JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

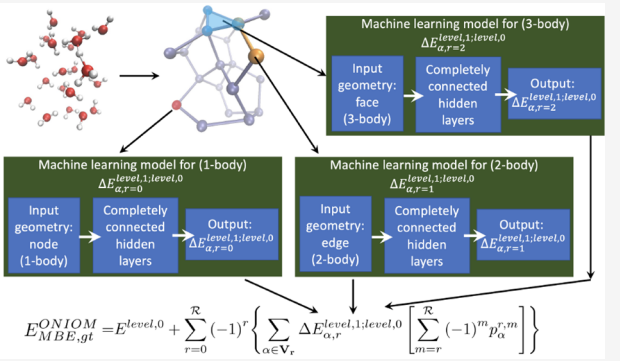
Graph Theoretic Molecular Fragmentation for Multidimensional Potential Energy Surfaces Yield an Adaptive and General Transfer Machine Learning Protocol

Xiao Zhu and Srinivasan S. Iyengar*

Cite This: J. Chem. Theory Comput. 2022, 18, 5125−5144 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Over a series of publications we have introduced a graph-theoretic description for molecular fragmentation. Here, a system is divided into a set of nodes, or vertices, that are then connected through edges, faces, and higher-order simplexes to represent a collection of spatially overlapping and locally interacting subsystems. Each such subsystem is treated at two levels of electronic structure theory, and the result is used to construct many-body expansions that are then embedded within an ONIOM-scheme. These expansions converge rapidly with many-body order (or graphical rank) of subsystems and have been previously used for ab initio molecular dynamics (AIMD) calculations and for computing multidimensional potential energy surfaces. Specifically, in all these cases we have shown that CCSD and MP2 level AIMD trajectories and potential surfaces may be obtained at density functional theory cost. The approach has been demonstrated for gas-phase studies, for condensed phase electronic structure, and also for basis set extrapolation-based AIMD. Recently, this approach has also been used to derive new quantum-computing algorithms that enormously reduce the quantum circuit depth in a circuit-based computation of correlated electronic structure. In this publication, we introduce (a) a *family of neural networks* that act in parallel to represent, efficiently, the post-Hartree−Fock electronic structure energy contributions for all simplexes (fragments), and (b) a new k-means-based tessellation strategy to glean training data for high-dimensional molecular spaces and minimize the extent of training needed to construct this family of neural networks. The approach is particularly useful when coupled cluster accuracy is desired and when fragment sizes grow in order to capture nonlocal interactions accurately. The unique multidimensional k-means tessellation/clustering algorithm used to determine our training data for all fragments is shown to be extremely efficient and reduces the needed training to only 10% of data for all fragments to obtain accurate neural networks for each fragment. These fully connected dense neural networks are then used to extrapolate the potential energy surface for all molecular fragments, and these are then combined as per our graph-theoretic procedure to *transfer* the learning process to a full system energy for the entire AIMD trajectory at less than one-tenth the cost as compared to a regular fragmentation-based AIMD calculation.

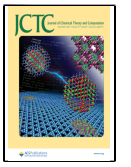


I. INTRODUCTION

Computing accurate potential energy surfaces for complex problems is challenging and at the forefront of modern computational chemical physics and computational chemistry.[1−9] Accurate potential surfaces are needed for predicting molecular spectroscopic properties beyond the harmonic approximation[8−12] and for studying most reactive processes.[13−20] Computing accurate potential energy surfaces is thought to be an exponential scaling problem. For example, for a system with N nuclear degrees of freedom, with M discretizations per dimension, determined based on the mass of the nuclear degrees of freedom, the complexity of the problem of computing potential surfaces may grow[2] as $O(M^N)$. Furthermore, each potential energy calculation needs the accurate quantum-mechanical treatment of electrons, and post-Hartree−Fock electron correlation methods such as CCSD(T) have a steep algebraic scaling. Thus, together, this problem of computing accurate potential surfaces presents a grand challenge for computational chemistry, and it is known to have critical applications in the areas of atmospheric,[21−24] biological,[13−20] and materials chemistry.[25−27]

To overcome the algebraic complexity of an electronic structure, many new fragmentation methods have been

Received: December 9, 2021
Published: August 22, 2022



https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

developed.[28−47] These methods essentially divide a molecular system or condensed-phase problem into orthogonal or overlapping sections, often determined based on chemical intuition. The associated individual fragments are then carefully arranged to obtain the overall energy, gradients, and higher-order properties of the entire system. The reference list cited above is not exhaustive but strongly suggests the deep impact these methods are having on the state of the art for electronic structure calculations.

In refs 48−57, the authors have introduced a graph theory-based approach to molecular fragmentation, where a system is divided into a set of orthogonal fragments that are treated as nodes within a graph. These nodes are then connected through edges, defined based on some predetermined distance cutoff, to capture two-body interactions. Faces obtained from three connected nodes then represent three-body interactions, and similarly high-order n-body interactions are captured via a set of n mutually connected nodes. Over a set of publications it has been established that, when the graph-theory-based many-body theory highlighted above is combined with Our own n-layered Integrated molecular Orbital and Molecular mechanics (ONIOM), the energy and the gradients obtained from such a formalism yield a convergent behavior with increasing orders in many-body ranks,[54−56] in addition to providing conservative molecular dynamics trajectories.[48−57] Furthermore, it has been shown[44,47,54,56] that there is a very strong connection between these graph theoretic methods,[55] many-body theory,[1,58−66] molecular fragmentation methods,[28−47] and ONIOM.[67−72] In refs 53 and 57 we have also shown that potential energy surfaces can be written as a linear combination of individual potential energy functions obtained from several such graph-theoretic descriptions of molecular fragmentation (i.e., by constructing a superposition of graphs). In essence, in refs 53 and 57, each molecular fragmentation protocol leads to a different graph, and the overall potential surface is assembled through a linear combination of the associated potential surface functions. An efficient approach for quantum computation has also been introduced from these methods in ref 73. Together these descriptions allow us to expand potential energy surfaces as functions of lower-dimensional subspaces of the nuclear degrees of freedom.[74] This is consistent with methods such as high-dimensional model representation,[3,75,76] weighted sum of products of approximations,[77,78] and the permutationally invariant polynomials[79] approach, where, influenced by the early work of Kolmogorov[80,81] and Sobol,[82] a high-dimensional function is written as a linear combination of lower-dimensional functions.

However, significant challenges remain with respect to the graph-theoretic molecular fragmentation methods described in refs 48−57. First, with the graph theory protocol, the number of electronic structure energy calculations to be performed for the fragments grows exponentially when potential surfaces are desired. These fragments are known as simplexes, as these are computed from the graph-theoretic description. Second, while the fragmentation procedure does alleviate the electron correlation scaling problem for the full system, the fragments may, in general, be large enough to also present a serious challenge for post-Hartree−Fock methods. This publication overcomes both challenges by using machine learning methods on top of graph-theoretic fragmentation protocols. In fact, here, a family of neural networks (NNs) is introduced, one neural network for each simplex within the graph, that helps us accurately predict the difference in energy between a high level of electronic structure and a low level of electronic structure, as needed within the ONIOM-type description of potential energy surfaces discussed here. Thus, the overall energy of the system becomes a linear combination of neural network energies.

The training data needed to construct the simplex (or fragment) neural networks in this paper are obtained from ab initio dynamics trajectories. The data are then classified using an unsupervised learning method called the mini-batch k-means clustering algorithm.[83,84] This algorithm is essentially a multidimensional tessellation[83,85−89] method that divides the multidimensional ab initio potential energy data space into regions of significance that yield the training set for the neutral network algorithm for each separate simplex. We benchmark our predictions using different numbers of such training sets and show that our method powerfully extrapolates both the simplex energy and the total energy of the system as a resummation based on graph theory.

The paper is organized as follows: In Section II, we present the graph-theoretic method for molecular fragmentation and further generalize the idea such that machine learning techniques can be used to enhance the efficiency of the high-level fragment calculations. As we note in Section III, this is particularly important for cases where larger fragment sizes and higher-rank many-body approximations may be necessary. At the end, Section III introduces a method where the electronic energy for a system is extrapolated using a parallel stream of independent neural networks. In Section IV IV.A, we introduce an efficient technique to obtain an adequate training set using k-means-based tessellation of the multidimensional nuclear coordinated space; the associated training data are then used in the construction of a family of neural network models. In Section V, the approach is demonstrated for protonated water clusters. Conclusions are given in Section VI.

II. GRAPH REPRESENTATIONS FOR LOCAL MANY-BODY INTERACTIONS

In a series of publications[48−57,73] graph theory-based techniques have been discussed to compute efficient and adaptive many-body expansions[1,58−66] that have strong connections to molecular fragmentation[28−47] and ONIOM.[67−72] The salient features of this approach are as follows: the molecular assembly is partitioned into a set of nodes, or vertices. These nodes are then connected through edges based on a chosen edge length threshold that captures the extent to which two-body bonded and nonbonded interactions are to be captured. Together the set of nodes and edges define a graph, $\mathcal{G}_{\bar{\mathbf{x}}} \equiv \{\mathbf{V}_0^{\bar{\mathbf{x}}}; \mathbf{V}_1^{\bar{\mathbf{x}}}\}$, which is obtained from the instantaneous molecular structure $\bar{\mathbf{x}}$. Here $\mathbf{V}_0^{\bar{\mathbf{x}}}$ is the set of vertices, and $\mathbf{V}_1^{\bar{\mathbf{x}}}$ is the set of edges for the molecular graph depicting the structure $\bar{\mathbf{x}}$. This graph is now said to represent local interactions, where through the presence of edges, two-body local interactions are captured. But inherently present in this graph are also higher-order interactions represented by triangles, tetrahedrons, and objects with five or more nodes. The critical aspect here, which makes the connections to many-body theory rigorous, is that these higher-order objects are completely connected. That is, here all pairs of nodes in the included higher-order objects are connected through edges, and such objects are known as affine simplexes;[90−92] furthermore, the resultant graphical representation is often referred to as a simplical complex.[91] Each set of rank-r simplexes, $\{\mathbf{V}_r | r = 0 \cdots \mathcal{R}\}$, thus arises from a (truncated) power-set of the elements within the graph, as follows.

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

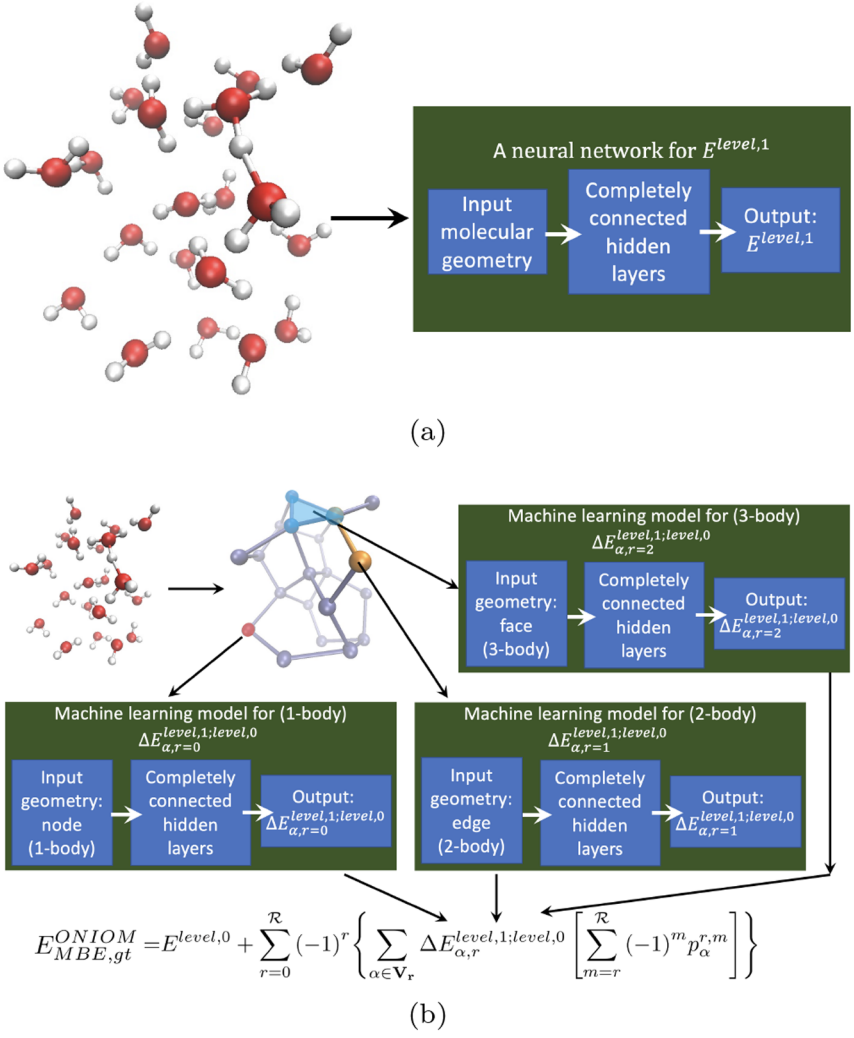


Figure 1. (a) The approach for constructing an NN for the full system. The learning space scales significantly with system size. (b) Our approach of computing $\Delta E^{level,1;level,0}_{\alpha,r}$ in eq 4 using NNs. There are multiple NNs in (b), one for each simplex. The complexity of these NNs in (b) are tremendously reduced by using eq 4, where only nodes, edges, and other higher-order simplexes may be used to compute machine learning models.

$$\{\mathbf{V_r} | r = 0, \cdots, \mathcal{R}\} \equiv \{\mathbf{V_0}, \mathbf{V_1}, \mathbf{V_2}, \cdots, \mathbf{V_r}, \cdots, \mathbf{V_\mathcal{R}}\} \quad (1)$$

The elements of such a power set provide a general and robust scheme to construct many-body expansions for arbitrary systems.

The above graphical description allows a dynamic and flexible representation of local many-body interactions. The energetic measure considered in refs 48−57 and 73 consists of a perturbative, ONIOM-type, correction to a result obtained at a lower level of theory, where the perturbative correction is the difference between two many-body expansions (replacing the standard "model-high" minus "model-low" portion in ONIOM) given by the graphical representation above. Consistent with the notions behind ONIOM,[67] the energy expression[48−57,73] that conveys this general idea is

$$E^{\mathrm{ONIOM}}_{\mathrm{MBE,gt}}(\bar{\mathbf{x}}) = E^{\mathrm{level},0}(\bar{\mathbf{x}}) + E^{\mathrm{level},1}_{\mathrm{MBE}}(\bar{\mathbf{x}}) - E^{\mathrm{level},0}_{\mathrm{MBE}}(\bar{\mathbf{x}}) \quad (2)$$

where the left side, $E^{\mathrm{ONIOM}}_{\mathrm{MBE,gt}}(\bar{\mathbf{x}})$, denotes the graph-theoretically obtained many-body correction to ONIOM, and the term $E^{\mathrm{level,I}}_{\mathrm{MBE}}(\bar{\mathbf{x}})$ on the right side may encompass the full system or a chosen "active site". Ab initio molecular dynamics (AIMD) trajectories have been studied with both options;[48−52,56] furthermore, quantum nuclear effects arising from detailed potential energy surface treatments have also been introduced based on generalizations[53,57] to eq 2. In addition to the extrapolatory, ONIOM-like form of eq 2, each term in the extrapolation is a many-body expansion that is now written in a general and computationally robust fashion up to order (or rank) $\mathcal{R}$ as

$$E^{\mathrm{level,I}}_{\mathrm{MBE}} = \sum_{r=0}^{\mathcal{R}} (-1)^r \sum_{\alpha \in \mathbf{V_r}} E^{\mathrm{level,I}}_{\alpha,r} \left[\sum_{m=r}^{\mathcal{R}} (-1)^m p_\alpha^{r,m} \right] \quad (3)$$

where $p_\alpha^{r,m}$ is the number of times the α-th $(r+1)$-body term (in set $\mathbf{V_r}$) appears in all $(m+1)$-body terms (in set $\mathbf{V_m}$ for $m \geq r$), and consequently $\left[\sum_{m=r}^{\mathcal{R}} (-1)^m p_\alpha^{r,m}\right]$ is the overcounting correction for the number of times the α-th $(r+1)$-body term appears in all objects of rank greater than or equal to r. It is important to emphasize that eq 3 is essentially identical to standard many-body expressions but presented now using graph theory. Thus, the full energy expression, which combines eqs 2 and 3, becomes

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

$$
\begin{aligned}
&E_{MBE,gt}^{ONIOM}\\
&=E^{level,0}+\sum_{r=0}^{\mathcal{R}}(-1)^r\left\{\sum_{\alpha\in\mathbf{V_r}}\left(E_{\alpha,r}^{level,1}-E_{\alpha,r}^{level,0}\right)\left[\sum_{m=r}^{\mathcal{R}}(-1)^m p_\alpha^{r,m}\right]\right\}\\
&=E^{level,0}+\sum_{r=0}^{\mathcal{R}}(-1)^r\left\{\sum_{\alpha\in\mathbf{V_r}}\Delta E_{\alpha,r}^{level,1;level,0}\left[\sum_{m=r}^{\mathcal{R}}(-1)^m p_\alpha^{r,m}\right]\right\}\\
&=E^{level,0}+\sum_{r=0}^{\mathcal{R}}(-1)^r\sum_{\alpha\in\mathbf{V_r}}\Delta E_{\alpha,r}^{level,1;level,0}\mathcal{M}_\alpha^r\\
&=E^{level,0}+\Delta\mathrm{MBE}_{\mathcal{R}}^{level,1;level,0}
\end{aligned}
\tag{4}
$$

where

$$
\Delta E_{\alpha,r}^{\mathrm{level},1;\mathrm{level},0}=\left(E_{\alpha,r}^{\mathrm{level},1}-E_{\alpha,r}^{\mathrm{level},0}\right)
\tag{5}
$$

and $\mathcal{M}_\alpha^r\equiv\left[\sum_{m=r}^{\mathcal{R}}(-1)^m p_\alpha^{r,m}\right]$, as noted above, is the number of times the α-th $(r+1)$-body term appears in all objects of rank greater than or equal to r. In practice, the individual simplex energy contributions, $\{\Delta E_{\alpha,r}^{\mathrm{level},1;\mathrm{level},0}\}$ in eq 4, are computed independently, asynchronously, and in parallel.[56] Furthermore, the implementation also allows one to use separate electronic structure packages for each level of treatment. Our asynchronous, Python/C++ graph-theoretic fragmentation implementation currently supports the simultaneous use of Gaussian,[93] ORCA,[94] Psi4,[95] Quantum Espresso,[96] and OpenMX[97] within a single electronic structure, dynamics, and potential surface calculation. In this paper, electronic structure calculations are computed using the Gaussian package.[93]

When potential surfaces are needed,[53,57,98] multiple graphical representations may contribute to a single nuclear geometry. In refs 53 and 57 we have introduced a variational procedure to obtain the potential surface from a linear combination of graph representations, that is

$$
\begin{aligned}
E_{\mathrm{MBE,gt}}^{\mathrm{ONIOM}}-E^{\mathrm{level},0}&=\sum_\beta \upsilon_\beta(\mathbf{R})\sum_{r=0}^{\mathcal{R}}(-1)^r\\
&\quad\times\sum_{\alpha\in\mathbf{V_r^\beta}}\Delta E_{\alpha,r,\beta}^{\mathrm{level},1;\mathrm{level},0}\mathcal{M}_{\alpha,\beta}^r
\end{aligned}
\tag{6}
$$

$$
=\sum_\beta\sum_{r=0}^{\mathcal{R}}\sum_{\alpha\in\mathbf{V_r^\beta}}\tilde{\mathcal{M}}_{\alpha,\beta}^r\,\Delta E_{\alpha,r,\beta}^{\mathrm{level},1;\mathrm{level},0}
\tag{7}
$$

where the quantities $\mathcal{M}_{\alpha,\beta}^r$, in eq 6, are the analogous extensions to $\mathcal{M}_\alpha^r$ in eq 4 for each graph $\mathcal{G}^\beta\equiv\{\mathbf{V_0^\beta};\mathbf{V_1^\beta}\}$ and

$$
\tilde{\mathcal{M}}_{\alpha,\beta}^r\equiv\left[(-1)^r\upsilon_\beta(\mathbf{R})\mathcal{M}_{\alpha,\beta}^r\right]
\tag{8}
$$

That is, now, $\mathcal{M}_{\alpha,\beta}^r$ is the number of times the α^{th} $(r+1)$-body term appears in all objects of rank greater than or equal to r in graph $\mathcal{G}^\beta$. The terms $\{\upsilon_\beta(\mathbf{R})\}$ are weights computed variationally[53,57,98] for graph $\mathcal{G}^\beta$, with simplexes given by $\mathbf{V_r^\beta}$. Thus, the weighted-graph approach to molecular potential surfaces simply changes the weights on the energy correction terms in eq 4, leaving intact the complexity of computing the terms in eq 5. In this paper, we overcome this challenge through the machine learning protocols discussed below. We show how a family of neural networks may be used to "learn" and "predict" the behavior of eq 5 for all values of r, which may, in turn, be used in eq 4 to provide improved efficiency. We discuss strategies to optimize the learning space, to reduce the overfitting problem and show that a suitable set of individual neural networks, each representing a different simplex, can be computed separately to obtain sub-kJ/mol accuracy. Our approach is depicted in Figure 1b. This may be contrasted from the traditional NN approach in Figure 1a, where the full system energy $E^{\mathrm{level},1}$ is to be extrapolated.

For the remaining part of the paper we drop the index β and formulate our approach for a single graph. Generalization to multiple graphs is implied through simple modification of weights as in eq 7 and will numerically be considered in future.

III. MACHINE LEARNING APPROXIMATIONS TO $\Delta E_{\alpha,R}^{\mathrm{LEVEL},1;\mathrm{LEVEL},0}$ IN EQ 4 PROVIDE A PATHWAY TO TRANSFER LEARNING

To quantify our algorithm, we begin with the assumption that there may exist a Δ-machine learning (ΔML)[99−102] algorithm, denoted here as Q, which, when applied to our input data of molecular geometries, provides the difference in energy of the system between level 1 electronic structure and a lower scaling level 0 electronic structure, as follows.

$$
\mathbf{R}\xrightarrow{Q}E^{\mathrm{level},1}(\mathbf{R})-E^{\mathrm{level},0}(\mathbf{R})\approx\Delta\mathrm{MBE}_{\mathcal{R}}^{\mathrm{level},1;\mathrm{level},0}
\tag{9}
$$

This idea is captured in Figure 1a, and such approaches are common in several ML-based potential surface methods.[99] However, the complexity of the algorithm Q, in terms of training, will grow rapidly with system size $\mathbf{R}$.[103] This is because the number of features and weights in an NN model increase with system size. Additionally, the training model constructed from such a full system study may have limited transferability to other related systems. By contrast, in the studies presented here, the nodes are chosen as single water molecules and hydronium ions, edges are chosen as water-dimers or Zundel cations, and higher-order simplexes are protonated and neutral water clusters that are also found in other problems. Similarly, in refs 51 and 52, we consider single amino acid groups in a polypeptide chain as nodes. These fundamental chemical units are prevalent in a wide range of problems. Thus, a learning protocol based on such fundamental individual constituents of a system, when combined with eq 4, we show, will have an impact on both reduction in complexity of neural networks and also in creating new transfer learning protocols.

Toward this, given the graphical representation of molecular structure described above as $\mathcal{G}\equiv\{\mathbf{V}_0;\mathbf{V}_1\}$, in a fashion commensurate with the inclusion exclusion principle of set-theory,[104] we introduce projection operators, $\mathcal{P}_{\alpha,r}$, the action of each on which on the full molecular system yield, for example, the α-th rank-r simplex. A simple Venn diagrammatic illustration of these projectors is provided in Section III III.A. This approach is then used to obtain a general partitioning scheme, based on graph-theoretic decomposition, that we use to create a family of orthogonal and potentially complete set of neural networks to compute $\{\Delta E_{\alpha,r}^{\mathrm{level},1;\mathrm{level},0}\}$.

III.A. Orthogonal Molecular Learning Spaces from Graph-Theoretic Partitioning. Let us begin with a Venn diagram that divides a coordinate representation $|x\rangle\langle x|$ into regions A, B, C, etc. The regions may intersect, and in Figure 2b we superimposed the Venn diagram on top of an Eigen cation, which is shown in Figure 2a. Thus, the sets divide the electronic domain of a molecular system into several regions. Using the principle of inclusion exclusion,[104] the resolution of the identity for the Hilbert space depicted using the Venn diagram may be written as

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

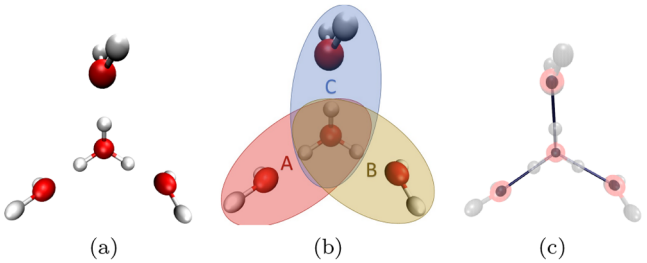


Figure 2. In Figure (b), we illustrate the sets *A*, *B*, and *C* for the system in Figure (a). Figure (b) is used to construct eq 10. In Figure (c) the sets are represented using a graph-theoretic form to make the transition from eq 10 to eq 14, which is a special case of eq 12.

$$\begin{aligned}\mathbf{I} &\equiv \int_{A\cup B\cup C} \mathrm{d}x|x\rangle\langle x| \\ &= \int_A \mathrm{d}x|x\rangle\langle x| + \int_B \mathrm{d}x|x\rangle\langle x| + \int_C \mathrm{d}x|x\rangle\langle x| \\ &\quad - \int_{A\cap B} \mathrm{d}x|x\rangle\langle x| - \int_{A\cap C} \mathrm{d}x|x\rangle\langle x| - \int_{B\cap C} \mathrm{d}x|x\rangle\langle x| \\ &\quad + \int_{A\cap B\cap C} \mathrm{d}x|x\rangle\langle x| \\ &= \mathcal{P}_A + \mathcal{P}_B + \mathcal{P}_C - \mathcal{P}_{A\cap B} - \mathcal{P}_{A\cap C} - \mathcal{P}_{B\cap C} + \mathcal{P}_{A\cap B\cap C}\end{aligned} \tag{10}$$

where the left side is the identity, since the right side sums over the entire Hilbert space represented here through integrals involving the dyads, $|x\rangle\langle x|$, defined within a chosen set obtained from the Venn diagram. Additionally, we also introduced projection operators

$$\mathcal{P}_A \equiv \int_A \mathrm{d}x|x\rangle\langle x| \tag{11}$$

that yield parts of the Hilbert space depicted in eq 10.

Equation 10 arises from the principle of inclusion exclusion[104] from set theory and may be generalized to an arbitrary number of sets. However, an alternate approach to divide the space represented by the identity operator $\mathbf{I}$ can be obtained by reintroducing the graph $\mathcal{G} \equiv \{\mathbf{V_0}; \mathbf{V_1}\}$ from Section II (see Figure 2c). As before, the graph comprises vertices, $\mathbf{V_0}$, edges, $\mathbf{V_1}$, and rank-*r* simplexes. An equivalent expression for the resolution of the identity in eq 10 may now be obtained in terms of projectors that yield domains for nodes, edges, and higher-order simplexes, and this is given by the following expression.

$$\mathbf{I} = \sum_r (-1)^r \sum_{\alpha\in\mathbf{V_r}} \mathcal{M}_\alpha^r \mathcal{P}_{\alpha,r} \tag{12}$$

To make the connections more apparent, we may rewrite eq 12 in decreasing order of rank, that is

$$\begin{aligned}\mathbf{I} = (-1)^{\mathcal{R}}\Bigg\{ &\sum_{\alpha\in\mathbf{V}_{\mathcal{R}}} \mathcal{P}_{\alpha,\mathcal{R}} - \sum_{\alpha\in\mathbf{V}_{\mathcal{R}-1}} \mathcal{M}_\alpha^{\mathcal{R}-1}\mathcal{P}_{\alpha,\mathcal{R}-1} \\ &+ \sum_{\alpha\in\mathbf{V}_{\mathcal{R}-2}} \mathcal{M}_\alpha^{\mathcal{R}-2}\mathcal{P}_{\alpha,\mathcal{R}-2} - \sum_{\alpha\in\mathbf{V}_{\mathcal{R}-3}} \mathcal{M}_\alpha^{\mathcal{R}-3}\mathcal{P}_{\alpha,\mathcal{R}-3} + \dots\Bigg\}\end{aligned} \tag{13}$$

where the appearance of alternating signs is clear and resembles that in eq 10. Additionally, for $\mathcal{R} = 1$, eq 13 becomes

$$\mathbf{I} = -\sum_{\alpha\in\mathbf{V}_1} \mathcal{P}_{\alpha,1} + \sum_{\alpha\in\mathbf{V}_0} \mathcal{M}_\alpha^0 \mathcal{P}_{\alpha,0} \tag{14}$$

which, for the graph in Figure 2c, leads to an identical result as in eq 10, constructed for Figure 2b.

However, the right side of eq 12 is only exactly the identity when the upper limit in the summation over "*r*" (represented as $\mathcal{R}$ below) tends to its maximum possible value, whereas eq 10 is always exact by definition. This is a critical difference between the top-down set-theoretic approach in eq 10 and the bottom-up graph-theoretic approach in eqs 12 and (13). For a completely connected graph the maximum possible value of $\mathcal{R}$ is equal to one less than the number of vertices in the graph. When the graph is not completely connected, the maximum value of $\mathcal{R}$ yields the problem commonly known as *MaxClique*,[105] which is considered to be NP-Complete (NP = nondeterministic polynomial time).[106]

III.B. Independent Machine Learning Models Educated from Graph-Theoretic Partitioning. Since the collection of all simplexes returns the graph, $\mathcal{G}$ (see eq 12), we envision the collective action of all these projection operators on the machine learning algorithm $\mathcal{Q}$ introduced in eq 9 to yield a family of fragment (or simplex) learning models, named $\mathcal{Q}_{\alpha,r}$, according to

$$\begin{aligned}\mathcal{P}_{\mathcal{R}}\mathcal{Q} &\equiv \left[\sum_{r=0}^{\mathcal{R}} (-1)^r \sum_{\alpha\in\mathbf{V_r}} \mathcal{M}_\alpha^r \mathcal{P}_{\alpha,r}\right]\mathcal{Q} \\ &= \sum_{r=0}^{\mathcal{R}} (-1)^r \sum_{\alpha\in\mathbf{V_r}} \mathcal{M}_\alpha^r (\mathcal{P}_{\alpha,r}\mathcal{Q}) \\ &= \sum_{r=0}^{\mathcal{R}} (-1)^r \sum_{\alpha\in\mathbf{V_r}} \mathcal{M}_\alpha^r \mathcal{Q}_{\alpha,r}\end{aligned} \tag{15}$$

where

$$\mathcal{P}_{\mathcal{R}} = \sum_{r=0}^{\mathcal{R}} (-1)^r \sum_{\alpha\in\mathbf{V_r}} \mathcal{M}_\alpha^r \mathcal{P}_{\alpha,r} \tag{16}$$

is the truncated (in summation over *r*) version of eq 12 and is educated by graphical decomposition consistent with the form of eq 4. In addition

$$\mathcal{P}_{\alpha,r}\mathcal{Q} \equiv \mathcal{Q}_{\alpha,r} \tag{17}$$

yields a family of machine learning algorithms, $\{\mathcal{Q}_{\alpha,r}\}$, representing one machine learning algorithm (or one neural network) for each fragment. This idea may be clear through the distinctions in Figure 1a,b. Furthermore, it is critical to highlight that the family of networks, $\{\mathcal{Q}_{\alpha,r}\}$, are not in any way overcomplete, since these resolve the identity as seen in eqs 12 and (16).

Thus, using eq 15 it is possible to independently construct machine learning algorithms, $\mathcal{Q}_{\alpha,r}$, one for each fragment

$$\mathbf{R} \xrightarrow{\mathcal{Q}_{\alpha,r}} \Delta E_{\alpha,r}^{\text{level},1;\text{level},0} \tag{18}$$

and use these in parallel to obtain the quantities in eq 4.

III.C. Accumulating the Graph-Theoretically Generated Neural Networks to Obtain a Transfer-Learning Approximation to $E_{\text{MBE,gt}}^{\text{ONIOM}}$ in eq 4. In this publication, we replace $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ in eq 4 with the ML approximations, $\Delta E_{\alpha,r,\text{ML}}^{\text{level},1;\text{level},0}$, to obtain the following relation.

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

$$E^{\text{ONIOM}}_{\text{MBE,gt,ML}} = E^{\text{level,0}} + \sum_{r=0}^{\mathcal{R}} (-1)^r \sum_{\alpha \in \mathbf{V}_r} \Delta E^{\text{level,1;level,0}}_{\alpha,r,\text{ML}} \mathcal{M}^r_\alpha = E^{\text{level,0}} + \Delta E^{\text{ML}} \quad (19)$$

Thus, neural networks are constructed for the fragments, labeled by the dual indices (α,r), and the resultant family of independent neural networks are combined based on eq 19 to *transfer* the fragment neural network predictions to a larger system potential energy prediction.

Compared to a direct learning paradigm constructed for the left side of eq 19, or the associated *level*, 1 energy value for the full system, the dimensionality of the feature space for fragment data are expected to be much smaller than a full system. Second, by creating neural networks for $\Delta E^{\text{level,1;level,0}}_{\alpha,r} \rightarrow \Delta E^{\text{level,1;level,0}}_{\alpha,r,\text{ML}}$, we extrapolate a function that has a smaller range as compared to $E^{\text{level,1}}_{\alpha,r}$. One would expect extrapolations to functions with a narrower range to be easier as compared to those with a wider range, and this aspect is shown in Appendix B. Finally, another advantage arises from the fact that the fragments may be universal and present in multiple systems. Hence fragment neural networks could be reused for other larger systems, or they could provide starting points to construct suitable networks for larger systems. This latter aspect is the essence of eq 19, which thus provides a new and robust paradigm to construct transfer learning protocols.

A few additional comments are needed with respect to Eq. 19, especially given the strong performance of neural network potentials (NNPs)[107] studied in the literature over the past several years.[108−116] In refs 108 and 109, AIMD data on small molecular systems, such as protonated water clusters, are used to construct "energies per atom" embedded within a cluster. These energies per atom are then used to additively transfer the learning paradigm to larger systems,[108−110] thus allowing a potentially general approach to obtain energies in larger molecular systems. While the approach discussed here, using Eq. 19, is complementary to these NNP methods, there are critical differences that must be noted here. First, given the training protocol to be introduced in Section IV.A, numerical advantages exist in regards to the extent of training needed. As shown in Table I, the amount of training needed is of the order of 5−10% of the total data, with 90−95% serving for testing purposes. Thus, our training protocols need far less training than is generally used in the neural networks community. For example, in ref 108, 90% of the AIMD data is used for training. In our case in Table I a far smaller percentage (∼5−10%) of data is used for training. Even with such stringent training guidelines, the accuracy from the fragment neural networks are on the order of 1/100th kcal/mol for the larger molecular fragments and 1/1000th kcal/mol for the smaller molecular fragments as seen on the rightmost column of Table I. This may be compared with NNPs where the accuracy is on the order of 0.01 kcal/mol but measured in terms of energy per atom.[108] This error in energy per atom when accumulated over the molecular system results in larger errors in NNPs as compared to those seen here, as will be discussed later in Section V.C. Additional analysis is provided in Section V.C and Appendix C. However, more benchmarking is necessary and will be carried out in future publications.

Second, achieving high levels of accuracy in our fragment neural networks is critical because, when these networks are transferred to larger systems, as allowed by Eq. 19, errors will accumulate as the number of fragments grows. A formal discussion on the upper bound to errors from Eq. 19 is presented in Appendix D. In contrast with the Behler-Parrinallo-type neural networks[115,117] commonly used for transfer learning-based potential surfaces,[109] Eq. 19 retains the chemical environment of subsystems represented as neural networks. By using fragments with size up to protonated quadrumers, we are able to construct a more accurate protonated heximer potential energy surface without systematic shift as compared to that in ref 109 and will be discussed in Section V.C.

Arising from such an analysis, in the Results section, we utilize two different strategies to prepare neural network models for the fragments that are then transferred to the full system. In one case we use a fixed maximum error bound in computing all fragment neural network models as enforced during the training process. For example, we construct neural network models for the fragments by increasing the amount of training data until a desired accuracy is obtained for the fragment testing data. A second approach we used in the Results section is to simply utilize a fixed fraction (e.g., 10%) for training purposes. We find that the fragment neural networks obtained from the latter approach work better in terms of extrapolation of energies to the full system, as allowed by Eq. 19. Additionally, given the lower levels of training allowed by the technique introduced in Section IV.A, the fragment training data can easily be expanded if needed.

Table I[a]

fragment type	training (test) set size	training (test) MAE in kcal/mol
H_2O	10 (38 910)	0.004 (0.007)
H_3O^+	10 (17 046)	0.015 (0.073)
H_4O_2	100 (62 380)	0.072 (0.090)
$H_5O_2^+$	300 (77 510)	0.054 (0.088)
H_6O_3	700 (46 940)	0.021 (0.087)
$H_7O_3^+$	2500 (139 580)	0.062 (0.094)
H_8O_4	460 (15 750)	0.021 (0.084)
$H_9O_4^+$	3500 (124 140)	0.068 (0.098)

[a]For each molecular fragment type, obtained from graph-theoretic decomposition (left column), the training set size and test set size are provided in the center column. The test set is provided within parentheses and is more than 10 times larger than the training set. The training data are gleaned using a multi-dimensional k-means-based clustering algorithm discussed in Section IV.A. The accuracy of the fragment neural networks is provided on the right column (in kcal/mol). For example, for H_4O_2, the training set size contains 100 geometries, whereas the test set size includes 62 380 geometries. The resultant MAE is 0.072 kcal/mol within the training set and 0.090 kcal/mol within the test set. The test set in parentheses is composed of all data obtained from AIMD including the training set. The test set errors are very close to the independent test set errors, which does not contain any training data for all fragments (<0.002 kcal/mol), so we use the above test set for all analysis in this paper.

IV. COMPUTATIONAL ASPECTS

We present a new approach to glean training data in Section IV.A and discuss our neural network topology in Section IV.B.

IV.A. Classifying the Training Space for Neural Networks. A sampling procedure, called the mini-batch-k-means clustering algorithm,[83] is used to obtain training data to define neural networks. The algorithm is an unsupervised machine learning method that returns a suitable sample set (referred to as centroids) from a given set of molecular geometries. The algorithm is a variant of the well-known k-means clustering algorithm,[84,118] which partitions any given data space into a set of mutually exclusive regions known as clusters.

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

The standard k-means procedure aims to divide the multi-dimensional data set, represented here as $\{\mathbf{r}_i\}$, into k clusters, named C_j, such that the cost function

$$\sum_{j=1}^{k} \sum_{\mathbf{r}_i \in \mathbf{C}_j} \left|\mathbf{r}_i - \bar{\mathbf{r}}_j\right|^2 \tag{20}$$

is minimized. Here $\bar{\mathbf{r}}_j$ is the centroid of C_j defined as the average value for all $\mathbf{r}_i \in C_j$. The quantity $\mathbf{r}_i$ in this publication is obtained from the distance matrix corresponding to molecular geometries, as will be discussed below, and these quantities are used as input vectors to construct the fragment neural networks presented in the next subsection. The cost function in Eq. 20 is frequently referred to as the within-cluster sum-of-square, and the minimization is usually achieved through an iterative procedure.

The mini-batch k-means algorithm begins with a random subset of b geometries. We then initialize the k centroid positions $\bar{\mathbf{r}}_j$ also randomly or by using a k-means algorithm on the data subset b. Following this another randomly selected additional data set of size b, from the original data set, is then placed, one geometry at a time, into the cluster spaces formed earlier from the centroids. Each newly added data point is included within the cluster with the closest centroid, and the position of the corresponding centroid position is recomputed as the mean value of all data within the cluster. This process is repeated until the position of all centroids converge and the resultant centroids are depicted using the symbol $\bar{\mathbf{r}}_j$, as noted above. When this happens, the data point closest to each centroid, namely

$$\mathbf{s}_j = \left\{ \mathbf{r}_i \middle| \min_{\mathbf{r}_i} \left|\mathbf{r}_i - \bar{\mathbf{r}}_j\right| \right\} \tag{21}$$

and its corresponding energy is included as part of the training set for the neural networks.

Thus, in our scheme, the choice of training data is determined based on the centroids computed as stated above, and it is a critical feature in determining an optimal training data set. The mini-batch k-means algorithm provides an approximation to the k-means algorithm at much reduced computational cost and yields a tessellation[9,85−88] of the multidimensional fragment conformational space to obtain the needed training data. The mini-batch k-means procedure is used, as highlighted above, to obtain the training set for all fragments with the goal to extrapolate $\{\Delta E_{\alpha,r,\mathrm{ML}}^{\mathrm{level,1;level,0}}\}$ in Eq. 19.

IV.B. Neural Network Arrays to Compute $\Delta E_{\alpha,r,\mathrm{ML}}^{\mathrm{level,1;level,0}}$ for Use in Eq. 19. The NN system that we construct to learn and predict the simplex structure energy contribution, $\Delta E_{\alpha,r,\mathrm{ML}}^{\mathrm{level,1;level,0}}$ in Eq. 19, is composed of a primary neural network and an arbitrary number of secondary neural networks arranged in series.[119] Each secondary neural network is used to learn and predict the *error* from the previous set of neural networks. Thus, for the l^{th} neural network, we minimize

$$\min_{\{W_{a,a+1;i}\}} \left| \Delta E_{\alpha,r}^{\mathrm{level,0;level,1}} - \sum_{i=1}^{l-1} \Delta E_{\alpha,r}^{ML,i} \right|^2 \tag{22}$$

where $\Delta E_{\alpha,r}^{\mathrm{ML},i}$ is the output from the i-neural network, and $\{W_{a,a+1;i}\}$ are the neural network weights for the i^{th} neural network to be explained below.

All neural networks use the same input: $x_{0;l} = \{1; \mathbf{s}_j\}$, where $l = 1 \cdots L$, for L neural networks connected in series. The quantity $\mathbf{s}_j$ is defined in Eq. 21 and represents the set of training data determined based on the k-means methodology described above. The number "1" is augmented to $x_{0;l}$ to provide arbitrary bias within the neural network. These are used to compute hidden and output layers for all neural networks according to

$$x_{a+1;l} = f_{a+1,l}(W_{a,a+1;l} \cdot x_{a;l}) \tag{23}$$

where $f_{a+1;k}$ represents the activation function for the $(a+1)$-th layer in the l-th neural network in the series, and $\{W_{a,a+1;l}\}$ are the weight matrices (including bias) connecting the a-th and $(a+1)$-th layers in the l-th network in the series. As seen in Eq. 22, each neural network is being trained for a residual error to increment the approximation from the previous neural networks. Thus, every additional neural network learns from an improved baseline provided by *all* the previous neural networks

$$\Delta E_{\alpha,r}^{\mathrm{ML},l} \approx \Delta E_{\alpha,r}^{\mathrm{level,0;level,1}} - \sum_{i=1}^{l-1} \Delta E_{\alpha,r}^{\mathrm{ML},i} \tag{24}$$

and the approximation to $\Delta E_{\alpha,r}^{\mathrm{level,0;level,1}}$ for a set of L neural networks connected in series is written as

$$\Delta E_{\alpha,r,\mathrm{ML}}^{\mathrm{level,0;level,1}} \equiv \sum_{l=1}^{L} \Delta E_{\alpha,r}^{\mathrm{ML},l} \tag{25}$$

In our demonstration of the models constructed here, two neural networks are placed in series (i.e., $L = 2$). Each of these neural networks is dense, where the nodes in each layer are connected to all nodes in the previous layer. Additionally, the number of nodes within a hidden layer is set to be 4 times the number of input features, which depends on fragment size. Such configurations provide flexibility by offering more hidden layer nodes to larger fragments. For each hidden layer, a standard rectified linear unit (RELU)[120] is used as activation function (f in Eq. 23) on every node. The epochs are set to 1000, which implies the training process is done 1000 times to update all weights. When the training is completed, an error evaluation is performed to verify the quality of the neural network models.

V. RESULTS AND DISCUSSION

Protonated water clusters have been extensively studied in order to understand proton transfer in membrane ion channels and in a wide range of biological processes and materials systems.[15,16,121−128] However, because of the mass of the transferring proton, these problems often display nontrivial quantum nuclear effects, such as tunneling and zero-point effects. These quantum effects are also throught to be significantly affected by the polarized nature of the electronic structure,[10,129−134] which often needs to be treated with accurate post-Hartree−Fock methods.[129,130,135−137] Hence, here we consider as part of our benchmarks the solvated Zundel water cluster $H_{13}O_6^+$, which is thought to play a significant role in the fundamental study of proton transfer in condensed-phase systems and a variety of chemical, biological, and materials problems.[13,15,16,18,21,22,138−142] Specifically, the solvated Zundel is the smallest protonated system that may contain both Eigen as well as Zundel moieties; the interplay between these two protonated water species is thought to have a significant role on proton migration.[131−134] We use a selected set of fragments obtained using the k-means tessellation of geometries obtained from an AIMD trajectory involving the solvated zundel cation to train a family of neural network models, one pertaining to each graphically determined fragment. The resultant family of neural

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

network models are then evaluated based on accuracy in determining the energies for all configurations within the AIMD trajectory.

This Results section is organized as follows. The AIMD data are discussed in Section V.A. In Section V.B we first discuss the accuracy and efficiency of the mini-batch k-means clustering algorithm in obtaining a multidimensional "gridification" or tessellation of all data. The goal here is to obtain the minimal data sets needed to accurately construct the machine learning fragment energy extrapolation protocol. Once this is done, in Section V.C, the neural network models are used to predict fragment energies $\Delta E^{\text{level,1;level,0}}_{\alpha,r,\text{ML}}$ to be used in Eq. 19. These are then used to predict the full system solvated Zundel potential energy at simplex ranks $\mathcal{R} = 3$, using Eq. 19, and the results are compared with those calculated by full system CCSD on the 9236 structures obtained from AIMD. At the end we note that the mean absolute errors (MAEs) between the solvated Zundel CCSD energies obtained using Eq. 19 are in agreement with the precomputed CCSD energies to within 1 kJ/mol.

V.A. Salient Features of the AIMD Training Data Set.

Ab initio molecular dynamics methods are used in a wide range of applications including the study of vibrational properties beyond harmonic approximation.[10,12,24,122,142−153] However, AIMD requires accurate and detailed information about the potential energy surface sampled by the system during dynamics. This is a challenging computational task that is prohibitive to perform globally for any system with more than a few degrees of freedom. The problem becomes especially serious when state-of-the-art electronic structure accuracy, such as coupled cluster theory, is also needed. Our goal here is to use training data obtained from an AIMD trajectory to create a multidimensional potential surface based on the graph-theoretic molecular fragmentation protocol amended by machine learning, as discussed above and shown in Eq. 19. Thus, our AIMD trajectory for the solvated Zundel cation is based on that obtained in ref 50 and utilizes the energy function in Eq. 4, along with associated gradients, to propagate the nuclear degrees of freedom. However, we note at the outset that it is not necessary to use an AIMD trajectory computed in this fashion and that a trajectory obtained from other levels of theory, or simply configurations sampled from a Monte Carlo simulation, may also be used as a starting point for our computational framework, and these aspects will be considered as part of future publications. Furthermore, it must also be highlighted that, to construct NNPs, in ref 108 ring-polymer-based AIMD methods were used with a goal to sample classically forbidden regions for the training set. This was not done here, and as can be seen below, the results are in excellent agreement with CCSD calculations. Future publications will further evaluate the need for such expensive ring-polymer approaches in determining the learning data samples, especially when combined with our k-means protocol.

In the trajectory obtained from studies in ref 50, CCSD was used as level, 1 electronic structure theory and B3LYP is used as level, 0 theory as needed in the equations presented above. The basis set used for both levels of theory was 6-31+g(d,p), whereas the initial geometry was obtained through optimization at the B3LYP/6-31+g(d,p) level of theory. The simulations were performed using the NVE ensemble, with total simulation time of 1.86 ps and time step of 0.2 fs, and the total energy was conserved to within a standard deviation of 0.01 kcal/mol with a drift of 0.02 kcal/mol over the length of the entire trajectory. The average average temperature for these simulations was 162 ± 27 K. This includes a set of 9326 solvated Zundel molecular geometries. An oxygen−oxygen radial distribution function for this trajectory is shown in Figure 3 and represents the range of structures obtained and used for our ML formalism.

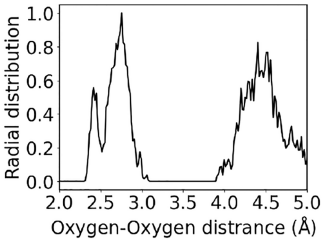


Figure 3. Oxygen−Oxygen radial distribution function computed from the AIMD trajectory used to construct the training set for the ML formalism.

Molecular fragmentation-based energy computations are performed for every structure obtained from AIMD, as per the previously discussed[48−52,54,55,154] protocols for protonated water systems. Specifically, we set each oxygen as a node and include within the node all hydrogen atoms that are within a 1.4 Å range from a given oxygen atom. Higher-rank simplexes are obtained by combining nodes within a node center distance cutoff of 7.5 Å. This distance cutoff allows all nodes in the solvated Zundel system to be fully connected, as shown in Figure 4. Hence the number of simplexes of rank r is $\binom{6}{r+1}$, as shown

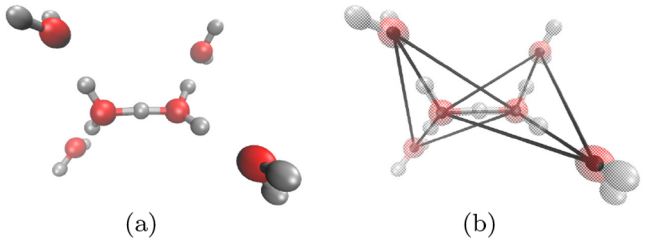


Figure 4. Graphical representation (b) of the solvated Zundel cations (a).

in Figure 5a. When simplexes for each rank are further divided into protonated fragments and neutral fragments, the total number of fragments from the entire trajectory are shown in Figure 5b. Note that, since hydrogen atoms are considered as part of nodes containing oxygens within 1.4 Å, some hydrogen atoms can be part of multiple nodes and more than one protonated fragment can be obtained in any given structure. For example, the number of water and hydronium fragments do not, in general, follow the ratio 5:1 (as would be the case if there is only one hydronium in each solvated Zundel). Instead, the ratio is found to be between 5:1 and 4:2, as shown in Figure 5b. This allows for the presence of both Zundel-like as well as Eigen-like substructures within the geometries as would result from the likelihood of proton hops sampled during dynamics.

In order to implement rotational and translational symmetry of fragments into the neural networks, the set of interatomic distances is used as input features in the machine learning models. Furthermore, atomic numbers are sorted in increasing order before constructing the distance matrix to ensure consistent feature types. For example, for a single water molecule, the input vector is formulated using hydrogen−

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

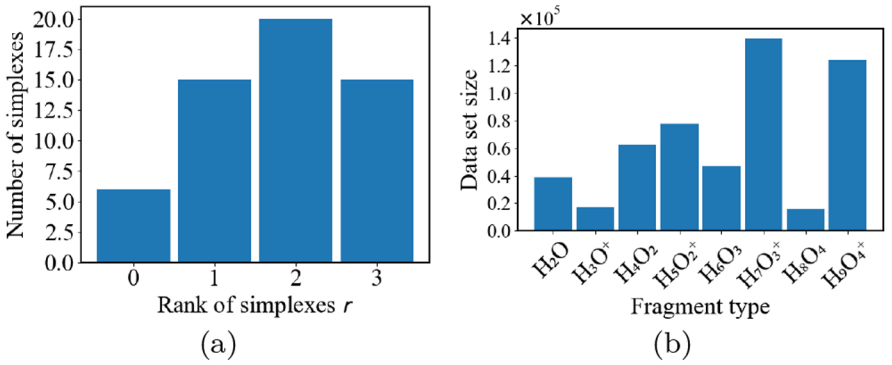


Figure 5. (a) The average number of simplexes for each rank obtained for each solvated Zundel structure in the AIMD trajectory. The trajectory contains 9326 different geometries. The $r = 3$ fragment calculations are approximately $4^6 \equiv 4096$ times more expensive as compared to the $r = 0$ fragment calculations for CCSD accuracy. This publication lowers this expense tremendously, by introducing machine learning techniques as seen in Eq. 19. The full data set size for each fragment is shown in (b).

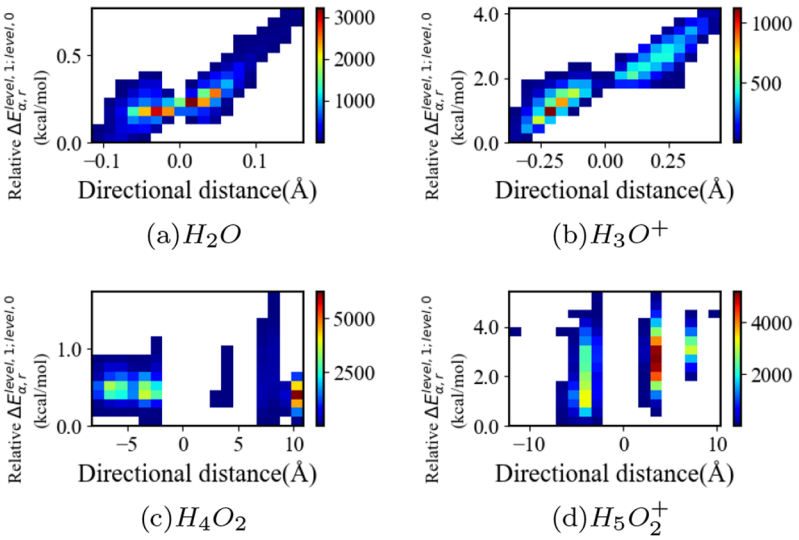


Figure 6. AIMD data distribution for H_2O, H_3O^+, water dimer, and protonated water dimer. In all cases the horizontal axis represents the directional distance in Eq. 28, the vertical axis shows the quantity $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ shifted with respect to its minimum value; the color map represents the density of data at the respective pair $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ and directional distance values.

hydrogen, hydrogen−oxygen, and hydrogen−oxygen distances. In this fashion, the fragment geometry **R** is mapped onto a one-dimensional distance vector from here on referred to as **r**.

V.B. Accuracy and Efficiency of the k-Means Clustering Strategy in Creating Training Data for Energy Prediction Based on an Ensemble of Neural Networks. We now inspect the numerical effectiveness of the k-means clustering based learning space characterization described in Section IV.A. Toward this, we first introduce a one-dimensional distance measure that gauges the spread of fragment geometries encountered and is introduced to distinguish the fragment structures from each other. To arrive at this one-dimensional measure we first compute an average structure from the set of distance vectors (represented as r_i in Sections V.A and IV.A) for each type of fragment as follows.

$$\langle \mathbf{r} \rangle = \frac{1}{\mathbf{N}} \sum_i \mathbf{r}_i \tag{26}$$

Here **N** is the total number of input geometries for a specific type of fragment. Note that this total number includes all geometries of a specific fragment type obtained from all configurations in the dynamics trajectory. Next, a reference structure, referred to here as $\mathbf{r}_{\max}$, is defined as the structure with the largest distance vector L^2 norm.

$$\mathbf{r}_{\max} = \arg\max_{\mathbf{r}_i} \|\mathbf{r}_i\|_2 \tag{27}$$

We then compute the distance of the i-th structure from the average structure in Eq. 26 to define the "directional distance" as

$$\text{sign}[(\mathbf{r}_{\max} - \langle \mathbf{r} \rangle)\cdot(\mathbf{r}_i - \langle \mathbf{r} \rangle)] \times |\mathbf{r}_i - \langle \mathbf{r} \rangle| \tag{28}$$

where sign [···] represents the "sign" of the term within parentheses.

In Figure 6, the horizontal axis in each figure is labeled as "directional distance", explained above. The vertical axis is the energy correction $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ relative to the geometry that has the lowest value of this quntity. The color represents the density of full data for each value of directional distance and $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ values. In Figure 6a,b we note that $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ increases as a function of the directional distance, whereas, in Figure 6c,d, the distribution appears in discrete groups. The reason for this is apparent upon inspection of Figures 3 and 7. In Figure 7 the solvated-Zundel nodes are labeled A,...,F. For example, for H_4O_2,

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

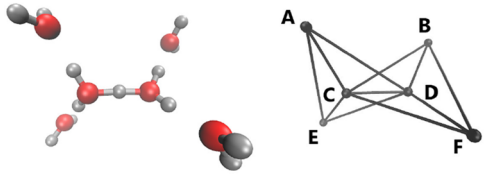


Figure 7. Graphical representation (b) of the solvated Zundel cation (a). The nodes of the graph are labeled with A to F. For clarity, the graph is not fully connected in the figure. In real calculation and analysis, AB, AF, BE, and EF are also connected.

there are five distinct possible combinations of dimer fragments, all of which are included in our simulations. These can be broadly categorized to include water dimer configurations AC, AD, AE, AB, and AF. The configuration CD is generally protonated and generally appears as an $H_5O_2^+$. Clearly the five dimer configurations are partly distinct, being organized roughly into three broad blocks as seen in Figure 6c. These three blocks are also reflected in Figure 3 where the oxygen−oxygen radial distribution function is provided. For example, as shown in Figure 6c, the combinations AC and AD together form the wide group of data on the negative side of directional distance, the combinations AB and AF form the rightmost group on the positive side, and AE provides a sharp distribution in the middle.

V.B.1. Performance of Mini-Batch k-Means in Constructing Fragment Training Data. Our next step is to gauge the effectiveness in appropriate coarse-graining of the distributions presented in Figure 6 through use of the k-means-based clustering algorithm. Our expectation is that the learning data constructed from k-means must appropriately model the diversity of data present in Figure 6 but also compress this data sufficiently so as to lead to efficient learning models. The results of our analysis are presented in Figures 8 and 9, where the data for each fragment are presented separately. In each case, we present three subfigures, where the bottom panel represents the behavior of the full data set for a specific fragment type. It must be noted that the smaller sized fragments are also shown in Figure 6. Thus the bottom panel in each case represents the range of structures present in the AIMD trajectory, for which we wish to construct a learning model. However, we wish to compress the amount of information present in this bottom panel and use only a representative sample to construct the fragment learning models. This is done using the mini-batch k-means clustering procedure described in Section IV.A.

The panels immediately above the full data set in Figures 8 and 9 represent the distribution of training samples obtained from the mini-batch k-means clustering procedure, with far fewer number of samples, as compared to total number of data points, as is clear from the distribution heights given in the key on the right side of each figure. For example, for the case of water, there is a 3 *orders of magnitude reduction in the number of training data set points* (k-means centroids) between the top and bottom panels in Figure 8a. The associated mean absolute errors obtained when the respective compressed data models are used to construct the neural networks are also shown in Figures 8 and 9. In Figure 8a, the 3 orders of magnitude reduction does not greatly affect the distribution of $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ as a function of directional distance, and the associated mean absolute errors noted inside each figure are well within the kJ/mol range. In all cases the middle panel provide less compression and more accuracy. For example, when these resultant centroids are used

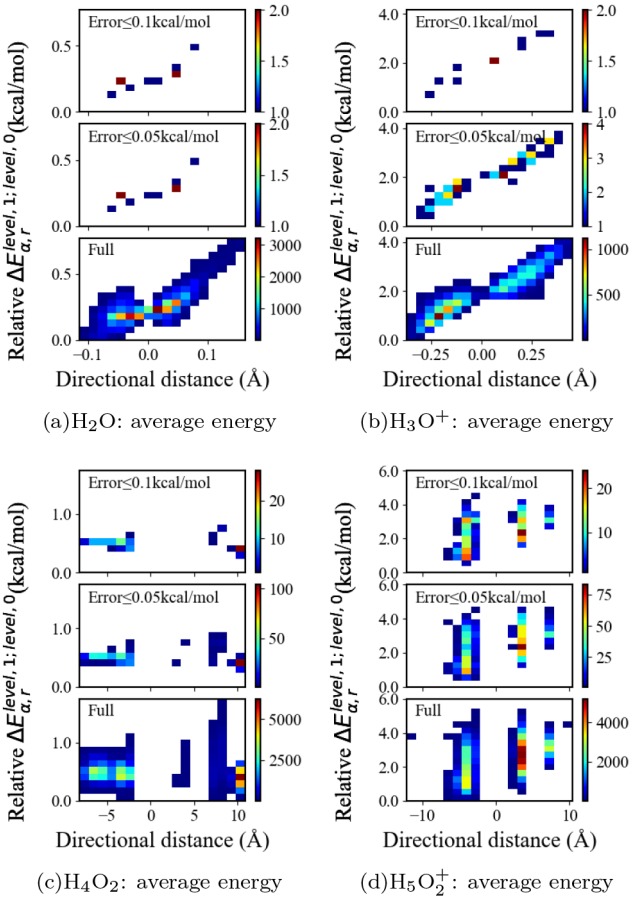
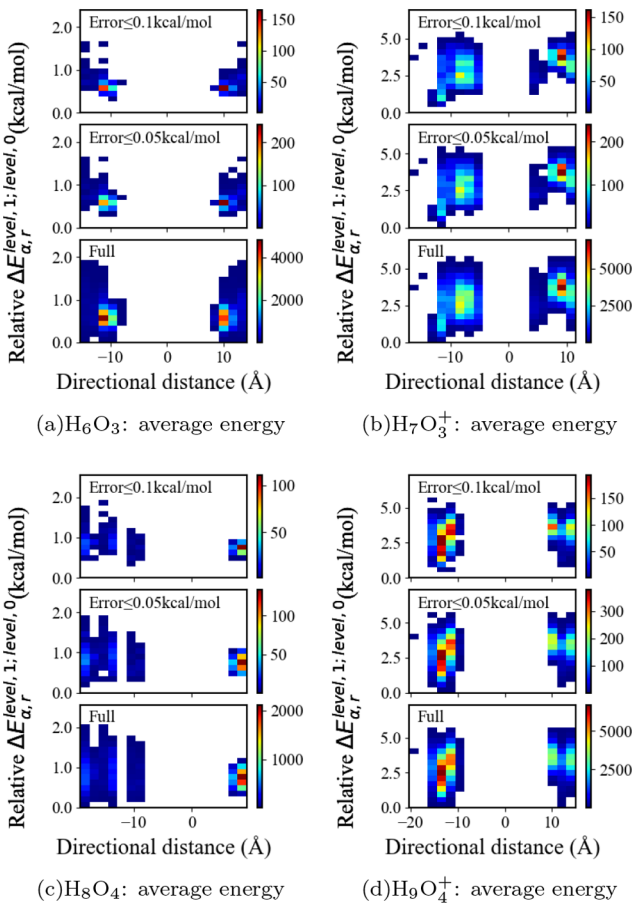


Figure 8. Training sample distributions for water, hydronium, water dimer, and protonated water dimer are displayed. As in Figure 6, each subfigure here shows the density of data (color map immediately to the right of each plot) as a function of direction distance (Eq. 28) along the horizontal axis and $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ along the vertical axis. Each panel represents a type of fragment and contains three plots: The bottom plot shows the distribution for the full data set (which would be the same as in Figure 6). Thus, by full data set we imply all similar fragments from all geometries. The two plots above in each panel show the distribution of data from the k-means generated learning data sets that use far fewer data points. The extrapolation error in $\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}$ for each case is noted in the figure. Importantly, as one rolls from the bottom panel to the top panel in each subfigure, the distribution gets coarser. Thus the k-means tessellation of learning data appropriately coarse grains the learning space, as expected, with well-controlled errors within 1 kJ/mol.

to determine training data, the network model yields an MAE for all water molecular fragments on the order of at most 0.01 kcal/mol as indicated in the figure legend. This is indeed remarkable and, as we will see, provides a stable algorithm to glean the necessary data from a large data set and thus greatly reduces the potential for overfitting. As seen in Figures 8 and 9, this behavior is rather uniform, and indeed, the k-means clustering procedure provides a stable algorithm to obtain the necessary data set to construct a learning procedure. This implementation is a critical difference in the way we obtain our training data here as compared to other neural network models.[108]

V.C. Extrapolating the Energies of the Solvated Zundel System, through Transfer Learning, using the Ensemble of Neural Network Approximations to Compute $\{\Delta E_{\alpha,r}^{\text{level},1;\text{level},0}\}$ in Eq. 19. We prepared two different kinds of learning configuration models to evaluate our transfer

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144



(a)H_6O_3: average energy (b)$H_7O_3^+$: average energy

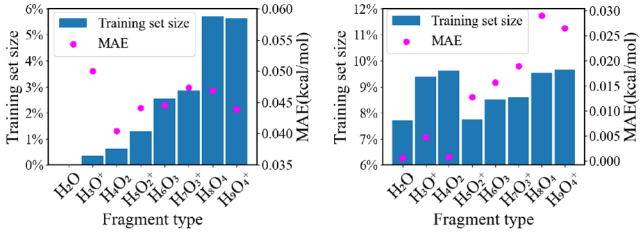
(c)H_8O_4: average energy (d)$H_9O_4^+$: average energy

Figure 9. Training sample distributions for water trimer, protonated water trimer, water quadrumer, and protonated water quadrumer are shown as in Figure 8.

learning protocol. In one case we enforce a fixed maximum error for each fragment neural network model during the training process for $\{\Delta E^{level,1;level,0}_{\alpha,r}\}$. Thus all k-means derived learning data are retained within a learning model until the extrapolation error falls below a chosen threshold for each fragment type. This scheme is referred to below as the fixed fragment error-threshold scheme (FES) and evaluates the impact on the transfer learning process when the neural network model error is uniform. For the second family of neural network models we use a fixed fraction of the entire data set for each fragment type during the training process. This second approach is referred to below as the fixed fragment-data fraction scheme (FFS). The analysis in Appendix D provides a rationale for the two types of protocols chosen here. This will also become clear from the discussion below.

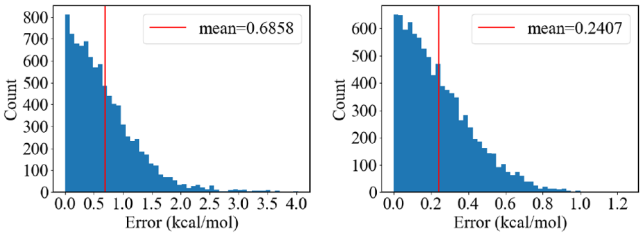
Figures 10 and 11 provide a summary of distinction in performance between the two approaches. In Figure 10a we present the mean absolute error for the FES learning model for each fragment (pink dots, right vertical axis) along with fraction of data used for training (blue histograms, left vertical axis), whereas the corresponding extrapolation error for the use of these learning models when used with Eq. 19 is presented in Figure 11a. Similarly, the associated information for the FFS scheme with fraction of fragment data being less than 10% is shown in Figure 10b, with the performance of these neural networks when used within Eq. 19 presented in Figure 11b.

In Figure 10a each fragment energy has an MAE less than 0.05 kcal/mol. In this case, the largest fraction of data used during the



(a)FES with MAE ≈ 0.05 kcal/mol (b)FFS ≈ 10% training data

Figure 10. Figure (a) depicts the errors (right vertical axis) and training data set size (left vertical axis) from the fixed fragment-data error (FES) protocol described in the text. Figure (b) represents the same for the fixed fragment-data fraction (FFS) learning protocol. In the FES protocol, we enforce a threshold for the fixed maximum error for each fragment neural network model during the training process. Thus all k-means derived learning data are retained within a learning model until when the extrapolation error falls below a chosen threshold. In FFS we use a fixed fraction of the data set for each fragment type during the training process. The errors are generally lower for FFS but require more training data as compared to FES.



(a)FES with MAE ≈ 0.05 kcal/mol (b)FFS ≈ 10% training data

Figure 11. MAE distributions for all solvated Zundel AIMD structures, when neural networks are used as allowed by Eq. 19 with $\mathcal{R}$ = 3. Here the two families of neural network models, FES and FFS, mentioned in Figure 10 and in the text, are used to to obtain (a) $\Delta E^{level,1;level,0}_{\alpha,r,\,ML}$ and (b) ΔE^{ML} in Eq. 19, and the distributions are provided. (a) The models are trained using the FES approach with samples shown in Figure 10a, and the fragment energy errors are retained below 0.05 kcal/mol. (b) The models are trained using 10% of all fragment data, as shown in Figure 10b. The FFS approach uses a larger training set as seen in Figure 10, where the training data are chosen using mini-batch k-means, thus allowing greater accuracy in extrapolating the full solvated Zundel energy. This essentially amounts to a factor of 10 reduction in computation effort with an MAE in the kJ/mol energy range.

training cycle is ∼6% for quartamers, and the training data distribution (centroids and errors) is shown in the middle panels of Figure 8 and 9. (Compare with the performance of other NNP work discussed in Appendix C.) The MAE for the full system (solvated Zundel) potential energy constructed from such FES models is ∼0.69 kcal/mol (Figure 11a), which is much larger than the error in any of the individual fragment models. This is caused by the large number of fragments from the graph representation in Eq. 4 as can be seen from the ML error estimate in Eq. D1. The combinatorial increase in number of fragments together with the weight $\mathcal{M}^r_\alpha$ produces hundreds of fragments of each kind on one solvated Zundel molecule. This can also be seen in Figure 5 where the number of fragments of each type in the overall training model are presented. Clearly the number of fragments grows, potentially in a prohibitive fashion as noted by the appearance of $\mathcal{N}_r$ in Eq. D1, thus affecting the accuracy in such fixed maximum error training models.

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

To reduce the full system prediction error, we further improve the accuracy for every individual fragment model using the FFS learning scheme. In Figure 11b, we raise the number of training samples to ∼10% of fragments and produce the second set of models. For these models, the accuracy of full system potential energy prediction reduces to ∼0.24 kcal/mol (i.e., on the order of 1 kJ/mol) and becomes comparable to the expected accuracy of standard electronic structure methods as also seen from Figure 14 (Also see discussion in Appendix C).

Thus, in summary, it appears that learning models can be constructed to reproduce AIMD data with ∼10% of the effort needed as compared to that for the full fragmentation calculations. A comprehensive analysis is provided in Figures 12 and 13, where in each case the horizontal axes refer to results

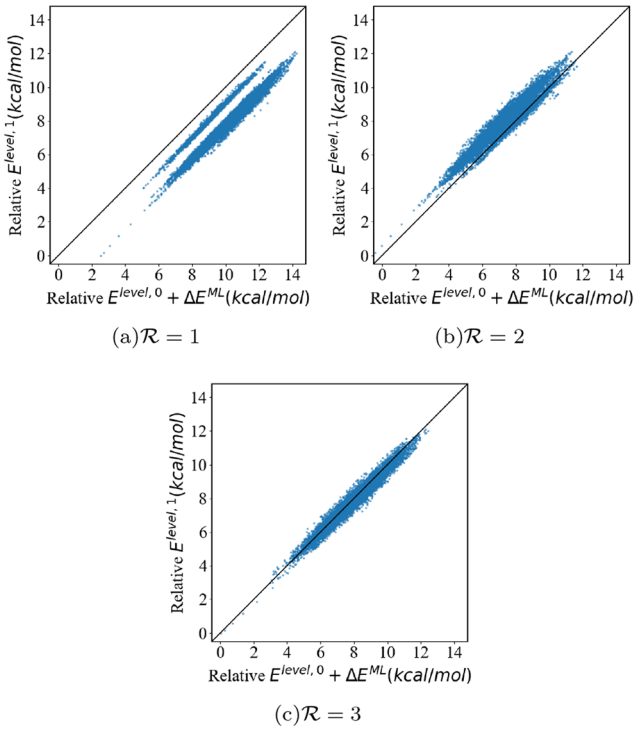


Figure 12. Correlation between machine learning approximations and graph-theoretic molecular fragmentation energy, i.e., Eq. 19 and the full system high-level energy for all solvated Zundel geometries. Both axes are shifted with respect to the same lowest $E^{level,1}$. The black line describes the maximum correlation between the two quantities. As the rank of fragments (represented by $\mathcal{R}$ in Eq. 19) increases, the agreement improves. However, even at $\mathcal{R} = 1$, there appears a simple constant shift, as is clear from the shifted plots in Figure 13.

from Eq. 19, whereas the vertical axes refer to the energy value at the higher level (CCSD) of theory for the full system. Clearly as the rank of the expansion in Eq. 19 increases the agreement is extremely high between the machine learning model and the more expensive higher-level electronic structure treatment. Specifically, while the $\mathcal{R} = 1$ results (including edges) are roughly systematically shifted down, $\mathcal{R} = 2$ (faces) does appropriately improve; the results from $\mathcal{R} = 3$ (rank-4 simplexes) are then in excellent agreement with the level, 1 results. Furthermore, including a constant shift to all these approximations in Figure 13 does suggest that even the lower values of $\mathcal{R}$ can provide reasonable agreement. These aspects will be further evaluated in future publications.

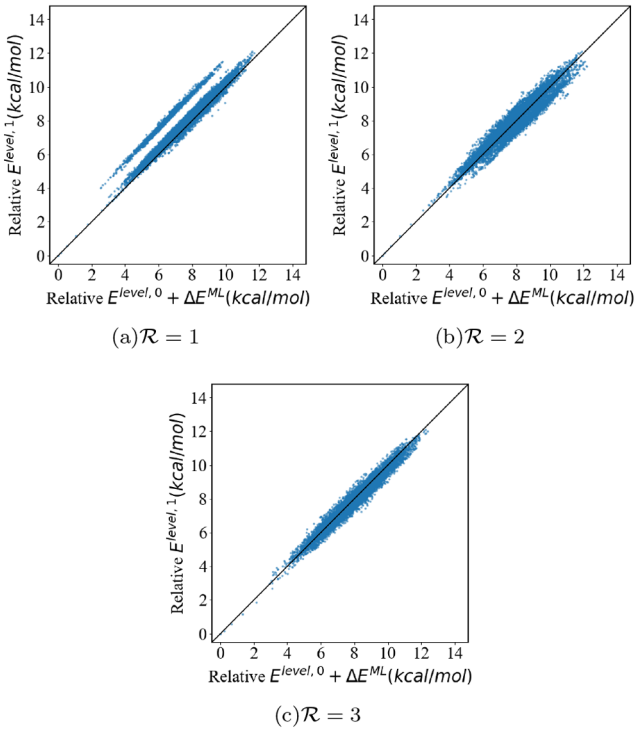


Figure 13. Similar to Figure 12, but with a constant shift added to all energy values from Eq. 19. Figure emphasizes the high degree of correlation between the machine-learning approximations and the high-level energy for all values of $\mathcal{R}$.

In Tables II and III we further analyze the performance of our approach with the goal of comparing it with the well-known

Table II[a]

fragment types	Training (test) set size	Training (test) set MAE in kcal/mol
H_2O	10 (38 910)	0.004 (0.007)
H_3O^+	60 (17 046)	0.039 (0.050)
H_4O_2	400 (62 380)	0.024 (0.040)
$H_5O_2^+$	1000 (77 510)	0.019 (0.044)
H_6O_3	1200 (46 940)	0.013 (0.045)
$H_7O_3^+$	4000 (139 580)	0.040 (0.047)
H_8O_4	900 (15 750)	0.038 (0.047)
$H_9O_4^+$	7000 (124 140)	0.041 (0.044)
$H_{13}O_6^+$	0 (9326)	− (0.686)

[a]Table complements Figure 10a and 11a and is presented in the same format as in Table I but also includes the full system extrapolation MAE from Eq. 19 on the last row. On the last row, the "0" and the "−" emphasize that no training was done on the full system, and in fact the fragment neural networks trained as per the data provided are tested both on fragments as well as on the full system, with testing errors noted.

NNP approach for potential energy surface extrapolation.[109] These tables complement Table I but now include performance data from the FES and FFS protocols. As already seen in Figure 10, Tables II and III show how the amount of training data increases between the FES and FFS protocols. However, in all cases the training data are less than 10%, which is in sharp contrast with other neural network models for constructing potential surfaces. Additionally, we emphasize in the bottom rows of Tables II and III that no training is done for the full

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

Table III[a]

fragment types	training (test) set size	training (test) set MAE in kcal/mol
H_2O	3000 (38 910)	0.001 (0.001)
H_3O^+	1700 (17 046)	0.004 (0.005)
H_4O_2	6000 (62 380)	0.008 (0.008)
$H_5O_2^+$	6000 (77 510)	0.012 (0.013)
H_6O_3	4000 (46 940)	0.015 (0.016)
$H_7O_3^+$	12000 (139 580)	0.018 (0.019)
H_8O_4	1500 (15 750)	0.029 (0.029)
$H_9O_4^+$	12000 (124 140)	0.026 (0.027)
$H_{13}O_6^+$	0 (9326)	− (0.241)

[a]Similar to Table II. Table complements Figure 10b and 11b and includes the full system extrapolation MAE from Eq. 19 on the last row. On the last row, the "0" and the "−" emphasize that no training was done on the full system, and in fact the fragment neural networks trained as per the data provided are tested both on fragments as well as on the full system, with testing errors noted.

system, as Eq. 19 provides a well-defined scheme to *transfer* the trained fragment neural networks to the full system. Again, the total energies are within the kJ/mol range when the FFS approach is used. This may be contrasted with the performance of NNPs in ref 109, where an MAE per atom on the order of 0.07 kcal/mol results in an error upward of 1 kcal/mol in the solvated Zundel CCSD(T) energies.

These results are particularly encouraging, since the fragmentation results already are several orders of magnitude lower in computational cost as compared to the regular high-level, post-Hartree−Fock electronic structure calculations. Thus, these studies bode well for future accurate and efficient AIMD studies at high levels of electronic structure. In future publications we will construct on-the-fly learning paradigms based on the above protocols to arrive at ML-enabled graph-theoretic fragmentation methods for AIMD.

VI. CONCLUSIONS

In a series of publications, we have shown how a graph-theory-based molecular fragmentation approach can be used to obtain smooth, post-Hartree−Fock AIMD trajectories and accurate post-Hartree−Fock molecular potential energy surfaces for fluxional systems at DFT cost. Here, a full system is divided into individual units that are treated as nodes in a graph. These nodes are then connected to form edges, faces, and higher-order simplexes as allowed by the graph-theoretic decomposition of molecular assemblies and condensed-phase systems. The energy and gradients obtained from the molecular fragments associated with these simplexes are then computed at two levels of theory, as in ONIOM, and then combined with the full-system energy (and gradients) obtained at a lower level of theory (DFT here) to obtain a graph-theoretic expression that has close connections to several molecular fragmentation approaches as well as many-body approximations.

However, for polarizable systems, such as protonated water clusters treated here, the graph-theoretic decomposition still requires higher-order interactions to be included. As per the graphical decomposition of molecular and condensed-phase systems used in our approach, the higher-order many-body interactions are captured from higher-order simplexes depicted within the graph. These involve larger-sized molecular fragments that need to be processed at post-Hartree−Fock levels of electronic structure theory. While these calculations do not directly affect the computational scaling of the graph-theoretic fragmentation method, since the size of these clusters is independent of the size of the overall system, these calculations can still be prohibitive when high-order interactions are needed to improve accuracy.

In this publication, we have introduced a machine learning approach to extrapolate these higher-level fragment energies. We introduce an approach involving an ensemble of neural networks, one set of neural networks for each fragment type, that helps us to extrapolate post-Hartree−Fock electronic energies for larger-size protonated water clusters. Specifically, the higher-level electronic structure energies for each graph-theoretically obtained fragment is extrapolated using a series of cascading neural networks, where each neural network provides an improvement to the electronic energy approximation arising from the earlier set of networks. The training data for these neural networks is determined using a multidimensional tessellation algorithm known as k-means clustering.

When the associated fragment neural networks are transferred into the original graph-theoretic molecular fragmentation expression, we find a general and robust scheme to extrapolate the total energy for a larger system where no explicit training has been conducted. As a consequence of this, we find that our extrapolation techniques combined with neural networks for fragments yield energies in agreement with CCSD results in the kJ/mol range and shows the potential to reduce the number of electronic structure calculations in AIMD trajectory by a factor of 10.

APPENDIX A. THE RATIONALE FOR MACHINE LEARNING APPROXIMATIONS TO EQ. 4

It is very clear from the above discussion that the complexity of the electronic structure problem could be tremendously reduced by using Eq. 4. For example, for a system with M basis functions a CCSD calculation would scale as $O(M^6)$. On the contrary, using Eq. 4, the computational complexity reduces to

$$O(M^{3.5}) + \sum_{r}^{\mathcal{R}} O([r \times M_F]^{3.5}) + O([r \times M_F]^6) \tag{A1}$$

where we have assumed that the electronic properties of the clusters within each node may be suitably represented using M_F basis functions within each node, and thus rank-r fragments will roughly need $[r * M_F]$ basis functions within rank-r fragments. The complexity of the $E^{\text{level},0}$ calculation in Eq. 4 is approximately $O(M^{3.5})$ (first term above), when the lower level of theory is DFT, and for rank-r objects the last two terms in Eq. A1 refer to the cost of computing the terms in Eq. 5.

For small values of $\mathcal{R}$, the terms in Eq. A1 are small as compared to $O(M^6)$, the computational cost when the full system is treated at the CCSD level of theory. But for cases where $\mathcal{R}$ is not negligible, the terms $\{O((r \times M_F)^6)\}$ may become a significant bottleneck. In refs 54 and 56 it has been shown that $\mathcal{R} = 2$ or $\mathcal{R} = 3$ (three-body and four-body interactions) may be necessary to obtain very high accuracy in AIMD trajectories and in periodic systems. For example, in Figure 5a, we show that the number of rank-3 simplexes may grow for strongly hydrogen-bonded systems, such as water clusters, where sub-kJ/mol accuracy is desired as seen in Figure 14. Figures 5 and 14 are obtained from a set of 9326 different geometries of the solvated Zundel $[H_2\ O_6\ H^+]$ system, obtained from an AIMD trajectory where Eq. 4 is computed at every step, and the mean absolute error is plotted in Figure 14 in comparison with the case

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

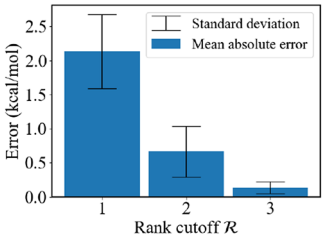


Figure 14. MAE between $E^{\mathrm{ONIOM}}{}_{\mathrm{MBE,gt}}$ in Eq. 4 and $E^{\mathrm{level,1}}$ for all the geometries obtained from an AIMD trajectory for the solvated Zundel $[H_2\ O_6\ H^+]$ system. (level,1 is CCSD all through this paper.) The trajectory contains 9326 different geometries.

where the full system energy is computed at the CCSD level of theory for every geometry. Similarly, Figure 5a shows the number of simplexes of each kind. Clearly, while Eq. 4 does reduce the complexity, as noted in ref 54, in general, the number of higher-rank simplexes may increase thus presenting a significant secondary computational challenge as quantified in the caption for Figure 5.

APPENDIX B. BENEFITS OF EXTRAPOLATING $\Delta E^{\mathrm{LEVEL,1;LEVEL,0}}_{\alpha,R}$ AS OPPOSED TO $E^{\mathrm{LEVEL,1}}_{\alpha,R}$ OR $E^{\mathrm{LEVEL,1}}$ FOR FULL SYSTEM

The machine learning algorithms for $\{Q_{\alpha,r}\}$ are expected to require less training as compared to that for Q, due to the likely exponential scaling of the number of representative nuclear configurations with number of nuclear degrees of freedom. That is, a higher-dimensional function would generally be harder to fit. This effect is shown in Figure 15. For example, in Figure 15,

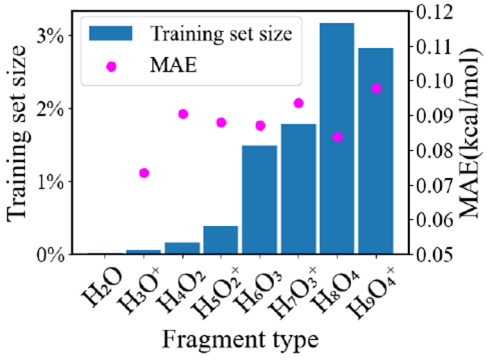


Figure 15. Percentage of training data used from within the full data set of fragments obtained from an AIMD trajectory referred to in Figure 5. The error in extrapolation from the neural network models constructed from such a training procedure, referred to as $Q_{\alpha,r}$ in Section III, is shown using pink dots. The figure shows that, to retain similar errors, the training data size grows in a steep fashion with fragment size.

the pink dots represent the error in extrapolation when training data of size corresponding to the blue histograms is used. If we compare the larger protonated water cluster with the smaller protonated water cluster, the fraction of data used increases with system size while the error remains relatively constant. This is also independently true for the neutral water clusters.

Additionally, one may be curious to ask why the difference in energy is used for extrapolation in Eqs. 9 and (18) as against the respective *level*, 1 energy values. In Figure 16 we provide the energy range and standard deviation for the difference in simplex energy and level, 1 energy. For every simplex generated from a solvated zundel system, the $E^{\mathrm{level,1}}$ energy range and fluctuation are much higher for all types of fragments (Figure 16a) as opposed to difference in energies (Figure 16b). The larger range and standard deviation in energy values make it harder to construct convergent learning paradigms to compute fragment $E^{\mathrm{level,1}}$ values.

We also note from Figure 16a,b that the protonated subsystems, H_3O^+, $H_6O_2{}^+$, $H_7O^+{}_3$, and $H_9O^+{}_4$, have a larger energy range for both $E^{\mathrm{level,1}}$ and $\Delta E^{\mathrm{level,1;level,0}}_{\alpha,r}$ values. This is because the AIMD trajectory used here involves proton hops as part of the trajectory. Hence higher-energy configurations are sampled as part of the trajectory by the protonated subsystems to facilitate proton hops. This is a critical feature of the study here as compared to other standard applications of machine learning in chemistry. It is common to use machine learning to study equilibrium properties. Here, the goal is to also include rare events such as proton hops within the learning protocol.

APPENDIX C. A DETAILED COMPARISON OF OUR TRANSFER LEARNING APPROACH IN EQ 19 AND OTHER NEURAL NETWORK POTENTIAL METHODS

We explicitly compare our results with those from refs 108 and 109 where (a) the same system considered here (the solvated Zundel cation) is studied using NNPs constructed for smaller protonated water clusters and (b) a range of organic molecules are studied. In refs 108 and 109, the NNPs are constructed with the goal to obtain energies per atom that are then combined together to obtain the total energies for any arbitrary systems. Tables IV and V provide a brief summary and contrast the quality of our results from those obtained previously. The first line in both tables includes results from the current work that is then contrasted from previous NNP work. We specifically distinguish the training data extrapolation error, the transfer data extrapolation error, and the fraction of data used for the training process.

First, as highlighted in Tables IV and V, our fragment extrapolations need only 10% training to obtain accuracy in the 1/100th kcal/mol range during the testing cycle. Furthermore, this error is for the entire training system energy and not error in per-atom energies as is normally reported in the refs 108 and 109. This aspect is to be contrasted with the results in ref 108 where the error in energy per atom error is on the order of 0.01 kcal/mol/atom during the testing cycle. See footnote b in Table IV. But this level of accuracy in refs 108−110 requires 80−90% of data during the training cycle and 10−20% data during the testing cycle that may be contrasted with our training efficiency. Since obtaining training data is normally a computational bottleneck and entails post-Hartree−Fock calculations, it appears that this is may be one key advantage in our approach.

Second, in our case, the transfer learning process occurs through the graph-theoretic fragmentation scheme, that is, eq 19, where the fragment neural networks appear on the right side (with final values given by $\Delta E^{\mathrm{level,1;level,0}}_{\alpha,r,\mathrm{ML}}$). It must also be noted that the graph-theoretic fragmentation scheme has independently been benchmarked over several years, and this NN extension to it provides an accelerated transfer-learning protocol. At the end, as shown in Table IV, the results for the left side of eq 19 have an accuracy in the kJ/mol range when ML models are used for the individual terms on the right side. This exceeds the accuracy shown by NNPs in ref 109 by nearly a factor of 5 and is also seen in the table above.

It is critical to note that, while all other methods listed in the Table I use NNPs to compute energies per atom that are then

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

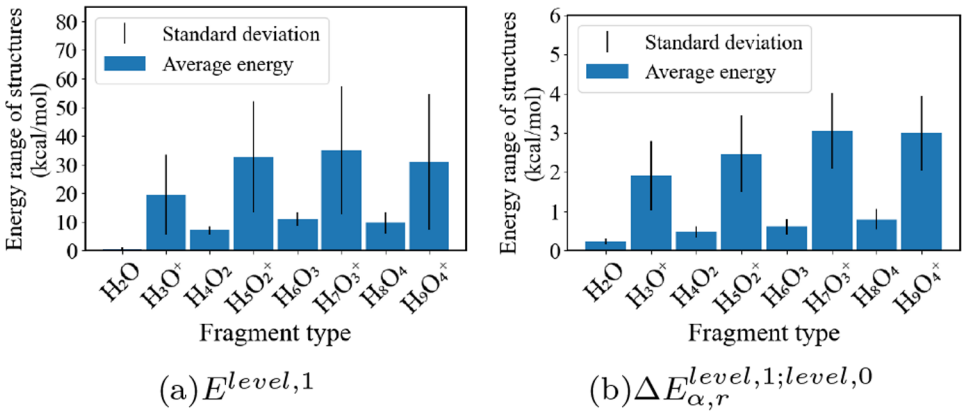


Figure 16. Average energy range and standard deviation for $E_{\alpha,r}^{\text{level,1}}$ (a) and $\Delta E_{\alpha,r}^{\text{level,1;level,0}}$ (b) for each type of fragment obtained from a solvated Zundel AIMD trajectory mentioned in Section V.A and in Figures 5 and 15. As we will see later the smaller range and standard deviation of $\Delta E_{\alpha,r}^{\text{level,1;level,0}}$ make extrapolation easier.

Table IV. Summary of Neural Network Potential Performance for Water Clusters with Goal for Coupled Cluster Accuracy

study	training systems	test set error (kcal/mol)	transfer systems	transfer set error (kcal/mol)	training data fraction
this work[a]	$(H_2O)_{1;2}$	0.001; 0.005	$H(H_2O)_6^+$	0.24	10%
	$H_2O)_{3;4}$	0.008; 0.013			
	$H(H_2O)_{1;2}^+$	0.016; 0.019			
	$H(H_2O)_{3;4}^+$	0.029; 0.027			
	H_2O	0.021			
ref 108[b]	$H(H_2O)_{1;2}^+$	0.048; 0.154			90%
	$H(H_2O)_{3;4}^+$	0.171; 0.312			
ref 109[c]			$H(H_2O)_6^+$	1.31	
ref 110[d]	$(H_2O)_{2;3}$	0.079; 0.063	$(H_2O)_{4-6}$	<0.3	81%

[a]Neural network MAEs for all fragments produced from graph-theoretic decomposition of solvated Zundel. Training data obtained using k-means clustering as discussed in the paper. [b]The work provides training data errors on "per atom basis". In this table we have converted this per atom error to total error for the system by simply multiplying by the number of atoms. The actual per atom root-mean-square error for the training systems H_2O, H_3O^+, $H_5O_2^+$, $H_7O_3^+$, and $H_9O_4^+$ are 0.007, 0.012, 0.022, 0.019, and 0.024 kcal/mol/atom, respectively. [c]This work uses NN models produced in ref 108 for transfer learning and provides the error in per atom basis as 0.069 kcal/mol/atom with a 0.038 kcal/mol/atom shift. Again, the error quoted in this table is obtained by multiplying the per atom error with the number of atoms. [d]As noted in the paper, neutral water clusters generally have a smaller energy range compared to protonated water clusters. Yet our work on protonated water clusters provides similar accuracy for transfer learning as for neutral clusters in ref 110.

Table V. A Summary of Neural Network Potential Performance on Organic Systems with a Goal for Coupled Cluster Accuracy[c]

study	training systems	test set error (kcal/mol)	transfer systems	transfer set error (kcal/mol)	training data fraction
our work	$(H_2O)_{1;2}$	0.001; 0.005	$H(H_2O)_6^+$	0.24	10%
	$H_2O)_{3;4}$	0.008; 0.013			
	$H(H_2O)_{1;2}^+$	0.016; 0.019			
	$H(H_2O)_{3;4}^+$	0.029; 0.027			
ref 111[a]	$H + CH_3OH$	(0.49)			90%
ref 112[a]	$H_2 + SH \rightarrow H + H_2S$	(0.07)[b]			90%
ref 113[a]	O_2CH, H_3S, H_4N,	(0.13), (0.07), (0.08),			90%
	H_3C, H_5CO	(0.12), (0.09)[b]			
ref 114	$H_2O^+ + H_2 \rightarrow$	(0.17)[b]			90%
	$H_3O^+ + H$				

[a]These studies provide comparison among multiple types of neural network models. Here we only provide the best error for each system. [b]Errors in parentheses are provided as root-mean-square error. By contrast, the errors in our work are presented as MAE. [c]Our work is provided here simply for completeness.

accumulated to obtain energies for any other system, our approach is fundamentally different in that the NNPs here provide fragment energies that are then used within the graph-theoretic expression to perform the transfer learning process. Thus, combined with the limited need for training and higher accuracy, we feel that the approach discussed here provides a competitive and novel option to creating NNPs.

APPENDIX D. UPPER BOUND TO ERRORS FROM EQ 19

By using the ML ideas proposed here, one can easily arrive at upper bounds to overall error as

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

$$|\epsilon_{\mathcal{R}}^{\mathrm{ML}}| \leq \sum_{r=0}^{\mathcal{R}} \sum_{\alpha \in \mathbf{V}_r} \mathcal{M}_{\alpha}^{r} |\epsilon_{\alpha,r}^{\mathrm{ML}}| \approx \sum_{r=0}^{\mathcal{R}} \mathcal{N}_r \mathcal{E}_r \tag{D1}$$

where the left side, $|\epsilon_{\mathcal{R}}^{\mathrm{ML}}|$, represents an estimate of the reconstructed full system mean absolute error. On the right side, the quantity $|\epsilon_{\alpha,r}^{\mathrm{ML}}|$ represents the MAE for the neural network constructed for a specific fragment type, depicted here using indices (α, r). Furthermore, an estimate of the total neural network error for a given rank is represented as $\mathcal{E}_r$ with $\mathcal{N}_r$ being the total number of fragments of a given type or given simplex rank, in one full system geometry. Clearly for complex systems, the error will be dominated by the size of $\mathcal{N}_r$ and there will be a need to drive down the errors $\mathcal{E}_r$ so as to maintain the products $\mathcal{N}_r\mathcal{E}_r$ to within acceptable accuracy. Furthermore, eq D1 only provides an estimate, and the true error for full system is also affected by the alternating signs of individual fragment errors as seen from eq 4, which may lead to some degree of error cancellation. Regardless, the error for the full system is expected to be far greater than individual neural-network model errors because of the number of fragments $\mathcal{N}_r$, as we will discuss later in Section V.

AUTHOR INFORMATION

Corresponding Author

Srinivasan S. Iyengar − *Department of Chemistry and Department of Physics, Indiana University, Bloomington 47405 Indiana, United States;* orcid.org/0000-0001-6526-2907; Email: iyengar@indiana.edu

Author

Xiao Zhu − *Department of Chemistry and Department of Physics, Indiana University, Bloomington 47405 Indiana, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c01241

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (Grant No. CHE-2102610 to S.S.I.).

REFERENCES

(1) Murrell, J.; Carter, S.; Farantos, S.; Huxley, P.; Varandas, A. *Molecular Potential Energy Functions;* Wiley: New York, 1984.

(2) Rabitz, H.; Alış, Ö. F. General foundations of high dimensional model representations. *J. Math. Chem.* **1999**, *25*, 197−233.

(3) Manzhos, S.; Carrington, T., Jr. Using neural networks, optimized coordinates, and high-dimensional model representations to obtain a vinyl bromide potential surface. *J. Chem. Phys.* **2008**, *129*, 224104.

(4) Braams, B. J.; Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577.

(5) Xie, Z.; Bowman, J. M. Permutationally Invariant Polynomial Basis for Molecular Energy Surface Fitting Via Monomial Symmetrization. *J. Chem. Theory and Comput.* **2010**, *6*, 26.

(6) Otto, F. Multi-layer Potfit: An accurate potential representation for efficient high-dimensional quantum dynamics. *J. Chem. Phys.* **2014**, *140*, 014106.

(7) Peláez, D.; Meyer, H.-D. The multigrid POTFIT (MGPF) method: Grid representations of potentials for quantum dynamics of large systems. *J. Chem. Phys.* **2013**, *138*, 014108.

(8) Sumner, I.; Iyengar, S. S. Quantum Wavepacket *Ab Initio* Molecular Dynamics: An Approach for Computing Dynamically Averaged Vibrational Spectra Including Critical Nuclear Quantum Effects. *J. Phys. Chem. A* **2007**, *111*, 10313.

(9) DeGregorio, N.; Iyengar, S. S. Efficient and Adaptive Methods for Computing Accurate Potential Surfaces for Quantum Nuclear Effects: Applications to Hydrogen-Transfer Reactions. *J. Chem. Theory Comput.h* **2018**, *14*, 30−47.

(10) Hammer, N. I.; Diken, E. G.; Roscioli, J. R.; Johnson, M. A.; Myshakin, E. M.; Jordan, K. D.; McCoy, A. B.; Huang, X.; Bowman, J. M.; Carter, S. The Vibrational Predissociation Spectra of the $H_5O_2^+$·RG_n(RG = Ar,Ne) clusters: Correlation of the solvent perturbations in the free OH and shared proton transitions of the Zundel ion. *J. Chem. Phys.* **2005**, *122*, 244301.

(11) Diken, E. G.; Headrick, J. M.; Roscioli, J. R.; Bopp, J. C.; Johnson, M. A.; McCoy, A. B. Fundamental Excitations of the Shared Proton in the $H_3O_2^-$ and $H_5O_2^+$ Complexes. *J. Phys. Chem. A* **2005**, *109*, 1487.

(12) Kaledin, M.; Kaledin, A. L.; Bowman, J. M. Vibrational Analysis of the $H_5O_2^+$ Infrared Spectrum Using Molecular and Driven Molecular Dynamics. *J. Phys. Chem. A* **2006**, *110*, 2933.

(13) Pomes, R.; Roux, B. Structure and Dynamics of a Proton Wire: A Theoretical Study of H. Translocation Along the Single-File Water Chain in the Gramicidin a Channel. *Biophys. J.* **1996**, *71*, 19.

(14) Pomes, R.; Roux, B. Theoretical Study of H^+ Translocation Along a Model Proton Wire. *J. Phys. Chem.* **1996**, *100*, 2519.

(15) Decornez, H.; Drukker, K.; Hammes-Schiffer, S. Solvation and Hydrogen-Bonding Effects on Proton Wires. *J. Phys. Chem. A* **1999**, *103*, 2891.

(16) Brewer, M. L.; Schmitt, U. W.; Voth, G. A. The Formation and Dynamics of Proton Wires in Channel Environments. *Biophys. J.* **2001**, *80*, 1691.

(17) Teeter, M. M. Water Structure of a Hydrophobic Protein at Atomic Resolution: Pentagon Rings of Water Molecules in Crystals of Crambin. *Proc. Natl. Acad. Sci. U.S.A* **1984**, *81*, 6014.

(18) Neidle, S.; Berman, H. M.; Shieh, H. S. Highly Structured Water Network in Crystals of a Deoxydinucleoside-Drug Complex. *Nature* **1980**, *288*, 129.

(19) Lipscomb, L. A.; Peek, M. E.; Zhou, F. X.; Bertrand, J. A.; VanDerveer, D.; Williams, L. D. Water Ring Structure at Dna Interfaces - Hydration and Dynamics of Dna Anthracycline Complexes. *Biochemistry* **1994**, *33*, 3649.

(20) Tu, C.; Rowlett, R. S.; Tripp, B. C.; Ferry, J. G.; Silverman, D. N. Chemical Rescue of Proton Transfer in Catalysis by Carbonic Anhydrases in the Beta- and Gamma-Class. *Biochemistry* **2002**, *41*, 15429.

(21) McEwan, M. J.; Phillips, L. F. *Chemistry of the Atmosphere;* Eward Arnold: London, England, 1975.

(22) Wayne, R. P. *Chemistry of the Atmosphere;* Clarendon Press: Oxford, England, 1994.

(23) Dietrick, S. M.; Pacheco, A. B.; Phatak, P.; Stevens, P. S.; Iyengar, S. S. The Influence of Water on Anharmonicity, Stability and Vibrational Energy Distribution of Hydrogen-Bonded Adducts in Atmospheric Reactions: Case Study of the OH + Isoprene Reaction Intermediate Using *Ab-Initio* Molecular Dynamics. *J. Phys. Chem. A* **2012**, *116*, 399.

(24) Iyengar, S. S. Dynamical Effects on Vibrational and Electronic Spectra of Hydroperoxyl Radical Water Clusters. *J. Chem. Phys.* **2005**, *123*, 084310.

(25) Iannuzzi, M.; Parrinello, M. Proton Transfer in Heterocycle Crystals. *Phys. Rev. Lett.* **2004**, *93*, 025901.

(26) Tse, Y.-L. S.; Herring, A. M.; Kim, K.; Voth, G. A. Molecular dynamics simulations of proton transport in 3M and nafion perfluorosulfonic acid membranes. *J. Phys. Chem. C* **2013**, *117*, 8079−8091.

(27) Lin, I-H.; Lu, Y.-H.; Chen, H.-T. Nitrogen-doped C_{60} as a robust catalyst for CO oxidation. *J. Comput. Chem.* **2017**, *38*, 2041−2046.

(28) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313*, 701.

(29) Hopkins, B. W.; Tschumper, G. S. A multicentered approach to integrated QM/QM calculations. Applications to multiply hydrogen bonded systems. *Journal of computational chemistry* **2003**, *24*, 1563.
(30) Zhang, D. W.; Zhang, J. Z. H. Molecular Fractionation with Conjugate Caps for Full Quantum Mechanical Calculation of Protein-molecule Interaction Energy. *J. Chem. Phys.* **2003**, *119*, 3599.
(31) Hopkins, B. W.; Tschumper, G. S. Multicentred QM/QM Methods for Overlapping Model Systems. *Mol. Phys.* **2005**, *103*, 309.
(32) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular Tailoring Approach for Geometry Optimization of Large Molecules: Energy Evaluation and Parallelization Strategies. *J. Chem. Phys.* **2006**, *125*, 104109.
(33) Wang, L.-W.; Zhao, Z.; Meza, J. Linear-scaling three-dimensional fragment method for large-scale electronic structure calculations. *Phys. Rev. B* **2008**, *77*, 165113.
(34) Gordon, M.; Mullin, J.; Pruitt, S.; Roskop, L.; Slipchenko, L.; Boatz, J. Accurate Methods for Large Molecular Systems. *J. Phys. Chem. B* **2009**, *113*, 9646.
(35) Mayhall, N. J.; Raghavachari, K. Molecules-In-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.* **2011**, *7*, 1336.
(36) Mayhall, N. J.; Raghavachari, K. Many-overlapping-body (MOB) expansion: A generalized many body expansion for nondisjoint monomers in molecular fragmentation calculations of covalent molecules. *J. Chem. Theory Comput.* **2012**, *8*, 2669−2675.
(37) Collins, M. A. Systematic Fragmentation of Large Molecules by Annihilation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7744.
(38) Wen, S.; Nanda, K.; Huang, Y.; Beran, G. J. Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7578.
(39) Le, H.-A.; Tan, H.-J.; Ouyang, J. F.; Bettens, R. P. A. Combined Fragmentation Method: A Simple Method for Fragmentation of Large Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 469.
(40) Frankcombe, T. J.; Collins, M. A. Growing Fragmented Potentials for Gas-Surface Reactions: The Reaction between Hydrogen Atoms and Hydrogen-Terminated Silicon (111). *J. Phys. Chem. C* **2012**, *116*, 7793−7802.
(41) Sahu, N.; Yeole, S. D.; Gadre, S. R. Appraisal of molecular tailoring approach for large clusters. *J. Chem. Phys.* **2013**, *138*, 104101.
(42) Lange, A. W.; Voth, G. A. Multi-State Approach to Chemical Reactivity in Fragment Based Quantum Chemistry Calculations. *J. Chem. Theory Comput.* **2013**, *9*, 4018.
(43) Li, S.; Li, W.; Ma, J. Generalized Energy-Based Fragmentation Approach and Its Applications to Macromolecules and Molecular Aggregates. *Acc. Chem. Res.* **2014**, *47*, 2712.
(44) Raghavachari, K.; Saha, A. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015**, *115*, 5643−5677.
(45) Saha, A.; Raghavachari, K. Analysis of different fragmentation strategies on a variety of large peptides: Implementation of a low level of theory in fragment-based methods can be a crucial factor. *J. Chem. Theory Comput.* **2015**, *11*, 2012.
(46) Liu, J.; Qi, L.-W.; Zhang, J. Z. H.; He, X. Fragment Quantum Mechanical Method for Large-Sized Ion − Water Clusters. *J. Chem. Theory Comput.* **2017**, *13*, 2021.
(47) Herbert, J. M. Fantasy versus reality in fragment-based quantum chemistry. *J. Chem. Phys.* **2019**, *151*, 170901.
(48) Li, J.; Iyengar, S. S. Ab initio Molecular Dynamics using Recursive, Spatially Separated, Overlapping Model Subsystems Mixed Within an ONIOM Based Fragmentation Energy Extrapolation Technique. *J. Chem. Theory Comput.* **2015**, *11*, 3978.
(49) Li, J.; Haycraft, C.; Iyengar, S. S. Hybrid extended Lagrangian, post-Hartree-Fock Born-Oppenheimer ab initio molecular dynamics using fragment-based electronic structure. *J. Chem. Theory Comput.* **2016**, *12*, 2493.
(50) Haycraft, C.; Li, J.; Iyengar, S. S. Efficient,"On-the-Fly", Born−Oppenheimer and Car-Parrinello−type Dynamics with Coupled Cluster Accuracy through Fragment Based Electronic Structure. *J. Chem. Theory Comput.* **2017**, *13*, 1887.
(51) Ricard, T. C.; Haycraft, C.; Iyengar, S. S. Adaptive, geometric networks for efficient coarse-grained ab initio molecular dynamics with post-Hartree-Fock accuracy. *J. Chem. Theory Comput.* **2018**, *14*, 2852.
(52) Ricard, T. C.; Iyengar, S. S. Efficiently capturing weak interactions in *ab initio* molecular dynamics through "on-the-fly" basis set extrapolation. *J. Chem. Theory Comput.* **2018**, *14*, 5535.
(53) Kumar, A.; Iyengar, S. S. Fragment-based electronic structure for potential energy surfaces using a superposition of fragmentation topologies. *J. Chem. Theory Comput.* **2019**, *15*, 5769.
(54) Ricard, T. C.; Iyengar, S. S. An efficient and accurate approach to estimate hybrid functional and large basis set contributions to condensed phase systems and molecule-surface interactions. *J. Chem. Theory Comput.* **2020**, *16*, 4790.
(55) Ricard, T. C.; Kumar, A.; Iyengar, S. S. Embedded, graph-theoretically defined many-body approximations for wavefunction-in-DFT and DFT-in-DFT: applications to gas- and condensed-phase AIMD, and potential surfaces for quantum nuclear effects. *Int. J. Quantum Chem.* **2020**, *120*, e26244.
(56) Zhang, J. H.; Ricard, T. C.; Haycraft, C.; Iyengar, S. S. Weighted-Graph-Theoretic Methods for Many-Body Corrections within ONIOM: Smooth AIMD and the Role of High-Order Many-Body Terms. *J. Chem. Theory Comput.* **2021**, *17*, 2672−2690.
(57) Kumar, A.; DeGregorio, N.; Iyengar, S. S. Graph-Theory-Based Molecular Fragmentation for Efficient and Accurate Potential Surface Calculations in Multiple Dimensions. *J. Chem. Theory Comput.* **2021**, *17*, 6671−6690.
(58) Varandas, A. J.; Murrell, J. N. A many-body expansion of polyatomic potential energy surfaces: application to H_n systems. *Faraday Discuss. Chem. Soc.* **1977**, *62*, 92.
(59) Varandas, A.; Pais, A. A realistic double many-body expansion (DMBE) potential energy surface for ground-state O_3 from a multiproperty fit to ab initio calculations, and to experimental spectroscopic, inelastic scattering, and kinetic isotope thermal rate data. *Mol. Phys.* **1988**, *65*, 843.
(60) Xantheas, S. S. *Ab Initio* Studies of Cyclic Water Clusters $(H_2O)_n$, N = 1−6. II. Analysis of Many-body Interactions. *J. Chem. Phys.* **1994**, *100*, 7523.
(61) Xantheas, S. S. *Ab Initio* Studies of Cyclic Water Clusters $(H_2O)_n$, N = 1−6. III. Comparison of Density Functional with MP2 Results. *J. Chem. Phys.* **1995**, *102*, 4505.
(62) Dahlke, E. E.; Truhlar, D. G. Electrostatically Embedded Many Body Expansion for Large Systems, with Applications to Water Clusters. *J. Chem. Theory Comput.* **2007**, *3*, 46.
(63) Bygrave, P. J.; Allan, N. L.; Manby, F. R. The embedded many-body expansion for energetics of molecular crystals. *J. Chem. Phys.* **2012**, *137*, 164102.
(64) Yang, J.; Hu, W.; Usvyat, D.; Matthews, D.; Schuetz, M.; Chan, G. K.-L. Ab initio determination of the crystalline benzene lattice energy to sub-kilojoule/mol accuracy. *Science* **2014**, *345*, 640.
(65) Yu, Q.; Bowman, J. M. Communication: VSCF/VCI vibrational spectroscopy of $H_7O_3^+$ and $H_9O_4^+$ using high-level, many-body potential energy surface and dipole moment surfaces. *J. Chem. Phys.* **2017**, *146*, 121102.
(66) Liu, K.-Y.; Herbert, J. M. Energy-Screened Many-Body Expansion: A Practical Yet Accurate Fragmentation Method for Quantum Chemistry. *J. Chem. Theory Comput.* **2020**, *16*, 475.
(67) Maseras, F.; Morokuma, K. IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.* **1995**, *16*, 1170.
(68) Svensson, M.; Humbel, S.; Froese, R. D.; Matsubara, T.; Sieber, S.; Morokuma, K. ONIOM: a multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and $Pt(P(t\text{-}Bu)_3)_2 + H_2$ oxidative addition. *J. Phys. Chem.* **1996**, *100*, 19357.
(69) Kerdcharoen, T.; Morokuma, K. ONIOM-XS: An Extension of the ONIOM Method for Molecular Simulation in Condensed Phase. *Chem. Phys. Lett.* **2002**, *355*, 257.

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

(70) Guo, W.; Wu, A.; Xu, X. XO: An Extended ONIOM Method for Accurate and Efficient Geometry Optimization of Large Molecules. *Chem. Phys. Lett.* **2010**, *498*, 203.
(71) Chung, L. W.; Hirao, H.; Li, X.; Morokuma, K. The ONIOM Method: Its Foundation and Applications to Metalloenzymes and Photobiology. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 327.
(72) Chung, L. W.; Sameera, W. M. C.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H.-B.; Ding, L.; Morokuma, K. The ONIOM Method and Its Applications. *Chem. Rev.* **2015**, *115*, 5678.
(73) Zhang, J. H.; Iyengar, S. S. Graph-|Q⟩ ⟨Cl: A Graph-based Quantum-classical algorithm for efficient electronic structure on hybrid quantum/classical hardware systems: Improved quantum circuit depth performance. *J. Chem. Theory Comput.* **2022**, *18*, 2885.
(74) Collins, M. A. Can Systematic Molecular Fragmentation Be Applied to Direct Ab Initio Molecular Dynamics? *J. Phys. Chem. A* **2016**, *120*, 9281.
(75) Alıs, O. F.; Rabitz, H. Efficient Implementation of High Dimensional Model Representations. *J. Math. Chem.* **2001**, *29*, 127−142.
(76) Alış, Ö. F.; Rabitz, H. Efficient implementation of high dimensional model representations. *J. Math. Chem.* **2001**, *29*, 127−142.
(77) Jäckle, A.; Meyer, H.-D. Product representation of potential energy surfaces. *J. Chem. Phys.* **1996**, *104*, 7974−7984.
(78) Jäckle, A.; Meyer, H.-D. Product representation of potential energy surfaces. II. *J. Chem. Phys.* **1998**, *109*, 3772−3779.
(79) Conte, R.; Qu, C.; Bowman, J. M. Permutationally invariant fitting of many-body, non-covalent interactions with application to three-body methane−water−water. *J. Chem. Theory Comput.* **2015**, *11*, 1631−1638.
(80) Lorentz, G.; Golitschek, M.; Makovoz, Y. *Constructive Approximation;* Springer: New York, 1996.
(81) Girosi, F.; Poggio, T. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neurocomputing* **1989**, *1*, 465−469.
(82) Sobol', I. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 86−112.
(83) Sculley, D. Web-scale k-means clustering. *WWW '10: Proceedings of the 19th international conference on World wide web;* Association for Computing Machinery: New York, NY, 2010; pp 1177−1178.
(84) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.
(85) Farin, G. Surfaces over Dirichlet Tessellations. *Computer Aided Geometric Design* **1990**, *7*, 281.
(86) Coffey, T. M.; Wyatt, R. E.; Schieve, W. C. Quantum Trajectories from Kinematic Considerations. *J. Phys. A: Mathematical and Theoretical* **2010**, *43*, 335301.
(87) Aurenhammer, F. Voronoi Diagrams — A survey of a fundamental geometric data structure. *ACM Comput. Survey* **1991**, *23*, 345.
(88) Okabe, A.; Boots, B.; Sugihara, K.; Chiu, S. N. *Spatial Tessellations—Concepts and applications of Voronoi diagrams;* John Wiley and Sons, 2000.
(89) Watson, D. F. Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *Comp. J.* **1981**, *24*, 167.
(90) Dey, T. K.; Shah, N. R. On the number of simplicial complexes in R^d. *Comput. Geom.* **1997**, *8*, 267.
(91) Adams, C. C.; Franzosa, R. D. *Introduction to topology: pure and applied;* Pearson, **2008**.
(92) Berger, M.; Pansu, P.; Berry, J.-P.; Saint-Raymond, X. Affine Spaces. In *Problems in Geometry;* Springer, 1984; p 11.
(93) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16,* rev. B.01; Gaussian Inc.: Wallingford, CT, 2016.
(94) Neese, F. The ORCA program system Wiley Interdiscip. *Rev. Comput. Mol. Sci.* **2012**, *2*, 73.
(95) Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F.; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. PSI4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.h* **2017**, *13*, 3185.
(96) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matt.* **2009**, *21*, 395502.
(97) Ozaki, T.; Kino, H.; Yu, J.; Han, M.; Ohfuchi, M.; Ishii, F.; Sawada, K.; Ohwaki, T.; Weng, H.; Toyoda, M.; Okuno, Y.; Perez, R.; Bell, P.; T, D.; Xiao, Y.; Ito, A.; Terakura, K. . *User's manual of OpenMX,* ver. 3.8; OpenMX, 2016.
(98) Zhang, J. H.; Ricard, T. C.; Haycraft, C.; Iyengar, S. S. Weighted-Graph-Theoretic Methods for Many-Body Corrections within ONIOM: Smooth AIMD and the Role of High-Order Many-Body Terms. *J. Chem. Theory. Comput.* **2021**, *17*, 2672−2690.
(99) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. Δ−machine learning for potential energy surfaces: A PIP approach to bring a DFT-based PES to CCSD(T) level of theory. *J. Chem. Phys.* **2021**, *154*, 051102.
(100) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(101) Chen, J.; Xu, W.; Zhang, R. -Machine learning-driven discovery of double hybrid organic-inorganic perovskites. *J. Mater. Chem. A* **2022**, *10*, 1402−1413.
(102) Unzueta, P. A.; Greenwell, C. S.; Beran, G. J. O. G. J. O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via -Machine Learning. *J. Chem. Theory Comput.* **2021**, *17*, 826−840.
(103) Haykin, S. S. Neural networks and learning machines, 3rd ed.; Pearson Education: Upper Saddle River, NJ, 2009.
(104) Björklund, A.; Husfeldt, T.; Koivisto, M. Set partitioning via inclusion-exclusion. *SIAM J. Comput.* **2009**, *39*, 546.
(105) Konc, J.; Janežić, D. An improved branch and bound algorithm for the maximum clique problem. *MATCH Commun. Math. Comput. Chem.* **2007**, *58*, 569−590.
(106) Karp, R. M. Reducibility among Combinatorial Problems. In *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations*; Yorktown Heights, NY, March 20−22, 1972; Miller, R. E., Thatcher, J. W., Bohlinger, J. D., Eds.; Springer US: Boston, MA, 1972; pp 85−103.
(107) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **1995**, *103*, 4129−4137.

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144

(108) Schran, C.; Behler, J.; Marx, D. Automated Fitting of Neural Network Potentials at Coupled Cluster Accuracy: Protonated Water Clusters as Testing Ground. *J. Chem. Theory Comput.* **2020**, *16*, 88−99.
(109) Schran, C.; Brieuc, F.; Marx, D. Transferability of machine learning potentials: Protonated water neural network potential applied to the protonated water hexamer. *J. Chem. Phys.* **2021**, *154*, 051101.
(110) Nguyen, T. T.; Szekely, E.; Imbalzano, G.; Behler, J.; Csanyi, G.; Ceriotti, M.; Goetz, A. W.; Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **2018**, *148*, 241725.
(111) Lu, D.; Behler, J.; Li, J. Accurate Global Potential Energy Surfaces for the H + CH_3OH Reaction by Neural Network Fitting with Permutation Invariance. *J. Phys. Chem. A* **2020**, *124*, 5737−5745.
(112) Lu, D.; Qi, J.; Yang, M.; Behler, J.; Song, H.; Li, J. Mode specific dynamics in the H_2 + SH → H + H_2S reaction. *Phys. Chem. Chem. Phys.* **2016**, *18*, 29113−29121.
(113) Li, J.; Song, K.; Behler, J. A critical comparison of neural network potentials for molecular reaction dynamics with exact permutation symmetry. *Phys. Chem. Chem. Phys.* **2019**, *21*, 9672−9682.
(114) Li, A.; Guo, H. A nine-dimensional ab initio global potential energy surface for the H_2O^+ + H_2 → H_3O^+ + H reaction. *J. Chem. Phys.* **2014**, *140*, 224313.
(115) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121*, 10037−10072.
(116) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schutt, K. T.; Tkatchenko, A.; Muller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121*, 10142−10186.
(117) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*. DOI: 10.1103/PhysRevLett.98.146401
(118) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. series c (applied statistics)* **1979**, *28*, 100−108.
(119) Manzhos, S.; Wang, X.; Dawes, R.; Carrington, T. A Nested Molecule-Independent Neural Network Approach for High-Quality Potential Fits. *J. Phys. Chem. A* **2006**, *110*, 5295−5304.
(120) Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). *arXiv*2018, abs/1803.08375. DOI: 10.48550/arXiv.1803.08375
(121) Yang, X.; Zhang, X.; Castleman, A. W., Jr Kinetics and Mechanism Studies of Large Protonated Water Clusters, $H^+(H2O)_N$,N = 1−60, at Thermal Energy. *Int. J. Mass Spectrom. Ion Proc.* **1991**, *109*, 339.
(122) Iyengar, S. S.; Petersen, M. K.; Day, T. J. F.; Burnham, C. J.; Teige, V. E.; Voth, G. A. The Properties of Ion-Water Clusters. I. the Protonated 21-Water Cluster. *J. Chem. Phys.* **2005**, *123*, 084309.
(123) Iyengar, S. S. Further Analysis of the Dynamically Averaged Vibrational Spectrum for the "magic" Protonated 21-Water Cluster. *J. Chem. Phys.* **2007**, *126*, 216101.
(124) Lynden-Bell, R. M.; Rasaiah, J. C. Mobility and Solvation of Ions in Channels. *J. Chem. Phys.* **1996**, *105*, 9266.
(125) Park, M.; Shin, I.; Singh, N. J.; Kim, K. S. Eigen and Zundel Forms of Small Protonated Water Clusters: Structures and Infrared Spectra. *J. Phys. Chem. A* **2007**, *111*, 10692−10702.
(126) Nagle, J. F.; Morowitz, H. J. Molecular mechanisms for proton transport in membranes. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 298−302.
(127) Ye, Y.-S.; Rick, J.; Hwang, B.-J. Water Soluble Polymers as Proton Exchange Membranes for Fuel Cells. *Polymers* **2012**, *4*, 913−963.
(128) Vener, M. V.; Rozanska, X.; Sauer, J. Protonation of water clusters in the cavities of acidic zeolites: (H2O)n·H-chabazite, n = 1−4. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1702−1712.
(129) Sadhukhan, S.; Munoz, D.; Adamo, C.; Scuseria, G. E. Predicting Proton Transfer Barriers with Density Functional Methods. *Chem. Phys. Lett.* **1999**, *306*, 83.
(130) Xie, Y.; Remington, R. B.; Schaefer, H. F., III The Protonated Water Dimer: Extensive Theoretical Studies of $H_5O_2^+$. *J. Chem. Phys.* **1994**, *101*, 4878.
(131) Tuckerman, M. E.; Ungar, P. J.; Vonrosenvinge, T.; Klein, M. L. *Ab Initio* Molecular Dynamics Simulations. *J. Phys. Chem.* **1996**, *100*, 12878.
(132) Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. *Ab Initio* Molecular Dynamics Simulation of the Solvation and Transport of H_3O^+ and OH^- Ions in Water. *J. Phys. Chem.* **1995**, *99*, 5749.
(133) Schmitt, U. W.; Voth, G. A. The computer simulation of proton transport in water. *J. Chem. Phys.* **1999**, *111*, 9361−9381.
(134) Schmitt, U. W.; Voth, G. A. Multistate empirical valence bond model for proton transport in water. *J. Phys. Chem. B* **1998**, *102*, 5547−5551.
(135) Klimes, J.; Michaelides, A. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.* **2012**, *137*, 120901.
(136) Peverati, R.; Truhlar, D. The Quest for a Universal Density Functional: The Accuracy of Density Functionals Across a Broad Spectrum of Databases in Chemistry and Physics. *Philos. Trans. Royal Soc. A* **2014**, *372*, 10120476.
(137) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289.
(138) Pomes, R.; Roux, B. Theoretical study of H+ translocation along a model proton wire. *J. Phys. Chem.* **1996**, *100*, 2519−2527.
(139) Teeter, M. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. *Proc. Natl. Acad. Sci. U. S. A.* **1984**, *81*, 6014−6018.
(140) Lipscomb, L. A.; Peek, M. E.; Zhou, F. X.; Bertrand, J. A.; VanDerveer, D.; Williams, L. D. Water ring structure at DNA interfaces: hydration and dynamics of DNA-anthracycline complexes. *Biochemistry* **1994**, *33*, 3649−3659.
(141) Tu, C.; Rowlett, R. S.; Tripp, B. C.; Ferry, J. G.; Silverman, D. N. Chemical rescue of proton transfer in catalysis by carbonic anhydrases in the β-and γ-class. *Biochemistry* **2002**, *41*, 15429−15435.
(142) Dietrick, S. M.; Iyengar, S. S. Constructing Periodic Phase Space Orbits from Ab Initio Molecular Dynamics Trajectories to Analyze Vibrational Spectra: Case Study of the Zundel (H5O2+) Cation. *J. Chem. Theory and Comput.* **2012**, *8*, 4876.
(143) Petersen, M. K.; Iyengar, S. S.; Day, T. J. F.; Voth, G. A. The Hydrated Proton at Water Liquid/Vapour Interfaces. *J. Phys. Chem. B* **2004**, *108*, 14804.
(144) Vendrell, O.; Gatti, F.; Meyer, H.-D. Dynamics and Infrared Spectroscopy of the Protonated Water Dimer. *Ang. Chem. Int. Ed.* **2007**, *46*, 6918.
(145) Headrick, J. M.; Diken, E. G.; Walters, R. S.; Hammer, N. I.; Christie, R. A.; Cui, J.; Myshakin, E. M.; Duncan, M. A.; Johnson, M. A.; Jordan, K. Spectral Signatures of Hydrated Proton Vibrations in Water Clusters. *Science* **2005**, *308*, 1765.
(146) Vendrell, O.; Gatti, F.; Meyer, H.-D. Dynamics and Infrared Spectroscopy of the Protonated Water Dimer. *Angew. Chemie Int. Ed.* **2007**, *46*, 6918.
(147) Li, X.; Moore, D. T.; Iyengar, S. S. Insights from First Principles Molecular Dynamics Studies Towards Infra-Red Multiple-Photon and Single-Photon Action Spectroscopy: Case Study of the Proton-Bound Di-Methyl Ether Dimer. *J. Chem. Phys.* **2008**, *128*, 184308.
(148) Li, X.; Oomens, J.; Eyler, J. R.; Moore, D. T.; Iyengar, S. S. Isotope Dependent, Temperature Regulated, Energy Repartitioning in a Low-Barrier, Short-Strong Hydrogen Bonded Cluster. *J. Chem. Phys.* **2010**, *132*, 244301.
(149) Li, X.; Teige, V. E.; Iyengar, S. S. Can the Four-Coordinated, Penta-Valent Oxygen in Hydroxide Water Clusters Be Detected Through Experimental Vibrational Spectroscopy? *J. Phys. Chem. A* **2007**, *111*, 4815.
(150) Li, J.; Pacheco, A. B.; Raghavachari, K.; Iyengar, S. S. A Grotthuss-like proton shuttle in the anomalous $C_2H_3^+$ Carbocation: Energetic and vibrational properties for isotopologues. *Phys. Chem. Chem. Phys.* **2016**, *18*, 29395.

(151) Sager, L. M.; Iyengar, S. S. Proton shuttle in the anomalous $C_2H_5^+$ Carbocation: Energetic and vibrational properties. *Phys. Chem. Chem. Phys.* **2017**, *19*, 27801.
(152) Li, X.; Oomens, J.; Eyler, J. R.; Moore, D. T.; Iyengar, S. S. Isotope dependent, temperature regulated, energy repartitioning in a low-barrier, short-strong hydrogen bonded cluster. *J. Chem. Phys.* **2010**, *132*, 244301.
(153) Li, J.; Pacheco, A. B.; Raghavachari, K.; Iyengar, S. S. A Grotthuss-like proton shuttle in the anomalous $C_2H_3^+$ carbocation: energetic and vibrational properties for isotopologues. *Phys. Chem. Chem. Phys.* **2016**, *18*, 29395.
(154) Kumar, A.; Iyengar, S. S. Fragment-based electronic structure for potential energy surfaces using a superposition of fragmentation topologies. *J. Chem. Theory Comput.* **2019**, *15*, 5769.

https://doi.org/10.1021/acs.jctc.1c01241
J. Chem. Theory Comput. 2022, 18, 5125−5144