Interpretation and Evaluation of the Predictive Power of a Continuous-Filter Convolutional Neural Network using Graph-Theoretical Descriptors for Learning the Potential Energy Surface of Water Clusters

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

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

(Dated: 6 March 2020)

A continuous-filter convolutional neural network (CF-CNN) was trained to predict the potential energy of water cluster networks, $(H_2O)_N$, ranging in size from N=10–30 taken from a newly published database containing 4.2 million unique water cluster networks. To understand the chemical space from which we were sampling, we characterized all 4.2 million structures using chemical descriptors derived from graph theory, which led to the discovery of interesting trends in the topology, connectivity, and ring structures within the clusters. The collection of graphs is available alongside the original database at `https://sites.uw.edu/wdbase/`. The CF-CNN trained on a subset of 500,000 clusters gave a mean absolute error per water molecule of 0.002±0.002 kcal/mol, giving our trained CF-CNN the highest accuracy of any neural network-based surrogate model to date. In addition, clusters of sizes not included in the training set (N=10, 30) showed the same low error rate, indicating that the CF-CNN architecture generalizes well and can accurately predict energies of smaller- and larger-sized clusters than seen during training. We then returned to the graph-theoretical descriptors developed to analyze the full database to interpret the predictive power of our CF-CNN. Using topology measures of the Wiener index and average shortest path length along with two similarity measures, we showed that all clusters from the test set were within the range of clusters from the training set, meaning the training set well covered the chemical space of interest. We found that the mean degree and number of cycles of clusters with larger errors tended to lie further from the mean than those of clusters with lower errors. Overall, we show the generalizability of CF-CNNs and present a graph-theoretical method to interpret the results.

I. INTRODUCTION

The use of artificial intelligence (AI) for scientific applications has rapidly increased over the past decade. Neural networks, in particular, have proven useful for tasks 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 the generation of potentials that have the accuracy of higher-level computational methods but are orders of magnitude less expensive. However, training of such models requires a large amount of data generated at the desired level of accuracy, and the more general the model, the larger the coverage of chemical space required.

Scalability is a key aspect of deep learning, and advancements in parallel and distributed computing have been critical to enabling the deep learning revolution. Newly available data sources, such as the atlas of low-lying energy networks of water clusters produced by Xantheas and co-workers,[7] motivate the exploration of methods that are both scalable and highly accurate. Such exploration led us to SchNet,[8,9] a novel deep learning architecture for modeling quantum interactions in molecules. SchNet is based off the convolutional neural network (CNN) architecture, which is often the first choice for euclidean data sources such as images and videos. Scalability is a major benefit of CNNs, as much work concerning this issue has been done on both parallel and distributed computing platforms. To maintain rotational, translational, and ordering invariances not present in traditional CNN architectures, SchNet applies continuous-filter convolutional layers that accept the precise location of each atom without discretization (via mapping to a grid) and learns the behavior of chemical interactions through local correlations. This unique architecture is described as a continuous-filter convolutional neural network (CF-CNN).

We were motivated to examine the predictive power of the CF-CNN architecture by access to the large database of water clusters produced by Xantheas and co-workers.[7] Water dissolves more substances than any other liquid, earning it the moniker of "universal solvent", and the vast majority of reactions necessary for life occur in aqueous environments. Therefore, to gain a full understanding of chemistry in aqueous environments, it is first necessary to understand the properties of water. In particular, clusters of water molecules allow quantitative examination of the nature and magnitude of intermolecular interactions within a water network.[10] Over the years, many important insights have been gained through the application of high-level *ab initio* methods, but the computa-

[a)]Electronic mail: jenna.bilbrey@pnnl.gov.

[b)]Electronic mail: heindelj@uw.edu.

tional cost of these methods often prohibit the study of large systems over long time scales. A number of flexible, polarizable classical potentials have been developed for water clusters, which have been shown to accurately reproduce macroscopic properties of water.[11,12] As a complementary approach, Behler and co-workers have shown that neural networks can also be used to provide high-quality potentials for water clusters, which then allow simulations of bulk liquid water with DFT-level accuracy.[13–15]

We show that by training on 500,000 unique water clusters– to our knowledge, the largest molecular training set applied to neural networks to date–the CF-CNN was able to produce energies with a mean absolute error per water molecule of 0.002 ± 0.002 kcal/mol. The CF-CNN was also able to accurately produce the energy of clusters both smaller and larger in size than those included in the training set. We then examined the chemical space learned by the CF-CNN using chemical descriptors derived from graph theory. We found that the network learned the behavior of a variety of different structures present in the full dataset. Using similarity measures and topological indexes, we showed that the training set covered the desired area of chemical space. We also showed that the CF-CNN gave poorer predictions for clusters with graph-based properties further from the mean.

This paper is organized as follows. In Section II, we detail prior work using neural networks to solve chemistry problems, provide a description of the learning objective, and discuss message-passing neural networks and the specific CF-CNN implementation. In Section III, we discuss previous work concerning the potential energy surface (PES) of water, the new atlas of low-lying energy networks of water clusters produced by Xantheas and co-workers, the graph-theoretical descriptors used to analyze this database and the CF-CNN results, and the analysis of the full database using these descriptors. Optimization of the CF-CNN is presented in Section IV, followed by the results of training the optimal CF-CNN on 500,000 unique water clusters, and finally an analysis of these results using the graph-theoretical descriptors described previously. We conclude the paper in Section V.

II. NEURAL NETWORKS IN CHEMISTRY

A. Prior Work

Neural networks are gaining ground in the chemistry and materials science communities as surrogate models for computationally expensive quantum-chemical methods as well as generative models to produce new compounds with desired properties. The majority of methods share the common goal of finding an easy-to-evaluate representation of the PES of the system in question. Common neural network architectures can be broadly defined as either descriptor-based or structure-based. Comparable to quantitative structure–property relationship (QSPR) models, descriptor-based architectures rely on feature engineering through the selection of appropriate input features, which often take the form of molecular-level descriptors, such as crystal space group and HOMO/LUMO energy, or atom-level descriptors, such as the ionization potential and Pauling electronegativity of atomic components.[16–18] Defining the proper features requires substantial domain knowledge, and features must be redefined for each problem. Structure-based architectures, in contrast, do not rely on feature engineering but simply consider the molecular structure itself. The molecular structure can be defined in a number of ways. Some researchers have found success using SMILES strings.[19–22] However, You et al. showed that molecular graphs generated from the atom connectivity are more robust than SMILES representations when implemented in a policy network for reinforcement learning.[23] Graphs also have the advantage of representing molecular substructures, such as functional groups, through partial generation of the molecular graph. However, property prediction from molecular graphs is only applicable to equilibrium structures, as only the connectivity, and not the atomic positions, are needed to create the graph.

Some structure-based neural network architectures accept input comparable to that of standard *ab initio* packages, such as nuclear charges (often defined by the atom type) and atomic coordinates. For this, the relevant symmetries must be obeyed in the representation if a neural network is to accurately learn the behavior of the system. As recognized in previous works, Cartesian coordinates alone are not a good representation for use with neural networks, as they change in value with translation and rotation. Noé et al. demonstrated the sampling of equilibrium states for a many-body system using a neural network, but found that training the network using the Cartesian coordinates led to unrealistic structures and large energies.[24] To solve this issue, they employed an internal coordinate system that is invariant rotation and translation. However, the choice of internal coordinate system is not unique, and choosing rational internal coordinates can become very difficult as the system size increases.

To overcome this difficulty, Behler and Parrinello transformed the Cartesian coordinates using a set of functions for each atom that obey the relevant physical symmetries.[25] With this unique symmetry description of the system, they were able to predict the energy and forces several orders of magnitude more quickly than with DFT at comparable accuracy. In 2011, Behler introduced atom-centered symmetry functions specifically for constructing high-dimensional neural networks that are applicable to a variety of systems, such as molecules, crystalline and amorphous solids, and liquids.[26] The resulting RuNNer code is available from Prof. Behler for training potential energy surfaces for general systems.[4,27] The RuNNer code does not parallelize training across multiple CPUs or run on GPUs, which could make training slow for large datasets. However, in 2017, Smith et al. introduced a neural network called Accurate NeurAl networK engINe for Molecular Energies (ANAKIN-ME or ANI for short) using a modified version of Behler's symmetry functions, which allows parallelization and can be run on GPUs.[5]

RuNNer has been applied to both water clusters and protonated water clusters. A water cluster potential was developed by training on approximately 40,000 dispersion-corrected DFT reference computations, which resulted in errors on the

order of 2 meV/H_2O.[13] More recently, RuNNer has been used to construct neural network-based PESs for protonated water clusters based on DFT reference data[14] and based on coupled-cluster reference data.[15,28] 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 choose to use RuNNer as a benchmark for accuracy in this study.

Schütt et al. built upon Behler's atom-centered symmetry function approach by using continuous-filter convolutional layers to model interactions between atoms, creating an CF-CNN architecture nicknamed SchNet.[8,9,29] Atom-based filters are formed from the vectors between atoms and their neighbors, which provides a unique internal representation that incorporates symmetry invariances. The filters are learned during training of the network, and each atom is represented by an array filters. The learned filters then act as features from which the continuous-filter 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. Multiple interaction layers can be stacked, and because the filter generator is contained within the interaction layer, the learned features will be different for each layer. The atom-wise contributions are then summed to give the value of the desired molecular-level property. SchNet 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, SchNet is able to pick up subtle changes in structure to produce continuous energy surfaces. Notably, SchNet has been shown to be applicable to a number of chemical properties, such as the absolute energy, HOMO/LUMO energies, heat of formation, and Gibbs free energy. Systems under periodic boundary conditions can also be represented with this filter-generation method due to the linearity of the convolution. Here, we use the codebase from[9] to train a CF-CNN that can predict the potential energy of water clusters.

B. Learning Objective

The many-body expansion is a well established way to decompose the energy of an atomistic system as an interaction between N_b atoms or molecules.

$$E(S) = \sum_{i=1}^{N_b} E^{(1)}(\mathbf{r}_i) + \sum_{i<j}^{N_b} E^{(2)}(\mathbf{r}_i,\mathbf{r}_j) + \sum_{i<j<k}^{N_b} E^{(3)}(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k) + \ldots \tag{1}$$

Alternatively, the total energy can be expressed as a summation over the contribution from N_b chemical environments, where *chemical environment* refers to an atom or a molecule. This can be accomplished by reordering the above equation as follows:

$$E(S) = \sum_{i=1}^{N_b} \Big[E^1(\mathbf{r}_i) + \frac{1}{2} \sum_{i\neq j}^{N_b} E^{(2)}(\mathbf{r}_i,\mathbf{r}_j) + \tag{2}$$

$$\frac{1}{3} \sum_{i\neq j}^{N_b} \sum_{i\neq k, j\neq k}^{N_b} E^{(2)}(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k) + \ldots \Big] \tag{3}$$

$$= \sum_{i=1}^{N_b} E_i(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_{N_b}) \tag{4}$$

The energy contributions E_i can be computed from the many-body interaction terms described above or approximated by a function. As an example of the former, Behler and Parrinello expressed radial functions for a given an atom i as the sum of the contribution of every neighboring atom j as follows:

$$G_i^2 = \sum_{j\neq i}^{N_b} e^{-\zeta(r_{ij}-r_s)^2} f_c(r_{ij}) \tag{5}$$

$$f_c(r_{ij}) = \begin{cases} \frac{1}{2} cos\left(\frac{\pi r_{ij}}{r_c} + \frac{1}{2}\right) & \text{for } r_{ij} \leq r_c \\ 0 & \text{for } r_{ij} \leq r_c \end{cases} \tag{6}$$

Using these functions, a neural network is trained for each element in the system to predict energy contribution from each atom (E_i), the collection of which are then summed to obtain the total energy. Thus, the energy contributions are (potentially hidden or latent) state representations learned from the training dataset. Notably, neural networks can learn the geometry of the underlying system through multiple layers or iterations in the learning process, as discussed next.

C. Continuous-Filter Convolutional Neural Networks

This section briefly describes the continuous-filter convolutional neural network developed by[8]. The computation graph in this architecture has two steps. The first step learns a atom-level representation of the energy function, followed by an aggregation step that combines all atom-level energies to predict the energy of the overall structure.

INTERACTION LAYER The atom-wise representation learning is expressed as an iterative computation similar to ones employed by message-passing neural networks. Each atom is assigned a hidden state representation referred to as h_v^t (for atom v, iteration t),

$$h_v^{t+1} = h_v^t + \sum_{w\in N(v)} f_{aggr}(v,w) \tag{7}$$

Where f_{aggr} is a learned differentiable function, and referred to as the continuous-convolution filter or filter-generation network component in[8]. The above computation is visually described in Figure 1. Given an atom v, we first identify it's neighbors within a cut-off (denoted as $N(v)$). The filter-generation network function is implemented as $f_{aggr}(v,w) = h_w \circledast W(\mathbf{r}_v - \mathbf{r}_w)$

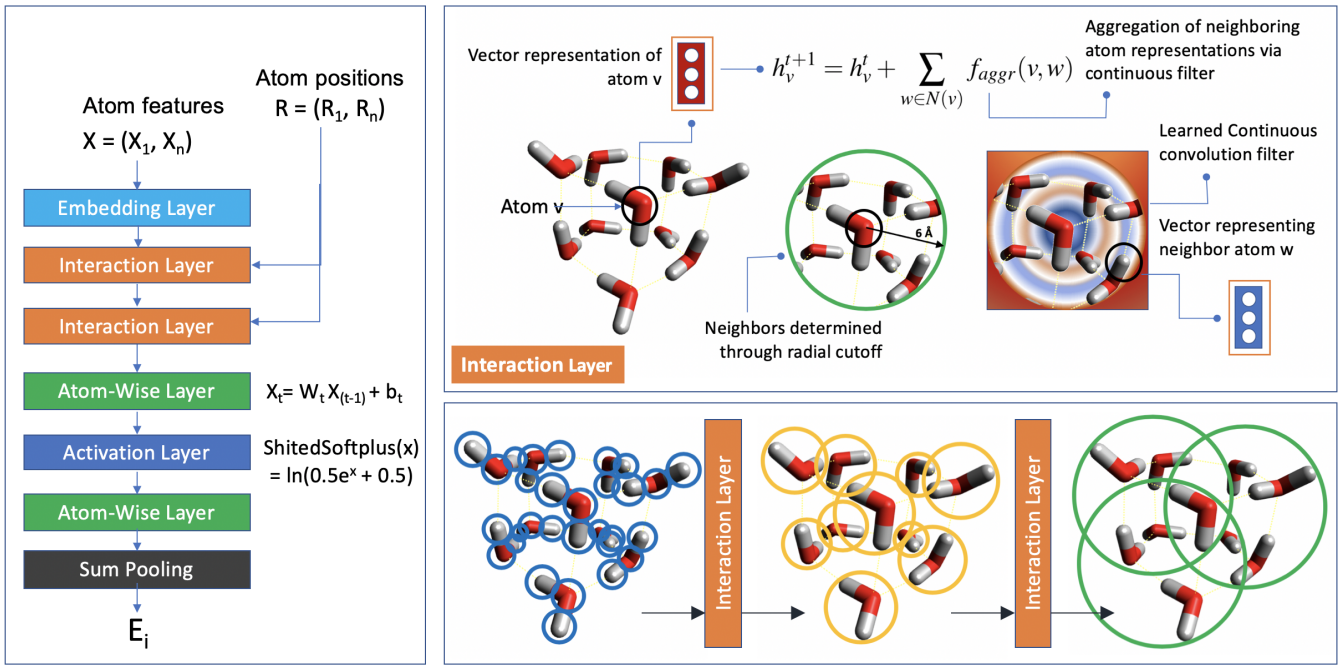

FIG. 1. Illustration of an atom's neighborhoods examined by the CF-CNN[8] during learning. From these neighborhoods arises a representation of the geometry of the underlying system.

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 fixed interval ranging from 0Å - r_{cutoff}Å.

ATOM-WISE REPRESENTATION LEARNING Finally, the interaction blocks are repeated a number of times (specified by a parameter N_{iter}, which allows the representation of an atom to be updated through propagating the influence of other atoms that are more than r_{cutoff} away. Figure 1 illustrates how this propagation expands in an atomistic system through each repetition of the interaction layer. We refer the reader to[8] for a description of the specific configurations of each layer in the neural architecture.

III. GRAPH-THEORETICAL DESCRIPTORS FOR WATER CLUSTERS

A. Prior Studies of Water

The local structure of water networks has long been studied in contexts ranging from clusters to liquid to ice. The natural approach to studying aqueous systems is to examine the connectivity of water molecules defined by the presence or lack of hydrogen bonds, which provides great detail about the local structure. The particular definition of a hydrogen bond has been explored thoroughly, and different methods appear to give similar results in most applications. Here, we use the definition from Kumar et al. in which the hydrogen bond is parameterized by a distance r and 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.[30] This definition comports with the idea that electron density is donated from an O−H bond into a π^* orbital of the receiving water molecule.

In an early study concerning the connectivity of liquid water that is still relevant today, Rahman and Stillinger quantified the presence of "non-short-circuited polygons" in the structure.[31] They demonstrated that liquid water is networked by hydrogen bonds forming large-size polygons and that individual molecules have the potential to be simultaneously connected to more than four hydrogen-bond partners. Though not explicitly stated at the time, this view of the structure of liquid water is, in essence, an analysis of a particular class of sub-graphs of the graph defining a snapshot of the structure of liquid water.

In fact, the use of graphs to represent water clusters has become increasingly common. Typically, a graph describing a water cluster is a directed graph (or digraph) in which the edges specify not only the connectivity but also the direction of each hydrogen bond. This type of graph fully details the structure of a water cluster up to the direction of the dangling hydrogen atoms (which, in our experience, almost always point out in space). Conversely, one could instead focus on the graph determined by the oxygen frame itself. This is a projection of the digraph described above. To be clear, we refer to the graph determined by the oxygen frame as the projected graph and the graph that contains individual hydrogen atoms as the all-atoms graph (Fig. 2).

Both of these graph representations have been used successfully in the last few decades for various applications. For instance, projected graphs can be employed for the complete enumeration of all digraphs corresponding to a particular oxygen frame. This calculation has been performed for all 30,026 hydrogen-bond networks of the pentagonal dodecahedron structure of $(H_2O)_{20}$.[32–34] Notably, this revealed 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 40 kcal/mol.[32,33] 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 increase 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. Although, due to the density of this manifold of states, it is not clear that the true global minimum is of utmost importance, as many structures will be thermally populated.

In global searches of water clusters, graphs can be used to condense the representation of water clusters such that one no longer needs to chronicle large lists of Cartesian or internal coordinates. Attempts to use the graph representation within the search itself have been undertaken. For example, a graph corresponding to a cluster of size N can be used to provide an initial guess structure for a cluster of size $N-1$.[35,36] Because one of the most difficult parts of the search is sifting through the many millions of hydrogen-bond arrangements, some authors include explicit "topology-altering" steps in their search for low-energy water clusters.[37,38] This type of search process is particularly important for locating certain low-energy structures. It follows that the explicit use of graph-theoretical descriptors can further improve our understanding of structure–property correlations as well as greatly assist in global optimization problems.

B. Generation of the Database

Xantheas and co-workers recently published a database of water clusters of sizes N=3–30 derived from the PES of the flexible, polarizable Thole-Type Model (TTM2.1-F, version 2.1) interaction potential for water using Monte Carlo temperature basin paving (MCTBP).[7] This potential is a fully many-body, flexible, and polarizable potential with parameters derived from high-level *ab initio* calculations.[39–41]

MCTBP is a global optimization method that aims to improve the convergence rate of global optimization techniques such as basin hopping,[42] and thorough discussions of the algorithmic details can be found elsewhere.[36,43] 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 that takes us 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

$$\min\left(1, \exp(-\beta_{\text{old}}(E_{\text{new}} - E_{\text{old}}))\right). \tag{8}$$

This method of searching 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.

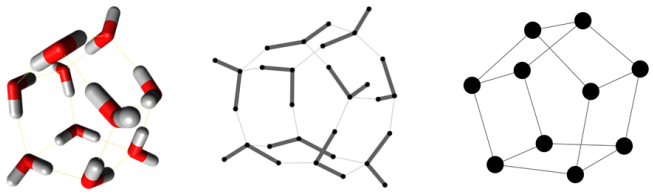

FIG. 2. Structure (left), all-atoms graph (middle), and projected graph (right) for the lowest-energy water cluster of size N=10.

C. Specific Graph-Theoretic Descriptors

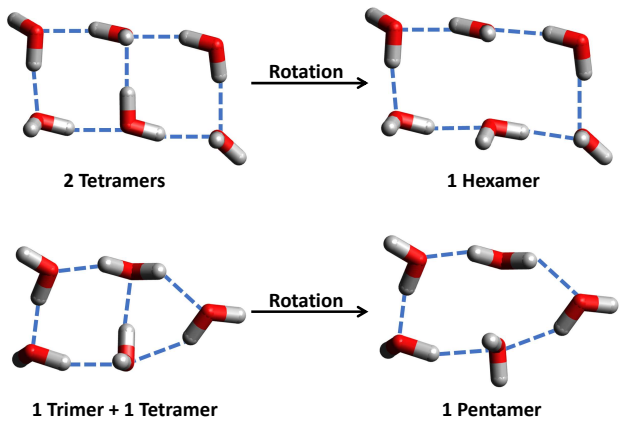

FIG. 3. Visual explanation of the definition of rings in a cluster. Fused rings (shown by the molecules on the left-hand side) are counted as the smallest component rings. In this work, we counted the number of 3–6-membered rings in the projected graph of each cluster.

Neural networks depend on exposure to numerous examples of the desired subset of chemical space to predict the behavior of new systems in that space. It follows that as the chemical space in question grows, the number of structures in the training set must also grow. The chemical space examined here is defined by hydrogen-bonded clusters of water molecules, which often show rich structural diversity. To gain an understanding of the chemical space of these water clus-

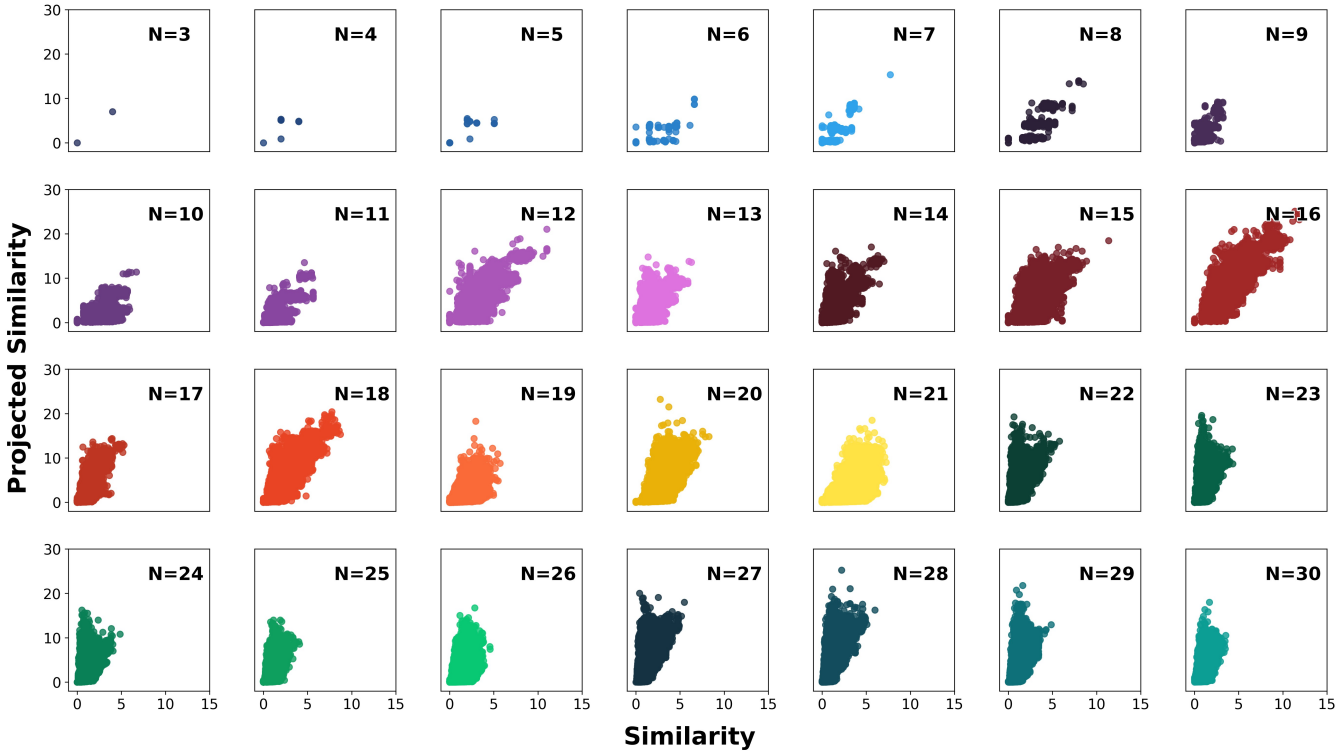

FIG. 4. Illustration of diversity in the training data as described by the similarity and projected similarity for each cluster size N in the database. The values for each cluster are computed against the lowest-energy structure of the same size N in the database. The similarity is computed for the all-atoms graphs, and the projected similarity is computed for the projected graph. All subplots share the same x- and y-axes for convenient visualization.

ters, we first characterize the full database using the graph-theoretical descriptors described below to gain an understanding of the chemical space covered by the database. The collection of graphs is available alongside the original database, which contains Cartesian coordinates and relative energies, at `https://sites.uw.edu/wdbase/`.

Each cluster is examined by both the all-atoms graph and the projected graph (see Fig. 2). Explicitly, in the all-atoms graph, each atom is a node and edges are represented by the covalent and hydrogen bonds present in the cluster; in the projected graph, each water molecule is a node and edges are represented by the hydrogen-bond network. The natural invariances present in chemistry (translational, rotational, and atom ordering) correspond to invariance under isomorphism of the corresponding graph.[44] 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 3D information is lost upon conversion to a graph-based representation.

Graph-based representations allow the speedy 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 one neighbor 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.[45] This metric relies on the eigenvalue (λ) of the laplacian of each graph, where the laplacian of the graph is 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

$$s = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2, \tag{9}$$

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, ∞), 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, though 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 the average shortest path length and 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.[7] 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.

We are also able to compute geometric shapes present in the water clusters through 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).

D. Characterization of the Full Database

The similarity of each graph of a certain cluster size N was computed against the lowest-energy cluster of that size present in the database. Examining the range of similarity values of the all-atoms graphs and the projected graphs of each cluster gives a sense of the structural diversity present in the dataset. As discussed above, the set of all-atoms 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 4 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,

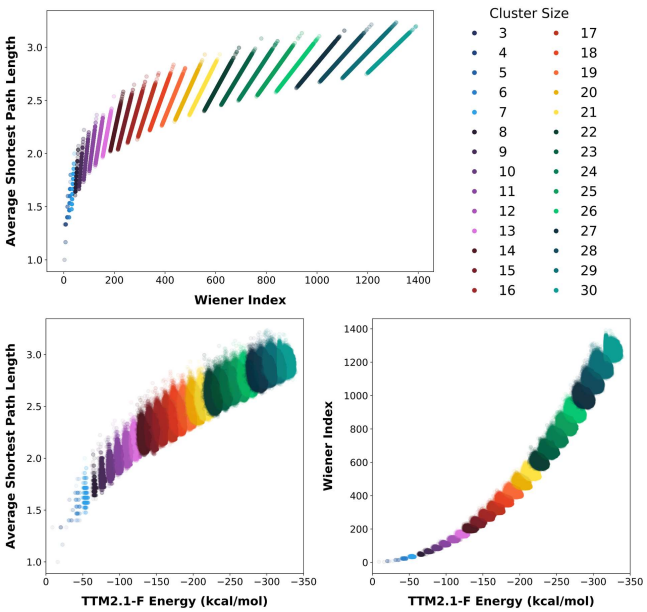

FIG. 5. (top) Plot of the Wiener index versus the average shortest path length for the full database. (bottom) Plots of the average shortest path length (left) and Wiener index (right) versus energy computed at the TTM2.1-F level. In all plots, each point is colored by cluster size and plotted with 0.2 opacity, such that the color value represents the density of structures.

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.

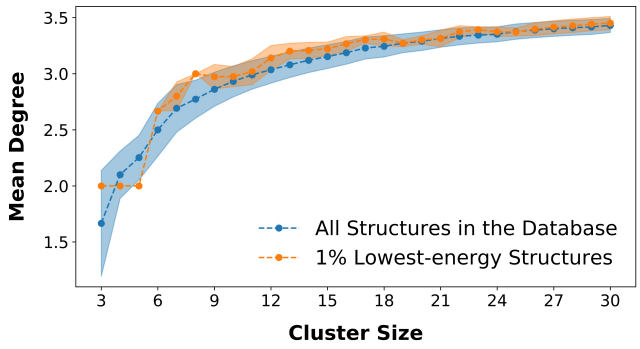

FIG. 6. Plot of the degree of the projected graph as the cluster size increases. The blue circles represent the mean degree, and the shaded area is the standard deviation. The orange diamonds mark the degree of the lowest-energy structure.

We next examined the topological diversity of clusters in the dataset. Figure 5 (top) shows a plot of the Wiener index versus average shortest path length computed on the projected graphs of all clusters in the database. A clear structure 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 notably 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. 5), 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. 5). Networks are considered "regular" when each node is connected to a fixed number of nodes, which is assumed to be the case in large, low-temperature water systems, where each water molecule has roughly tetrahedral symmetry. 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.[46] 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 small-worldedness, 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 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 6 shows the mean degree for the projected graph of all clusters of a certain size, with the standard deviation shown by the shaded region. The connectivity begins quite low due to the limited geometries available to small clusters and increases in a logarithmic fashion until around 3.5. This trend comports to the average connectivity in liquid water, which is approximately 3.8 hydrogen bonds per water molecule[30]. Interestingly, even though the database contains a greater number of large-size clusters, the standard deviation in the mean degree (shown by the shaded region in Fig. 6) narrows as the cluster size increases. We believe this is due to the increasingly cage-like structure of large-size clusters, which tend to show fully connected water molecules in the internal region. The lowest-energy structures of N=3–5 have a degree of 2, following the known stability of homodromic trimer, tetramer, and pentamer clusters.[47] The mean degree of the 1% lowest-energy clusters then rises and tends to have slightly higher than average values until 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 is explicitly shown here through graph descriptors.

Another interesting property of water cluster networks is the presence of rings. The number of trimers, tetramers, pentamers, and hexamers were quantified from the projected graph of each cluster. The mean and standard deviation of these values for each cluster size is plotted in Figure 7. The mean number of trimers is low for all cluster sizes and further declines as N increases. The number of tetramers, pentamers, and hexamers increase as N increases, as expected. Consistently, there are, on average, more pentamers than hexamers present in the clusters. To our knowledge, the only study which quantifies the relative number of 5-membered versus 6-membered rings in liquid water is that of Rahman and Stillinger[31], in which the number of each type of ring is essentially the same, and which ring is a maximum depends on the chosen hydrogen-bond definition. In ice, hexamers are the dominant ring structure, rather than pentamers. At 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 minimum. 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.

IV. DATA-DRIVEN PES OF WATER CLUSTERS

A. CF-CNN Setup

SchNet implementations are provided in both Tensorflow[8] and Pytorch[9] 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. Here, we use SchNetPack to train all CF-CNNs.

To optimize the model, we examined training parameters 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 approximately 100,000 water clusters 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 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 parameters, 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,[8,29] 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 end of the second solvation shell in liquid water based on the O−O radial distribution function. Beyond this distance, water molecules are essentially de-correlated, indicating that their interactions are negligible. The batch size was set to 50, and the maximum number of epochs was set to 7,500, though all trainings 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 parameters 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.

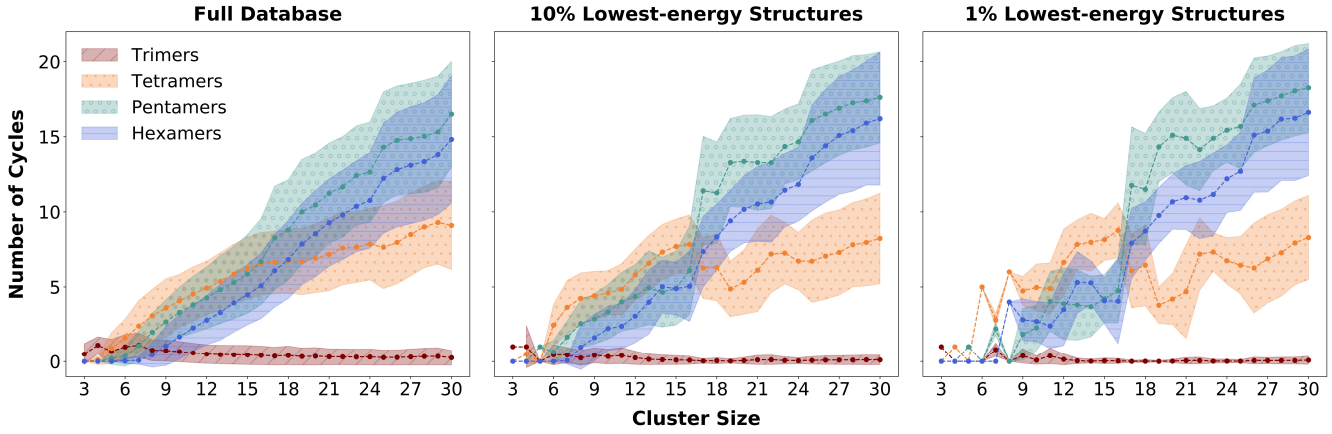

FIG. 7. Plots of the number of trimers, tetramers, pentamers, and hexamers per cluster size N in all clusters for the full database (left) and 10% of clusters with the lowest energy (middle), and 1% lowest-energy structures (right). The mean is given as the circular marker and the shaded area represents the standard deviation.

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 approximately 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) during validation of the final epoch are given, along with the mean error and standard deviation of the test set predictions.

Interaction blocks	Atom-wise features	Sampling strategy	Training loss	Validation loss	MAE	RMSE	Test error
1	100	even	0.02807	0.02903	0.12820	0.17039	-0.0055±0.17
2	100	even	0.00612	0.00908	0.06896	0.09527	0.0008±0.09
3	100	even	0.00351	0.00824	0.06499	0.09075	-0.0017±0.09
2	50	even	0.01722	0.01885	0.10127	0.13729	-0.0013±0.14
2	100	even	0.00612	0.00908	0.06896	0.09527	0.0008±0.09
2	200	even	0.00196	0.00964	0.07045	0.09820	-0.0007±0.10
2	500	even	0.00017	0.02011	0.10136	0.14182	0.0001±0.14
2	100	even	0.00612	0.00908	0.06896	0.09527	0.0008±0.09
2	100	linear	0.00793	0.01123	0.07920	0.10598	-0.0029±0.10
2	100	exponential	0.00876	0.01324	0.08693	0.11507	-0.0020±0.14

B. Optimization of CF-CNN Training

Before performing a large-scale training to learn the optimal PES, we first explored properties of the CF-CNN that affect 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 validation of the final epoch. The MAE is 0.10127 when 50 features are used and decreases to 0.06896 when increasing the number of features to 100. Further increasing the number of features to 200 slightly increases the MAE to 0.07045, while greatly increasing the number of features to 500 increases the MAE to 0.10136, which is higher than that 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 (see Fig. S2). Similar to the validation MAEs, the distribution in the errors is wider when too few features are used. 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 PES 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: evenly, linearly increasing as the cluster size increased, or exponentially increasing as the cluster size increased (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 analysis of the sampling strategy.

As seen in the comparison of errors in Table I and Figure S3, the method of evenly sampling from each cluster size provides the smallest distribution of errors most centered around zero (0.0008$\pm$0.09 kcal/mol). The linear sampling method also gives a narrow distribution, but is more offset towards overestimating the energy (-0.0029$\pm$0.10 kcal/mol). The exponential sampling method gives a wider distribution of errors with a similar offset towards overestimation (-0.0020$\pm$0.14 kcal/mol). In addition, the model trained with data exponentially sampled from the database has the largest outlying error: the maximum absolute error predicted by the network with evenly sampled data was 0.57 kcal/mol, that by the model with linearly sampled data was 1.19 kcal/mol, and that by the model with exponentially sampled data was 10.69 kcal/mol. The linear and exponential sampling strategies, by definition, contain larger proportions of larger sized clusters. However, in our analysis of the full database, we found that structural diversity decreased as N increased, as measured by the similarity and projected similarity. This reduced structural diversity led to training sets sampled with these two strategies not covering as much of the chemical 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 non-*ab initio* methods is considered to be 1 kcal/mol. Therefore, if our network produces estimates of the energy of less than 1 kcal/mol from the predictions from the TTM2.1-F potential, we can say chemical accuracy was achieved. Notably, all predictions with the even sampling strategy were within 1 kcal/mol of the computed value, 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 (8,153 of 10,500) were within 0.1 kcal/mol of the computed value, 77% (8,060 of 10,500) were when using the linear sampling strategy, and 76% (8,014 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 chosen from the database randomly; 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 parameters 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 0.009 to 0.10 kcal/mol, but with slightly different average error values, ranging from -0.0054 to 0.0025 kcal/mol. Therefore, the specific clusters 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$\pm$0.10 kcal/mol to 0.0010$\pm$0.10 kcal/mol. This shows that the training/validation split and the initial weights (two factors controlled by the random seed) do not greatly affect training of the CF-CNN.

Figure 8 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 (and 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 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 can be observed as the cluster size increases. Outliers can be observed in each cluster size, with the two largest outliers of approximately 0.025 kcal/mol belonging to clusters of size N=13 and 16. Again, no clear trend in the number or magnitude of outliers can be observed as the cluster size increases. Typically, neural networks are poor at 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 equivalent accuracy. The lack of 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 the network is learning the interactions between water clusters within this range.

We next compared the results from the optimized CF-CNN

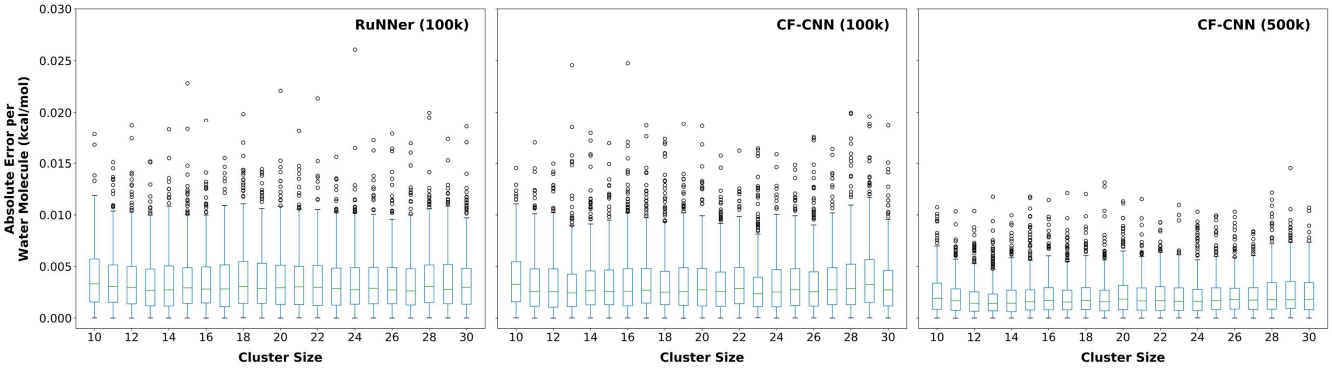

FIG. 8. Box and whisker plots showing the absolute error per water molecule in kcal/mol on the test set for the network trained using the even sampling strategy trained on 100,000 clusters using RuNNer (left), on 100,000 clusters using the CF-CNN (middle), and on 500,000 (right) clusters using the CF-CNN. 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.

with those obtained using the Behler-Parinello 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 errors to the CF-CNN trained on the same dataset (-0.0006±0.10 kcal/mol). Again, the network generalized well to system sizes not included in the training set (see Fig. 8). The median absolute error per water molecule ranged from 0.0026 to 0.0033 kcal/mol for each cluster size (compared to 0.0024 to 0.0033 kcal/mol for the CF-CNN), with the largest outlier of 0.026 kcal/mol belonging to cluster size N=24. Again, no general trend was observed as the cluster size increased. This favorable comparison indicates that the CF-CNN is capable of producing a high-quality PES. Namely, the errors achieved with both RuNNer and the CF-CNN are smaller than the intrinsic accuracy of essentially any method for which reference training data could be generated. Notably, the CF-CNN is easily parallelized across multiple GPUs, which allows us to train this network on a much larger portion of the database.

C. Large-Scale Training

Using the optimized parameters discussed above, 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 approximately 4.5 minutes, and the training required 873 epochs to converge. Overall, the network required 2.7 days to train. A final training loss of 0.00297 was achieved, and the final validation loss was 0.00349. The similar loss values indicate that the model was not overfitting. The MAE of the validation set was 0.04261, and the RMSE was 0.05906. 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 training set size, indicating that there may be a size limit after which the network will no longer appreciably improve.

Figure 8 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. Though 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 TTM2.1-F potential.

D. Interpretation of the CF-CNN Predictions

To obtain a surrogate model that is generalizable to water clusters of many different structures and sizes, adequate coverage of the chemical space must be achieved. We quantified the chemical space of our 500,000 cluster 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 clusters as a node and each bond (covalent and hydrogen) 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 lapla-

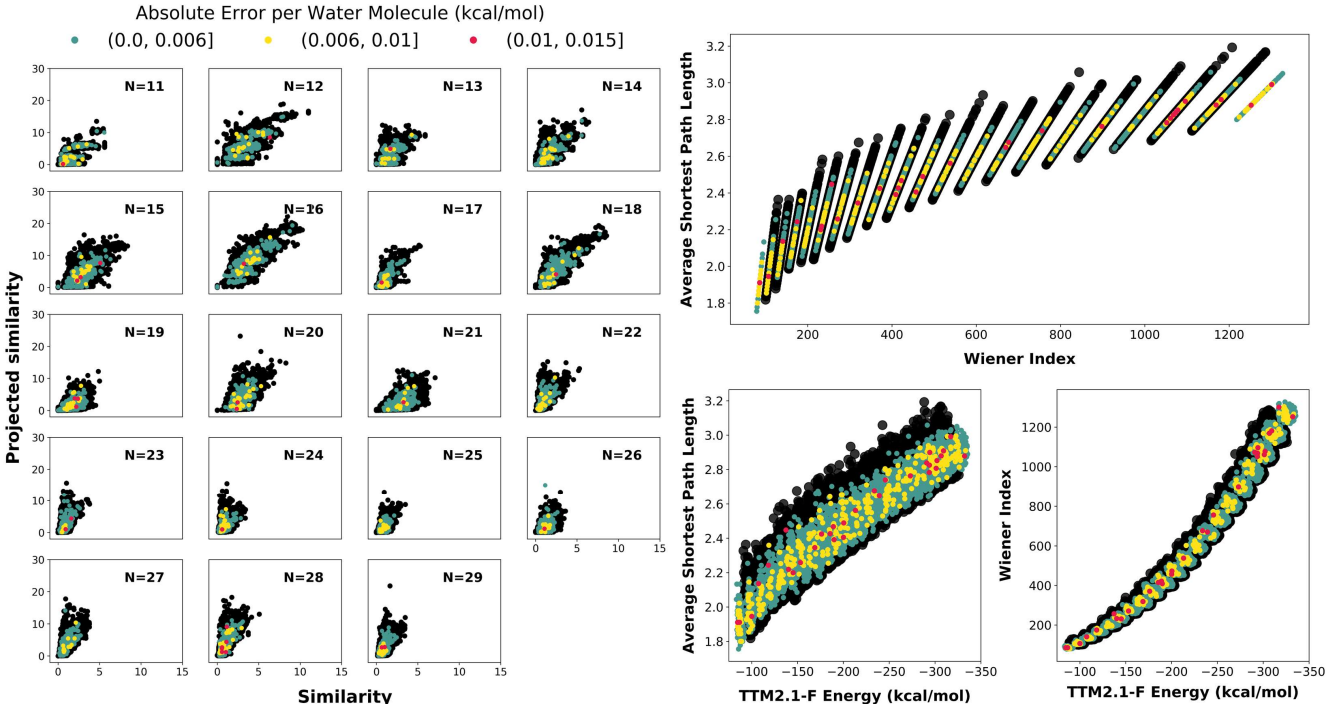

FIG. 9. (left) Plots of similarity versus projected similarity for each cluster size for the training and test sets used for the CF-CNN trained on 500,000 and tested on 10,500 clusters. The similarity is computed on the all-atoms graph, while the projected similarity is computed on the projected graph. (right) Plot of the Wiener index versus 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 versus 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.

cian of the lowest-energy cluster in the database, regardless of whether it was included in the training or 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 9 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 encapsulate a larger amount of structural diversity, likely due 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 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 wholly unexpected, as the database, and thus 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 9 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 and 0.006 kcal/mol, 404 had errors between 0.006 and 0.010 kcal/mol, and 32 had errors between 0.010 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 to any similarity or topological metrics, indicating that the CF-CNN was able to

accurately learn the examined area of chemical space.

We then examined the mean degree of each cluster size in the training and test sets, which are compared in Figure 10. 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 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 0.010–0.015 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 Figure 10. 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 the error was further increased, the mean number of cycles highly deviated from that of the training set for all size cycles.

These analyses 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. Though we showed above that the chemical space represented by the training set encompasses that of the test set, meaning 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.

V. CONCLUSION

We leveraged a database of 4.2 million unique water clusters of size N=3–30 recently published by Xantheas and coworkers to train a neural network with very high accuracy. First, to understand the data, we examined the full database using descriptors derived from graph theory. We computed two similarity indexes, one for the all-atoms graph and one for the projected graph, which together indicate the structural diversity present for each cluster size. We observed the largest diversity for N=16, with decreasing 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 the mean degree, showed an approach towards a maximum connectivity, which is supported in previous works on ice and liquid water. We also observed the structural shift of more symmetric structures to more cage-like 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.

We then trained a CF-CNN to learn the PES of water clusters of various sizes. We first optimized the training parameters 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 larger role during training of the CF-CNN. Varying the random seed, which affects the initial weights and training/validation split, did not have a large effect on training, indicating that the CF-CNN is stable. We then used the optimized parameters 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. This similar errors indicates that the network has the ability to accurately predict the energy of clusters with both fewer and more atoms than contained in clusters in the training set.

Clusters in the training and test sets were analyzed using the same graph-theoretical descriptors used in the analysis of the full database. The range of similarity and topological indexes present in the training set encompassed those in the test set, indicating that the full range of chemical space in question was contained in the training set. Errors in the energy prediction did not depend on similarity or topological metrics; therefore, we can say our network accurately learned the chemical space of hydrogen-bonded water clusters. We observed that structures with mean degrees and number of cycles that deviated from the mean values in the training set tended to have larger errors, indicating that clusters that deviated in from the mean of the training set–though they were within the chemical space learned by the CF-CNN–were less well learned than more average-values structures.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

We wish to thank Prof. Jörg Behler from the University of Göttingen for making the RuNNer code available to us. J.A.B., S.C., and M.S. were supported by the DOE Exascale Computing Project, ExaLearn Co-design Center. J.P.H. and S.S.X. acknowledge support from the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences. Pacific Northwest National Laboratory (PNNL) is a multiprogram national laboratory operated for 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. PB??

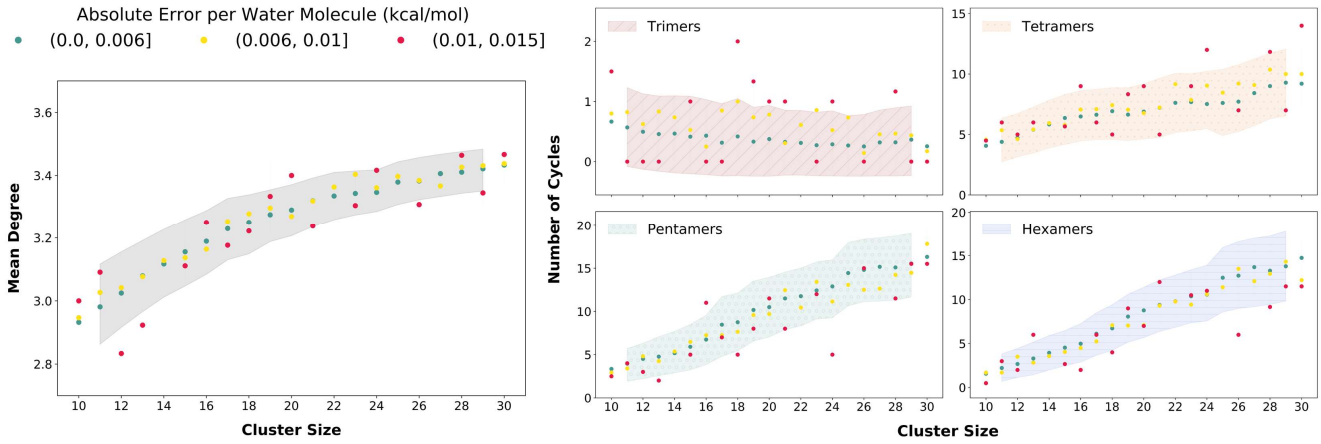

FIG. 10. (left) The mean degree calculated from the projected graph of clusters in the training and test sets used for the CF-CNN trained on 500,000 clusters. (right) The mean count of trimers, tetramers, pentamers, and hexamers in the training and test sets. The shaded regions shows one standard deviation 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.

AIP style for references–remove titles

REFERENCES

[1] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of artificial intelligence for computer-assisted drug discovery," Chemical Reviews **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), https://science.sciencemag.org/content/361/6400/360.full.pdf.

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

[4] J. Behler, "First principles neural network potentials for reactive simulations of large molecular and condensed systems," Angewandte Chemie International Edition **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," The Journal of Chemical Physics **148**, 241715 (2018), https://doi.org/10.1063/1.5011181.

[7] 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," The Journal of Chemical Physics **151**, 214307 (2019), https://doi.org/10.1063/1.5128378.

[8] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet – a deep learning architecture for molecules and materials," The Journal of Chemical Physics **148**, 241722 (2018), https://doi.org/10.1063/1.5019779.

[9] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, "Schnetpack: A deep learning toolbox for atomistic systems," Journal of Chemical Theory and Computation **15**, 448–455 (2019), https://doi.org/10.1021/acs.jctc.8b00908.

[10] 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.

[11] G. S. Fanourgakis, G. K. Schenter, and S. S. Xantheas, "A quantitative account of quantum effects in liquid water," The Journal of Chemical Physics **125**, 141102 (2006), https://doi.org/10.1063/1.2358137.

[12] F. Paesani, S. Iuchi, and G. A. Voth, "Quantum effects in liquid water from an ab initio-based polarizable force field," The Journal of Chemical Physics **127**, 074506 (2007), https://doi.org/10.1063/1.2759484.

[13] T. Morawietz and J. Behler, "A density-functional theory-based neural network potential for water clusters including van der waals corrections," The Journal of Physical Chemistry A **117**, 7356–7366 (2013).

[14] 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).

[15] C. Schran, F. Uhl, J. Behler, and D. Marx, "High-dimensional neural network potentials for solvation: The case of protonated water clusters in helium," The Journal of Chemical Physics **148**, 102310 (2018), https://doi.org/10.1063/1.4996819.

[16] K. Choudhary, B. DeCost, and F. Tavazza, "Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape," Phys. Rev. Materials **2**, 083801 (2018).

[17] J. Im, S. Lee, T.-W. Ko, H. W. Kim, Y. Hyon, and H. Chang, "Identifying pb-free perovskites for solar cells by machine learning," npj Computational Materials **5**, 37 (2019).

[18] M. L. Agiorgousis, Y.-Y. Sun, D.-H. Choe, D. West, and S. Zhang, "Machine learning augmented discovery of chalcogenide double perovskites for photovoltaics," Advanced Theory and Simulations **2**, 1800173 (2019), https://onlinelibrary.wiley.com/doi/pdf/10.1002/adts.201800173.

[19] E. J. Bjerrum and R. Threlfall, "Molecular generation with recurrent neural networks (rnns)," CoRR **abs/1705.04612** (2017), arXiv:1705.04612.

[20] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, and A. Aspuru-Guzik, "Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (organic)," (2017).

[21] R. Gomez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernandez-Lobato, B. Sanchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," ACS Cent Sci **4**, 268–276 (2018).

[22] M. H. S. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," ACS Central Science **4**, 120–131 (2018), pMID: 29392184, https://doi.org/10.1021/acscentsci.7b00512.

[23] J. You, B. Liu, R. Ying, V. Pande, and J. Leskovec, "Graph convolutional policy network for goal-directed molecular graph generation," in *Proceedings of the 32Nd International Conference on Neural Information Process-*

ing Systems, NIPS'18 (Curran Associates Inc., USA, 2018) pp. 6412–6422.
[24]F. Noé, S. Olsson, J. Köhler, and H. Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," Science **365** (2019), 10.1126/science.aaw1147, https://science.sciencemag.org/content/365/6457/eaaw1147.full.pdf.
[25]J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[26]J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," The Journal of Chemical Physics **134**, 074106 (2011), https://doi.org/10.1063/1.3553717.
[27]J. Behler, "Constructing high-dimensional neural network potentials: A tutorial review," International Journal of Quantum Chemistry **115**, 1032–1050 (2015).
[28]C. Schran, J. Behler, and D. Marx, "Automated fitting of neural network potentials at coupled cluster accuracy: Protonated water clusters as testing ground," arXiv preprint arXiv:1908.08734 (2019).
[29]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nature Communications **8**, 13890 (2017).
[30]R. Kumar, J. Schmidt, and J. Skinner, "Hydrogen bonding definitions and dynamics in liquid water," The Journal of Chemical Physics **126**, 05B611 (2007).
[31]A. Rahman and F. Stillinger, "Hydrogen-bond patterns in liquid water," Journal of the American Chemical Society **95**, 7943–7948 (1973).
[32]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," The Journal of Chemical Physics **114**, 2527–2540 (2001).
[33]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).
[34]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.
[35]R. Shrivastava, A. Rakshit, S. Shanker, L. Vig, and P. Bandyopadhyay, "A combination of monte carlo temperature basin paving and graph theory: Water cluster low energy structures and completeness of search," Journal of Chemical Sciences **128**, 1507–1516 (2016).
[36]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 Advances **7**, 18401–18417 (2017).
[37]S. Kazachenko and A. J. Thakkar, "Improved minima-hopping. tip4p water clusters,$(H_2O)_n$ with $n \leq 37$," Chemical Physics Letters **476**, 120–124 (2009).
[38]S. Kazachenko and A. J. Thakkar, "Water nanodroplets: Predictions of five model potentials," The Journal of Chemical Physics **138**, 194302 (2013).
[39]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," The Journal of Chemical Physics **116**, 5115–5124 (2002).
[40]G. S. Fanourgakis and S. S. Xantheas, "The flexible, polarizable, thole-type interaction potential for water (ttm2-f) revisited," The Journal of Physical Chemistry A **110**, 4100–4106 (2006).
[41]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," The Journal of Chemical Physics **128**, 074506 (2008).
[42]D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," The Journal of Physical Chemistry A **101**, 5111–5116 (1997).
[43]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," Computational and Theoretical Chemistry **1021**, 206–214 (2013).
[44]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 (2018).
[45]D. Koutra, A. Parikh, A. Ramdas, and J. Xiang, "Algorithms for graph similarity and subgraph matching," in *Proc. Ecol. Inference Conf*, Vol. 17 (2011).
[46]D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature **393**, 440–442 (1998).
[47]S. S. Xantheas, "Cooperativity and hydrogen bonding network in water clusters," Chemical Physics **258**, 225 – 231 (2000).

SUPPORTING INFORMATION

Interpretation and Evaluation of the Predictive Power of a Continuous-Filter Convolutional Neural Network using Graph-Theoretical Descriptors in Learning the Potential Energy Surface of Water Clusters

Jenna A. Bilbrey,[1*] Joseph Heindel,[2*] Sutanay Choudhury,[1] Malachi Schram,[1] Pradipta Bandyopadyay,[3] and Sotiris S. Xantheas,[1,2]

[1)] *Pacific Northwest National Laboratory, 902 Battelle Boulevard, P.O. Box 999, Richland, Washington 99352, USA.*
[2)] *Department of Chemistry, University of Washington, Seattle, WA, 98195, USA.*
[3)] *School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India.*

J.A.B.: jenna.bilbrey@pnnl.gov; J.H.: heindelj@uw.edu

TABLE S1. Number of structures from each cluster size used in the training set for each sampling strategy. From the full set, 90,000 clusters were used for training and the remainder for validation.

	Sampling Strategy		
*C*luster size	Even	Linear	Exponential
11	5263	1384	530
12	5263	1815	646
13	5263	2246	786
14	5263	2677	957
15	5263	3108	1165
16	5263	3539	1418
17	5263	3970	1726
18	5263	4401	2102
19	5263	4832	2559
20	5263	5263	3115
21	5263	5694	3792
22	5263	6125	4616
23	5263	6556	5620
24	5263	6987	6841
25	5263	7418	8328
26	5263	7849	10139
27	5263	8280	12343
28	5263	8711	15026
29	5263	9142	18292
TOTAL	99997	99997	100001
train	90000	90000	90000
validation	9997	9997	10001

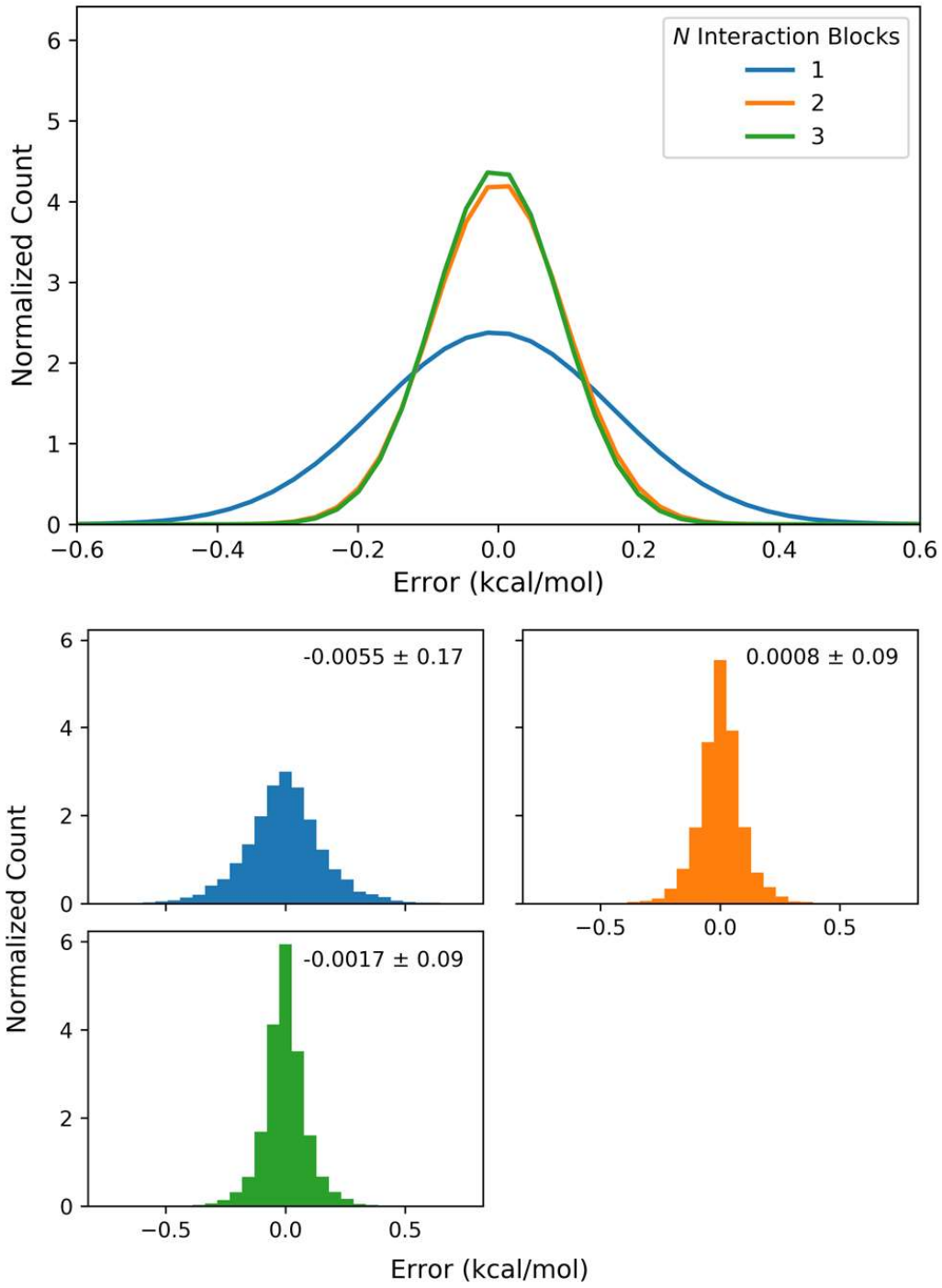

FIG. S1. Comparison of errors (in kcal/mol) produced by the CF-CNN for different numbers of interaction blocks. The mean and standard deviation of each distribution is listed in the upper right of the respective histogram. Each network was trained on 99,997 water clusters, with 5,263 water clusters taken from each cluster size (N=11–29). The test set consisted of 10,500 water clusters, with 500 taken from each cluster size (N=10–30). Though the developers of SchNet recommend using 3 interaction blocks, we find that in this case 2 interaction blocks is adequate. Training with 3 interaction blocks took approximately 20 hours, compared with approximately 16 hours when training with 2 interaction blocks.

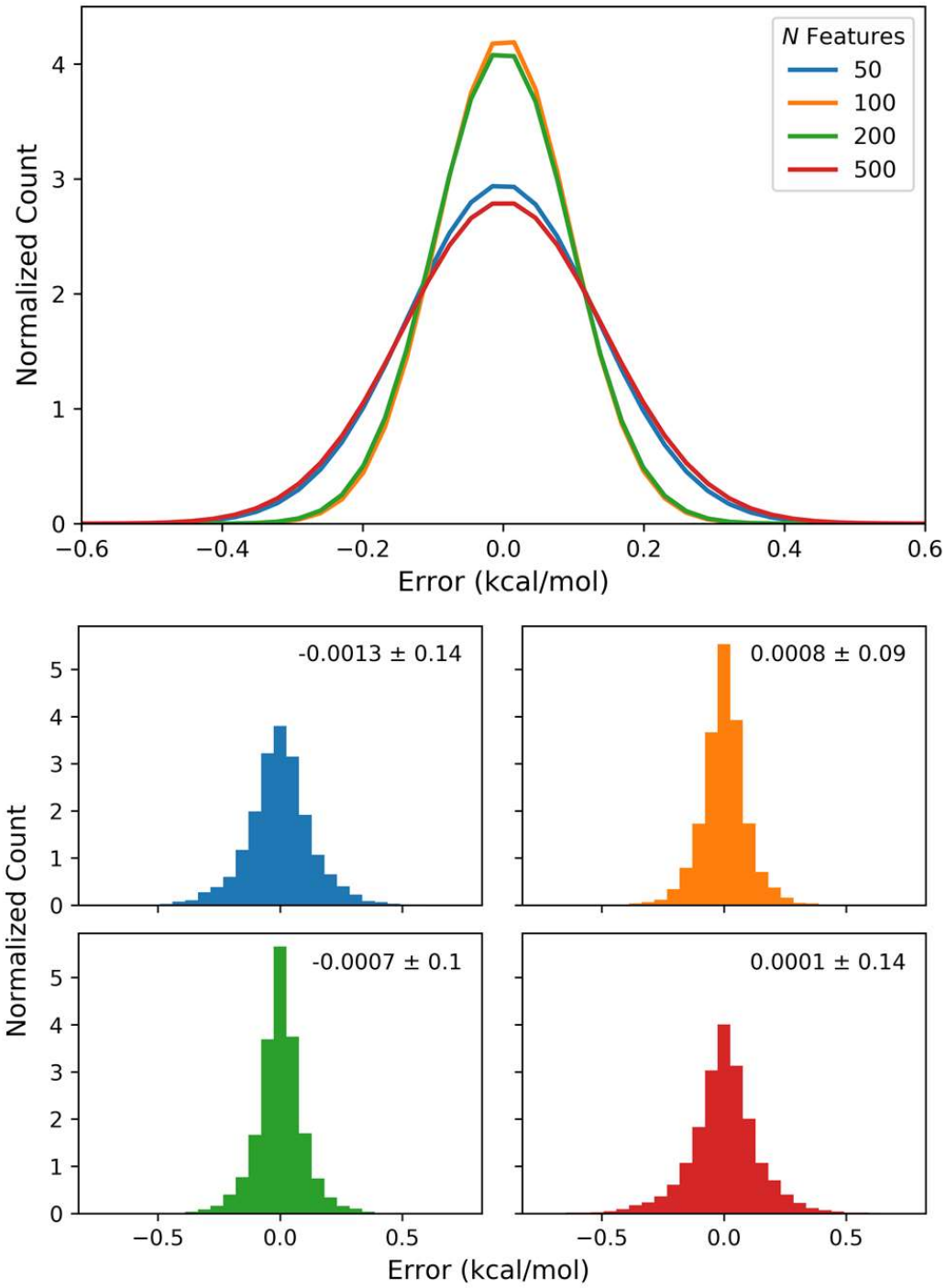

FIG. S2. Comparison of errors (in kcal/mol) for different numbers of atom-wise features in SchNet. The mean and standard deviation of each distribution is listed in the upper right of the respective histogram. Each network was trained on 99,997 water clusters, with 5,263 water clusters taken from each cluster size (N=11–29). The test set consisted of 10,000 water clusters, with 500 taken from each cluster size (N=10–30).

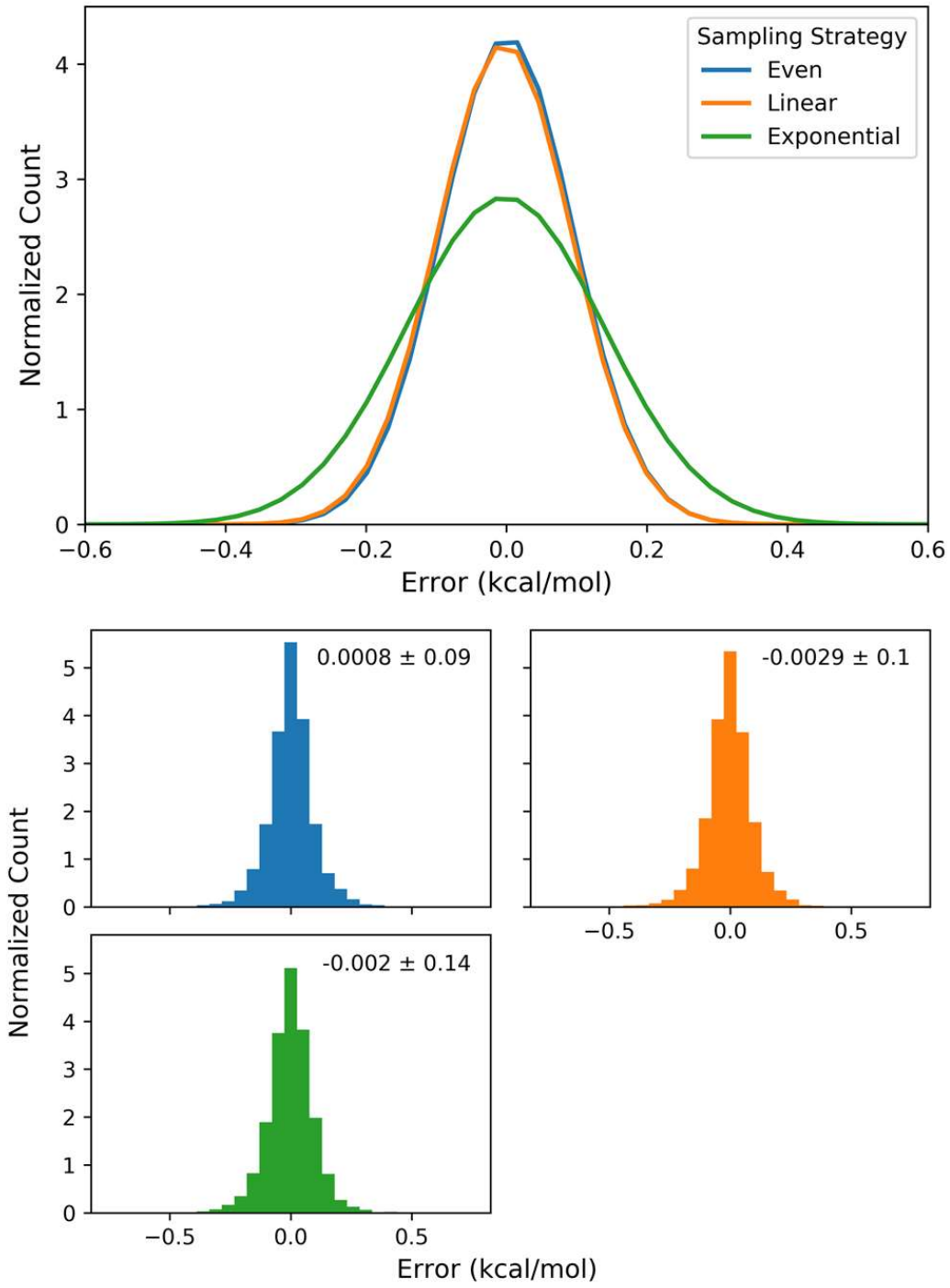

FIG. S3. Comparison of errors (in kcal/mol) for different sampling techniques used to compile the training set. The mean and standard deviation of each distribution is listed in the upper right of the respective histogram. Each network was trained on 99,997 water clusters (N=11–29), using the value given in Table S1. The test set consisted of 10,500 water clusters, with 500 taken from each cluster size (N=10–30). Blue corresponds to the even sampling, orange to the linear sampling, and green to the exponential sampling.

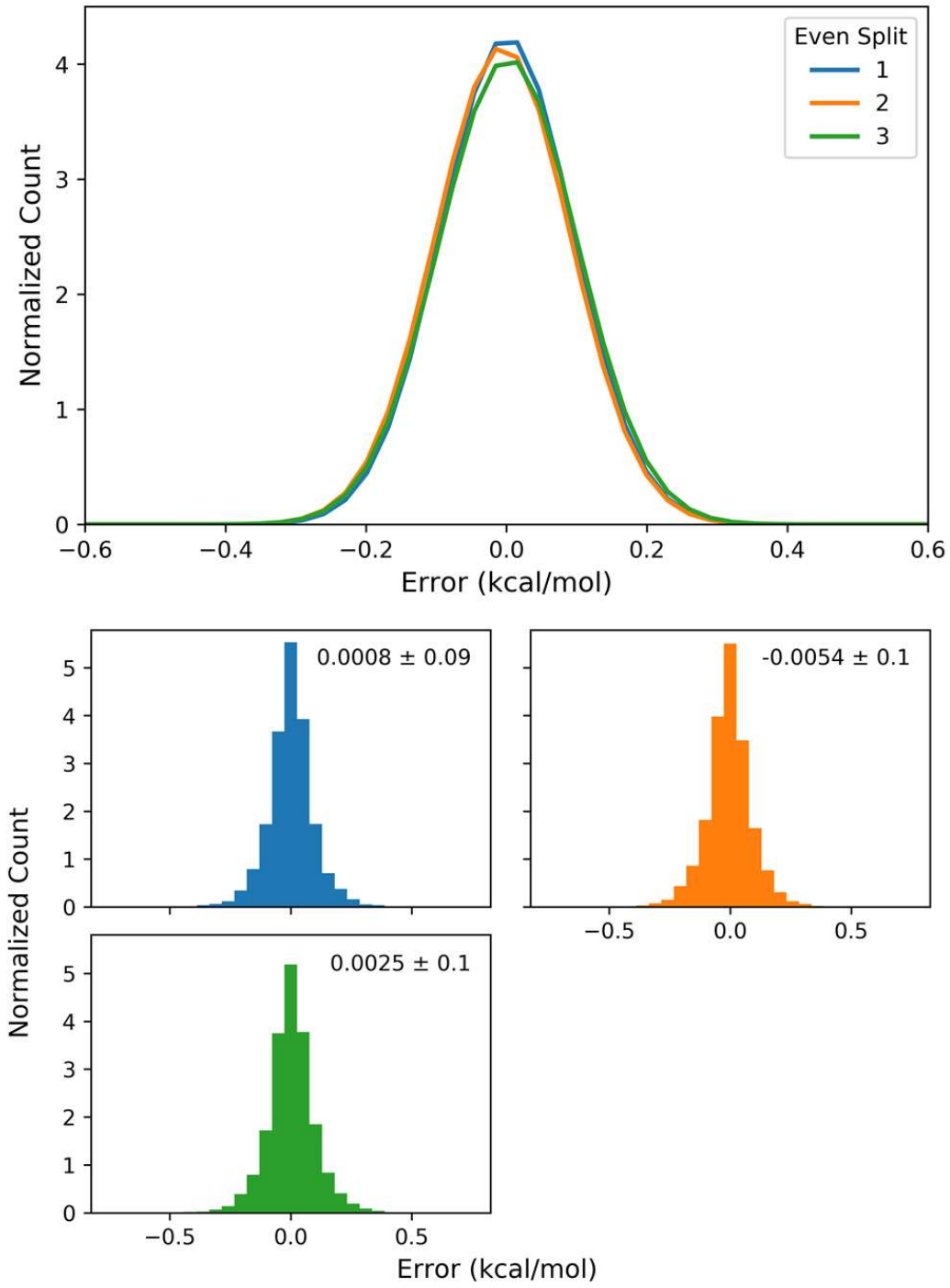

FIG. S4. Comparison of errors (in kcal/mol) produced by the CF-CNN for different samplings of clusters from the full dataset. The mean and standard deviation of each distribution is listed in the upper right of the respective histogram. Each network was trained on 99,997 water clusters, with 5,263 water clusters taken from each cluster size (N=11–29). The test set consisted of 10,500 water clusters, with 500 taken from each cluster size (N=10–30).

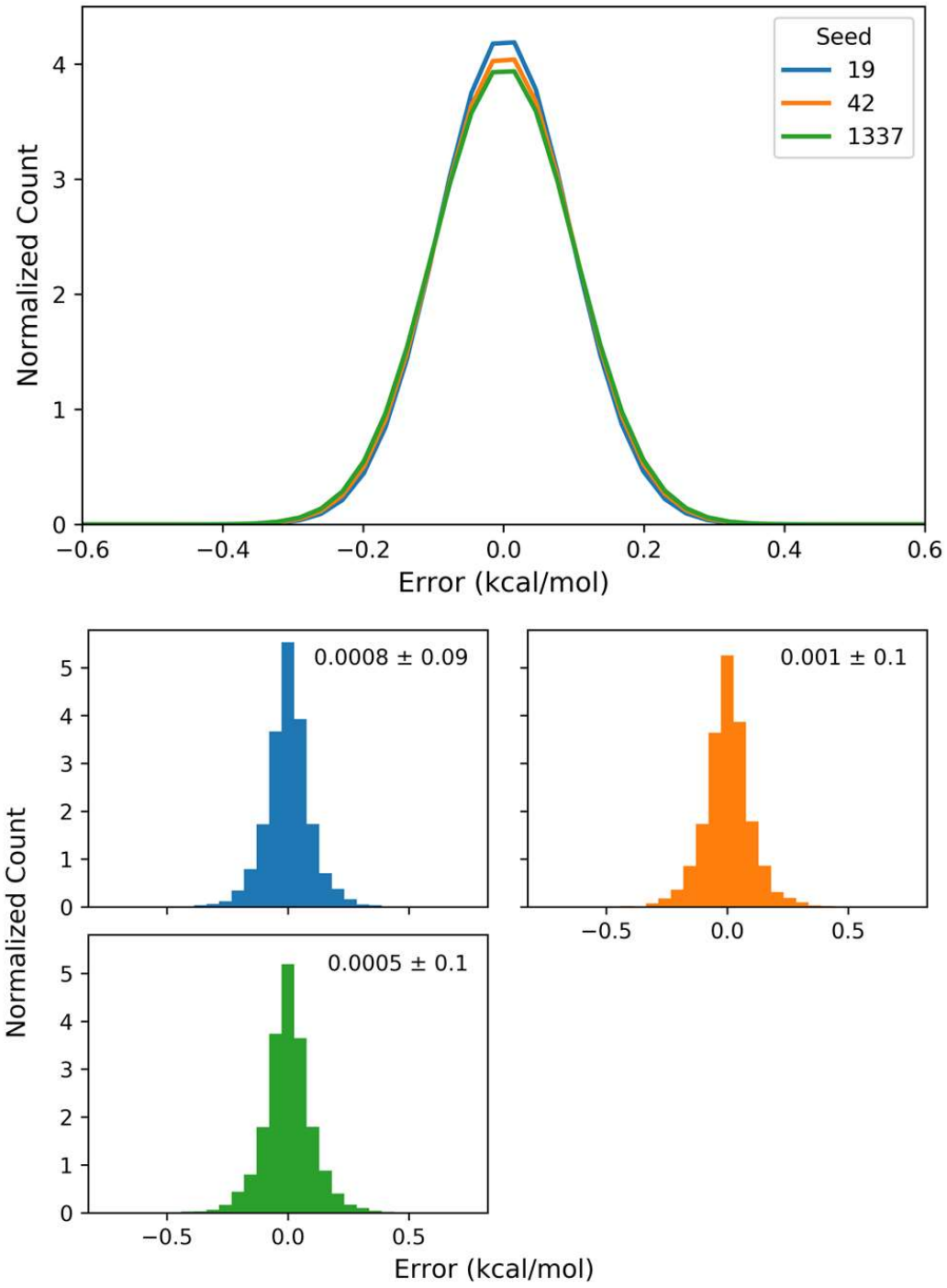

FIG. S5. Comparison of errors (in kcal/mol) produced by the CF-CNN when training with different seeds for sampling 1 from Fig. S4. The mean and standard deviation of each distribution is listed in the upper right of the respective histogram. Each network was trained on 99,997 water clusters, with 5,263 water clusters taken from each cluster size (N=11–29). The test set consisted of 10,500 water clusters, with 500 taken from each cluster size (N=10–30).

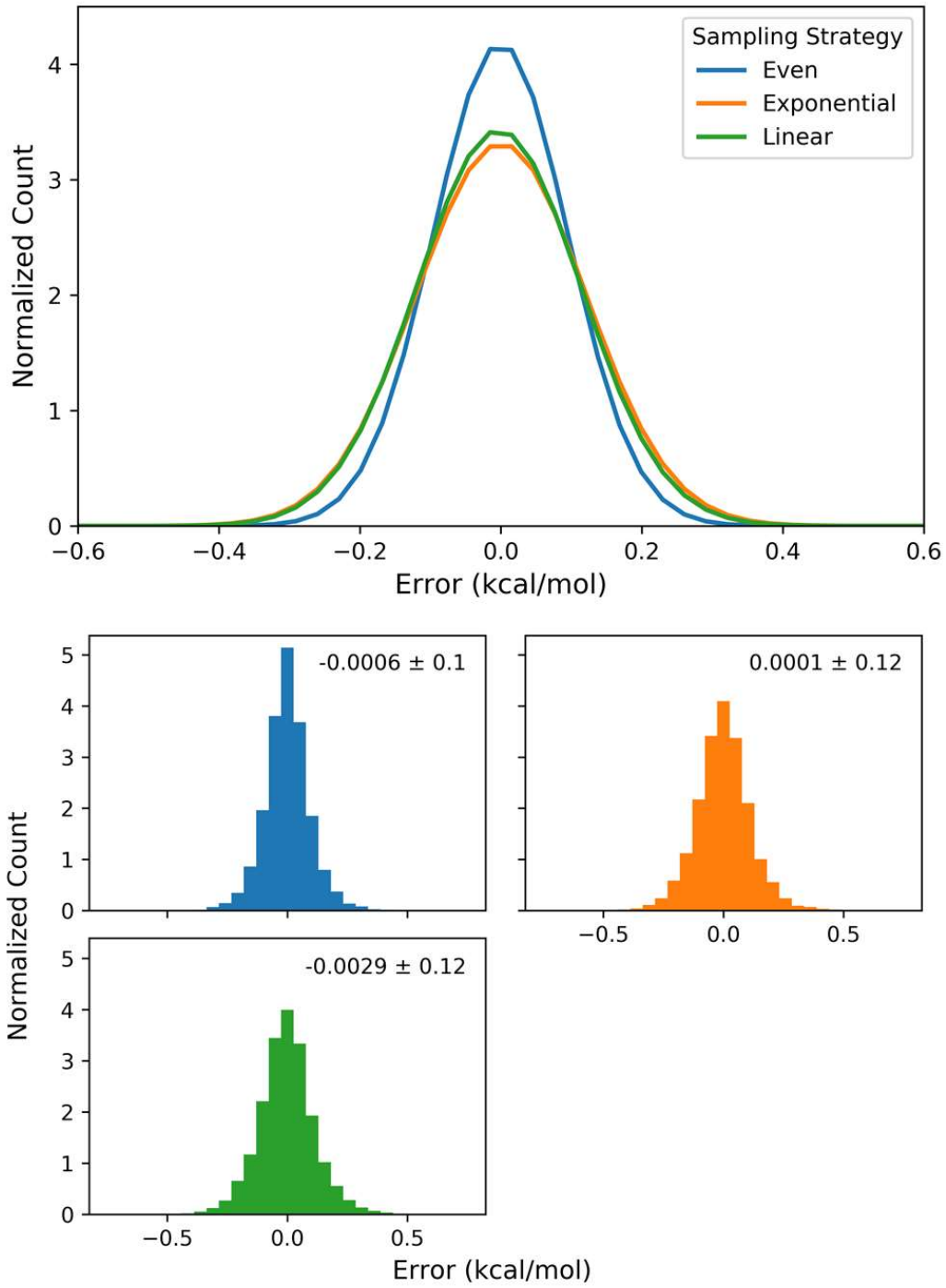

FIG. S6. Comparison of errors (in kcal/mol) for different sampling strategies using RuNNer. The mean and standard deviation of each distribution is listed in the upper right of the respective histogram. For details of each sampling strategy used to create the training set see Table S1. Blue corresponds to the even sampling, green to the linear sampling, and orange to the exponential sampling. The test set consisted of 10,500 water clusters, with 500 taken from each cluster size (N=10–30).