THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA
Article

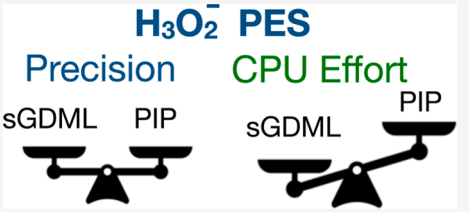
Assessing Permutationally Invariant Polynomial and Symmetric Gradient Domain Machine Learning Potential Energy Surfaces for $H_3O_2^-$

Published as part of The Journal of Physical Chemistry A virtual special issue "Xueming Yang Festschrift".

Priyanka Pandey, Mrinal Arandhara, Paul L. Houston,* Chen Qu, Riccardo Conte, Joel M. Bowman,* and Sai G. Ramesh*

Cite This: J. Phys. Chem. A 2024, 128, 3212−3219

ABSTRACT: The singly hydrated hydroxide anion $OH^-(H_2O)$ is of central importance to a detailed molecular understanding of water; therefore, there is strong motivation to develop a highly accurate potential to describe this anion. While this is a small molecule, it is necessary to have an extensive data set of energies and, if possible, forces to span several important stationary points. Here, we assess two machine-learned potentials, one using the symmetric gradient domain machine learning (sGDML) method and one based on permutationally invariant polynomials (PIPs). These are successors to a PIP potential energy surface (PES) reported in 2004. We describe the details of both fitting methods and then compare the two PESs with respect to precision, properties, and speed of evaluation. While the precision of the potentials is similar, the PIP PES is much faster to evaluate for energies and energies plus gradient than the sGDML one. Diffusion Monte Carlo calculations of the ground vibrational state, using both potentials, produce similar large anharmonic downshift of the zero-point energy compared to the harmonic approximation of the PIP and sGDML potentials. The computational time for these calculations using the sGDML PES is roughly 300 times greater than using the PIP one.



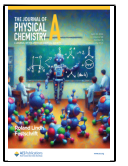
INTRODUCTION

The singly hydrated hydroxide anion $OH^-(H_2O)$ has long been of interest to theorists and experimentalists.[1−12] The first ab initio-based, full-dimensional, machine-learned potential energy (MLP) was reported in 2004[3,5] using permutationally invariant polynomials (PIPs) in terms of primary and secondary PIPs.[13] In brief, this potential energy surface (PES) was a least-squares fit to almost 67,000 ab initio energies[3] (later updated to a fit to about 23,000 energies[5]), obtained with the CCSD(T) method with an aug-cc-pVTZ basis. The variables of the fit are the ten internuclear distances, and the polynomial basis is constructed to be permutationally invariant with respect to the permutation of like atoms. This PIP PES was used in VSCF/VCI (reaction path) and fixed-node diffusion Monte Carlo (DMC) calculations of the vibrational energies. While this PES was successful in obtaining these energies and making insightful comparisons with experiments, it did not have extensive coverage of the high-energy saddle point for the exchange of the shared H atom with the terminal one, also referred to as the bifurcation saddle point. The structure of this saddle point as well as the global minimum and H atom transfer saddle point are shown in Figure 1 below.

This lack of coverage was remedied very recently by two of us (MA and SGR) who calculated 15,024 energies and gradients at the CCSD(T)/aug-cc-pVTZ level of theory using CFOUR.[15] They trained a symmetric gradient domain machine learning (sGDML)[16−20] PES on a subset of 3000 energies and gradients, with validation (hyperparameter adjustment) using another 3000 data points and then tested it on the remaining data. The PES was then used in path integral simulations of the temperature dependence of the motion along the so-called bifurcation pathway, whose TS is shown in Figure 1.

This data set and the sGDML PES provide an opportunity to assess that PES and a PIP PES trained on this data set. We do that here. In the next section, we provide details of the sGDML and PIP approaches and the specifics for this particular data set. The performance of the two fits is examined in detail in the Results and Discussion Section. Results of DMC calculations are also presented in that section. The final section contains a summary and conclusions.

Received: February 16, 2024
Revised: March 15, 2024
Accepted: March 20, 2024
Published: April 16, 2024



https://doi.org/10.1021/acs.jpca.4c01044
J. Phys. Chem. A 2024, 128, 3212−3219

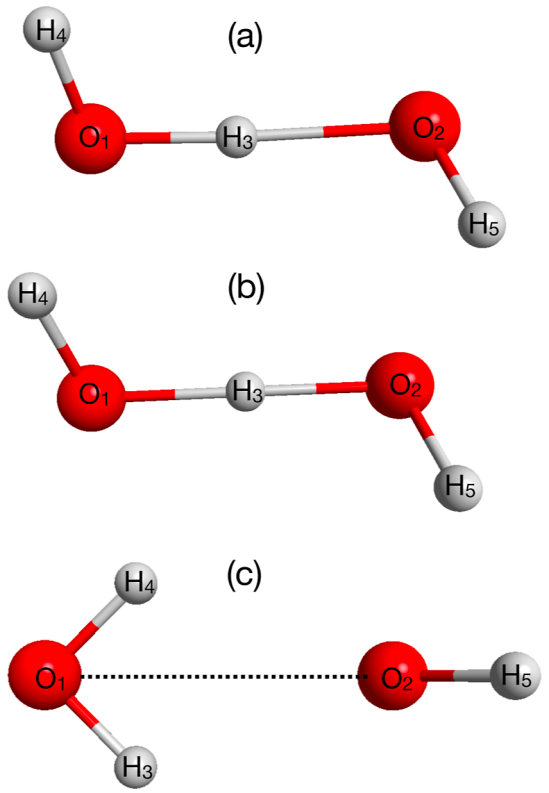


Figure 1. Structures of three stationary points of $H_3O_2^-$: (a) global minimum, (b) H-transfer TS, and (c) bifurcation TS.

FITTING METHODS

PIPs and MSA Software. MLPs using a basis of PIPs have been reported for nearly 20 years, with the PES for $H_3O_2^-$ in 2004 being one of the first such MLPs. The expression for the PIP potential is given by

$$V(\mathbf{y}) = \sum_{i=1}^{n_p} c_i\, p_i(\mathbf{y}) \tag{1}$$

where c_i are linear coefficients, p_i are PIPs, n_p is the total number of polynomials (and linear coefficients c_i) for a given maximum polynomial order, and $\mathbf{y}$ are transformed internuclear distances. We have used the following 3 transformations: $y_{ij} = \exp(-r_{ij}/a)$, $y_{ij} = \exp(-r_{ij}/a)/r_{ij}$,[13] and $y_{ij} = 1/r_{ij}$.[21] PIPs are polynomials that are invariant with respect to permutations of like atoms.

Our current software to generate PIPs and perform least-squares fitting[22,23] is based on monomial symmetrization.[13,24] The first part of MSA creates the PIP basis and writes it to a text file named "basis.f90", where the file has the suffix .f90 for later use in Fortran. Analytical gradients are also provided in this code. In a second step, fast gradient evaluation is available, based on reverse differentiation algorithms and Mathematica scripts. These are described in detail elsewhere[23] and are used here. The final code is written in Fortran 90.

To complete this short review, we note that the linear coefficients c_i are optimized to minimize the L2 loss, i.e., the sum of the square of the differences between the data and fit $V(\mathbf{y}; \mathbf{c})$, where we explicitly indicate the parametric dependence on the coefficients $\mathbf{c}$. The standard approach leads to the matrix equation

$$\mathbf{Ac} = \mathbf{d} \tag{2}$$

where the matrix $\mathbf{A}$, elements of which are given by $A_{i,j} = p_j(\mathbf{y}_i)$, is $N \times n_p$, where N is the size of the data set of energies plus gradients (if they are used), $\mathbf{c}$ is the column vector of length n_p and $\mathbf{d}$ is the column vector length N and consists of these data. In general, $n_p \ll N$, and so this is an overdetermined set of linear equations. The solution to this least-squares problem is given formally by

$$\mathbf{c} = (\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{d}) \tag{3}$$

There are several ways to proceed; we use singular value decomposition of the matrix $\mathbf{A} = \mathbf{U\Sigma V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of size $N \times N$ and $n_p \times n_p$, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix of n_p singular values in descending order with zeros below the diagonal element σ_{n_p}. $\mathbf{U}$ can be partitioned into two blocks, $\mathbf{U}_1$ and $\mathbf{U}_2$, where $\mathbf{U}_1$ is $N \times n_p$. The final expression for the coefficients is

$$\mathbf{c} = (\mathbf{V\Sigma}^{-1}\mathbf{U}_1^T)\mathbf{d} \tag{4}$$

We use dgelss.f90 for this analysis.

In general, we fit an entire data set (including gradients if available). This is because the linear regression method depends on the size of the PIP basis and not the size of the data set. So, bases with n_p of the order of thousands present no difficulties for data sets that are an order of magnitude bigger. Of course testing of the fit is done, generally on out-of-sample data. This protocol is not the usual split-train-test protocol. Methods based on Kernel Ridge Regression and Gaussian Process Regression are "trained" directly on a data set and the resulting linear algebra problem, i.e., a matrix inverse of the kernel at the configurations of the data set, is limited to data sets of the order of thousands, much smaller than the data sets used in PIP Linear Regression.

The overall work flow of the MSA software to obtain a PIP PES is given in Figure 2. In the present work, energies and gradients are included in the data set. DMC calculations are run to locate large negative regions of the fit, called "holes". In general, the holes occur at high energies, which (not surprisingly) were not sampled in the data set. Additional data are added at the hole configuration, and a new fit is done. This is repeated until there are no holes or very few holes.

We provide specific details of this approach below where we discuss the fit to the $H_3O_2^-$ data set.

sGDML. sGDML refers to a kernel ridge method that fits the gradient of the potential while identifying and incorporating symmetries of the molecule.[16−20] This approach has been widely used for numerous applications and the Python code for usage is available.[25] Two of us (MA and SR), who developed the sGDML PES for $H_3O_2^-$, trained the model using the Python code but wrote a Fortran 90 code to evaluate the gradient and energy. The sGDML potential and Fortan code are available on Github (https://github.com/arandharamrinal/H3O2M).

We briefly describe how sGDML works. The input to sGDML is a set of molecular geometries, their ab initio energies, and atomic forces obtained from a high-level quantum chemistry calculation. As per the choice of the user, this data set is split into training, validation, and test points. The points in each set are chosen so that the energy distribution in each of the three data sets is consistent with the distribution of the full data set. From (a subset of) the set of M training points, sGDML identifies the molecular symmetries in a data-driven manner,[17,18] producing atomic permutation

https://doi.org/10.1021/acs.jpca.4c01044
J. Phys. Chem. A 2024, 128, 3212−3219

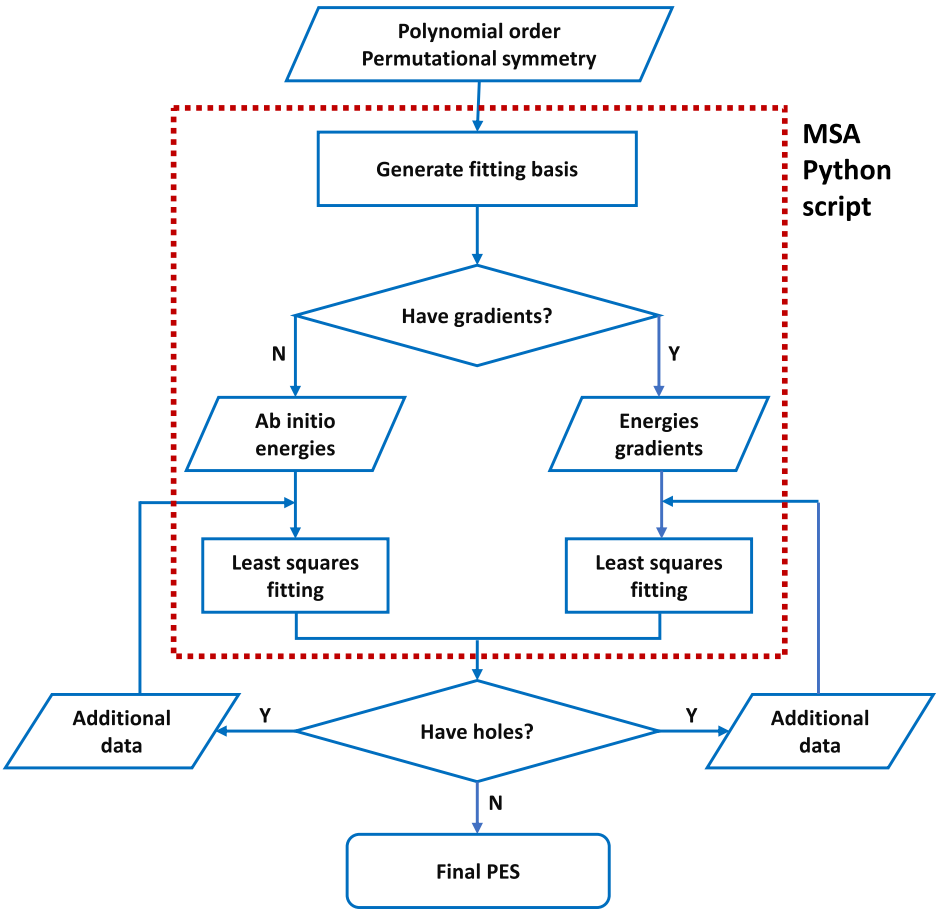


Figure 2. Flowchart of the PES fitting procedures. The procedures in the red rectangle are now integrated in one single Python script.

matrices. In the case of $H_3O_2^-$, all 12 = 2!3! permutations are found. A set of descriptors comprising all $n(n-1)/2$ inverse pairwise interatomic distances, where n is the number of atoms, are constructed for each training set geometry (10 in the present case). Using this, a kernel matrix is constructed in descriptor space where all identified permutational symmetries are summed over. This is finally transformed back to Cartesian space yielding a 3 nm × 3 nM matrix, **K**. Using the atomic forces as a vector **f** of length 3 nM, the equation

$$(\mathbf{K} + \lambda \mathbf{I})\boldsymbol{\alpha} = -\mathbf{f} \tag{5}$$

is solved, where λ is a regularization parameter and **I** is the identity matrix.

The coefficient vector $\boldsymbol{\alpha}$ is the main result of the training. The values of the coefficients depend on a hyperparameter σ that is used in the Matérn kernel. In order to choose the optimal σ, the error is evaluated using the validation data set (rather than the training set) and σ is changed on a grid (e.g., in units of 1), and eq 5 is solved again with each new σ until the error is the least. The optimal σ and $\boldsymbol{\alpha}$ are then used for testing and for prediction.

At a new geometry **x**, the $3n$ forces $\hat{\mathbf{f}}(\mathbf{x})$ at a new configuration are evaluated using

$$\hat{\mathbf{f}}(\mathbf{x}) = \sum_i \mathbf{K}(\mathbf{x}, \mathbf{x}_i)\alpha_i \tag{6}$$

where the summation runs over all the training points and their replicas through permutations. The fitting coefficients are also suitably permuted; they carry the same symmetry as the data points. A technical aspect is that for the prediction stage, the $\boldsymbol{\alpha}$ are saved in descriptor space, and the kernel matrix between the query and training points is also prepared in this space. Hence, the forces are first obtained in the descriptors (inverse distances) and then transformed by the chain rule to Cartesian space. Equation 6 is used to both obtain force and energy errors in the test data set as well as predict them at a queried geometry. The energies are obtained as an integral over the forces.

We note in passing that a possible alternative to the inverse distance descriptors is the use of PIPs in sGDML, which builds in the symmetry aspect of the ML potential. This approach has been used with great success for Neural Network potentials[26,27] and also for Gaussian Approximation potentials.[28]

RESULTS AND DISCUSSION

$H_3O_2^-$ Data Set. A detailed description of the data set of 15,024 CCSD(T) energies and gradients has been reported,[14] so we just briefly summarize it here. The energies extend to 54,000 cm^{-1} with a concentration of energies at roughly 10,000 cm^{-1}. The distribution of the energies is shown in Figure 3. As seen, most of the energies are below 20,000 cm^{-1} with a small number extending to 50,000 cm^{-1}, beyond the range of the abscissa.

The complete data set of energies and gradients is not practical for training and prediction with the sGDML method. Thus, 3000 configurations and 45,000 gradients were selected for training, with another 3000 points used for validation through which the hyperparameter is adjusted. Testing was

https://doi.org/10.1021/acs.jpca.4c01044
J. Phys. Chem. A 2024, 128, 3212−3219

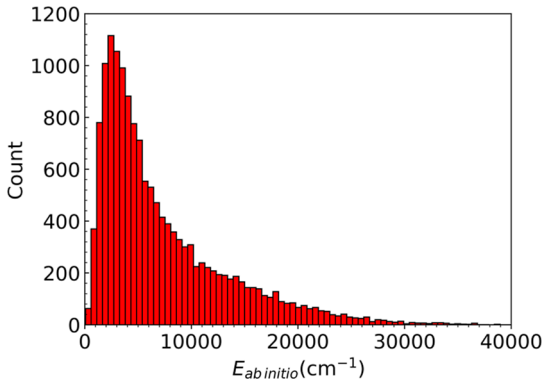


Figure 3. Histogram of the ab initio energies in the full data set. See ref 14 for the construction of the data set.

performed using the remaining 9024 points. Further details of the sGDML fit have been given previously[14] and so we do not repeat those here. For the PIP fit, the total data set was used with no weighting of the data. Another fit was done using 6000 configurations, so a total data size of 96,000. These are the same points used in the sGDML training and validation steps.

Precision and Performance of the Fits. The new PIP PESs use full permutational symmetry, i.e., 3!2! = 12 and a maximum polynomial order of seven. This results in a basis size of 2022 PIPs. The generation of this basis and the fitting are both fast (about 5 min of wall-clock time) using MSA software.[22] The energies and gradients are not weighted, as noted already, and the Morse range parameter, a, equals 3 bohr. A short video showing the interactive steps to do this on a Linux workstation in command line mode can be found here (https://scholarblogs.emory.edu/bowman/msa/).

To begin the assessment of the new PIP PES, we show in Figure 4 the correlation plot of the PIP fit and the eRMSE vs energy for all 15,024 energies. Also shown is the eRMSE for the PIP PES trained on 6000 configurations, denoted PIP^b, and the sGDML PES. As seen, the PIP PES eRMSEs are about half those of sGDML. Very high precision is seen for energies up to 20,000 cm^{-1}, which is sufficient even for quantum studies of the dynamics of this complex.

The energy and force RMSEs, denoted eRMSE and fRMSE, respectively, for the PIP and sGDML PESs are given in Table 1. The RMSEs for the PIP fits using the full and a subset of the data are almost the same and are smaller than the corresponding ones already reported for sGDML.[14] It is perhaps noteworthy that the PIP PES fRMSE is lower than the sGDML one.

Table 1. RMS Errors in Fitted Energies and Forces Computed Using All the CCSD(T)/aug-cc-pVTZ Energies (E) and Forces (F)

error	units	sGDML	PIP^a	PIP^b
eRMSE	cm^{-1}	78.8	42.4	42.2
fRMSE	cm^{-1} Å^{-1} dof^{-1}	194.1	141.4	131.5

^{a}Fit using all data. ^{b}Fit using data at 6000 configurations.

Next, we consider some properties of the various PESs. Hereafter, we consider only the PIP PES fit to all of the data. First, we show the internal coordinates of the global minimum in Table 2. As can be seen, both PESs are in good agreement with the direct ab initio values. This is gratifying since the ab initio minimum configuration is not included in the training data set.

Table 2. Optimized Structure of $H_3O_2^-$ Global Minimum^a

	ab initio	sGDML	PIP
$R(O_1O_2)$	2.4887	2.4836	2.4840
$R(O_1H_3)$	1.0904	1.0844	1.0842
$R(O_1H_4)$	0.9614	0.9588	0.9588
$R(O_2H_5)$	0.9641	0.9612	0.9611
$\theta(H_3O_1O_2)$	1.48	1.50	1.47
$\theta(H_4O_1O_2)$	100.44	100.46	100.41
$\theta(H_5O_2O_1)$	105.72	105.83	105.81
$\phi(H_3O_1O_2H_5)$	−61.66	−62.18	−62.83
$\phi(H_4O_1O_2H_5)$	102.26	101.37	100.85

^{a}All bond lengths are in Å, while the bond angles θ and dihedral angles ϕ are in degrees.

Next consider harmonic frequencies at the global minimum; these are given in Table 3. Here again, the two PESs perform almost equally, with the exception of lowest frequency mode, where the PIP PES is more precise.

Next, we considered the various saddle points. A comparison of ab initio normal-mode frequencies of $H_3O_2^-$ with those from both PESs at the various stationary points, along with their energies, is presented in the Tables 4 and 5. Note that the energies are relative to the global minimum energy. The Cartesian coordinates of these saddle points along with the

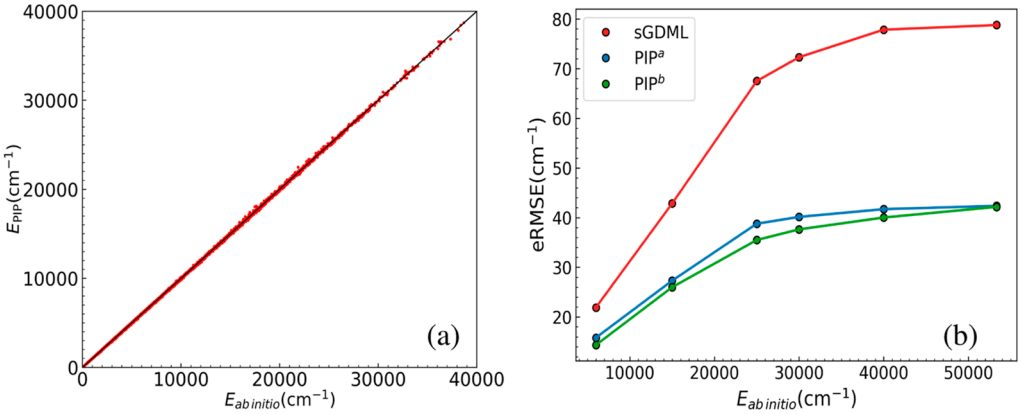


Figure 4. (a) Correlation between the PIP PES fit to all data and CCSD(T)/aVTZ energies. (b) eRMSE for PIP^a, fit to the full data set, PIP^b, fit to data at 6000 configurations, and sGDML PES as a function of the ab initio energy.

https://doi.org/10.1021/acs.jpca.4c01044
J. Phys. Chem. A 2024, 128, 3212−3219

Table 3. Comparison of Harmonic Frequencies (in cm^{-1}) at the Optimized Global Minimum Geometry from Ab Initio Calculations at CCSD(T)/aug-cc-pVTZ Level of Theory Using CFOUR 2.1,[15] the sGDML and PIP PESs

mode	ab initio	sGDML	PIP
HOOH dih	203.0	158.4	193.0
O−O str	326.1	316.5	322.3
OH^- bend	470.4	467.6	482.3
H_2O rock	585.9	583.6	590.9
OH_{in} oop bend	1354.9	1353.9	1366.3
OH_{in} str	1605.1	1627.4	1598.8
OH_{in} i.p. bend	1739.2	1756.7	1734.9
OH^- str	3815.0	3804.4	3814.6
OH_{out} str	3866.3	3863.1	3866.3

Table 4. Comparison of Normal Mode Frequencies and Energies, E, (in cm^{-1}) of $H_3O_2^-$ at the Bifurcation TS and the Shared Proton Transfer TS, from Ab Initio Calculations at the CCSD(T)/aug-cc-pVTZ Level Using CFOUR 2.1, sGDML PES and PIP PES

mode	TS bifurcation			TS H-transfer		
	ab initio	sGDML	PIP	ab initio	sGDML	PIP
Q_1	443.7i	509.5i	471.9i	667.8i	705.1i	642.6i
Q_2	259.6	156.9	171.5	210.6	171.8	202.3
Q_3	287.9	275.4	289.1	568.5	566.0	576.43
Q_4	390.7	295.8	346.5	577.4	575.7	579.41
Q_5	880.2	853.4	850.9	632.2	630.4	636.6
Q_6	1658.8	1651.6	1659.1	1528.1	1524.2	1532.7
Q_7	3613.1	3653.2	3685.1	1626.6	1643.4	1630.4
Q_8	3676.4	3749.8	3736.4	3840.0	3842.4	3840.6
Q_9	3757.9	3810.1	3799.8	3840.6	3847.3	3840.9
E	2521.1	2552.9	2478.4	81.0	85.3	74.8

Table 5. Comparison of Normal Mode Frequencies and Energies, E, (in cm^{-1}) of $H_3O_2^-$ at the Cis and Trans HO−OH Torsion Barriers, from Ab Initio Calculations at the CCSD(T)/aug-cc-pVTZ Level Using CFOUR 2.1, sGDML PES and PIP PES

mode	TS cis			TS trans		
	ab initio	sGDML	PIP	ab initio	sGDML	PIP
Q_1	229.2i	205.9i	224.3i	182.4i	122.2i	161.2i
Q_2	322.5	326.0	321.4	312.1	311.2	311.9
Q_3	437.8	461.2	454.3	413.7	417.5	425.2
Q_4	676.2	691.0	690.4	690.7	695.3	700.8
Q_5	1178.7	1179.7	1182.1	1181.5	1180.0	1193.6
Q_6	1713.1	1729.0	1725.5	1696.9	1726.5	1688.4
Q_7	1838.1	1838.2	1842.2	1817.1	1834.5	1808.3
Q_8	3809.5	3805.9	3810.8	3819.0	3799.8	3820.0
Q_9	3868.2	3860.9	3874.0	3867.8	3867.3	3869.9
E	373.6	309.0	357.1	165.4	75.2	130.2

global minimum, obtained from the PIP PES optimizations, are given in the Supporting Information.

Timing Comparisons. Having established that the sGDML and PIP PES provide precise fits to the CCSD(T) data set, we consider the speed of evaluation of the PESs. The timing was done on the same workstation and, in both cases, using Fortran 90 software. Results for 100,000 evaluations of energy and energy plus gradient are given in Table 6, relative to the PIP time for energy only. First, note that sGDML is

Table 6. Time Per 100,000 Calls

time taken (s)	sGDML	PIP
energy		1
energy + gradient	206	13
energy + fast reverse derivative		3

trained only for gradients, and since the energy is obtained from gradients, we leave the entry blank for the sGDML energy. The timing for the energy plus gradient is 13× the time for energy only for the PIP PES. This is as expected for a standard analytical (forward) gradient evaluation as is done in MSA. Note that this time is much faster (a factor of 15.5) than the time for the sGDML PES. As described in detail elsewhere,[23,29] fast reverse differentiation has been implemented in the Fortran software via a Mathematica script. As seen, there is a substantial speedup in the gradient evaluation (roughly a factor of 4). Thus, the final ratio between the sGDML:PIP timing for the energy plus gradient is 69. For energy evaluation only relevant to quantum calculations, including the DMC ones reported below, the factor is 206. We discuss this large difference in speed (in line with a similar factor for ethanol[29]) below, where we summarize the assessments of the two fitting methods as applied here. Before doing that, we present some results from a DMC calculation of the ground state wave function.

DMC Calculations of the Zero Point Wave Function. Next, we present the results obtained through the DMC calculations[30−32] using our in-house software, as described in our recent paper on using DMC to locate "holes" in a PES.[33] For each PES, we performed five DMC calculations initiated at the global minimum, with 20,000 random walkers and an imaginary time step of $\Delta\tau$ = 5 au for 30,000 time steps. Upon completion of the unconstrained DMC, both PIP and sGDML PESs are identified as "hole-free" surfaces, signifying the absence of any configuration with unphysical negative energies. However, the PIP PES exhibits a speed advantage in this computation, performing 360 times faster than the sGDML PES. The average ZPE values obtained from five independent DMC simulations for PIP and sGDML PES are 6641 ± 2 and 6612 ± 4 cm^{-1}, respectively, the uncertainties given are the statistical ones from the DMC calculations. These ZPEs are substantially lower than the harmonic ZPEs of 6965 and 6983 cm^{-1}, respectively, for the PIP and sGDML PESs.

We now present several plots of 1d wave functions from the DMC wave function obtained with the PIP PES. These 1d wave functions are obtained from histograms of walkers for selected variables at the last time step of the DMC trajectory. Of major interest is the wave function of the shared H atom, H_3 in Figure 1. This is shown using the difference variable $R_{O_2H_3} - R_{O_1H_3}$ in Figure 5, panel (a). As seen, the peak is at zero and this signifies an equal sharing of the hydrogen atom between the two oxygen atoms, and consequently corresponds to the H-transfer TS. This is in agreement with earlier DMC studies of the ground state wave function using an earlier PIP PES.[5] As noted, this symmetric delocalization of the shared H atom is due to the low potential energy (ca. 80 cm^{-1}) of the H-transfer TS. This delocalization was noted in prior path integral,[1,7,9,14,34] DMC,[5] and MCTDH[8] studies.

Next, we investigate a 1d wave function that provides information about the bifurcation TS. This is shown in Figure 5b in the difference variable $R_{O_2H_4} - R_{O_2H_3}$. This variable is zero at the bifurcation TS. As seen, unlike the result in panel

https://doi.org/10.1021/acs.jpca.4c01044
J. Phys. Chem. A 2024, 128, 3212−3219

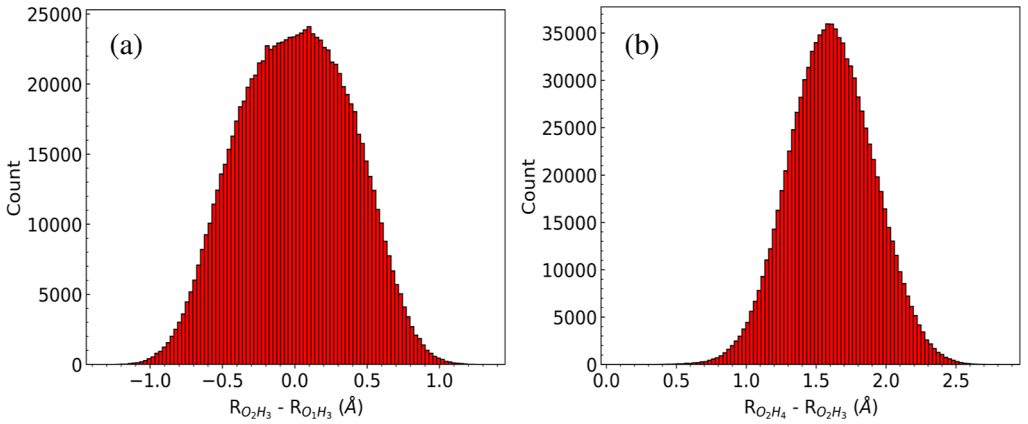


Figure 5. Histogram of cuts of the ground state DMC wave function vs the difference of the bond lengths of (a) $R_{O_2H_3}$ and $R_{O_1H_3}$ and (b) $R_{O_2H_4}$ and $R_{O_2H_3}$ (in Å).

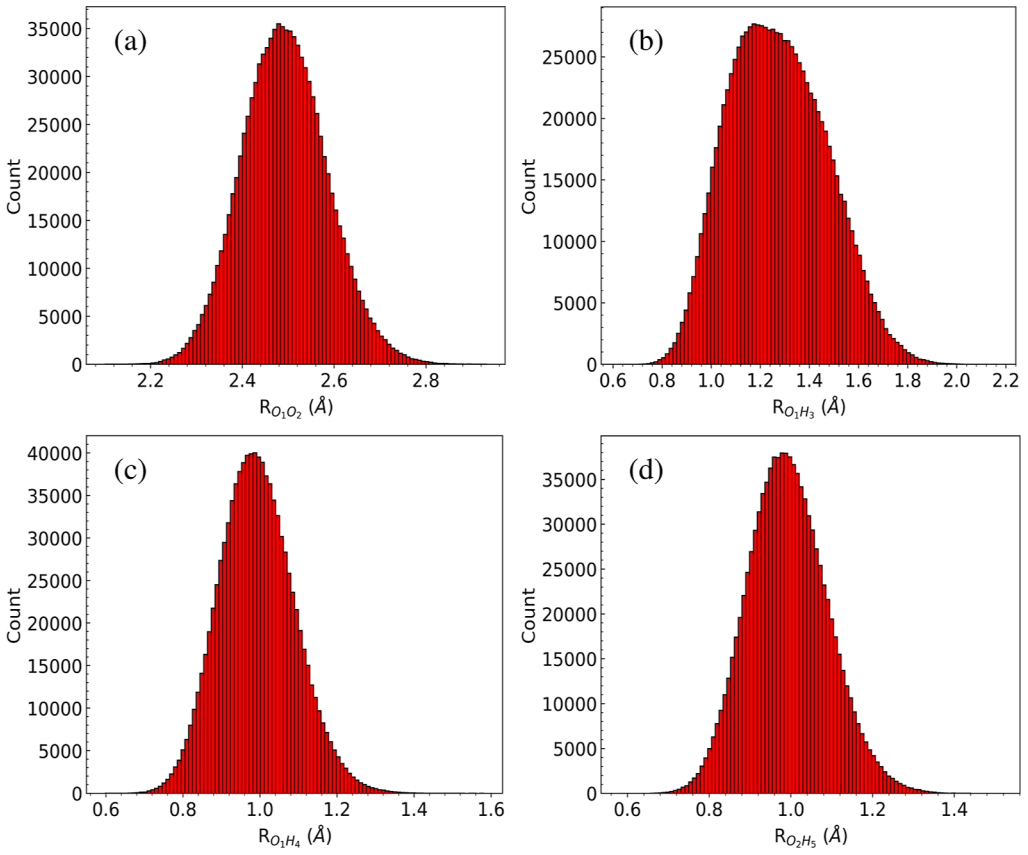


Figure 6. Histogram of cuts of the ground state DMC wave function vs the bond lengths of $H_3O_2^-$ system (in Å): (a) $R_{O_1O_2}$, (b) $R_{O_1H_3}$, (c) $R_{O_1H_4}$, and (d) $R_{O_2H_5}$.

(a), the histogram peaks at 1.58 Å. This difference indicates that one hydrogen atom of the water molecule occupies the space between two oxygen atoms, while the other hydrogen atom remains farther from the oxygen atom of OH^-. Note that the difference of the bond lengths between O_2H_4 and O_2H_3 for the global minimum and the H-transfer TS are 1.42 and 1.59 Å, respectively. On this plot, the wave function is essentially zero at the bifurcation TS. This is not surprising given that the potential energy of this TS is roughly 2500 cm^{-1}. Finally, we show 1d wave functions for the indicated bond lengths in Figure 6. Panel (a) shows the expected Gaussian shape for the O−O bond length extended over a range of 0.6 Å. Panel (b) shows the large range of the shared H atom motion with respect to the O atom. The wave functions in panels (c) and (d) for the O_1H_4 and O_2H_5, respectively, are virtually identical, as expected from the structure in panel (c) of Figure 1.

SUMMARY AND CONCLUSIONS

We presented assessments of two PESs of $H_3O_2^-$. One is a sGDML PES, and the other is a new PIP one. These are successors to the earlier PIP PES reported in 2004. We described the details of both fitting methods and then compared the two PESs with respect to the precision, properties, and speed of evaluation. The two methods

https://doi.org/10.1021/acs.jpca.4c01044
J. Phys. Chem. A 2024, 128, 3212−3219

approach training differently. sGDML uses a subset of the data, which are exclusively 45,000 gradients. While this conforms to the usual training-test protocol, it is numerically necessary for sGDML and all kernel methods, which scale steeply in cost with the training data size. PIPs use all the data, consisting of 15,024 energies and gradients, for a total data size of 240,384. This data size is easily managed in the least-squares fitting performed in the PIPs approach.

The two PESs are similarly precise; however, the PIP PES is much faster to evaluate for energies and energies plus gradient than the sGDML one, with factors of 200 and 70, respectively. This factor of roughly 2 orders of magnitude is consistent with a similar factor found in a previous assessment for ethanol[29] and a very new one for 21-atom aspirin.[35] DMC calculations of the ground vibrational state wave function were done using both PESs. Since these require just energies, the calculation using the PIP PES took roughly 300 times less CPU time than the sGDML one. Analysis of the DMC wave function from the PIP PES calculation indicates that the shared proton is symmetrically located between OH groups but has near zero amplitude at the bifurcation saddle point. As noted previously[3] and also recently,[14] the bifurcation TS energy is much higher than the H atom transfer one. Here, we see that this results in delocalization of the shared H atom between two OH groups but significant localization of the H atom with respect to the bifurcation pathway. This is consistent with results using PIMD simulations using the sGDML PES at low temperatures.[14] It would be interesting to investigate tunneling splittings, where the delocalized shared H atom is replaced by a different H atom via the bifurcation pathway. The present fast PIP PES should enable future studies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.4c01044.

Cartesian coordinates of the various stationary points of $H_3O_2^-$, optimized with the PIP PES, are provided (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Paul L. Houston − *Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, United States; Department of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0003-2566-9539; Email: plh2@cornell.edu

Joel M. Bowman − *Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322, United States;* orcid.org/0000-0001-9692-2672; Email: jmbowma@emory.edu

Sai G. Ramesh − *Department of Inorganic and Physical Chemistry, Indian Institute of Science, Bangalore 560012, India;* orcid.org/0000-0001-6848-867X; Email: sairamesh@iisc.ac.in

Authors

Priyanka Pandey − *Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322, United States;* orcid.org/0000-0002-6930-792X

Mrinal Arandhara − *Department of Inorganic and Physical Chemistry, Indian Institute of Science, Bangalore 560012, India*

Chen Qu − *Independent Researcher, Toronto, Ontario M9B0E3, Canada;* orcid.org/0000-0001-8889-4851

Riccardo Conte − *Dipartimento di Chimica, Università degli Studi di Milano, Milano 20133, Italy;* orcid.org/0000-0003-3026-3875

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.4c01044

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.M.B. thanks the Army Research Office, DURIP grant (W911NF-14-1-0471), for funding a computer cluster where most of the calculations were performed. J.M.B. and P.P. acknowledge current support from the NASA grant (80NSSC22K1167). M.A. and S.G.R. thank the Science and Engineering Research Board of India (grant EMR/2017/003881) for support with computational resources.

■ REFERENCES

(1) Tuckerman, M. E.; Marx, D.; Parrinello, M. The nature and transport mechanism of hydrated hydroxide ions in aqueous solution. *Nature* **2002**, *417*, 925−929.

(2) M. Neumark, D. Spectroscopy of reactive potential energy surfaces. *PhysChemComm* **2002**, *5*, 76−81.

(3) Huang, X.; Braams, B. J.; Carter, S.; Bowman, J. M. Quantum calculations of vibrational energies of H_3O_2-on an ab initio potential. *J. Am. Chem. Soc.* **2004**, *126*, 5042−5043.

(4) Lee, H. M.; Tarkeshwar, P.; Kim, K. S. Structures, energetics, and spectra of hydrated hydroxide anion clusters. *J. Chem. Phys.* **2004**, *121*, 4657−4664.

(5) McCoy, A. B.; Huang, X.; Carter, S.; Bowman, J. M. Quantum studies of the vibrations in H_3O_2− and D_3O_2−. *J. Chem. Phys.* **2005**, *123*, 064317.

(6) 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*, 1487−1490.

(7) Tachikawa, M.; Shiga, M. Geometrical H/D Isotope Effect on Hydrogen Bonds in Charged Water Clusters. *J. Am. Chem. Soc.* **2005**, *127*, 11908−11909.

(8) Yang, Y.; Kühn, O. A full-dimensional quantum dynamical study of the vibrational ground state of $H_3O_2^-$ and its isotopomers. *Z. Phys. Chem.* **2008**, *222*, 1375−1387.

(9) Suzuki, K.; Shiga, M.; Tachikawa, M. Temperature and isotope effects on water cluster ions with path integral molecular dynamics based on the fourth order Trotter expansion. *J. Chem. Phys.* **2008**, *129*, 144310.

(10) McCoy, A. B.; Diken, E. G.; Johnson, M. A. Generating Spectra from Ground-State Wave Functions: Unraveling Anharmonic Effects in the $OH^-{\cdot}H_2O$ Vibrational Predissociation Spectrum. *J. Phys. Chem. A* **2009**, *113*, 7346−7352.

(11) Peláez, D.; Meyer, H.-D. On the infrared absorption spectrum of the hydrated hydroxide (H_3O_2-) cluster anion. *Chem. Phys.* **2017**, *482*, 100−105.

(12) Ludwig, R. New insight into the transport mechanism of hydrated hydroxide ions in water. *Angew. Chem., Int. Ed.* **2003**, *42*, 258−260.

(13) Braams, B. J.; Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577−606.
(14) Arandhara, M.; Ramesh, S. G. Nuclear Quantum Effects in Hydroxide Hydrate Along the H-Bond Bifurcation Pathway. *J. Phys. Chem. A* **2024**, *128*, 1600−1610.
(15) Matthews, D. A.; Cheng, L.; Harding, M. E.; Lipparini, F.; Stopkowicz, S.; Jagau, T.-C.; Szalay, P. G.; Gauss, J.; Stanton, J. F. Coupled-cluster techniques for computational chemistry: The CFOUR program package. *J. Chem. Phys.* **2020**, *152*, 214108.
(16) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K. R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
(17) Chmiela, S.; Sauceda, H. E.; Müller, K. R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.
(18) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K. R.; Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.* **2019**, *240*, 38−45.
(19) Sauceda, H. E.; Chmiela, S.; Poltavsky, I.; Müller, K. R.; Tkatchenko, A. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J. Chem. Phys.* **2019**, *150*, 114102.
(20) Sauceda, H. E.; Chmiela, S.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. *Machine Learning Meets Quantum Physics*; Springer International Publishing, 2020; pp 277−307.
(21) Wang, Y.; Bowman, J. M.; Kamarchik, E. Five ab initio potential energy and dipole moment surfaces for hydrated NaCl and NaF. I. Two-body interactions. *J. Chem. Phys.* **2016**, *144*, 114311.
(22) MSA 2.0 Software with Gradients. 2019. https://github.com/szquchen/MSA-2.0 (accessed Jan 19, 2020).
(23) Houston, P. L.; Qu, C.; Yu, Q.; Conte, R.; Nandi, A.; Li, J. K.; Bowman, J. M. PESPIP: Software to fit complex molecular and many-body potential energy surfaces with permutationally invariant polynomials. *J. Chem. Phys.* **2023**, *158*, 044109.
(24) Xie, Z.; Bowman, J. M. Permutationally invariant polynomial basis for molecular energy surface fitting via monomial symmetrization. *J. Chem. Theory Comput.* **2010**, *6*, 26−34.
(25) http://www.sgdml.org/, symmetric Gradient Domain Machine Learninwebg.
(26) Jiang, B.; Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J. Chem. Phys.* **2013**, *139*, 054112.
(27) Shao, K.; Chen, J.; Zhao, Z.; Zhang, D. H. Communication: Fitting potential energy surfaces with fundamental invariant neural network. *J. Chem. Phys.* **2016**, *145*, 071101.
(28) Qu, C.; Yu, Q.; Van Hoozen, B. L.; Bowman, J. M.; Vargas-Hernández, R. A. Assessing Gaussian Process Regression and Permutationally Invariant Polynomial Approaches To Represent High-Dimensional Potential Energy Surfaces. *J. Chem. Theory Comput.* **2018**, *14*, 3381−3396.
(29) Houston, P. L.; Qu, C.; Nandi, A.; Conte, R.; Yu, Q.; Bowman, J. M. Permutationally invariant polynomial regression for energies and gradients, using reverse differentiation, achieves orders of magnitude speed-up with high precision compared to other machine learning methods. *J. Chem. Phys.* **2022**, *156*, 044120.
(30) Anderson, J. B. A random-walk simulation of the Schrödinger equation: H_3^+. *J. Chem. Phys.* **1975**, *63*, 1499−1503.
(31) Kosztin, I.; Faber, B.; Schulten, K. Introduction to the diffusion Monte Carlo method. *Am. J. Phys.* **1996**, *64*, 633−644.
(32) McCoy, A. B. Diffusion Monte Carlo approaches for investigating the structure and vibrational spectra of fluxional systems. *Int. Rev. Phys. Chem.* **2006**, *25*, 77−107.
(33) Li, J.; Qu, C.; Bowman, J. M. Diffusion Monte Carlo with fictitious masses finds holes in potential energy surfaces. *Mol. Phys.* **2021**, *119*, No. e1976426.
(34) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. On the Quantum Nature of the Shared Proton in Hydrogen Bonds. *Science* **1997**, *275*, 817−820.
(35) Houston, P. L.; Qu, C.; Yu, Q.; Pandey, P.; Conte, R.; Nandi, A.; Bowman, J. M. No Headache for PIPs: A PIP Potential for Aspirin Outperforms Other Machine-Learned Potentials. *J. Chem. Theory Comput.* **2024**, *20*, 3008−3018.

https://doi.org/10.1021/acs.jpca.4c01044
J. Phys. Chem. A 2024, 128, 3212−3219