$ and $N = 26$ groups contain the largest number of unique clusters, their structural diversity is comparatively low. The exponential increase in unique digraphs for each oxygen frame may explain the reduced diversity as N increases, indicating that certain oxygen frame families are more energetically stable.

We next examine the topological diversity of clusters in the dataset. Figure 6 (top) shows a plot of the Wiener index vs the average shortest path length computed on the projected graphs of all clusters in the database. A clear pattern emerges in which each cluster size falls along a separate line with a distinct slope. Because these two metrics have similar forms, the slope is $2/N(N - 1)$, which is the inverse of the number of pairs in the system. Therefore, the slope decreases as the cluster size increases until reaching 0 for an infinite-sized system. The Wiener index is an extensive property that increases exponentially with cluster size (shown in the bottom right of Fig. 6), while the average shortest path length is not an extensive property and, in this system, appears to be converging to a maximum value (bottom left of Fig. 6). Networks are considered "regular" when each node is connected to a fixed number of nodes, a scenario that is assumed to be the case in large, low-temperature water clusters, in which each water molecule has a roughly tetrahedral arrangement. Small-world networks lie between random networks and regular networks; in other words, small-world networks are regular graphs in which an amount of disorder has been introduced.[50] Such systems have the high clustering characteristics of regular lattices but the small path lengths of random graphs. This plot appears to show characteristics of a small-world pattern in which the typical distance between two nodes (quantified by the average shortest path length) grows proportionally to the logarithm of the number of nodes in the network.

Another useful measure of connectivity in a graph is the mean degree. The degree is computed for each node in the graph, and the mean gives a single descriptor for the full graph. Figure 7 shows the mean degree for the projected graph of all clusters of a certain size, with the standard deviation indicated by the shaded region. The

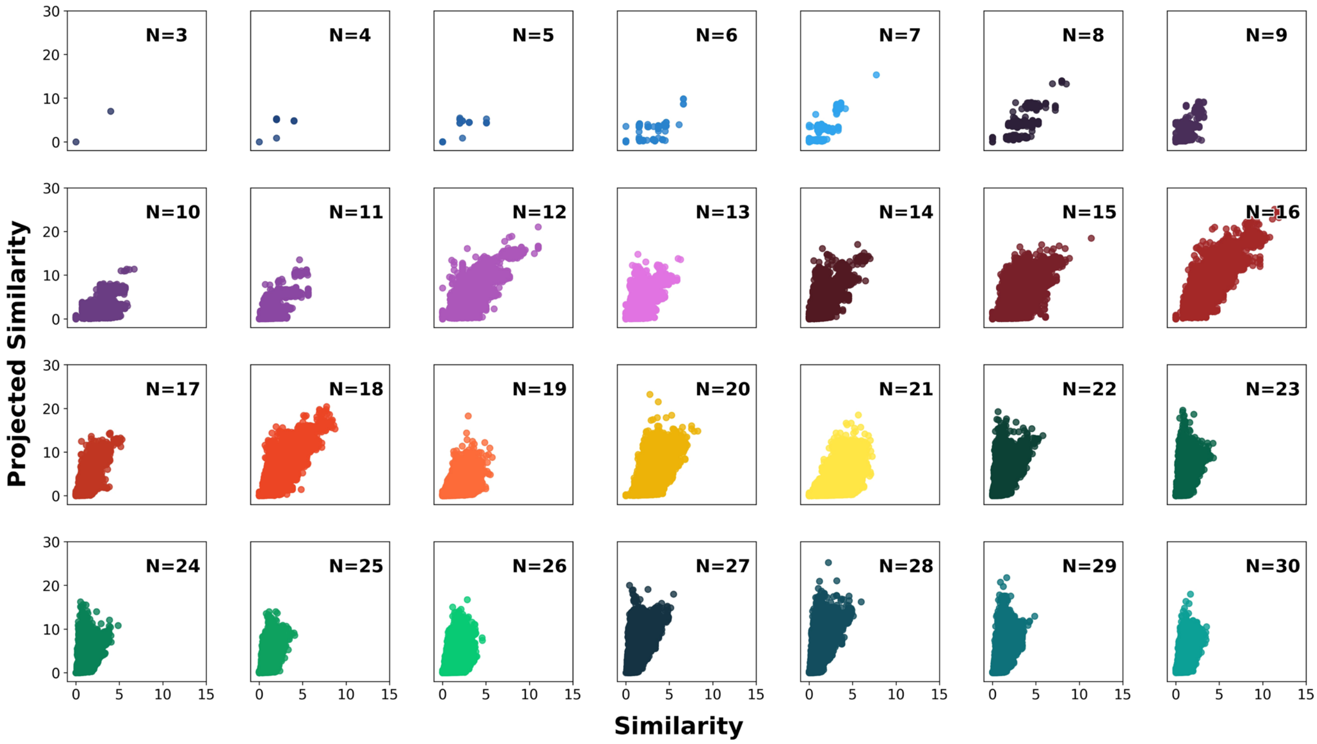


FIG. 5. Illustration of the structural diversity in the full database as described by the similarity and projected similarity for each cluster size N in the database. The values for each cluster are computed against the lowest-energy structure of the same size N. The similarity is computed for the all-atoms graphs, and the projected similarity is computed for the projected graph. All subplots share the same x- and y-axes for convenient visualization.

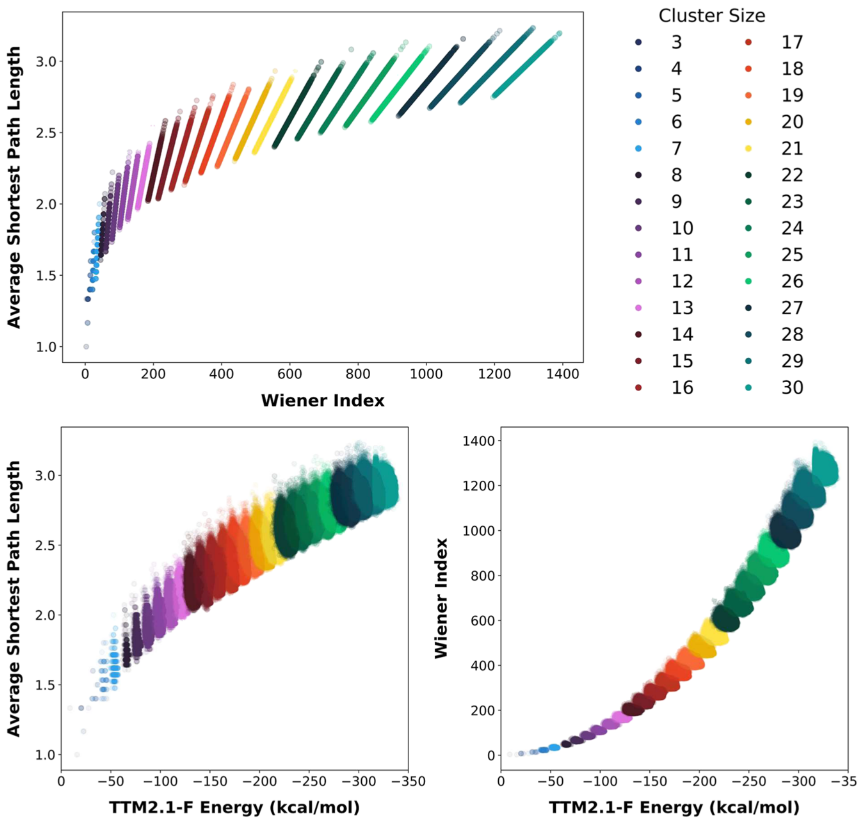


FIG. 6. Top panel: plot of the Wiener index vs the average shortest path length for the full database. Bottom panel: plots of the average shortest path length (left) and Wiener index (right) vs energy computed at the TTM2.1-F level. In all plots, each point is colored by cluster size and plotted with 0.2 opacity such that the color value represents the density of structures.

connectivity starts quite low due to the limited geometries available in small clusters and increases in a logarithmic fashion until around 3.5. This trend comports to the average connectivity in liquid water, which is ~3.8 hydrogen bonds per water molecule.[44] Interestingly, even though the database contains a greater number of large-size clusters, the standard deviation in the mean degree (denoted by the shaded region in Fig. 7) narrows as the cluster size increases. We believe that this is due to the increasingly "liquid-like" networks in large size clusters, which exhibit fully connected networks. The lowest energy structures of N = 3–5 have a degree of 2, following the known stability of the homodromic trimer, tetramer, and pentamer clusters.[51] After that cluster size, the mean degree of the 1% lowest-energy clusters increases and tends to have slightly higher-than-average values up to N = 18, after which the mean degrees of the 1% lowest-energy structures are similar to those of the full database. This indicates that increased connectivity plays a role in stabilizing water clusters—a fact that is well known but it is explicitly shown here through graph descriptors.

Finally, an interesting property of the water cluster networks is the number of polygons in the network. The number of trimers, tetramers, pentamers, and hexamers was quantified from the projected graph of each cluster. The mean and standard deviation of these values for each cluster size is plotted in Fig. 8. The mean number of trimers is quite low for all cluster sizes and further declines as N increases. The number of tetramers, pentamers, and hexamers increases with N as expected. Consistently, there are, on average, more pentamers than hexamers present in the clusters. To the best of our knowledge, the earliest study that quantifies the relative number of 5-membered vs 6-membered rings in liquid water is that of Rahman and Stillinger.[52] In that work, the number of each type of pentamer and hexamer rings was essentially the same, and

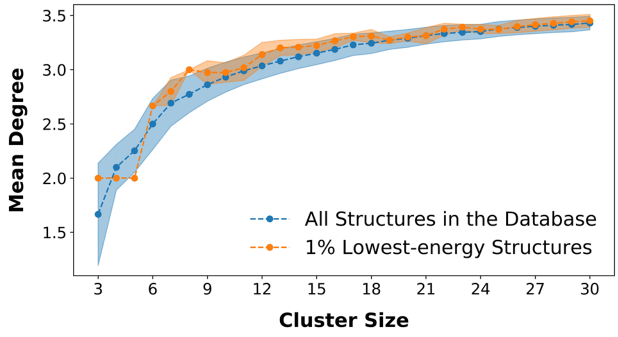


FIG. 7. Plot of the degree of the projected graph as the cluster size increases for the full database. The blue circles represent the mean degree, and the shaded area is the standard deviation. The orange diamonds mark the degree of the lowest-energy structure.

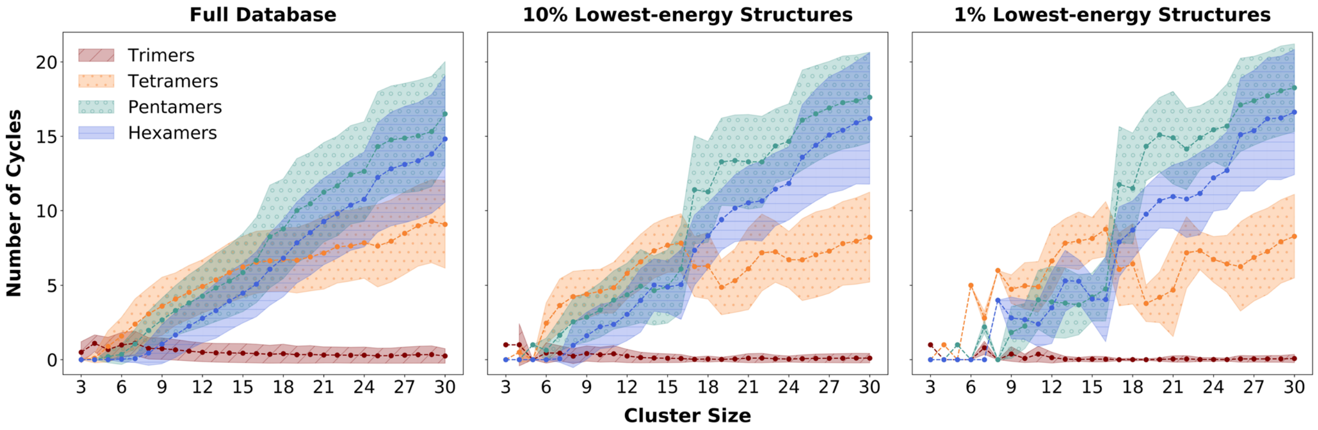


FIG. 8. Plots of the number of trimers, tetramers, pentamers, and hexamers per cluster size *N* in all clusters for the full database (left) and 10% of clusters with the lowest energy (middle) and 1% lowest-energy structures (right). The mean is given as the circular marker, and the shaded area represents the standard deviation.

which ring was the maximum depended on the chosen hydrogen-bond definition. Hexamers, rather than pentamers, are the dominant ring structure in ice. For low *N*, the clusters show a propensity for tetramers over both pentamers and hexamers. However, at $N = 17$, a distinct switch occurs, which is even more prominent when examining the 10% and 1% lowest-energy clusters. At this size, the propensity for tetramers decreases and their numbers grow at a slower rate than those of pentamers and hexamers as *N* increases. We observed that at $N = 17$, the clusters became more cage-like and highly symmetric structures were no longer the putative minima. This behavior is reflected in the number of cycles, as the added internal geometry in cage-like structures leads to the formation of additional five- and six-membered rings.

B. Optimization of the CF-CNN training

Before performing a large-scale training, we first explored the properties of the CF-CNN that affect its training, such as the number of atom-wise features, the strategy for sampling from the full database to create the training set, the variance in this sampling, and the variance in the network itself.

We found that training was highly sensitive to the number of atom-wise features used to describe the atomic environment. When too few features are used, the network does not have the capacity to learn the full system; conversely, if too many features are used, the network overfits to the training data and gives poor predictions on the test set. Table I shows training metrics when 50, 100, 200, or 500 features are used. The training loss decreases as the number of features increases, while the validation loss initially decreases and then increases as the network begins to overfit. This behavior is also reflected in the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) during the validation of the final epoch. The validation MAE and RMSE are 0.1013 and 0.1373 when 50 features are used, which decrease to 0.0690 and 0.0953 when the number of features increases to 100. Further increasing the number of features to 200 slightly increases the validation MAE and RMSE to 0.0705 and 0.0982, while greatly increasing the number of features to 500 increases the validation MAE and RMSE to 0.1014 and 0.1418, values that are higher than those when too few features are used.

We also examined the error in the predictions on the test set made by the networks trained with different numbers of features. The test MAEs and RMSEs follow the same pattern and are very close in value to the validation MAEs and RMSEs. Notably, the error distribution is wider when too few features are used (see Fig. S2). The distribution narrows, giving a mean of 0.0008 kcal/mol, when 100 features are used. The distribution remains narrow but shifts in the negative direction to −0.0007 when 200 features are used and widens when 500 features are used. This increase in error with 500 features again indicates that the network is overfitting to the training data when a large number of features are learned. Because the network trained using 100 features gave the best validation and test scores, we use 100 features in all further studies presented in this work.

Next, we examined the strategy for sampling from the database to build our training set. When building training sets for neural networks, the data must be well distributed over the entirety of the chemical space under examination; otherwise, the network will not have learned the behavior of systems in the absent region and will not produce suitable estimations of the potential energy in that region. As the cluster size increases past $N = 17$, the clusters begin to resemble cage-like structures, and the number of possible variants with an energy of less than 5 kcal/mol from the putative minimum increases. To examine the effect of the training set, we sampled from each cluster size bin using three different strategies: (i) evenly, (ii) linearly increasing as the cluster size increases, and (iii) exponentially increasing as the cluster size increases (see Table S1 for the exact count of clusters of each size used in each sampling strategy). Each strategy produced a training set of 100 000 clusters, divided in the manner discussed above. The test set of 10 500 clusters with 500 of each size was again used for the analysis of the sampling strategy.

As seen in the comparison of errors in Table I and Fig. S3, the method of evenly sampling from each cluster size provides the

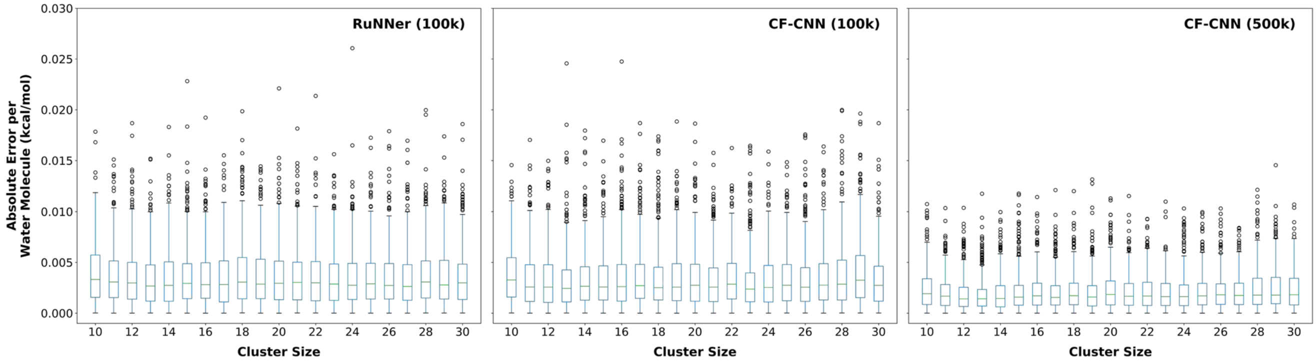


FIG. 9. Box and whisker plots showing the absolute error per water molecule in kcal/mol on the test set for the network trained using the even sampling strategy trained on 100 000 clusters using RuNNer (left), on 100 000 clusters using the CF-CNN (middle), and on 500 000 (right) clusters using the CF-CNN. Boxes extend from the lower to upper quartiles with a line at the median value. Bars extend to 1.5 times the interquartile range, with points beyond this being plotted individually to show outliers explicitly. All plots have the same *y*-axis for direct comparison.

smallest distribution of errors mostly centered around zero (0.0008 ± 0.09 kcal/mol) and the lowest test MAE and RMSE (0.0682 and 0.0941, respectively). The linear sampling method also gives a narrow distribution, but it is more offset toward overestimating the energy (−0.0029 ± 0.10 kcal/mol), and the test MAE and RMSE and are slightly larger (0.0695 and 0.0955, respectively). The exponential sampling method gives a wider distribution of errors with a similar offset toward overestimation (−0.0020 ± 0.14 kcal/mol), and a notably larger test MAE and RMSE (0.0708 and 0.1404, respectively). This trend is amplified in the examination of the largest outlying error: the maximum absolute error predicted by the network with evenly sampled data is 0.57 kcal/mol, that by the model with linearly sampled data is 1.19 kcal/mol, and that by the model with exponentially sampled data is 10.69 kcal/mol. By definition, the linear and exponential sampling strategies contain larger proportions of larger sized clusters. However, in our analysis of the full database, we found that structural diversity decreases as N increases, as measured by the similarity and the projected similarity metrics. This reduced structural diversity led to training sets sampled with these two strategies not covering as much of the structural space of interest as the training set sampled evenly, which is likely the cause of the increased error in the predictions of CF-CNNs trained on data sampled linearly and exponentially.

The colloquial "chemical accuracy" of *ab initio* methods is considered to be 1 kcal/mol. The TTM2.1-F potential has been shown to reproduce the total binding energies of clusters with sizes up to $N = 20$ within <1% from the second order perturbation theory (MP2) complete basis set estimates.[53,54] This is superior to a DFT-based result using density functionals that are customarily used for bulk water simulations. Notably, all predictions with the even sampling strategy were in excellent agreement with the values obtained with the TTM2.1-F potential, and both the linear and exponential sampling strategies produced only 1 prediction above 1 kcal/mol. In fact, using the even sampling strategy, 78% of predictions (8153 of 10 500) were within 0.10 kcal/mol of the computed value, 77% (8060 of 10 500) were when using the linear sampling strategy, and 76% (8014 of 10 500) were when using the exponential sampling strategy. Therefore, all three sampling strategies produce highly accurate results. For the remainder of our studies, we used the even sampling strategy, as the network trained using this strategy gave an error distribution most centered around zero and showed the lowest number of outliers.

The clusters in the training set were randomly chosen from the database; the only consideration was given to cluster size. Therefore, we also compared the results of training on three different random samplings using the same sampling strategy and training hyperparameters described above. Figure S4 shows the error on the test set for the three different CF-CNNs. All CF-CNNs give similar standard deviations of 0.09 kcal/mol–0.10 kcal/mol, but with slightly different average error values ranging from −0.0054 kcal/mol to 0.0025 kcal/mol. The test MAEs range from 0.0682 to 0.0717, while the test RMSEs range from 0.0941 to 0.0985. Therefore, the specific networks sampled from each cluster size have a noticeable effect on the trained CF-CNN. We also examined the variance in the CF-CNN itself by using a single training set but varying the random seed (see Fig. S5). Among three different seeds, the errors on the test set ranged from 0.0005 ± 0.10 kcal/mol to 0.0010 ± 0.10 kcal/mol. The test MAEs ranged from 0.0682 to 0.0729, while the test RMSEs ranged from 0.0941 to 0.1003. This shows that the training/validation split and the initial weights (two factors controlled by the random seed) also have a noticeable effect on CF-CNN training.

Figure 9 shows a box and whisker plot of the absolute error per water molecule in kcal/mol for each cluster size in the test set. The absolute error per water molecule provides a normalized view of the error, as the energy (thus the potential error) becomes larger in absolute value as the cluster size increases. As seen in the figure, the median absolute error per water molecule (denoted by green lines) ranges from 0.0024 kcal/mol to 0.0033 kcal/mol for each cluster size, and no general trend can be observed as the cluster size increases. The boxes extend from the lower quartile to the upper quartile of each cluster size, ranging from 0.0011 to 0.0055, and again no clear trend is observed as the cluster size increases. Outliers are present for each cluster size, with the two largest outliers of ~0.025 kcal/mol belonging to clusters of size $N = 13$ and 16. Again, no clear trend in the number or magnitude of outliers was observed as the cluster size increases. Typically, NNs perform poorly when extrapolating past the bounds of their training set. However, even though the CF-CNN was trained on clusters of size $N = 11–29$, the trained network was able to predict clusters of size $N = 10$ and 30 with similar accuracy. The lack of a trend in the error and the ability to accurately predict the energy of clusters of smaller and larger size than those in the training set indicate that the localized bonding pattern, rather than a global pattern, is being learned by the network. This idea is supported by the architecture of the CF-CNN, which learns an energy representation for each atom in the system and then sums over these representations to produce the global energy. Because we set the nearest-neighbor cutoff to 6 Å, it is reasonable to assume that the network is learning the interactions between water clusters within this range.

C. Comparison between RuNNer and SchNet

We next compare the results from the optimized CF-CNN with those obtained using the Behler–Parrinello RuNNer network. Figure S6 shows the test set errors for RuNNer trained on the training sets generated for the examination of the three sampling strategies described above. The even sampling strategy provided the best results, showing similar test errors to the CF-CNN trained on the same dataset (−0.0006 ± 0.10 kcal/mol, MAE of 0.0725, and RMSE of 0.0971). Again, the network generalized well to system sizes not included in the training set (see Fig. 9). The median absolute error per water molecule ranged from 0.0026 kcal/mol to 0.0033 kcal/mol for each cluster size (compared to 0.0024 kcal/mol–0.0033 kcal/mol for the CF-CNN), with the largest outlier of 0.0260 kcal/mol belonging to cluster size $N = 24$. Again, no general trend was observed as the cluster size increases. This favorable comparison indicates that the CF-CNN is capable of producing high-quality predictions of the potential energy. Notably, the errors achieved with both RuNNer and CF-CNN are smaller than the intrinsic accuracy of essentially any method for which reference training data could be generated.

D. Large-scale training

From the above analysis on models trained on 100 000 clusters, we determined the ideal training hyperparameters to be 2 interaction blocks, 100 atom-wise features, and an even sampling strategy. Using these optimized hyperparameters, we then undertook a large-scale training using a training set composed of half a million clusters with 26 316 clusters taken from each cluster size $N = 11–29$. Of this training set, 450 000 were used to learn the weights and 50 000 to validate those weights at each epoch. Again, a test set composed of 10 500 clusters with 500 taken from each cluster size $N = 10–30$ was used to evaluate the trained network.

Each epoch took ∼4.5 min, and the training required 873 epochs to converge. Overall, the network required 2.7 days to train. A final training loss of 0.0030 was achieved, and the final validation loss was 0.0035. The similar loss values indicate that the model was not overfitting. The validation MAE and RMSE were 0.0426 and 0.0591, respectively, while the test MAE and RMSE were 0.0427 and 0.0593, respectively. These values are improved over those of the optimal CF-CNN trained on 100 000 clusters. However, we note that the improvement scales less than linearly with the training set size, indicating that there may be a size limit after which the network will no longer appreciably improve.

Figure 9 shows the box and whisker plot of the absolute error per water molecule for each cluster in the test set. The mean absolute error per water molecule was 0.0021 ± 0.0018 kcal/mol for clusters of size $N = 11–29$, 0.0024 ± 0.0020 kcal/mol for $N = 10$, and 0.0023 ± 0.0019 kcal/mol for $N = 30$. These similar errors indicate that the network can accurately predict the energy of clusters with fewer or more atoms than contained in the clusters in the training set. The reduction in error again did not scale with the increase in the size of the training set. Although the training set was increased by 500%, from 100 000 to 500 000, the reduction in absolute error per water molecule was only 38%. It seems that a point of diminishing returns was reached in which the environment within a 6 Å radius of each atom was well learned and additional data did not add new information. Nonetheless, the network trained on 500 000 water clusters was able to predict the energy of hydrogen-bonded water clusters to a high degree of accuracy to the values obtained with the TTM2.1-F potential.

E. Interpretation of the CF-CNN predictions

To obtain a surrogate model that is generalized to water clusters of many different structures and sizes, adequate coverage of the configuration space must be achieved. We quantified the configuration space spanned by the 500 000 cluster structures in the training set by computing the similarity and projected similarity of the clusters in training set. The similarity is computed by considering each atom in the cluster as a node and each bond (covalent and/or hydrogen bond) as an edge, while the projected similarity is computed by considering each water molecule as a node and each hydrogen bond as an edge. In both cases, the Laplacian of the cluster under consideration is compared against the Laplacian of the lowest-energy cluster in the database, regardless of whether it was included in the training or the test set. The similarity value is unique to each cluster, while projected similarity values are the same for clusters with the same oxygen frame. Therefore, the projected similarity is a convenient metric to group clusters within the same oxygen frame family. Moreover, the similarity value can be determined within a single oxygen frame family to examine the diversity of the hydrogen atom arrangement.

Figure 10 shows the similarity and projected similarity of the clusters in the training and test sets for each cluster size. The black dots represent clusters in the training set, while the colored dots represent clusters in the test set, colored by the corresponding absolute error per water molecule. There are no black dots for cluster sizes $N = 10$ and 30 because clusters of those sizes were not included in the training set but were included in the test set. It is immediately clear that clusters with smaller N are associated with a larger amount of structural diversity, likely owing to the combinatorial increase in hydrogen positions within an oxygen frame family as N increases, as discussed in our analysis of the full database. The clusters in the test set were similar to those in the training set. Only a single cluster of $N = 26$ had a similarity value not covered by a cluster of that size in the training set, and its energy was

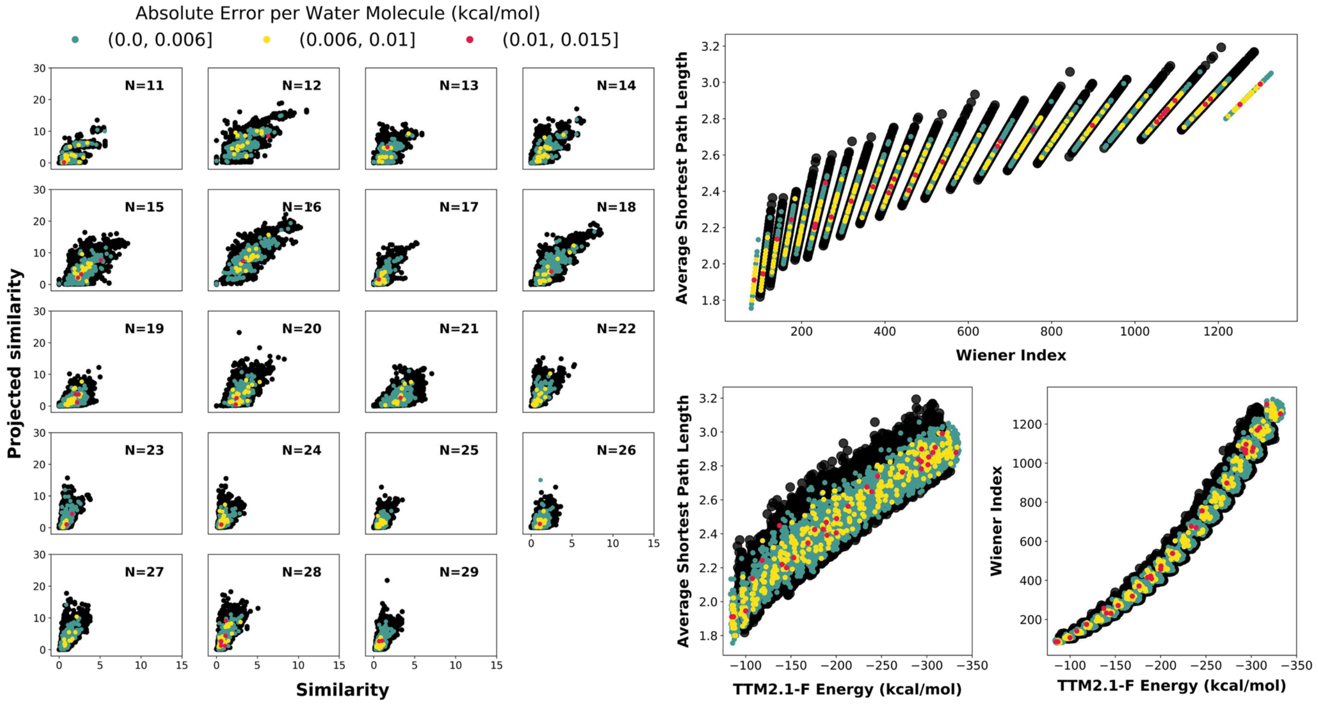


FIG. 10. Left: plots of similarity vs projected similarity for each cluster size for the training and test sets used for the CF-CNN trained on 500 000 and tested on 10 500 clusters. The similarity is computed on the all-atoms graph, while the projected similarity is computed on the projected graph. Right: plot of the Wiener index vs average shortest path length derived from the projected graph of clusters in the training and test sets along with plots of the Wiener index and average shortest path length vs energy computed at the TTM2.1-F level. Clusters in the training set are shown in black; clusters in the test set are colored according to the absolute error per water molecule (in kcal/mol) in the CF-CNN prediction.

well predicted by the trained CF-CNN. Interestingly, clusters with larger absolute error per water molecule tended to be closer in similarity to the lowest-energy cluster in the database. This indicates that our network is better at learning structures higher in energy than the putative minimum. This behavior is not entirely unexpected, as the database, and consequently the training set, contains a large number of structures within 5 kcal/mol from the putative minimum.

The average shortest path length and Wiener index are complementary topological metrics derived from the projected graph. Figure 10 shows these two metrics plotted against each other for clusters in the 500 000 cluster training set and the corresponding 10 500 cluster test set. The test set did not contain any clusters with a Wiener index or average shortest path length outside of the bounds of the training set. In the test set, 10 064 clusters had an absolute error per water molecule between 0 kcal/mol and 0.006 kcal/mol, 404 had errors between 0.006 kcal/mol and 0.010 kcal/mol, and 32 had errors between 0.010 kcal/mol and 0.015 kcal/mol. The mean average shortest path length of clusters in each error category slightly increased from 2.44 to 2.46 to 2.51 as the error increased. The mean Wiener index increased from 525 to 558 to 607 as the error increased. Larger errors tended to be located toward the middle of the range of values.

These two analyses indicate that clusters in the test set had similar structures and topological metrics as clusters in the training set. Therefore, the CF-CNN was exposed to a wide range of potential bonding structures in water cluster networks, and thus, all error analyses were interpolative in nature. The errors in prediction did not correlate with any similarity or topological metrics, indicating that the CF-CNN was able to accurately learn the examined area of configuration space.

We then examined the mean degree of each cluster size in the training and test sets, which are compared in Fig. 11. Clusters with errors of less than 0.006 kcal/mol had very similar mean degrees to those in the training set of the same cluster size. As the error increased to between 0.006 kcal/mol and 0.010 kcal/mol, the mean degrees began to deviate from those in the training set. Finally, the degrees of clusters with the largest errors of 0.010 kcal/mol–0.015 kcal/mol showed large deviations from those in the training set, occasionally lying outside of the standard deviation of the mean degrees in the training set for that particular cluster size.

We applied the same analysis to the mean number of cycles in the training and test sets, as shown in Fig. 11. We enumerated the number of trimers, tetramers, pentamers, and hexamers present in each cluster in the same manner as described in our analysis of the full database. A similar trend to that of the mean degree is seen in which the mean number of cycles for all cycle types is similar between clusters in the training set and test set clusters that showed the lowest errors. As the error increased, the mean number of cycles began to deviate from that of the training set, and when

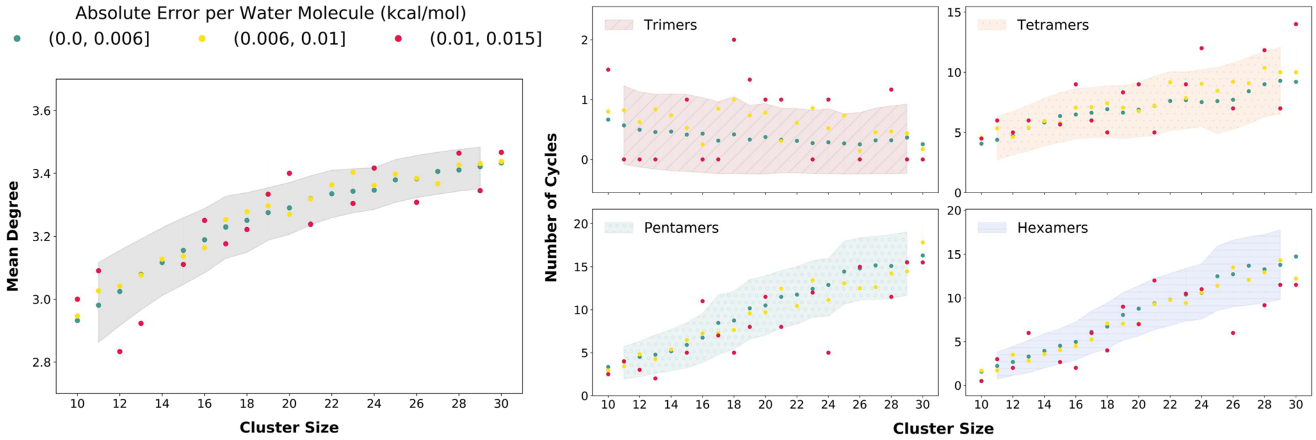


FIG. 11. Left: the mean degree calculated from the projected graph of clusters in the training and test sets used for the CF-CNN trained on 500 000 clusters. Right: the mean count of trimers, tetramers, pentamers, and hexamers in the training and test sets. The shaded regions shows one standard deviation of the mean for clusters in the training set; clusters in the test set are represented as points colored according to the absolute error per water molecule (in kcal/mol) in the CF-CNN prediction.

the error was further increased, the mean number of cycles deviated from that of the training set for all size cycles.

These results indicate that clusters that deviate from the mean of the training set are more poorly predicted by the CF-CNN than clusters that are more similar to those in the training set. Although we showed above that the configuration space represented by the training set encompasses that of the test set, meaning that the CF-CNN is not exposed to new environments when making predictions on the test set, the CF-CNN had lower predictive power for clusters that showed geometric deviations from the training set mean.

IV. CONCLUSIONS

We used 500 000 structures from a recently published database of over 5×10^6 unique water networks corresponding to local minima of cluster sizes $N = 3$–30 to train a neural network that predicts the potential energy with very high accuracy. In order to understand the structural features of the networks in the database, we examined the full database using descriptors derived from graph theory. We computed two similarity indexes, one for the all-atoms graph and the other for the projected graph, which together indicate the structural diversity present for each cluster size. We observed the largest diversity for $N = 16$, with the decrease in diversity as N increased, likely due to the combinatorial increase in all-atoms graphs within each oxygen frame family of projected graphs. Complementary measurements of connectivity, the average shortest path length, and mean degree showed an approach toward a maximum connectivity, which is supported in previous studies on ice and liquid water. We also observed the structural shift of more symmetric structures to more cage-like, fully connected structures at $N = 17$ by examining the number of cycles present in the cluster. At $N = 17$, the prevalence of tetramers decreased, while that of pentamers and hexamers increased.

The clusters in the database were used to train a CF-CNN to learn the potential energy of water clusters of various sizes. We first optimized the training hyperparameters using a 100 000 cluster subset from the database. For our dataset, 2 interaction blocks, 100 atom-wise features, and even sampling from each cluster size $N = 11$–29 gave the lowest error on a test set of 10 500 clusters of size $N = 10$–30. Multiple samplings from the database gave similar results, indicating that the sampling strategy rather than the actual sampling played a role during training of the CF-CNN. Varying the random seed, which affects the initial weights and training/validation split, had only a minor effect on training, indicating that the CF-CNN is stable. We then used the optimized hyperparameters to train a network on 500 000 clusters of size $N = 11$–29. In a test set of 10 500 clusters of size $N = 10$–30, the mean absolute error per water molecule was 0.0021 ± 0.0018 kcal/mol for clusters of size $N = 11$–29, 0.0024 ± 0.0020 kcal/mol for clusters of size $N = 10$, and 0.0023 ± 0.0019 kcal/mol for $N = 30$. These similar errors indicate that the network has the ability to accurately predict the energy of clusters with both fewer and more molecules than the ones contained in the clusters used in the training set.

The structural patterns of the clusters in the training and test sets were analyzed using the same graph-theoretical descriptors used in the analysis of the full database in order to provide a post-hoc interpretation of the predictive power of the trained CF-CNN. The range of similarity and topological indexes present in the training set encompassed those in the test set, indicating that the full range of configuration space in question was contained in the training set. The errors in the energy prediction were not found to depend on the similarity or topological metrics. However, structures with mean degrees and number of cycles that deviated from the mean values in the training set were associated with larger errors. This indicates that clusters that deviated from the mean of the training set—although they were within the configuration space learned by the CF-CNN—were less well learned than structures with values close to the mean. Therefore, the structural space covered by the training set influences the predictions of the resulting neural network and must be considered alongside the chemical space when developing training sets to encompass a compositionally diverse test set.

In summary, we developed descriptors to analyze the structural patterns of a large (over 5 × 10^6) database of water cluster minimum structures lying within 5 kcal/mol of the putative minima of the $N = 3$–30 clusters. Using up to 500 000 structures from this database, we demonstrated a scalable approach of training a CF-CNN model using large volumes of data. Finally, we validated the effectiveness of the developed descriptors in explaining the results of the CF-CNN (a.k.a. the "black box") model. To this end, our study should not be seen as a fit of an Machine Learning Potential (MLP) to the TTM2.1-F model, as only the minimum energy structures are used as part of the training set and the TTM2.1-F potential was only used to obtain the minimum energy structures because of its speed and accuracy. In principle, this procedure can be used with any other model, *ab initio* or classical. The work presented here provides the foundation for future analysis of the structural patterns presented in more complex hydrogen-bonded networks, such as liquid water and ice.

SUPPLEMENTARY MATERIAL

See the supplementary material for exact counts of clusters in the training sets generated by different sampling strategies and error plots generated during the optimization of the CF-CNN.

ACKNOWLEDGMENTS

The authors thank Professor Jörg Behler from the University of Göttingen for making the RuNNer code available to us. J.A.B., M.S., and S.C. were supported by the DOE Exascale Computing Project, ExaLearn Co-design Center. J.P.H. and S.S.X. acknowledge support from the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences. P.B. acknowledges DST-PURSE, UPE-II, and JNU for funding. Pacific Northwest National Laboratory (PNNL) is a multi-program national laboratory operated for the DOE by Battelle. This research also used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

DATA AVAILABILITY

The data that support the findings of this study are freely available at https://sites.uw.edu/wdbase/.[55]

REFERENCES

[1] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of artificial intelligence for computer-assisted drug discovery," Chem. Rev. **119**, 10520–10594 (2019).

[2] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," Science **361**, 360–365 (2018).

[3] C. W. Coley, W. H. Green, and K. F. Jensen, "Machine learning in computer-aided synthesis planning," Acc. Chem. Res. **51**, 1281–1289 (2018).

[4] J. Behler, "First principles neural network potentials for reactive simulations of large molecular and condensed systems," Angew. Chem., Int. Ed. **56**, 12828–12840 (2017).

[5] 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).

[6] N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network," J. Chem. Phys. **148**, 241715 (2018).

[7] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in Workshop at International Conference on Learning Representations, 2014.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), pp. 1135–1144.

[9] Q. Zhang, Y. Nian Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2018).

[10] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," J. Comput. Phys. **378**, 686–707 (2019).

[11] A. Rakshit, P. Bandyopadhyay, J. P. Heindel, and S. S. Xantheas, "Atlas of putative minima and low-lying energy networks of water clusters n = 3–25," J. Chem. Phys. **151**, 214307 (2019).

[12] E. Aprà, A. P. Rendell, R. J. Harrison, V. Tipparaju, W. A. deJong, and S. S. Xantheas, "Liquid water: Obtaining the right answer for the right reasons," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC'09* (ACM, New York, NY, USA, 2009), pp. 66-1–66-7.

[13] S. Yoo and S. S. Xantheas, "Structures, energetics and spectroscopic fingerprints of water clusters n = 2–24," in *Handbook of Computational Chemistry*, 2nd ed., edited by J. Leszczynski (Springer International Publishing Switzerland, 2017), Chap. 26, pp. 1139–1175.

[14] C. J. Burnham and S. S. Xantheas, "Development of transferable interaction models for water. IV. A flexible, all-atom polarizable potential (TTM2-F) based on geometry dependent charges derived from an *ab initio* monomer dipole moment surface," J. Chem. Phys. **116**, 5115–5124 (2002).

[15] G. S. Fanourgakis and S. S. Xantheas, "The flexible, polarizable, thole-type interaction potential for water (TTM2-F) revisited," J. Phys. Chem. A **110**, 4100–4106 (2006).

[16] G. S. Fanourgakis and S. S. Xantheas, "Development of transferable interaction potentials for water. V. Extension of the flexible, polarizable, thole-type model potential (TTM3-F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water," J. Chem. Phys. **128**, 074506 (2008).

[17] Y. Wang and J. M. Bowman, "Towards an *ab initio* flexible potential for water, and post-harmonic quantum vibrational analysis of water clusters," Chem. Phys. Lett. **491**, 1–10 (2010).

[18] 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 Comput. **10**, 2906–2910 (2014).

[19] G. S. Fanourgakis, G. K. Schenter, and S. S. Xantheas, "A quantitative account of quantum effects in liquid water," J. Chem. Phys. **125**, 141102 (2006).

[20] F. Paesani, S. Iuchi, and G. A. Voth, "Quantum effects in liquid water from an *ab initio*-based polarizable force field," J. Chem. Phys. **127**, 074506 (2007).

[21] F. Paesani, S. S. Xantheas, and G. A. Voth, "Infrared spectroscopy and hydrogen-bond dynamics of liquid water from centroid molecular dynamics with an *ab initio*-based force field," J. Phys. Chem. B **113**, 13118–13130 (2009).

[22] F. Paesani, S. Yoo, H. J. Bakker, and S. S. Xantheas, "Nuclear quantum effects in the reorientation of water," J. Phys. Chem. Lett. **1**, 2316 (2010).

[23] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, "Effective force fields for condensed phase systems from *ab initio* molecular dynamics simulation: A new method for force-matching," J. Chem. Phys. **120**, 10896–10913 (2004).

[24] A. Davtyan, J. F. Dama, G. A. Voth, and H. C. Andersen, "Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence," J. Chem. Phys. **142**, 154104 (2015).

[25] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).

[26] P. O. Dral, A. Owens, S. N. Yurchenko, and W. Thiel, "Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels," J. Chem. Phys. **146**, 244108 (2017).

[27] N. Bernstein, G. Csányi, and V. L. Deringer, "De novo exploration and self-guided learning of potential-energy surfaces," npj Comput. Mater. **5**, 99 (2019).

[28] T. Morawietz and J. Behler, "A density-functional theory-based neural network potential for water clusters including van der Waals corrections," J. Phys. Chem. A **117**, 7356–7366 (2013).

[29] S. Kondati Natarajan, T. Morawietz, and J. Behler, "Representing the potential-energy surface of protonated water clusters by high-dimensional neural network potentials," Phys. Chem. Chem. Phys. **17**, 8356–8371 (2015).

[30] C. Schran, F. Uhl, J. Behler, and D. Marx, "High-dimensional neural network potentials for solvation: The case of protonated water clusters in helium," J. Chem. Phys. **148**, 102310 (2018).

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

[32] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).

[33] A. Singraber, J. Behler, and C. Dellago, "Library-based lammps implementation of high-dimensional neural network potentials," J. Chem. Theory Comput. **15**, 1827–1840 (2019).

[34] A. Singraber, T. Morawietz, J. Behler, and C. Dellago, "Parallel multistream training of high-dimensional neural network potentials," J. Chem. Theory Comput. **15**, 3075–3092 (2019).

[35] C. Schran, J. Behler, and D. Marx, "Automated fitting of neural network potentials at coupled cluster accuracy: Protonated water clusters as testing ground," J. Chem. Theor. Comput. **16**(1) , 88–99 (2019).

[36] J. Behler, "Constructing high-dimensional neural network potentials: A tutorial review," Int. J. Quantum Chem. **115**, 1032–1050 (2015).

[37] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).

[38] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).

[39] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, "SchNetPack: A deep learning toolbox for atomistic systems," J. Chem. Theory Comput. **15**, 448–455 (2019).

[40] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms," J. Phys. Chem. A **101**, 5111–5116 (1997).

[41] A. Rakshit and P. Bandyopadhyay, "Finding low energy minima of $(H_2O)_{25}$ and $(H_2O)_{30}$ with temperature basin paving Monte Carlo method with effective fragment potential: New 'global minimum'and graph theoretical characterization of low energy structures," Comput. Theor. Chem. **1021**, 206–214 (2013).

[42] A. Rakshit, T. Yamaguchi, T. Asada, and P. Bandyopadhyay, "Understanding the structure and hydrogen bonding network of $(H_2O)_{32}$ and $(H_2O)_{33}$: An improved Monte Carlo temperature basin paving (MCTBP) method and quantum theory of atoms in molecules (QTAIM) analysis," RSC Adv. **7**, 18401–18417 (2017).

[43] P. Battaglia, J. B. C. Hamrick, V. Bapst, A. Sanchez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. Allen, C. Nash, V. J. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, "Relational inductive biases, deep learning, and graph networks," arXiv:1806.01261 (2018).

[44] R. Kumar, J. R. Schmidt, and J. L. Skinner, "Hydrogen bonding definitions and dynamics in liquid water," J. Chem. Phys. **126**, 204107 (2007).

[45] J.-L. Kuo, J. V. Coe, S. J. Singer, Y. B. Band, and L. Ojamäe, "On the use of graph invariants for efficiently generating hydrogen bond topologies and predicting physical properties of water clusters and ice," J. Chem. Phys. **114**, 2527–2540 (2001).

[46] A. M. Tokmachev, A. L. Tchougréeff, and R. Dronskowski, "Hydrogen-bond networks in water clusters $(H_2O)_{20}$: An exhaustive quantum-chemical analysis," ChemPhysChem **11**, 384–388 (2010).

[47] A. M. Tokmachev, A. L. Tchougréeff, and R. Dronskowski, "Benchmarks of graph invariants for hydrogen-bond networks in water clusters of different topology," in *Péter R. Surján* (Springer, 2016), pp. 157–164.

[48] S. Yoo, M. V. Kirov, and S. S. Xantheas, "Lowest energy networks of the T-cage $(H_2O)_{24}$ cluster and their use in constructing unit cells of the structure I (sI) hydrate lattice," J. Am. Chem. Soc. **131**, 7564 (2009).

[49] D. Koutra, A. Parikh, A. Ramdas, and J. Xiang, "Algorithms for graph similarity and subgraph matching," in *Proceedings of Ecology Inference Conference* (2011), Vol. 17.

[50] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature **393**, 440–442 (1998).

[51] S. S. Xantheas, "Cooperativity and hydrogen bonding network in water clusters," Chem. Phys. **258**, 225–231 (2000).

[52] A. Rahman and F. H. Stillinger, "Hydrogen-bond patterns in liquid water," J. Am. Chem. Soc. **95**, 7943–7948 (1973).

[53] T. H. Dunning, Jr., R. J. Harrison, D. Feller, and S. S. Xantheas, "Promise and challenge of high-performance computing, with examples from molecular modeling," Philos. Trans. R. Soc. London, Ser. A **360**, 1079 (2002).

[54] G. S. Fanourgakis, E. Aprà, and S. S. Xantheas, "High-level *ab initio* calculations for the four low-lying families of minima of $(H_2O)_{20}$. I. Estimates of MP2/CBS binding energies and comparison with empirical potentials," J. Chem. Phys. **121**, 2655 (2004).

[55] J. P. Heindel and S. S. Xantheas, "Database of water cluster minima," https://sites.uw.edu/wdbase/, 2019.