Fast Near *Ab Initio* Potential Energy Surfaces using Machine Learning[†]

Fenris Lu,[‡] Lixue Cheng,[¶] Ryan J. DiRisio,[‡] Jacob M. Finney,[‡] Mark A. Boyer,[‡] Pattarapon Moonkaen,[‡] Jiace Sun,[¶] Sebastian J. R. Lee,[¶] J. Emiliano Deustua,[¶] Thomas F. Miller, III,[*,¶] and Anne B. McCoy [*,‡]

‡*Department of Chemistry, University of Washington, Seattle, WA 98195, USA*

¶*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States*

E-mail: tfm@caltech.edu; abmccoy@uw.edu

Phone: 206-543-7464

[†] FL, LC and RJD contributed equally to this work

Abstract

A machine-learning based approach for evaluating potential energies for quantum mechanical studies of properties of the ground and excited vibrational states of small molecules is developed. This approach uses the molecular-orbital-based machine learning (MOB-ML) method to generate electronic energies with the accuracy of CCSD(T) calculations at the same cost as a Hartree-Fock calculation. To further reduce the computational cost of the potential energy evaluations without sacrificing the CCSD(T) level accuracy, GPU-accelerated Neural Network Potential Energy Surfaces (NN-PES) are trained to geometries and energies that are collected from small-scale Diffusion Monte Carlo (DMC) simulations, which were run using energies evaluated using the MOB-ML model. The combined **NN+(MOB-ML)** approach is used in variational calculations of the ground and low-lying vibrational excited states of water, and DMC calculations of the ground states of water, CH_5^+ and its deuterated analogues. For both of these molecules, comparisons are made to the results obtained using potentials that were fit to much larger sets of electronic energies than were required to train the MOB-ML models. The **NN+(MOB-ML)** approach is also used to obtain a potential surface for $C_2H_5^+$, which is a carbocation with a non-classical equilibrium structure for which there is currently no available potential surface. This potential is used to explore the CH stretching vibrations, focusing on those of the bridging hydrogen atom. For both CH_5^+ and $C_2H_5^+$ the MOB-ML model is trained using geometries that were sampled from an AIMD trajectory, which was run at 350 K. By comparison, the structures sampled in the ground state calculations can have energies that are as much as ten times larger than those used to train the MOB-ML model. For water a higher temperature AIMD trajectory is needed to obtain accurate results due to the smaller thermal energy. A second MOB-ML model for $C_2H_5^+$ was developed with additional higher energy structures in the training set. The two models are found to provide nearly identical descriptions of the ground state of $C_2H_5^+$.

Introduction

Quantum descriptions of molecular vibrations require an accurate representation of the potential energy surface (PES) for the system of interest. For systems where the vibrational ground state is localized near the potential minimum, and which undergo small amplitude vibrational motions, harmonic treatments of the potential may be sufficient.[2] Such a description can be readily achieved at a broad range of levels of electronic structure theory and basis sets using electronic structure packages, as all that is required is the optimized geometry and Hessian. Significant insights may also be obtained from quartic expansions of the potential about the minimum, as this forms the basis for second-order perturbation theory calculations.[3] Unfortunately, there are many problems for which such low-order expansions of the potential are insufficient, and it becomes desirable to be able to evaluate the potential at arbitrary molecular configurations.

A common strategy for developing potentials for molecular spectroscopy, quantum dynamics or other non-local quantum applications is to evaluate the electronic energies over a broad range of geometries, and fit this data to a potential function. This has often involved fitting the electronic energies to functional forms that reflect the expected physics.[4–7] Consider, for example, two of the systems explored in the present study: H_2O and ${CH_5}^+$. Partridge and Schwenke fit a potential surface for H_2O based on 1056 internal contraction multireference configuration-interaction (ICMRCI) energies, and adjusted the parameters to include Born-Oppenheimer corrections and to match experimental data.[4] This surface will be referred to as the PS potential in the remainder of this paper. Jin, Braams and Bowman (JBB) fit a potential surface for ${CH_5}^+$ to more than 35 000 energies, which were evaluated at the CCSD(T)/aug-cc-pVTZ level of theory/basis. To allow the ion to properly dissociate to ${CH_3}^+$ + H_2, they extrapolated their fit surface to long range by splicing a long-range CH_3^+ + H_2 potential onto the fit surface using a switching function in the ${CH_3}^+$-H_2 distance.[8]

In recent years, the fitting of potential surfaces has been facilitated by the introduction of machine learning (ML). Two general strategies that are relevant to the present work are those that use these approaches to fit the calculated electronic energies and those that focus on providing ways to correct to energies that are calculated at a lower-level of electronic structure theory to provide energies with the accuracy of a high-level electronic structure calculation.[9–12] Examples of the first approach include studies that obtain high-dimensional potential energy surfaces by mapping atomic and geometric information including coordinates, bond length and bond angles to molecular energies that have been obtained from DFT calculations.[13–20] Other studies have used machine-learning approaches to correct energies from lower-level electronic structure theories to higher levels of theories, for example predicting the energy differences between results from DFT or Hartree–Fock (HF) or MP2 calculations and to those obtained from CCSD(T)[12,21–25] or variational quantum Monte Carlo[26] calculations. To include more information and to improve the learning efficiency, this second class of machine-learning approaches usually use descriptors to represent quantum level properties such as Slater determinants,[26] electron densities,[27] atomic orbitals,[28] or molecular orbitals,[24,29–31] which are computed using a lower-level and less-expensive approach. Although the prediction accuracies and learning efficiencies of this class of approaches are much better than the ones from direct learning approaches, their evaluation costs are also much higher due to their dependence on the results of low-level electronic structure computations.

Molecular-orbital-based machine learning (MOB-ML) is an example of this second type of approach.[24,29–31] By exploiting localized molecular orbitals obtained from HF calculations, MOB-ML is able to reproduce highly accurate potential energy surfaces at a fraction of the cost incurred by the target high-level methods, even when small data sets are employed during training of the MOB-ML models.[24,29,31] This approach is highly efficient. For example, a MOB-ML model for water that is trained using a single water structure is able to achieve chemical accuracy of 1 kcal mol^{-1},[29] while ML methods that directly fit energies can require hundreds of thousands of structures to achieve chemical accuracy.[32]

In the present study, we combined the MOB-ML approach with the first strategy and explore the accuracy and utility of the MOB-ML approach for generating potential energy surfaces that can be used to study ground and excited state vibrational states and energies of small molecules. Even though MOB-ML replaces the N^7 scaling of CCSD(T) with the N^3 scaling of HF, the cost of the Hartree-Fock calculations can still be prohibitive when a large number of potential energy evaluations are needed. This problem is particularly severe for diffusion Monte Carlo studies[33–37] where a typical simuation requires on the order of 10^8 single-point energy evaluations. Similar problems will occur for other approaches for studying molecular vibrations as energies will need to be evaluated at all of the grid or quadrature points used in the calculations, and these can change as the basis set is modified. While direct use of MOB-ML models for such calculations may still be prohibitive, the reduced expense compared to the underlying electronic structure calculations makes this approach more favorable for generating the large number of points needed to use ML to map energies to molecular structures. Additionally, such a study introduces a more rigorous test on the accuracy of the MOB-ML models as they will require errors no larger than several cm^{-1} which are two orders of magnitude smaller than 1 kcal mol^{-1} chemical accuracies.

In a recent study,[1] some of us developed a strategy for developing fitting potentials using a GPU-based neural network regression scheme, and the resulting potentials will be referred to as **NN-PES** in the following discussion. The approach relies on small-scale DMC simulations, which provide a set of molecular geometries that are sampled by the ground state wave function. By running simulations, with both natural masses and where all of the masses are reduced by a constant factor, we are able to greatly increase the sampling of higher energy configurations. Such simulations, while expensive when the MOB-ML model is used to evaluate the energies, can be implemented in high-performance computing (HPC) environments. This generates a large set of training data for the **NN-PES**, and the resulting **NN-PES** can be used in subsequent large scale calculations. In a recent study, we focused on applying this approach for DMC calculations, but the approach is much more general.

In the present study, we will focus on developing potentials for water, CH_5^+ and $C_2H_5^+$. To explore the efficacy of the combined **NN+(MOB-ML)** approach results will be compared to those obtained using the well-established PS and JBB potential surfaces for H_2O and CH_5^+, respectively. These are two systems that we previously studied using **NN-PES** based on the JBB and PS potential surfaces.[1] Water provides an example of a molecule that is straightforward to study by a variety of approaches, but where the relatively large amplitude OH stretching vibrations sample the dissociative part of the potential even in the ground vibrational state. $CH_5{}^+$, on the other hand, is an ion that undergoes large amplitude vibrational motions in its ground vibrational state. In fact the ground state wave function has been shown to have comparable amplitude at the 120 equivalent minima on the potential and the 180 low-energy saddle points that connect these minima.[38] Here we compare the results obtained from the MOB-ML model to those obtained from earlier fits to the electronic energies. Finally, we use this approach to explore the consequences of the large amplitude motion of the excess proton in $C_2H_5{}^+$. The minimum energy structure of this ion has the excess proton in a bridging position, with the classical $H_2C\text{-}CH_3^+$ structure corresponding to a saddle point.[39] Previous spectroscopic studies support this structure, but have only been performed above 2000 cm^{-1},[39,40] and do not appear to be sensitive to the motions of the excess proton parallel to the C-C bond. There is also no available full-dimensional potential for this ion. In the final part of this study, we will use the **NN+(MOB-ML)** approach to explore this large amplitude motion along with possible spectral consequences. For all three of these systems we will explore the range of energies that need to be accessed by the training data for the MOB-ML model in order to obtain accurate **NN-PES** for subsequent calculations of the vibrational ground state.

Theory/Methods

Molecular-Orbital-Based Machine Learning (MOB-ML)

MOB-ML is a method for accurately predicting high-level molecular energies, such as those provided by CC, Møller–Plesset (MP) perturbation theory, and other wave-function-based electronic structure theories, by using only molecular orbital information obtained from HF computations with much reduced costs. The main idea behind MOB-ML is rooted in Nesbet's theorem,[41,42]

$$E^{\text{corr}} = \sum_{ij}^{\text{occ}} \varepsilon_{ij}. \tag{1}$$

which ensures that the correlation energy of an N-electron system, E^{corr}, can always be expressed as the sum over energy contributions evaluated from pairs of occupied orbital. The pair energies, ε_{ij}, take various functional forms, which can be readily defined for the specific electronic structure theory of choice, such as CCSD(T) or MP2.[31] Indeed, computing pair energies is oftentimes computationally intractable since the high-order polynomial costs associated with CC, MP, and other theories far exceed those of HF calculations. MOB-ML is designed to alleviate this issue by approximating the pair energy contributions via the general ML mapping

$$\varepsilon_{ij} \approx \varepsilon \left[\{\phi_p\}^{ij} \right], \tag{2}$$

which associates pair energies to molecular orbitals (MOs) directly, bypassing high-level calculations altogether.

This general MOB-ML approach can be imbued with any particular ML methodology to define the mapping and trained to approximate energies of virtually any wave-function-based electronic structure method. We employ Gaussian process regression (GPR) to fit pair energies computed at the CCSD(T) level of theory. We do this by first subdividing Eq. 2 into diagonal and off-diagonal contributions

$$\varepsilon_{ij} \approx \begin{cases} \varepsilon_{\mathrm{d}}^{\mathrm{ML}} \left[\mathbf{f}_i \right] & \text{if } i = j \\ \varepsilon_{\mathrm{o}}^{\mathrm{ML}} \left[\mathbf{f}_{ij} \right] & \text{if } i \neq j \end{cases} \tag{3}$$

which separates the different character of both types of pair energies and improves on the accuracy of the machine-learned models. The feature vectors $\mathbf{f}_i$ and $\mathbf{f}_{ij}$ are constructed from consistently ordered Fock, Coulomb, and exchange interaction matrix elements using localized HF molecular orbitals. We employ the IBO or Boys localization procedures to guarantee the transferability of MOB-ML models across different chemical systems and conformations.[24,29,31] Generated in this way, the MOB-ML model ensures that the calculated potential energies are invariant under permutation of identical particles. The details of MOB feature designs have been fully described in our previous studies.[24,29,31]

Developing Neural Network (NN) Potentials Based on MOB-ML Models

Although the use of the MOB-ML energies reflects a significant savings over CCSD(T) calculations, the large number of potential energy evaluations required for many calculations can make it impractical to use the MOB-ML energies directly in studies of nuclear quantum effects. In a recent study, we showed that we could train a **NN-PES** using geometries collected from a small-scale DMC simulation, which could then be used for large-scale, production run DMC simulations as well as other types of calculations.[1] In the first part of this study, we apply the same strategy to studies of CH_5^+ and water to explore whether a MOB-ML model that is based on relatively low-energy structures can effectively describe the ground and low-energy excited state wave functions, which sample configurations with energies that are as much as ten times larger than those used to develop the MOB-ML model. This allows us to explore how well the MOB-ML models extrapolate to the higher energy regions of the potential, which are explored by the ground state wave function. For the calculations of $C_2H_5^+$, we introduced some modifications to the previously described approach for obtaining the **NN-PES** in order to improve the efficiency.

Specifically, to obtain the **NN-PES**, we use a Feedforward NN to evaluate the potential energies. Implementing this approach on GPUs, we are able to achieve a high level of parallelization, which results in a roughly 10^5 fold acceleration compared to evaluations of the MOB-ML energies. To obtain the energies used to train the NN, we developed a parallel implementation of DMC.[43,44] Because a constant simulation size simplifies the MPI communication, these DMC calculations are run using the continuous weighting scheme described in the Supporting Information. The MOB-ML surfaces used in the study were accessed through the ENTOS QCORE software package,[45] a more detailed discussion of the DMC approach can be found elsewhere,[37] and a brief description of the DMC procedure is provided in the Supporting Information.

While we were able to use this approach for DMC simulations of H_2O and CH_5^+ that use small ensembles of walkers, it does not provide a practical approach for full-scale DMC simulations or for simulations of larger molecules. For example a simulation of $C_2H_5^+$ with an ensemble size of 7168 walkers, which was run for 2500 time steps required 931 CPU hours on 28 cores on our local cluster. On the other hand, this approach allows us to generate data for training the NN.

The fundamental algorithm of Feedforward NN can be expressed as

$$y = f_n(W_n^T(f_{n-1}(W_{n-1}^T(...f_1(W_1^T x + b_1)...) + b_{n-1})) + b_n) \tag{4}$$

where x and y represent the input and output vectors, respectively. The number of layers is represented by n, and f_j, W_j and b_j represent the activation function, weight matrix and bias vector for jth layer, respectively. The weight matrices and bias vectors are updated during training to minimize the prediction error of the training set, and the error is evaluated using mean squared error. The output of the NN, which is the potential energy of the input geometry in cm^{-1}, is shifted and scaled using

$$y' = \ln\left(\frac{y}{1000} + 1\right) \tag{5}$$

In order to train the NN, the molecular geometries, which are represented by the Cartesian coordinates of the atoms, need to be converted into descriptors, which are vectors that encode the molecular geometries and provide the input for the NN.[46] The descriptors must be translationally and rotationally invariant. Ideally they should also be permutationally invariant.

In this study, we employ two descriptors, the Coulomb Matrix,[47] which we have used in our earlier study

$$M_{ij}^{\text{Coulomb}} = \begin{cases} 0.5Z_i^{2.4} \text{ for } i = j \\ \frac{Z_i Z_j}{R_{ij}} \text{ if } i \neq j \end{cases} \tag{6}$$

and Simons-Parr-Finlan(SPF) Matrix[48]

$$M_{ij}^{\mathrm{SPF}} = \begin{cases} 0 \text{ for } i = j \\ \frac{R_{ij}-R_{ij}^{e}}{R_{ij}} \text{ if } i \neq j \end{cases} \tag{7}$$

In the Coulomb Matrix descriptor, Z_i represents the nuclear charge of the ith atom. For both descriptors R_{ij} is the distance between the ith and jth atoms and for the SPF descriptor, R_{ij}^{e} provides the corresponding distance at minimum energy geometry. Notice that in both cases, the descriptor matrices are symmetric, and the diagonal elements do not contain any geometrical information. Therefore, the upper triangle excluding the diagonal elements of the matrices can be extracted and flattened into a vector, which provides the input to the NN. As both descriptors rely on the atom-atom distances, they are translationally and rotationally invariant. Explicit permutation invariance can be achieved by sorting the rows and columns of each matrix based on their norm before they are turned into vectors, although this comes at the cost of increased computation time and the possible introduction of discontinuities in the fitted PES.[1] For molecules like CH_5^+ that have very high symmetry, the introduction of this permutational invariance is essential. For most other molecules, we have found that training the NN without explicitly including the permutation symmetry works quite well as the NN is able to learn this property. Thus, the choice of method to achieve permutation invariance should be carefully evaluated based on the system of interest. In the present study, we use an unsorted Coulomb Matrix for H_2O, a sorted Coulomb Matrix for CH_5^+, and an unsorted SPF Matrix for $C_2H_5^+$.

In Figure 1, we compare the evolution of the mean absolute error (MAE) of the **NN-PES** for 100 epochs when the Coulomb (solid blue line) and SPF (dashed orange line) matrix descriptors are used. As can be seen, the use of the SPF matrix leads to much more efficient training. This is because at the equilibrium geometry the elements of the descriptor vector and the energy are all zero, removing the need to introduce a bias vector in the model. This reduces the number of trainable parameters.

As in the previous study,[1] the acitvation functions used in this work are the Swish function, developed by Google:[49]

$$\text{swish}(x) = \frac{x}{1 - e^{-\beta x}} \tag{8}$$

where β is a tunable parameter and typically set to 1. It was shown to outperform conventional activation functions like sigmoid or tanh by avoiding the vanishing gradient problem while providing a completely smooth and differentiable fitting function. This contrasts the

$$\text{ReLU}(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ if } x \geq 0 \end{cases} \tag{9}$$

activation function, which is used for the last layer. Using the RuLU activation function for the last layer ensures that the output of the NN will be non-negative. This helps to prevent nonphysical energy predictions even in regions not well-learned by the NN.

The **NN-PES** is trained over large number of epochs to achieve the desired accuracy. This can result in over-fitting. Over-fitting is problematic as it can lead to the generation of holes in the PES, where unphysically small energy predictions are made in regions of configuration space that are not well-sampled by the training data, or not well-learned by the NN. We have found that as the training progresses the entries in the weight matrices, W^T in Eq. 4, in the NN tend to increase, and the numerical instabilities that result from these large weights are correlated to potentials being overfit.[50] Thus, a simple and practical way to circumvent the problem of over-fitting is to constrain the maximum value of the norm of each column in the W^T-matrices in Eq. 4 to a pre-chosen value. In situations where the norm exceeds this value, the elements in the associated column are scaled so the norm of the column is equal to the maximum allowed value. In the present work, we have found that a maximum value of 3.5 works well.

Results and Discussion

Since we will be using several approaches for evaluating energies, before discussing the results we define the notation used to indicate how energies are evaluated. As noted in the Introduction, we will refer to the H_2O potential energy surface generated by Partridge and Schwenke as the **PS** surface and the global CH_5^+ potential energy surface generated by Jin, Braams, and Bowman as the **JBB** surface. We will refer to surfaces generated using the MOB-ML technique, as the **MOB-ML** surfaces, and we will refer to the neural network generated potential energy surface that was trained using the **MOB-ML** energies as the **NN+(MOB-ML)** surface. The energies obtained by using these surfaces will be denoted as $\mathrm{E}^{\mathrm{potential}}_{\mathrm{system}}$, where potential is replaced by **JBB**, **PS**, **MOB-ML** or **NN+(MOB-ML)**, while system is replaced with either H_2O, CH_5^+ or $C_2H_5^+$.

Validation and Comparison of the MOB-ML Potential Energy Surfaces to Previous Work

We use the mean absolute error (MAE) of the predicted CCSD(T)/aug-cc-pVTZ energies of a test set consisting of a set of geometries that were collected from a thermalized AIMD trajectory to assess the quality of the MOB-ML model. In our previous study,[24] MOB-ML models that were based on small numbers of training configurations were shown to accurately predict the energies of geometries sampled at room temperature. The MAE for these models are provided in the right most column of Table 1. Before applying this approach to studies of H_2O, CH_5^+, and $C_2H_5^+$, we explore the accuracy obtainable by MOB-ML protocols described in Ref. 31 when they are applied to the eight small molecules considered in Ref. 24. These molecules will be referred to as the validation molecules. In this application, the training and test configurations are sampled from thermalized geometries at 3000 K. The results of this analysis are provided in the second through fifth columns of Table 1.

As the results reported in Table 1 show, by using the revised protocols, the MAE of the MOB-ML model remain below 2.5 cm^{-1} for all eight molecules when the MOB-ML model has been trained to only on 500 configurations. This reflects as much as an order of magnitude improvement over the previously described approaches.[24] Of equal importance for the DMC calculations is the fact that the errors are uniformly distributed, as indicated by the mean signed errors (MSE) reported in Table 1 being less than 1 cm^{-1} for all but HNC, where the MSE is 1.034 cm^{-1}. It is notable that when there are the same number of electrons in the molecular system, the accuracy decreases as the number of vibrational degrees of freedom increases. This can be seen by comparing the MAE for HF, NH_3 and CH_4. All three of these molecules have the same number of electrons as H_2O and $CH_5{}^+$.

Based on this analysis, for the development of the MOB-ML models used in this study, we employ slightly larger training sets. For H_2O 3000 geometries were sampled from a thermalized trajectory at 6003 K, while for the other two ions, the 3000 geometries were obtained from trajectories that were run at 350 K. The higher temperature trajectory for water was used because we found that when we used configurations that were evaluated based on the lower temperature trajectory the zero-point energy was more than 25 cm^{-1} higher than the zero-point energy obtained when the potential was developed from CCSD(T) energies directly. Such problems do not appear to affect the MOB-ML models for $CH_5{}^+$ and $C_2H_5{}^+$, which is likely due to the larger total energy that is sampled by the 350 K thermalized AIMD trajectory due to the larger number of vibrational degrees of freedom in these ions compared to water.

Training sets are selected from the 3000 geometries from the AIMD trajectory, while the remaining geometries form the test set. The maximum sizes of the training sets are 1000 for H_2O and $CH_5{}^+$, leaving 2000 structures to form the test set. For the original $C_2H_5{}^+$ model, we use a training set composed of 2500 structures. A second MOB-ML model is trained for $C_2H_5{}^+$, which uses up to 2000 of the 2500 configurations that were used to train the original model. These structures are supplemented with up to 500 geometries in which the CH distances are replaced by randomly selected values between 0.8 and 1.3 Å so the training set contains structures from the AIMD trajctory and stretched structures in a 4:1 ratio (see the Supporting Information for additional details).

Table 2 provides the mean absolute errors and the mean signed errors for the MOB-ML models for H_2O, CH_5^+ and $C_2H_5^+$. We are able to obtain MAE's of 1.04 and 1.38 cm^{-1} for H_2O and CH_5^+, respectively. Comparable MAE's are obtained for $C_2H_5^+$ when only the structures from the AIMD trajectory are included in the test set. Importantly, adding the stretched structures to the training set does not lead to a deterioration of the description of these lower-energy structures. On the other hand, when only the stretched geometries are considered, the MAE for the original model is nearly an order of magnitude larger than the MAE for the stretched model.

To further explore the learning behavior of these potentials, in Figure 2 we report the MAEs of predicted energies for the test data as functions of total number of training configurations on a log-log scale for the four MOB-ML models developed in this study. Such curves are commonly referred to as learning curves.[51] By comparing the slopes of learning curves, we find that the MOB-ML approach has a slightly better learning efficiency for H_2O than for CH_5^+. This observation is consistent with the expectation that the larger number of vibrational degrees of freedom in CH_5^+ should make its potential surface a more difficult learning problem compared to H_2O. In both cases, high accuracies comprising MAEs below $5\,cm^{-1}$ are attainable when only 200 configurations are included in the training set. For both molecules, the error is uniformly distributed about zero, as indicated in the histograms of the errors from the test set for the four MOB-ML models, plotted in Figure S1.

Compared with the number of configurations used in the traditional parametric PESs, for instance, 1056 configurations in **PS**, and 36 173 configurations in **JBB**, **MOB-ML** requires significantly smaller number of configurations to provide high quality energies, closely resembling the ones provided by CCSD(T) calculations. Even when considering the most accurate **MOB-ML** models for the prediction of H_2O and CH_5^+ energies, which were trained on 1000 configurations each, **MOB-ML** achieves high accuracies with MAEs of only 1.04 and $1.38\,cm^{-1}$, respectively. Throughout this work, we utilize these high-accuracy models to predict H_2O and CH_5^+ energies. development of these models can be found in the Supporting Information.

The high accuracy of our **MOB-ML** model in describing the H_2O PES is further supported by comparing our **MOB-ML** predictions to a set of 2000 data points calculated at the CCSD(T)/aug-cc-pVTZ level of theory. As can be seen in Figure 3A, 96% of the **MOB-ML**-predicted energies lie within $0.5\,\mathrm{cm}^{-1}$ of the corresponding CCSD(T) energies, while including all points only increases this value to $4\,\mathrm{cm}^{-1}$. Nevertheless, **MOB-ML** accuracy is only as good as the underlying CCSD(T) level of theory. By comparing single-point energies obtained from **MOB-ML** predictions, $\mathrm{E}_{\mathrm{H_2O}}^{\text{MOB-ML}}$, and the **PS** PES, $\mathrm{E}_{\mathrm{H_2O}}^{\mathrm{PS}}$, as shown 3B, C and D, we immediately notice a large discrepancy over an order of magnitute larger than errors between **MOB-ML** and CCSD(T). These differences are also non-uniform, with a mean signed error (MSE) of $-130\,\mathrm{cm}^{-1}$. The large errors can be attributed to the failures of CCSD(T), and other CC methodologies based on perturbative energy corrections, in describing non-dynamical correlation effects, such as those dominating the symmetrically stretched geometries of the water molecule.[52,53]

In Figure 4, we make an analogous comparison for $\mathrm{CH_5}^+$. By comparing the **MOB-ML** and CCSD(T) energies computed for a combined selection of 3000 molecular geometries, containing both training and test set configurations, 99.5 % of the **MOB-ML** predictions show energy errors smaller than $25\,\mathrm{cm}^{-1}$, and 97 % are within $10\ \mathrm{cm}^{-1}$ of $\mathrm{E}_{\mathrm{CH_5^+}}^{\mathrm{CCSD(T)}}$. Similarly, when we compare **MOB-ML** energies to those coming from the **JBB** PES, we find that 88 % of the energy differences are smaller than $25\,\mathrm{cm}^{-1}$, with the remaining higher energy configurations showing slightly larger errors. The MSE for configurations with calculated energies below $1500\,\mathrm{cm}^{-1}$ is $0.1\,\mathrm{cm}^{-1}$, while for geometries with energies above $1500\,\mathrm{cm}^{-1}$ the MSE increases to $12.6\,\mathrm{cm}^{-1}$. These differences mirror the root-mean squared fitting error (RMSE) for the **JBB** surface, which the authors report as approximately $10\,\mathrm{cm}^{-1}$ for energies below $1500\,\mathrm{cm}^{-1}$ and approximately $17\,\mathrm{cm}^{-1}$ for energies between 1500 and $4500\,\mathrm{cm}^{-1}$.[8]

Calculating Vibrational Wave Functions and Energies for Water and $\mathrm{CH_5}^+$ Using MOB-ML Surfaces

While comparing single-point energies between **MOB-ML**, CCSD(T), and other previously reported sources provides a strong sense of the accuracy attainable by **MOB-ML** energy predictions, a more demanding task is to compute accurate molecular properties, such as vibrational energies and wave functions. To this end, we employ two different approaches combining **MOB-ML**-generated PESs and DMC simulations. In the first, the energies are evaluated using the **MOB-ML** surface directly. Even with a parallel implementation of DMC, these calculations are expensive. Therefore, to make this approach tractable, we performed the smallest calculations that are expected to provide reliable results. The parameters for these calculations were based on a previous DMC study performed using the **PS** PES for water,[54] and the **JBB** surface for CH_5^+,[55] and are provided in the Supporting Information. While the parameters for these calculations were chosen to be as small as possible, they are still expensive. In order to perform larger DMC calculations, we used the NN-DMC approach.[1] Finally, variational calculations were performed to obtain excited state energies for H_2O.

We start by considering the ground state of H_2O. As shown in Table 3, the calculation based on the **MOB-ML** energies gives a zero-point energy of 4616(2) cm^{-1}, which is roughly 20 cm^{-1} lower than the corresponding zero-point energy obtained by performing a variational calculation using the **PS** potential. The smaller zero-point energy is consistent with the results plotted in Figure 3B, which show that the energies obtained from the **MOB-ML** surface are generally smaller than those obtained from the **PS** surface. It is also consistent with the 24 cm^{-1} lower harmonic zero-point energy obtained at the CCSD(T) level compared to the MRCI calculations used to generate the **PS** surface (see Table S1). On the other hand, this result is based on a small DMC calculation. To verify this zero-point energy, we have performed a larger NN-DMC calculation, which gives a zero-point energy of 4615(1) cm^{-1}. This energy agrees with the results of the smaller calculation. While these results are promising, to further ensure that the **NN+(MOB-ML)** technique is adequately learning the **MOB-ML** surface for the purposes of DMC, we provide comparisons of the single point **NN+(MOB-ML)** energies to **MOB-ML** energies in Figure S2 in the Supporting Information. Based on these comparisons, the **NN+(MOB-ML)** surface provides a similar level of accuracy when compared to our previous work, where the same neural network structure was used to learn the **PS** surface.[1] This gives us confidence in applying the neural network method to the **MOB-ML** surface beyond the ground state of H_2O.

To this end, we performed a variational calculation of the vibrational energies of water. The details of this calculation are reported in a previous study[1] and reproduced in the Supporting Information. As can be seen in the results reported in Table 4, the energies obtained from the variational calculation using the **MOB-ML** surface and the **NN+(MOB-ML)** surface are in very good agreement, further validating the **NN+(MOB-ML)** PES. When we compare the energies based on the **MOB-ML** and **PS** potentials, larger differences are observed. The anharmonic zero-point energy evaluated using the **MOB-ML** surface is approximately 20 cm^{-1} lower than the **PS** surface, and the energies of the levels with one quantum of excitation in the OH stretches each deviate by an additional 20 cm^{-1}. As mentioned above, the harmonic zero-point energies obtained using these two surfaces differ by around 24 cm^{-1}, and the deviation can be traced to a 20 cm^{-1} discrepancy in each of the OH stretch frequencies. Finally, the difference between the energies of the bend states, calculated using these two potentials, differ by 1 to 4 cm^{-1}.

We also calculated the ground state energy and wave function for CH_5^+ based on the **MOB-ML** potential. Due to its larger number of vibrational degrees of freedom, two of which are large-amplitude vibrations, we have only performed ground state DMC calculations for this ion. Additionally, the increased dimensionality makes the evaluation of the **MOB-ML** potential approximately twice as expensive, and the minimum number of walkers needed to obtain a reliable ground state wave function and energy are roughly twice as large as for H_2O. This makes DMC calculations based on the **MOB-ML** potential barely feasible. Using this approach, we obtain a zero-point energy of 10 912(15) cm^{-1}. When we use the NN-DMC approach, the zero-point energy becomes 10 909(2) cm^{-1}. While both values are slightly lower than the energies reported based on the global **JBB** surface, they are in excellent agreement with the DMC zero-point energy of 10 908(5) reported by Johnson and McCoy using the CCSD(T)-based surface (**JBB:CC**) from which the global surface was developed.[56] These results are summarized in Table 3. This level of agreement of the zero-point energies suggests that obtaining training geometries for the MOB-ML model from a 350 K AIMD trajectory for CH_5^+ is sufficient to generate energies of configurations with significantly more energy. This is in contrast to water, where a model based on a comparable AIMD simulation resulted in errors in the ZPE of roughly 25 cm^{-1}. The improved results for CH_5^+ reflect the increase in the total thermal energy with increased vibrational degrees of freedom, and the ergodicity of the classical sampling of the potential. As a result, the 350 K trajectory will sample a much broader range of CH displacements in CH_5^+ compared to the sampling of OH bond lengths in water.

CH_5^+ is an unusual ion in that it exhibits two large amplitude motions, which result in low barriers for permutation of the hydrogen atoms. While isomerization is facile, the five CH bonds are not equivalent at any of the low-energy stationary points. This is illustrated by the harmonic frequencies for the CH stretches, which range from 2400 to 3250 cm^{-1}.[8,56] As a result, when one or more of the hydrogen atoms is replaced by a deuterium atom, the ground state probability amplitude is no longer equally distributed among the 120 minima on the potential surface. This can be seen in the plots of the projection of the probability amplitude onto the HH distances, shown in Figure 5. In this figure, we compare the distributions obtained using NN-DMC calculations based on the **NN+(MOB-ML)** potential to results obtained running the analogous unguided DMC calculations on the **JBB** potential. The distributions change as hydrogen atoms are replaced with deuterium atoms, and the evolution of the distributions with deuteration reflects the localization described above. This effect has been discussed previously,[57,58] and the important observation for the current study is that calculations of the ground state probability amplitude based on both the **NN+(MOB-ML)** potential and the **JBB** potential yield nearly identical distributions. Analogous distributions for the HD and DD distances show similar agreement, and are provided in Figure S6 in the Supporting Information. For all isotopomers, the difference in the zero-point energies calculated using the **JBB** and the **NN+(MOB-ML)** potentials remain smaller than 15 cm^{-1}. The deviations in the energies among isotopomers reflect a sensitivity of this quantity to small differences among the potentials. As mentioned above, the primary source of these differences in the calculated zero-point energies is most likely from the introduction of a switching function that allows the **JBB** surface to dissociate properly. When that correction is not included, the differences between the zero-point energies reported in Ref. 56, and reproduced in Table 3, and those obtained using the **NN+(MOB-ML)** surface are less than 6 cm^{-1}.

The above agreement between the results of these two sets of calculations should not be surprising, as both the **MOB-ML** and the **JBB** surface are based on the same levels of electronic structure theory. On the other hand, whereas the earlier surface is based on fitting more than 35 000 electronic energies with energies up to 150 000 cm^{-1} to a potential function with 2300 coefficients,[57] the **MOB-ML** potential is based on 1000 electronic energies with energies below 4500 cm^{-1}. The similarity between the calculated properties based on these two surfaces provides an illustration of the power of the MOB-ML approach.

Extensions to $C_2H_5^+$

While the results for H_2O and CH_5^+ are promising, in most cases, though we will not have other surfaces to compare to, and the power of this approach is not in reproducing previous work, but in the capability of developing potentials for new molecules on ions. For this purpose, we explore the evaluation of the potential surface for $C_2H_5^+$. Initially, the training of the MOB-ML model followed the procedure used for CH_5^+, while the **NN+(MOB-ML)** surface was generated using the modified procedure, described above. Based on these calculations, we obtained a zero-point energy of 13 172 cm^{-1} based on a set of DMC simulations with 1 000 000 walkers, which were run for 50 000 time steps. This energy is substantially larger than the energies of the structures used to train the model, which were all below 6000 cm^{-1}. In this way, this study of $C_2H_5^+$ allows us to further explore the sensitivity of the accuracy of a MOB-ML model to the range of the potential that is sampled by the training data.

To explore the validity of the MOB-ML model for $C_2H_5^+$ we randomly selected 1000 configurations of $C_2H_5^+$ from the DMC simulation and evaluated the energies at the CCSD(T) level of electronic theory. These structures had energies as high as 60 000 cm^{-1}. The results are shown in Figure S5, and the MAE between the MOB-ML energies and the CCSD(T) ones for these geometries is 84 cm^{-1}, and errors exceeding 1000 cm^{-1} for structures with energies of 20 000 cm^{-1} are observed. Based on the size of these errrors, we re-trained the MOB-ML model using up to 2000 structures from the 350 K AIMD simulation along with 500 stretched structures. The procedure used to generate the stretched structures is described in the Supporting Information. With the same total number of training structures, the inclusion of the stretched structures leads to a reduction of the MAE to 58 cm^{-1}, while the calculated zero-point energy changes by only 2 cm^{-1}. This gives us confidence that despite the relative low energy of the structures that are used to train the original MOB-ML model, the zero-point energy and ground state wave function are well described by this model. Since we have a model that is trained on higher energy structures, we use that model in the analysis described below.

With the potential developed and validated, we turn to the question of the amplitude of the motions in the ground state of $C_2H_5{}^+$, which has an equilibrium structure in which the extra proton equidistant from the two carbon atoms, as is illustrated in the inset to Figure 6. Projecting the probability amplitude onto the six CH distances, we find that the projections for all of the ethylenic CH bond lengths are essentially identical as are the two projections onto the distance between the excess proton and the two carbon atoms. Based on the analysis of these projections of the probability amplitude, the average CH bond length for the ethylenic CH bonds is 1.12263(8) Å, which is slightly longer than the value for ethylene. It is also notable that the breadth of the projection of the ground state probability amplitude onto the distance between the bridging hydrogen atom and the two carbon atoms is roughly 50% wider than the projection onto the outer CH distances. This is consistent with the larger amplitude motion experienced by this hydrogen atom.

To further explore the amplitude of this motion, we also plot projections of the ground state probability amplitude onto the Cartesian coordinates of the bridging hydrogen atom in Figure 7. To obtain these projections, we define a coordinate system by embedding the molecule in a body-fixed axis system using the Eckart conditions,[59,60] based on the equilibrium geometry of the ion. The procedures follow those used in a recent study of protonated water clusters.[61] In Figure 7A, we project the probability amplitude onto the plane that contains the two carbon atoms and which bisects the two HCH angles. In this projection, we can see the very large amplitude motion of the bridging hydrogen atom along the CC bond axis, although there is no amplitude in structures that could be considered as a $H_2C{=}CH_3{}^+$ structure. In contrast, the amplitude of the motion perpendicular to the CC bond axis is much smaller, which is consistent with the 2158 cm^{-1} assigned vibrational frequency.[39] It is also clear from the projections that these motions are highly coupled. In Figure 7B we also plot the projections of the probability amplitude onto the plane that contains the four other hydrogen atoms, in the equilibrium structure. This projection shows much less coupling than the one plotted in Figure 7A, although again the amplitude of the motion along the CC bond is much larger than the motion that is perpendicular to the CC bond axis. The above observations signal that $C_2H_5{}^+$ will be an interesting ion for further investigation.

Conclusion

In this work, we introduced a general approach for generating efficient and highly accurate potential energy surfaces for use in large-scale molecular simulations. Specifically, we take advantage of the MOB-ML approach to generate CCSD(T)-quality potential energy surfaces for H_2O, $CH_5{}^+$ and $C_2H_5{}^+$, at a small fraction of the computational cost relative to CCSD(T). Furthermore, we demonstrate that by employing a NN approach to refit the MOB-ML energies, we can increase the computational efficiency of the MOB-ML approach by exploiting GPU technology, and achieve large scale DMC simulations while maintaining high accuracy.

The approach was applied to three molecules. We began with water as a small molecule where the evaluation of excited states is readily available. This enabled us to explore whether a potential that was fit based on simulations that sampled ground state properties could be used in studies of excited states. $CH_5{}^+$ provided an ion where the large amplitude vibrations makes fitting the potential challenging, but which is well-suited for DMC approaches and for the combined MOB-ML/NN-DMC approach. Through our exploration of $CH_5{}^+$ we showed that a MOB-ML model that was trained on structures with energies below 4200 cm^{-1} could reproduce the ground state wave function of CH_5^+ and its deuterated analogues, which have energies approaching twice that value. Finally, we used this approach to develop a potential for $C_2H_5{}^+$ and explored the amplitude of the motion of the bridging hydrogen atom in the ground state of this ion. We also explored the sensitivity of the results to the range of the energies of the configurations that were used to train the MOB-ML model. We found that when the MOB-ML model was trained using configurations below 7000 cm^{-1} and geometries that exceeded 43 000 cm^{-1} the final ground state probability amplitude and zero-point energy showed negligible differences.

Acknowledgement

ABM acknowledges support from the Chemistry Division of the National Science Foundation (CHE-1856125 for scientific work and OAC-1663636 for HPC code development). The development of PyVibDMC was supported by a fellowship to RJD from The Molecular Sciences Software Institute under NSF grant OAC-1547580. TFM also thanks the support and funds from the U.S. Army Research Laboratory (W911NF-12-2-0023), the U.S. Department of Energy (DE-SC0019390), the Caltech DeLogi Fund, and the Camille and Henry Dreyfus Foundation (Award ML-20-196). Parts of this work were performed using the Ilahie cluster at the University of Washington, which was purchased using funds from a MRI grant from the National Science Foundation (CHE-1624430). This work was also facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington as well as resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

Supporting Information Available

Description of DMC and its implementation; Numerical details; Histograms of the test errors for the MOB-ML modesl; Comparisons of **NN+MOB-ML** test energies to **MOB-ML** energies for H_2O, CH_5^+ and $C_2H_5^+$; Comparison of the energies obtained from he two **NN+(MOB-ML)** models for $C_2H_5^+$ and at the CCSD(T) level of theory; Comparisons of the DMC probability amplitude onto HD and DD distances for isotopomers of CH_5^+; Comparison of harmonic frequencies for H_2O using the underlying electronic structure on which the **PS** and **MOB-ML** surfaces are based; Mean absolute errors of the **NN+MOB-ML** training and test sets for H_2O and CH_5^+; Data used to generate the learning curves in Figure 2.

References

(1) DiRisio, R. J.; Lu, F.; McCoy, A. B. GPU-accelerated Neural Network Potential Energy Surfaces for Diffusion Monte Carlo. *J. Phys. Chem. A* **2021**, *125*, 5849–5859.

(2) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; Dover: New York, 1955.

(3) Nielsen, H. H. The Vibration-Rotation Energies of Molecules. *Rev. Mod. Phys.* **1951**, *23*, 90–136.

(4) Partridge, H.; Schwenke, D. W. The Determination of an Accurate Isotope Dependent Potential Energy Surface for Water from Extensive ab Initio Calculations and Experimental Data. *J. Chem. Phys.* **1997**, *106*, 4618–4639.

(5) Babin, V.; Leforestier, C.; Paesani, F. Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient. *J. Chem. Theory. Comput.* **2013**, *9*, 5395–5403, PMID: 26592277.

(6) Schatz, G. C. The Analytical Representation of Electronic Potential-energy Surfaces. *Rev. Mod. Phys.* **1989**, *61*, 669–688.

(7) Huang, X.; Schwenke, D. W.; Lee, T. J. Rovibrational Spectra of Ammonia. I. Unprecedented Accuracy of a Potential Energy Surface Used with Nonadiabatic Corrections. *J. Chem. Phys.* **2011**, *134*, 044320.

(8) Jin, Z.; Braams, B. J.; Bowman, J. M. An ab Initio Based Global Potential Energy Surface Describing $CH_5^+ \to CH_3^+ + H_2$. *J. Phys. Chem. A* **2006**, *110*, 1569–1574.

(9) Nguyen, T. T.; Székely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Götz, 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.

(10) Liu, Y.; Li, J.; Felker, P. M.; Bacic, Z. HCl-H_2O Dimer: An Accurate Full-dimensional Potential Energy Surface and Fully Coupled Quantum Calculations of Intra- and Intermolecular Vibrational States and Frequency Shifts. *Phys. Chem. Chem. Phys.* **2021**, *23*, 7101–7114.

(11) Kondati Natarajan, S.; Morawietz, T.; Behler, J. Representing the Potential-energy Surface of Protonated Water Clusters by High-dimensional Neural Network Potentials. *Phys. Chem. Chem. Phys.* **2015**, *17*, 8356–8371.

(12) Schmitz, G.; Godtliebsen, I. H.; Christiansen, O. Machine Learning for Potential Energy Surfaces: An Extensive Database and Assessment of Methods. *J. Chem. Phys.* **2019**, *150*, 244113.

(13) Behler, J.; Parrinello, M. Generalized Neural-network Representation of High-dimensional Potential-energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(14) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(15) Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153–1173.

(16) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(17) Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-filter Convolutional Neural Network for Modeling Quantum Interactions. 2017; `http://arxiv.org/abs/1706.08566`.

(18) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(19) Profitt, T. A.; Pearson, J. K. A Shared-weight Neural Network Architecture for Predicting Molecular Properties. *Phys. Chem. Chem. Phys.* **2019**, *21*, 26175–26183.

(20) Park, C. W.; Kornbluth, M.; Vandermause, J.; Wolverton, C.; Kozinsky, B.; Mailoa, J. P. Accurate and Scalable Graph Neural Network Force Field and Molecular Dynamics with Direct Force architecture. *Npj Comput. Mater.* **2021**, *7*, 1–9.

(21) Sparta, M.; Toffoli, D.; Christiansen, O. An Adaptive Density-guided Approach for the Generation of Potential Energy Surfaces of Polyatomic Molecules. *Theor. Chem. Acc.* **2009**, *123*, 413–429.

(22) 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.

(23) FCHL Revisited: Faster and More Accurate Quantum Machine Learning. *J. Chem. Phys.* **2020**, *152*, 044107.

(24) Welborn, M.; Cheng, L.; Miller III, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.

(25) 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.

(26) Hermann, J.; Schätzle, Z.; Noé, F. Deep-neural-network Solution of the Electronic Schrödinger Equation. *Nat. Chem.* **2020**, *12*, 891–897.

(27) Dick, S.; Fernandez-Serra, M. Machine Learning Accurate Exchange and Correlation Functionals of the Electronic Density. *Nat. Comm.* **2020**, *11*, 3509.

(28) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller III, T. F. OrbNet: Deep Learning for Quantum Chemistry Using Symmetry-adapted Atomic-orbital Features. *J. Chem. Phys.* **2020**, *153*, 124111.

(29) Cheng, L.; Welborn, M.; Christensen, A. S.; Miller III, T. F. A Universal Density Matrix Functional from Molecular Orbital-based Machine Learning: Transferability across Organic Molecules. *J. Chem. Phys.* **2019**, *150*, 131103.

(30) Cheng, L.; Kovachki, N. B.; Welborn, M.; Miller III, T. F. Regression Clustering for Improved Accuracy and Training Costs with Molecular-orbital-based Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 6668–6677.

(31) Husch, T.; Sun, J.; Cheng, L.; Lee, S. J.; Miller III, T. F. Improved Accuracy and Transferability of Molecular-orbital-based Machine Learning: Organics, Transition-metal Complexes, Non-covalent Interactions, and Transition States. *J. Chem. Phys.* **2021**, *154*, 064108.

(32) Lee, S. J.; Husch, T.; Ding, F.; Miller III, T. F. Analytical gradients for molecular-orbital-based machine learning. *J. Chem. Phys.* **2021**, *154*, 124120.

(33) Metropolis, N.; Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**, *44*, 334–341.

(34) Anderson, J. B. A Random-Walk Simulation of the Schrödinger Equation: H_3^+. *J. Chem. Phys.* **1975**, *63*, 1499–1503.

(35) Anderson, J. B. Quantum Chemistry by Random Walk. H ^2P, H_3^+ D_{3h} $^1A_1'$, H_2 $^3\Sigma_u^+$, H_4 $^1\Sigma_g^+$, Be ^1S. *J. Chem. Phys.* **1976**, *65*, 4121–4127.

(36) Suhm, M. A.; Watts, R. O. Quantum Monte Carlo Studies of Vibrational States in Molecules and Clusters. *Phys. Rep* **1991**, *204*, 293 – 329.

(37) DiRisio, R. J.; Finney, J. M.; McCoy, A. B. Diffusion Monte Carlo Approaches for Wtudying Nuclear Quantum Effects in Fluxional Molecules. *WIREs Computational Molecular Science* **2022**,

(38) Huang, X.; Johnson, L. M.; Bowman, J. M.; McCoy, A. B. Deuteration Effects on the Structure and Infra-Red Spectrum of CH_5^+. *J. Am. Chem. Soc.* **2006**, *128*, 3478–3479.

(39) Ricks, A. M.; Douberly, G. E.; v.R. Schleyer, P.; Duncan, M. A. Infrared Spectroscopy of Protonated Ethylene: The Nature of Proton Binding in the Non-Classical Structure. *Chem. Phys. Lett.* **2009**, *480*, 17–20.

(40) Andrei, H.-S.; Solcà, N.; Dopfer, O. IR Spectrum of the Ethyl Cation: Evidence for the Nonclassical Structure. *Angew. Chem. Int. Ed.* **2008**, *47*, 395–397.

(41) Nesbet, R. K. Brueckner's Theory and the Method of Superposition of Configurations. *Phys. Rev.* **1958**, *109*, 1632.

(42) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover: Mineola, 1996; pp 231–239.

(43) DiRisio, R. J.; McCoy, A. B. rjdirisio/pyvibdmc:1.1.8. 2021; `https://doi.org/10.5281/zenodo.4695231`.

(44) Boyer, M. A.; DiRisio, R. J.; Finney, J. M.; McCoy, A. B. McCoyGroup/PyHPCDMC. 2021; `https://doi.org/10.5281/zenodo.4739301`.

(45) Manby, F. R.; Miller III, T. F.; Bygrave, P.; Ding, F.; Dresselhaus, T.; Batista-Romero, F.; Buccheri, A.; Bungey, C.; Lee, S. J. R.; Meli, R. et al. entos: A Quantum Molecular Simulation Package. **2019**,

(46) Jiang, B.; Guo, H. Permutation Invariant Polynomial Neural Network Approach to Fitting Potential Energy Surfaces. *J. Chem. Phys.* **2013**, *139*, 054112.

(47) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 58301.

(48) Molski, M.; Konarski, J. Extended Simons-Parr-Finlan Approach to the Analytical Calculation of the Rotational-Vibrational Energy of Diatomic Molecules. *Physical review. A* **1993**, *47*, 711–714.

(49) Ramachandran, P.; Zoph, B.; Le, Q. V. Searching for Activation Functions. *ArXiv* **2018**, *abs/1710.05941*.

(50) Lawrence, S.; Giles, C. Overfitting and Neural Networks: Vonjugate Gradient and Backpropagation. **2000**, *1*, 114–119 vol.1.

(51) Cortes, C.; Jackel, L. D.; Solla, S. A.; Vapnik, V.; Denker, J. S. Learning Curves: Asymptotic Values and Rate of Convergence. In *Advances in Neural Information Processing Systems 6*; Cowan, J. D., Tesauro, G., Alspector, J., Eds.; Morgan-Kaufmann, 1994; pp 327–334.

(52) Bauman, N. P.; Shen, J.; Piecuch, P. Combining Active-space Coupled-cluster Approaches with Moment Energy Corrections via the CC (P; Q) Methodology: Connected Quadruple Excitations. *Mol. Phys.* **2017**, *115*, 2860–2891.

(53) Eriksen, J. J.; Matthews, D. A.; Jørgensen, P.; Gauss, J. Assessment of the Accuracy of Coupled Cluster Perturbation Theory for Open-shell systems. I. Triples Expansions. *J. Chem. Phys.* **2016**, *144*, 194102.

(54) Lee, V. G. M.; McCoy, A. B. An Efficient Approach for Studies of Water Clusters Using Diffusion Monte Carlo. *J. Phys. Chem. A* **2019**, *123*, 8063–8070.

(55) Finney, J. M.; DiRisio, R. J.; McCoy, A. B. Guided Diffusion Monte Carlo: A Method for Studying Molecules and Ions that Display Large Amplitude Vibrational Motions. *J. Phys. Chem. A* **2020**, *124*, 9567–9577.

(56) Johnson, L. M.; McCoy, A. B. Evolution of Structure in CH_5^+ and Its Deuterated Analogs. *J. Phys. Chem. A* **2006**, *110*, 8213–8220.

(57) McCoy, A. B.; Braams, B. J.; Brown, A.; Huang, X.; Jin, Z.; Bowman, J. M. Ab Initio Diffusion Monte Carlo Calculations of the Quantum Behavior of CH_5^+ in Full Dimensionality. *J. Phys. Chem. A* **2004**, *108*, 4991–4994.

(58) Fore, M. E.; McCoy, A. B. Statistical Analysis of the Effect of Deuteration on Quantum Delocalization in CH_5^+. *J. Phys. Chem. A* **2019**, *123*, 4623–4631.

(59) Eckart, C. Rotating Axes and Polyatomic Molecules. *Phys. Rev.* **1935**, *47*, 552–558.

(60) Louck, J. D.; Galbraith, H. W. Eckart Vectors, Eckart Frames, and Polyatomic Molecules. *Rev. Mod. Phys.* **1976**, *48*, 69–106.

(61) DiRisio, R. J.; Finney, J. M.; Dzugan, L. C.; Madison, L. R.; McCoy, A. B. Using Diffusion Monte Carlo Wave Functions to Analyze the Vibrational Spectra of $H_7O_3^+$ and $H_9O_4^+$. *J. Phys. Chem. A* **2021**, *125*, 7185–7197.

Table 1: Predicted Error of the MOB-ML Model Relative to CCSD(T)/aug-cc-pVTZ Energies.[a]

System	MAE[b]	MSE[c]	RMSE[d]	Max[e]	MAE (Ref. 24)
CH_4	1.80	-0.035	4.16	87.21	6.58
NH_3	1.05	0.11	3.17	45.97	35.12
HF	0.014	-0.008	0.19	3.44	6.58
CO	0.006	-0.004	0.041	0.23	6.58
N_2	0.028	0.026	0.85	13.12	13.17
F_2	0.54	-0.52	10.06	224.13	6.58
HCN	1.92	-0.81	16.47	303.12	8.78
HNC	2.39	1.03	23.56	191.07	19.75

[a] The models are trained on 500 configurations and tested on the remaining 500 configurations.

[b] Mean Absolute Error in cm^{-1}.

[c] Mean Signed Error in cm^{-1}.

[d] Root Mean Square Error in cm^{-1}.

[e] Maximum Error in cm^{-1}.

Table 2: Predicted Error of the MOB-ML Model Relative to CCSD(T)/aug-cc-pVTZ Energies.

Model	Training[a]	Test[b]	MAE[c]	MSE[d]	Max[e]
H_2O[f]	1000	2000	1.04	0.09	3.72
CH_5^+[g]	1000	2000	1.38	-0.25	67.31
$C_2H_5^+$[g]	2500	1500[h]	60.95	49.60	539.77
		500	1.91	0.21	45.95
		1000[h]	90.47	79.01	539.77
$C_2H_5^+$[g]	2500[i]	1500[i]	5.07	-0.46	96.97
		1000	2.13	-0.21	57.19
		500[i]	10.93	-0.96	96.97

[a] Number of structures used to train the MOB-ML model.

[b] Number of test structures used in this analysis.

[c] Mean Absolute Error in cm^{-1}.

[d] Mean Signed Error in cm^{-1}.

[e] Maximum Absolute Error in cm^{-1}.

[f] Structures are extracted from an AIMD trajectory at 6003 K.

[g] Structures are extracted from an AIMD trajectory at 350 K.

[h] 1000 of the structures contain stretched CH bond lengths, as described in the text.

[i] 500 of the structures contain stretched CH bond lengths, as described in the text.

Table 3: Calculated Zero-point Energies Obtained Using DMC (cm^{-1})

System	**MOB-ML**	**NN+(MOB-ML)**	**PS[a]/JBB[b]**	**JBB:CC[c]**
H_2O	4616 (2)	4615 (1)	4637 (2)	–
CH_5^+	10 912 (15)	10 908 (2)	10 917 (5)	10 908 (5)
CH_4D^+	–	10 301 (2)	10 303 (4)	10 298 (5)
$CH_3D_2^+$	–	9689 (4)	9698 (7)	9690 (5)
$CH_2D_3^+$	–	9086 (3)	9010 (3)	9090 (5)
CHD_4^+	–	8553 (2)	8565 (3)	8559 (5)
CD_5^+	–	8040 (3)	8044 (2)	8039 (5)
$C_2H_5^+$	–	13174 (1)[d]	-	–
$C_2H_5^+$	–	13172 (1)[e]	-	–

[a] Results of DMC simulations using the Partridge-Schwenke surface.[4]

[b] Results of DMC simulations using the Jin, Braams, and Bowman surface.[8]

[c] Results of DMC simulations on the CCSD(T) surface on which the **JBB** potential is based.[56]

[d] MOB-ML trained to 2500 structures extracted from a 350 K AIMD trajectory.

[e] MOB-ML trained to 2000 structures extracted from a 350 K AIMD trajectory and 500 structures that contain stretched CH bond lengths.

Table 4: Calculated Ground and Excited State Vibrational Energies[a] for $\mathbf{H_2O}$ ($\mathbf{cm^{-1}}$)

v_s^b	v_b	v_a	**MOB-ML**	**MOB-ML − NN+(MOB-ML)**	**PS[c]**
0	0	0	4614.6	−0.01	4636.8
0	1	0	1594.1	−0.01	1594.4
0	2	0	3151.8	0.8	3150.8
1	0	0	3638.8	−0.3	3656.2
0	0	1	3734.5	−0.2	3755.1
0	3	0	4669.5	−0.3	4665.7
1	1	0	5216.3	0.2	5233.8
0	1	1	5308.9	−0.5	5330.0

[a] The first row corresponds to the calculated zero-point energy E_0, and all subsequent rows correspond to $E - E_0$.

[b] v_s, v_b, and v_a correspond to the number of quanta in the symmetric OH stretch, HOH bend, and antisymmetric OH stretch, respectively.

[c] Ref. 1.

FIGURES

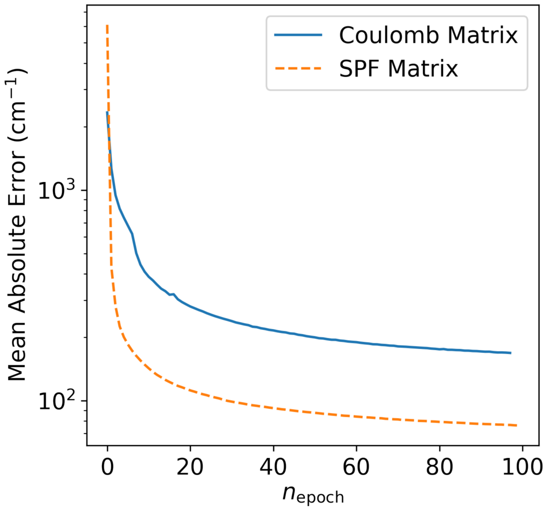

Figure 1: Mean absolute errors (MAE) of the validation set, plotted as a function of the number of epochs, n_{epoch}, when the descriptor is based on a Coulomb matrix (solid blue line) and an SPF matrix (dashed orange line), as described in Eqs. 7 for the **NN+(MOB-ML)** surface for $C_2H_5^+$ based on the MOB-ML model that was trained with stretched CH bond lengths.

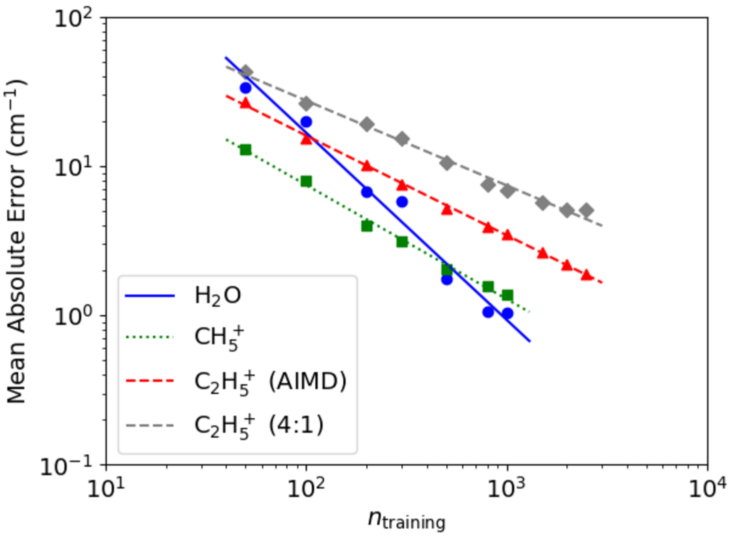

Figure 2: Prediction mean absolute errors (MAE) for total energies as a function of the number of training configurations, n_{training} (learning curves) of H_2O (blue circles, solid line), CH_5^+ (green squares, dotted line), and $C_2H_5^+$ plotted on a logarithm scale. The results for two models of $C_2H_5^+$ are presented. The red triangles, dashed line show the results when all of the structures are taken from the AIMD trajectory, while the grey diamonds, dashed line provide results when structures are taken from the AIMD trajectory and from stretched configurations in a 4:1 ratio, as described in the text. The slopes of learning curves represent the learnability of the MOB-ML model for H_2O, CH_5^+, and $C_2H_5^+$ energies, and steeper learning curve suggests a higher learning efficiency. The data that is plotted are provided in Table S3.

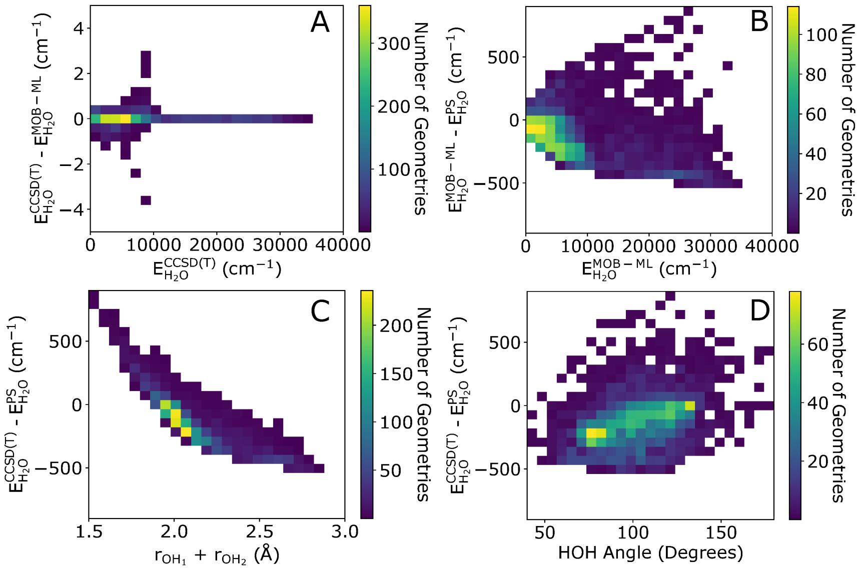

Figure 3: Comparison of the calculated energies of MOB-ML training and test set data for H_2O. (A) The number of geometries plotted as a function of the CCSD(T) energy and the difference between the calculated **MOB-ML** and CCSD(T) energies. (B) The number of geometries plotted as a function of the **MOB-ML** energies and the difference between the **PS**[4] and the **MOB-ML** energies. (C) The number of geometries plotted as a function of the difference between the **MOB-ML** and **PS** energies and the sum of r_{OH} distances and (D) the HOH angle.

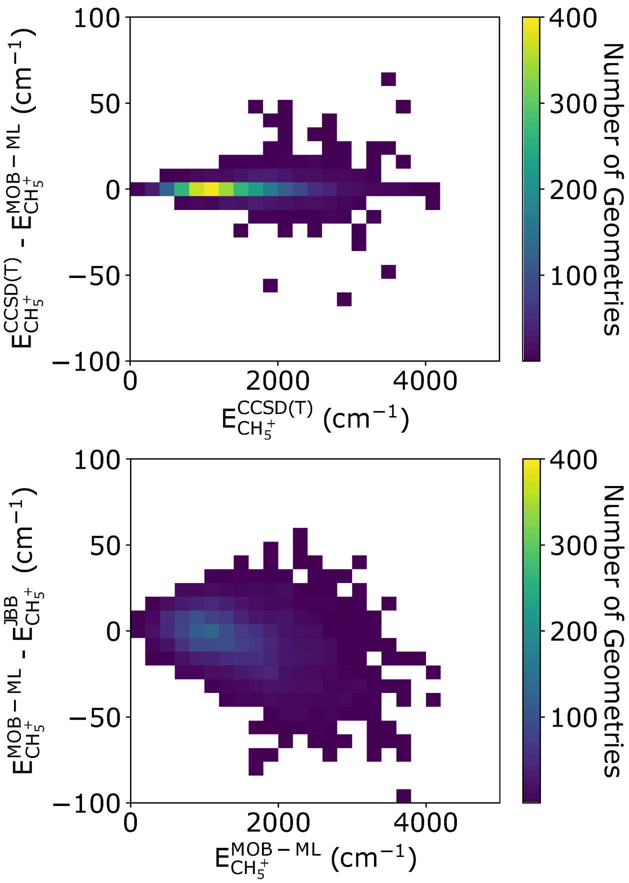

Figure 4: The comparison of the training and test geometries used to generate the CH_5^+ **MOB-ML** surface. The number of geometries plotted a function of the CCSD(T) energy and the difference between the **MOB-ML** and CCSD(T) energies (top), and the number of geometries plotted as a function of the **MOB-ML** energies and the difference between the **MOB-ML** and **JBB**[8] energies (bottom).

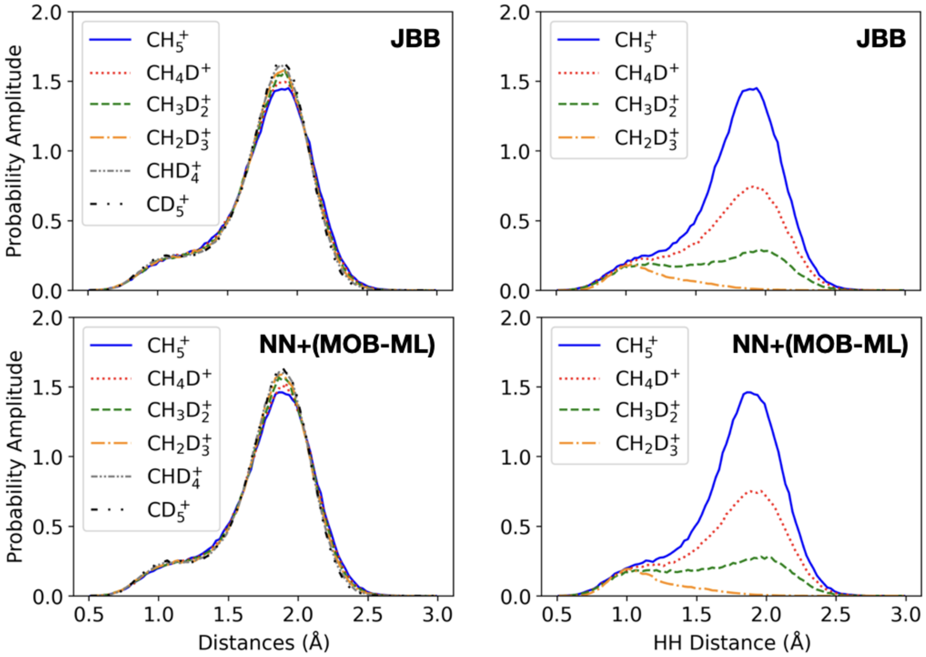

Figure 5: The calculated DMC probability amplitude projected onto the distances between all of the pairs of hydrogen and deuterium atoms (left) and HH distances (right) for the appropriate isotopologues of CH_5^+. The top two panels show the DMC probability amplitude using the **JBB** potential energy surface,[8] where the bottom two are using the **NN+(MOB-ML)** surface.

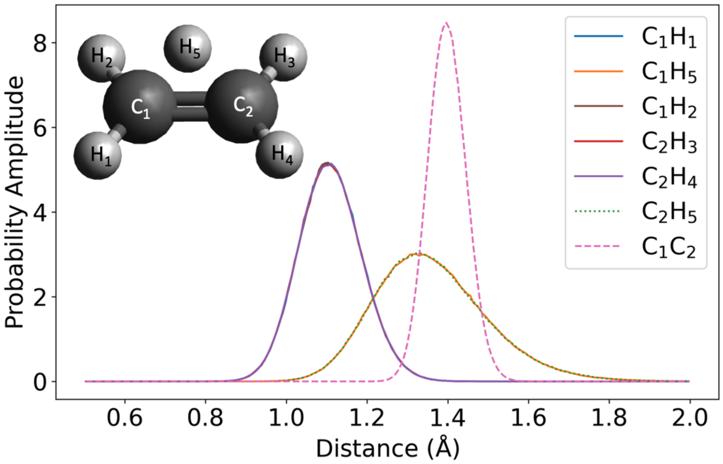

Figure 6: Projection of the ground state probability amplitude for $C_2H_5{}^+$ onto the four CH distances for the ethylenic CH bonds, the CC distance and the two distances between the bridging hydrogen atom and the carbon atoms. The equilibrium structure of $C_2H_5{}^+$ is shown in the inset. The peaks centered at 1.12 Å correspond to the four ethylenic CH distances. The peaks centered at 1.37 Å correspond to the two CH distancse between the bridging hydrogen atom and the carbon atoms. The peak centered at 1.40 Å corresponds to the CC distance.

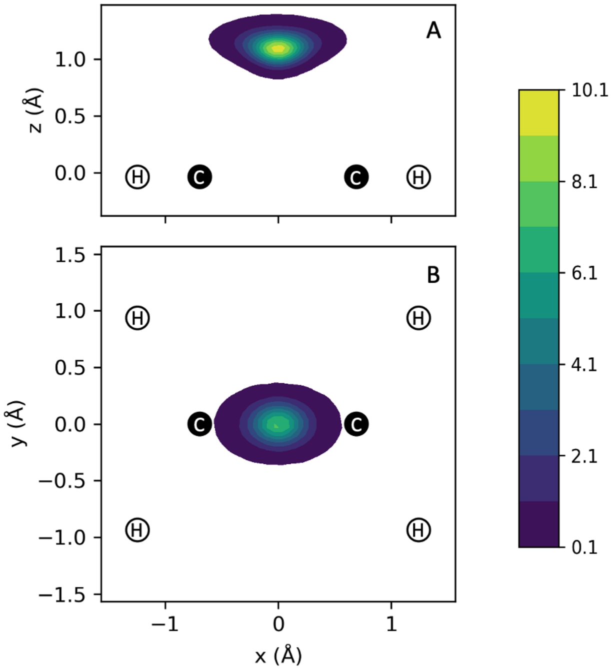

Figure 7: Projections of the ground state probability amplitude for $C_2H_5^+$ onto A) the x and z and B) the x and y Cartesian coordinates of the bridging hydrogen atom when the structures are rotated into an Eckart frame.[60] The black and white circles represent the positions of the carbon and ethylenic hydrogen atoms in the equilibrium structure of $C_2H_5^+$.