The Journal of Chemical Physics

The MD17 datasets from the perspective of datasets for gas-phase "small" molecule potentials

Cite as: J. Chem. Phys. **156**, 240901 (2022); https://doi.org/10.1063/5.0089200
Submitted: 23 February 2022 • Accepted: 22 May 2022 • Accepted Manuscript Online: 26 May 2022 • Published Online: 22 June 2022

Joel M. Bowman, Chen Qu, Riccardo Conte, et al.

The Journal of Chemical Physics PERSPECTIVE scitation.org/journal/jcp

The MD17 datasets from the perspective of datasets for gas-phase "small" molecule potentials

Cite as: J. Chem. Phys. 156, 240901 (2022); doi: 10.1063/5.0089200
Submitted: 23 February 2022 • Accepted: 22 May 2022 •
Published Online: 22 June 2022

View Online Export Citation CrossMark

Joel M. Bowman,[1,a)] Chen Qu,[2] Riccardo Conte,[3,b)] Apurba Nandi,[1] Paul L. Houston,[4,5,c)] and Qi Yu[6]

AFFILIATIONS

[1] Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322, USA
[2] Independent Researcher, Toronto, Canada
[3] Dipartimento di Chimica, Università Degli Studi di Milano, via Golgi 19, 20133 Milano, Italy
[4] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA
[5] Department of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
[6] Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA

[a)] **Author to whom correspondence should be addressed:** jmbowma@emory.edu
[b)] **Electronic mail:** riccardo.conte1@unimi.it
[c)] **Electronic mail:** plh2@cornell.edu

ABSTRACT

There has been great progress in developing methods for machine-learned potential energy surfaces. There have also been important assessments of these methods by comparing so-called learning curves on datasets of electronic energies and forces, notably the MD17 database. The dataset for each molecule in this database generally consists of tens of thousands of energies and forces obtained from DFT direct dynamics at 500 K. We contrast the datasets from this database for three "small" molecules, ethanol, malonaldehyde, and glycine, with datasets we have generated with specific targets for the potential energy surfaces (PESs) in mind: a rigorous calculation of the zero-point energy and wavefunction, the tunneling splitting in malonaldehyde, and, in the case of glycine, a description of all eight low-lying conformers. We found that the MD17 datasets are too limited for these targets. We also examine recent datasets for several PESs that describe small-molecule but complex chemical reactions. Finally, we introduce a new database, "QM-22," which contains datasets of molecules ranging from 4 to 15 atoms that extend to high energies and a large span of configurations.

I. INTRODUCTION

There has been dramatic progress in Machine-Learned Potentials (MLPs) over the past 10–15 years with many reviews and perspectives, and five perspectives in J. Chem. Phys.[1–5] The most recent one[5] is a survey of numerous approaches for MLPs for materials research. To quote from that paper, "Our discussion will be primarily about the use of interatomic potential models to model materials, ... We will focus on models that construct a continuous potential energy surface that allows for the calculation of forces and molecular dynamics (MD) simulations." As stated, the focus is on MD simulations, and indeed that has been the major focus of the work in materials. In addition, note a very recent perspective on methods and datasets for many molecules and materials with a focus on structures.[6]

There has also been major work developing MLPs for isolated molecules in the gas-phase, the vast majority of which contain hydrogens atoms and, thus, require some treatment of nuclear

quantum effects. The focus is also on molecular properties, e.g., isomerization, conformers, reaction dynamics, spectroscopy, etc. This contrasts to the focus of materials research, much of which deals with heavy atoms and is aimed at bulk thermal, structural, etc. properties and others for which classical MD simulations are suitable.

The focus of this Perspective is to examine the datasets typically seen in these two areas. For this purpose, we consider several from the MD17 datasets[7,8] and ones from our work for the same molecules.

The molecules at the center of much of the gas-phase work are sometime referred to as "small," and so, we adopt that terminology here and in the title. However, as we will show the datasets for this class of molecules can be large. We will also point out that this class of molecules now routinely includes ones with 15 atoms. In this regard, the spheres of materials and isolated molecules are beginning to overlap, and so there is an opportunity to look at datasets from these two spheres. We do that here.

A secondary but important common focus of both areas worth mentioning is the assessment of MLP methods in terms of precision, computational efficiency, etc. Assessments of these various methods have begun to appear.[9–13] The paper by Pinheiro *et al.*[12] is particularly noteworthy as it examines the performance of ML methods, such as GAP-SOAP,[14] ANI,[15] DPMD,[16] sGDML,[11] PhysNet,[17] KREG,[18] pKREG,[19] and KRR-CM,[18] for fits for ethanol and other molecules, all using the extended MD17 dataset of energies and forces.[8] These popular datasets provide energies and forces for 10 molecules obtained from direct-(classical)dynamics calculations at 500 K. We recently assessed the Permutationally Invariant Polynomial method for ethanol using the MD17 dataset,[13] and that is how we came to use that dataset.

Specifically, we examine datasets for ethanol, malonaldeyde, and glycine. These are nine and ten-atom molecules, so glycine could be considered borderline "large," but that is not of importance. We note that the glycine dataset that we label as "MD17" is not in the MD17 database. It was taken from Ref. 20, where the protocol for MD17 datasets was used with an enhancement of considering two conformers and a path between them. More details are given below.

Before proceeding, and at the risk of stating the obvious, it is worth clarifying what is meant here by a potential energy surface (PES). By this, we mean a full-dimensional, faithful, and precise representation of the adiabatic electronic energy (including nuclear repulsion) as a function of nuclear configuration.

Our group has been active for some time developing MLPs using Permutationally Invariant Polynomials (PIPs)[13,20–27] for "small molecules;" some early examples include CH_5^+ and $H_5O_2^+$, malonaldehyde, and the ten-atom formic acid dimer.[29] More recently, we have reported PIP PESs for 10–15 atom molecules, listed in Table I, using enhancements to the PIP basis.[30–32] The PIPs method was recently evaluated[13] against the ML methods assessed in Ref. 12 (PIPs were shown to be as precise as the most precise methods but to run much faster).

The general target of our PESs is to be accurate at large enough energies to enable both rigorous diffusion Monte Carlo (DMC) calculations of the zero-point energy and wavefunction[33–36] as well as general quantum vibrational and scattering calculations and also quasiclassical trajectory (QCT) calculations of chemical reaction dynamics.[37,38]

TABLE I. Quantum zero-point energies (ZPEs) in kcal/mol and cm^{-1}) of indicated molecules from diffusion Monte Carlo calculation on PIP PESs. In bold are the systems discussed in the text.

Molecule	ZPE (kcal/mol)	ZPE (cm^{-1})
Methane CH_4	27.82	9730
Methonium CH_5^+	31.38	10 975
Water dimer $(H_2O)_2$	28.30	9898
Zundel cation $H_5O_2^+$	35.43	12 393
Ethanol CH_3CH_2OH	49.58	17 339
Malonaldehyde $C_3H_4O_2$	41.97	14 678
Glycine $C_2H_5NO_2$	49.04	17 151
N-methyl acetamide C_3H_7NO	62.63	21 905
Tropolone $C_7H_6O_2$	71.77	25 100

Focusing on DMC calculations, we give in Table I DMC zero-point energies (ZPEs) of a subset of molecules for which PIP PESs have been reported. The energies span a large range (from around 28 to 72 kcal/mol) depending on the size of the molecule, although the variation is not monotonic with the number of atoms. Note also that all of these molecules have H atoms as the most common atom and this accounts in a major way for the large the large ZPEs. The DMC calculation also provides the zero-point wavefunction from which all observable properties of this 0 K state can be derived. Before proceeding, it is perhaps worth commenting on these rigorous 0 K energies and wavefunction and those from a normal-mode analysis. Of course, a full-dimensional PES is not needed for this analysis, which actually provides a good approximation to these DMC energies. However, the separable harmonic-oscillator model of the molecular motion has many well-known limitations, which can only be surmounted with a realistic PES that includes anharmonicity and mode-coupling. Of course, this is a motivation for MLPs.

In Sec. II, we discuss a recent PIP PES for CH_4[39] as a primer for the investigations that follow for some of the larger molecules included in Table I.

II. A PES PRIMER: METHANE

An instructive place to begin is with a rigorous DMC calculation of the zero-point energy of CH_4. We have done this using a demonstration PES and B3LYP density functional theory energies and gradients for the fit.[39] As usual, we probed for "holes," i.e., regions of unphysically low energy, using DMC calculations. These were found in initial fits and then ultimately eliminated by adding data in the dataset to be fitted. The ZPE obtained with this method is 27.8 kcal/mol (9730 cm^{-1}) with a statistical uncertainty of several wavenumbers. (We recently reported DMC as general method, possibly with fictitious masses, to explore a PES for holes.[40]) An interesting question to ask is: What is the distribution of potential energies sampled by the corresponding DMC wavefunction? The answer is shown in Fig. 1. What we see is that this distribution is broad and peaks at a value near the ZPE, but extends significantly beyond that value. This is exactly the qualitative behavior we expect to see for a wavefunction dominated by H atoms — that is, extension

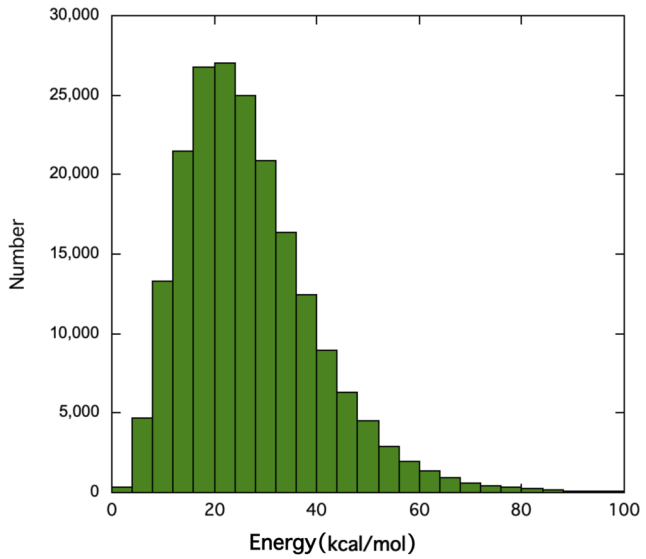

FIG. 1. Histogram of methane potential energies sampled by the DMC zero-point wavefunction.

of the wavefunction significantly into the classically forbidden region.

So clearly the PES used for this calculation needs to extend (faithfully) to energies higher than the ZPE. In fact, the PES used was a fit to data up to roughly 43 kcal/mol (15 000 cm^{-1}), which is about 1.5 times the ZPE. The potential values above 43 kcal/mol shown in Fig. 1 are thus extrapolations of the PES to energies higher than the data used in the fit. With this example in mind, which has a focus on the rigorous ground vibrational state wavefunction, we examine a number of larger examples and datasets for PES fits, which are taken from the literature.

III. CASE STUDIES OF THREE DATASETS FROM MD17, AND OUR WORK

Now we present three case studies that illustrate the points we are making. In all cases the dataset from our group is denoted by acronyms of the authors' last names. The data from MD17, obtained from DFT direct-dynamics run at 500 K, are labeled using that term. This approach, i.e., using DFT direct-dynamics at thermal energies, perhaps as high as 1000 K, is commonly done in the field to generate data for MLPs of a given molecule. We also use direct-dynamics as one means of generating configurations, however, at a number of total energies, including high energies.

A. Malonaldehyde

The tunneling splitting of the proton transfer in malonaldehyde is perhaps the most studied example of proton transfer governed by a barrier. The literature is extensive but of not direct relevance here articles. However, the interested reader is directed to some recent reviewing that literature.[40–42] A PIP ML PES to describe the proton transfer quantitatively (barrier height of 4.1 kcal/mol) used an extensive dataset, which exceeded the zero-point energy of roughly 42 kcal/mol by a factor of about 1.5 and was reported in 2008.[41] A newer ML (LASSO) PES was reported in Ref. 43 with a slightly different barrier height. The dataset for that PES extended to roughly 72 kcal/mol, which again is significantly above the ZPE.

The distributions of energies from the MD17 dataset and Wang *et al.*[41] are given in Fig. 2. As seen, the former one is limited to a maximum energy of around 30 kcal/mol whereas the latter extends to much higher energies. The structure in this dataset comes from several samples, details of which are given elsewhere.[41] This is an important aspect of generating datasets for MLPs, and we defer comments on this to Sec. IV. As seen, the range for most of the energies extends to roughly 100 kcal/mol, which is 2.4 times the ZPE. There is a smaller set of energies that extend to 140 kcal/mol. These high energies beyond the ZPE were added so that quantum calculations based on the Vibrational Self-Consistent Field and Configuration Interaction (VSCF/VCI)[41] and Multi-Configuration Time-Dependent Hartree (MCTDH)[44,45] methods could be done without encountering holes. The methods were used to finally obtain the "right answers," i.e., accurate tunneling splittings for malonaldeyde and d_1-malonaldehyde "for the right reasons."

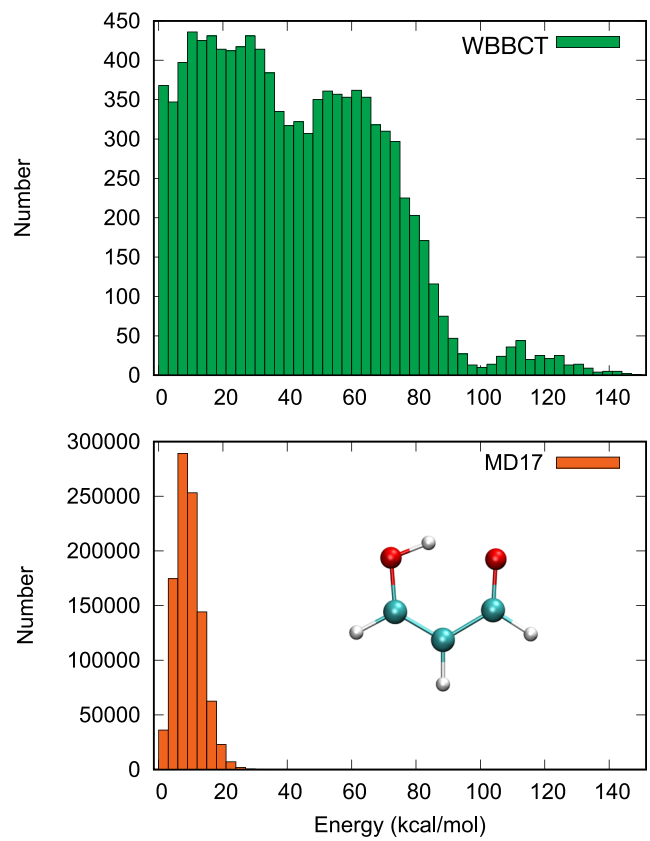

FIG. 2. Histogram of malonaldehyde electronic energies from Ref. 41 and MD17[8] datasets. The WBBCT dataset is reproduced from Wang *et al.* J. Chem. Phys. **128**, 224314 (2008) with the permission of AIP Publishing.

B. Ethanol

Ethanol is an important case to consider next for several reasons. First, it was the focus of two studies assessing numerous MLP methods, mentioned already.[12,13] These studies were based on the MD17 dataset. Second, it is scientifically of fundamental interest as it has two nearly isoenergetic conformers (trans and gauche) and two different methyl rotors. It is also of great applied interest in fields as diverse as combustion and astrochemistry.

In the recent paper from our group,[13] we examined learning curves for several PIP fits to the MD17 dataset following the approach of an earlier study,[12] where learning curves for all methods except the PIP one were reported. One selected plot of these is given Fig. 3. As seen, the performance of most methods is very good with the PIP one showing excellent performance. In addition, the computational speed of the PIP fits was shown to be factors of around 10 to 100 faster than other methods. Details are given in Ref. 13.

The MD17 energy distribution for ethanol is shown in Fig. 4 along with the dataset we recently reported.[13] We were motivated to extend the MD17 dataset after we found many holes in the precise PIP fit to that dataset in DMC calculations. This new dataset was generated at the B3LYP/6-311+G(d,p) level of theory using our usual protocol, i.e., direct-dynamics at a number of total energies and having much larger coverage of configuration space and energies than the MD17 set. These trajectories were propagated for 20 000 time steps of 5.0 a.u. (about 0.12 fs) and with total energies of 1000, 5000, 10 000, 20 000, 30 000, and 40 000 cm^{-1}. A total of 11 trajectories were calculated; one trajectory at the total energy of 1000 cm^{-1} and two trajectories for each remaining total energies. The final dataset consists of 11 000 energies and corresponding 297 000 forces. The extended dataset was fit with the same large PIP basis used for the assessment of PESs based on the MD17 dataset. This new PES was used successfully in DMC and semi-classical (SC) calculations of the ZPE. Thus, this new PES is mostly an example of the ease with which PESs for a molecule with nine atoms and two methyl rotors can be developed.

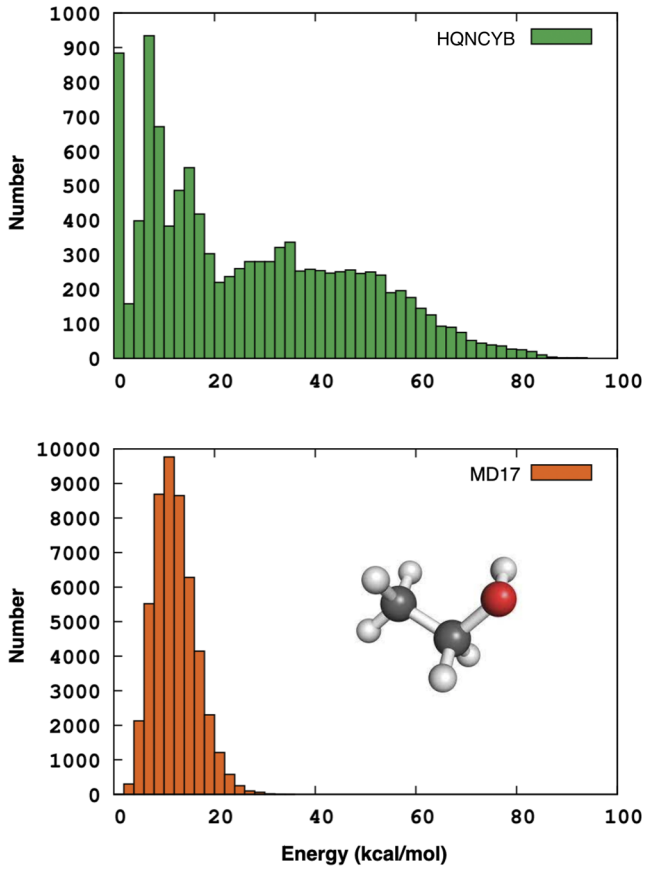

FIG. 4. Histograms of ethanol electronic energies (kcal/mol) from Ref. 13 and MD17 (Ref. 8) datasets.

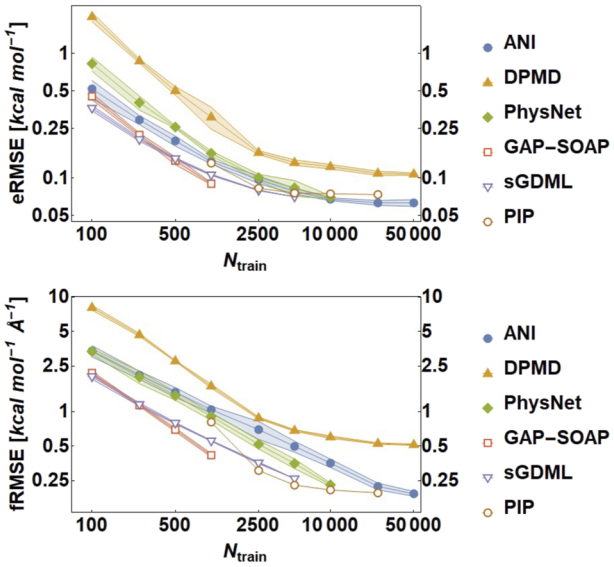

FIG. 3. Ethanol PES learning curves. Reproduced from Houston *et al.* J. Chem. Phys. **156**, 044120 (2022), with the permission of AIP Publishing.

C. Glycine

Next, we consider ten-atom glycine. The distributions of MD17 energies and the ones we recently reported[46] are shown in Fig. 5. We do note that in this case the "MD17" dataset was expanded to include conformer 3 and a reaction path between the lowest energy conformer 1 and conformer 3. No other conformers were included in the database.

Once again there is major difference between the two datasets. Our database for the glycine PES was constructed using DFT with B3LYP hybrid density functional and an aug-cc-pVDZ basis set.[46] Both the energies and gradients were calculated and then fit using the PIP method. The final database included 70 099 geometries and the fit was performed on all energies, and on those gradients associated with the 20 000 geometries with lower energies. The fit was also inverse-energy weighted and gradients were given 1/3 of the weights of the energies.

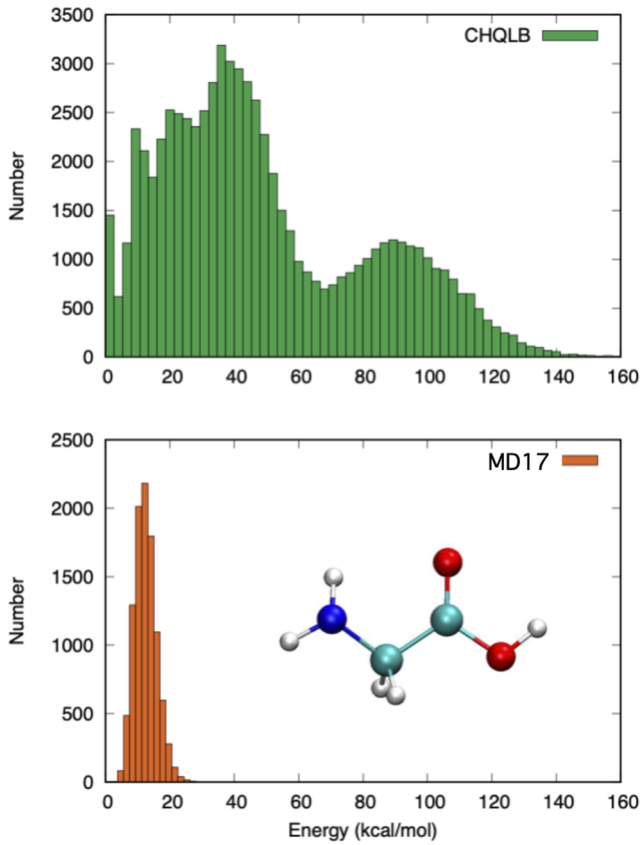

FIG. 5. Histogram of glycine electronic energies (kcal/mol) from Ref. 45 and MD17, Ref. 8, datasets.

Geometries for the database were chosen by an iterative process by first undertaking classical sampling using direct-dynamics, then performing a preliminary fit, and finally by adding points from randomly generated grids centered on the stationary points of the preliminary surface. To assess the reliability of the surface produced by this classical sampling, we used quantum DMC, which frequently revealed holes in the surface. These were removed by adding configurations at the hole configurations. Ultimately, the dataset consisted of 70 000 configurations, including eight conformers and 15 saddle points between them. Figure 6 shows how geometries in the database are distributed according to their nearest conformer. From the figure, it is evident that the global minimum (Conformer 1) has been sampled in more detail but the coverage is pretty uniform for all conformers including higher-energy ones. Table II details the energetics of the eight conformers.

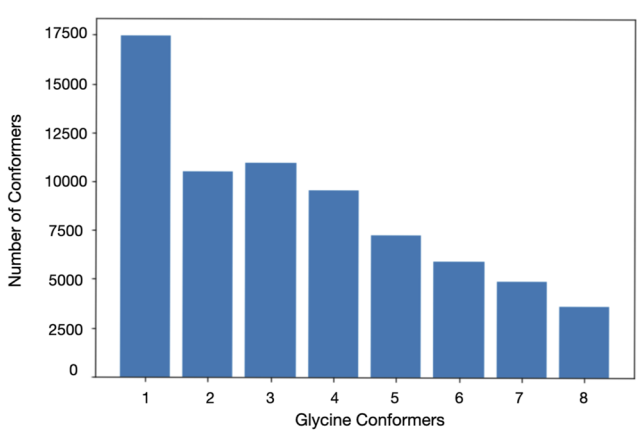

FIG. 6. Histogram of distribution of conformers of glycine for the dataset based on assigning a given configuration to the closest conformer as explained in Ref. 46.

TABLE II. Energies relative to the global minimum (in kcal/mol unless otherwise indicated in the table) of each conformer from the PES or indicated electronic structure theories. All data are from Ref. 45 except the CCSD(T)-F12 numbers that are from Ref. 46.

Conformer	PES (cm^{-1})	PES	B3LYP	CCSD(T)-F12
Conf 1	0.0	0.00	0.00	0.00
Conf 2	205.1	0.59	0.58	0.68
Conf 3	577.1	1.65	1.64	1.73
Conf 4	450.4	1.29	1.27	1.23
Conf 5	945.9	2.70	2.61	2.62
Conf 6	1719.4	4.92	4.91	4.80
Conf 7	2043.6	5.84	5.84	5.89
Conf 8	2174.3	6.22	6.25	6.06

DMC calculations were performed by initiating walkers at the minimum of each conformer. The zero-point energies and wavefunctions were determined.[48] Isosurface plots of the wavefunctions are shown in Fig. 7 where it is clear that the eight conformers can be grouped into pairs described by energetically isolated asymmetric double wells. The corresponding ZPEs were obtained and indeed these group into four values instead of eight. This important finding, i.e., that there are in fact four distinct conformers at 0 K and not eight, required the use of unrestricted DMC calculations.

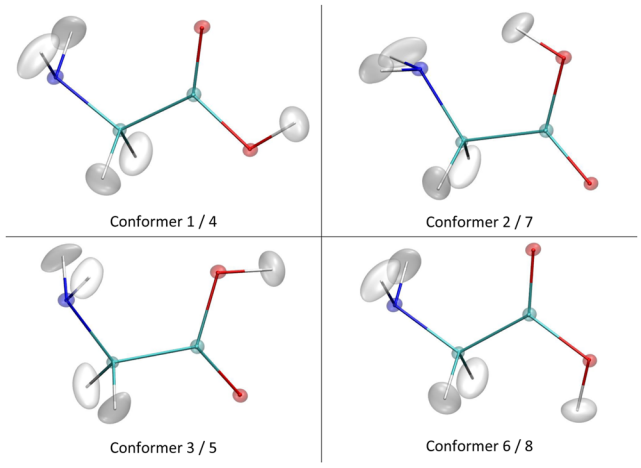

FIG. 7. Isosurface plots of diffusion Monte Carlo ground-state vibrational wavefunctions for glycine. Reproduced from Conte *et al.*, J. Chem. Phys. **153**, 244301 (2020), with the permission of AIP Publishing.

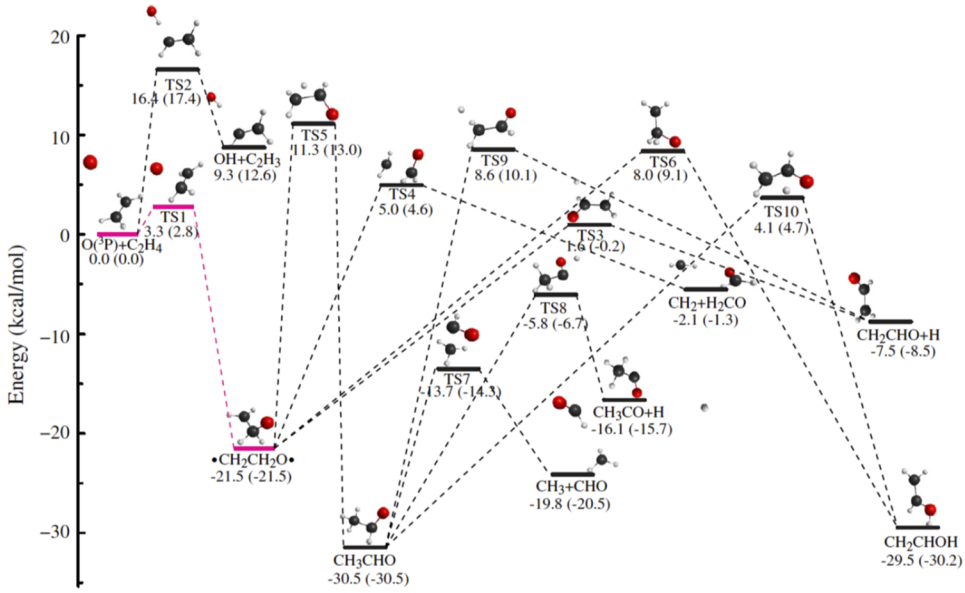

FIG. 8. Schematic showing stationary points for the singlet OC_2H_4 PIP-PES. Taken from Ref. 51. The energies given in pairs refer to the PES and (*ab initio*) ones.

MULTIMODE VSCF/VCI and semiclassical calculations were performed on the PES.[47] The latter were done using the adiabatically switched semiclassical initial value representation method (AS SCIVR)[49–51] to obtain an estimate of the ZPE energy of the several conformers to be compared with DMC values.[48] Results were often in excellent agreement with DMC ones, sometimes showing a difference of less than 5 cm^{-1}. Further details may be found in Ref. 45 and 47.

IV. GAS-PHASE REACTION POTENTIALS

This section examines several MLPs describing chemical reactions in the gas phase with 6 to 9 atoms.

First, consider the singlet PES for acetaldehyde, CH_3CHO. A schematic of a PIP PES, fit to roughly 200 000 (CCSD(T) plus MRCI) energies[52] showing various stationary points and energies, is given in Fig. 8. (This PES together with one for the triplet state[52] were used in QCT calculations in a joint experiment-theory paper.[52]) Here, we show the singlet PES as it illustrates the complex landscape of the PES. The dataset used in the fit spanned the energy range up to roughly 140 kcal/mol and all the stationary points. This is a complex energy landscape that required an extensive dataset in energies and configurations.

An analogous schematic is shown for the recent MLP describing the six-atom $O(^1D)$ + CH_4 reaction.[52] Even though this is a smaller number of atoms, the landscape is complex, and the dataset,

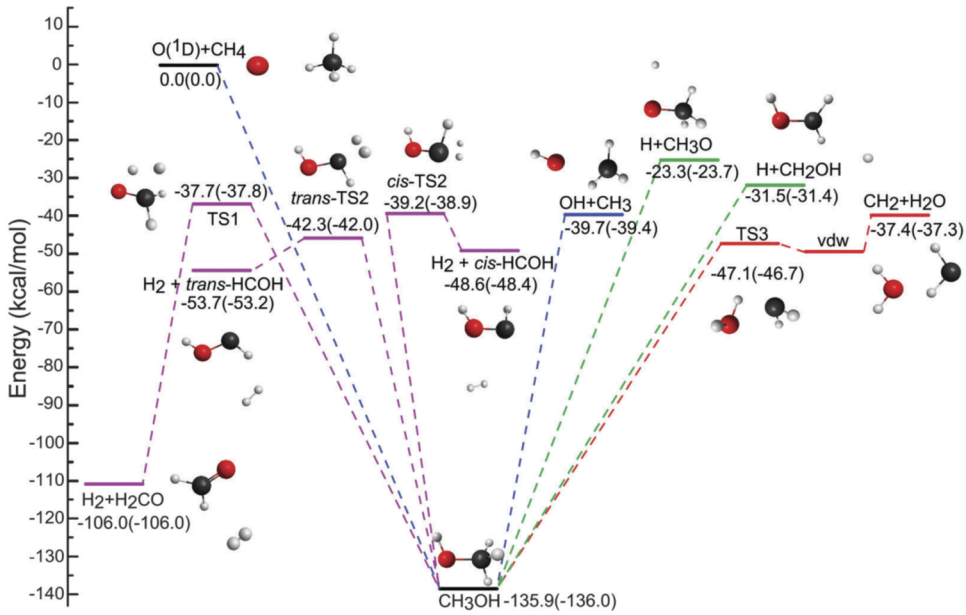

FIG. 9. Schematic showing stationary points on the MLP for the reaction $O(^1D)$ + CH_4. Taken from Ref. 52. The energies given in pairs refer to the PES and (*ab initio*) ones.

which consisted of roughly 340 000 MRCI + Q/aug-cc-pVTZ electronic energies, is both large and extensive. This PES was also used in extensive QCT calculations.[52]

The next example is for the unimolecular dissociation of *syn*-Criegee(*syn*-CH_3CHOO) to OH + CH_2CHO. This is an 8-atom reaction for which we developed a PIP-PES using a dataset of 157 278 energies (CCSD(T) plus MRCI) up to 70 kcal/mol relative to the *syn*-CH_3CHOO minimum.[53] Very recently, Upadhyay and Meuwly reported a PhysNet PES for this reaction[54] and used it in trajectory calculations. The fit was trained on a dataset of 84 322 MP2 energies and associated forces. As with the datasets for the above reaction PESs, this dataset contained a large sample of configurations that spanned the reactant, products, stable intermediates including the important vinyl hydroperoxide (VHP), and van der Waals complexes. Contour plots of important parts of the PES are given in Fig. 10. Note that beyond the VHP minimum, the CH and OO separations are coupled and not useful as the progression coordinate. The right-hand panel shows the O–O bond breaking in VHP. O–O bond-breaking involves multi-reference character which leads to further lowering of the energy beyond the transition state at an O–O separation of 2.2 Å. This PES is reliable for determining rates but not for final state distributions of the OH product.

The final example is the nine-atom reaction $F^- + CH_3CH_2Cl$. The details of an MLP for this reaction were recently reported[55] and used in joint theory (QCT) and experimental study.[56] The PES was fit to 35 474 CCSD(T) energies. The complexity of the PES, i.e., the numerous stationary points and labels of each, is shown in Fig. 11. Details of the labels of the stationary points for the interested reader can be found in Ref. 55. The dataset for this MLP is extended in energy and configuration space, as were the datasets shown above. The relatively small size of the final dataset was certainly a major accomplishment of this recent work, and more details are given in Ref. 55.

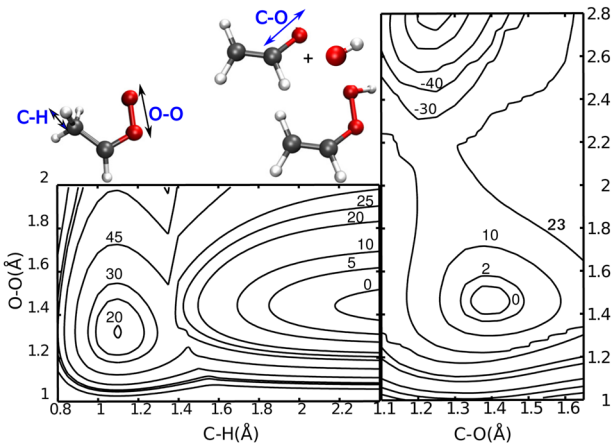

FIG. 10. Two-dimensional cuts through the full-dimensional PES for the *syn*-CH_3CHOO → VHP → CH_2CHO + OH decomposition. The PES is relaxed on each grid point. For the first step (left panel), the CH (proton transfer) and OO coordinates are the driving coordinates whereas for the second step the CO and OO separation are used, respectively. All energies are given in kcal/mol and the minimum of the VHP state is the zero of energy. The left hand panel represents the hydrogen transfer in *syn*-CH_3CHOO to form VHP. See text for more details.

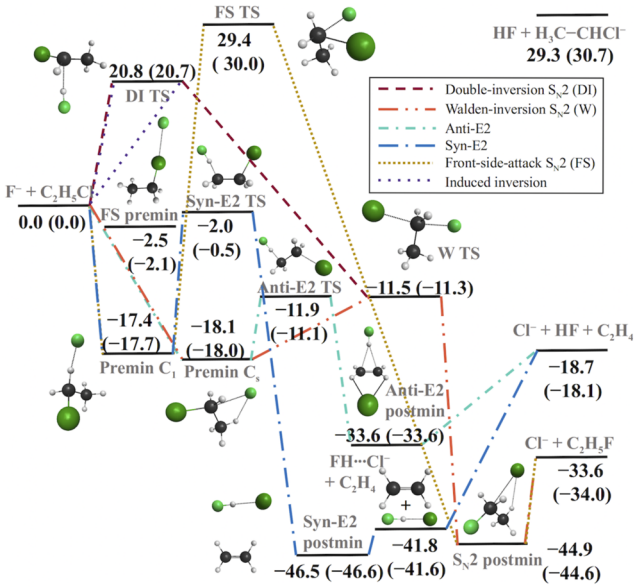

FIG. 11. Schematic from Ref. 55 showing stationary points on the ML PES for reaction $F^- + CH_3CH_2Cl$.

V. DISCUSSION

To begin this section, we return to the MD17 datasets. As noted, these are obtained from direct DFT dynamics at 500 K and so the energy distribution is a single classical one at this temperature. For a molecule of N atoms, the average of this distribution is in the harmonic approximation (k_BT = 500/2 K) (3N-6) ≈ 0.5(3N-6) kcal/mol. So, for malonaldehyde, ethanol, and glycine, these averages are 10.5, 10.5, and 12 kcal/mol, respectively. These are all in accord with the classical distributions shown above for the MD17 datasets.

This approach to generate datasets is certainly a reasonable way to create numerous datasets for a series of molecules which may be used primarily for testing and benchmarking MLP methods. Indeed, that is primarily what those datasets have been used for. The resulting MLPs can certainly be used in classical MD simulations at thermal energies. However, they would almost certainly not be suitable for rigorous quantum, semi-classical, and probably even QCT calculations. In contrast, the datasets we and others have generated are molecule-specific and account for specific target properties. This was described in detail for ethanol, malonaldehyde, and glycine. The datasets for ethanol and glycine, such as those in MD17, use DFT direct-dynamics for both energies and forces. However, they are run at several total energies, all much higher than 500 K, and in the example of glycine starting from the various conformers. The different starting conditions are responsible for the shapes seen, i.e., multiple maxima, in the energy distributions shown above. This sampling includes data at or near the various saddle points separating the eight low-lying conformers. These datasets (and the ones for ethanol, malonaldehyde, and many other molecules) had

quantum DMC calculations of the zero-point state as one target. Indeed, as stated in a recent review "Nuclear Quantum Effects Enter the Mainstream,"[58] the need for MLPs that are robust for quantum effects is well established. Even for MD calculations of reactions, MLPs require big datasets.

The range of nuclear configurations is an important aspect of the PES because this is determined by the target of the PES. Targets can range from spanning a single minimum of a PES to describe anaharmonicty and mode-coupling to studies that require a treatment of large amplitude motion perhaps across several minima to the extreme of reaction dynamics with many reaction channels, minima, and saddle points. There is a rich history of such studies in the gas phase for small molecules. In addition, many reactions involve numerous H atoms. Of course, the light H atoms present challenges ranging simply from large zero-point energies to massive tunneling effects, and these are additions to the complexity of the MLPs, especially for reactive MLPs. Quantum dynamics, which is ideally the approach of choice, makes the most demands on MLPs.[58,59] For reactions with more than five atoms, the challenges of rigorous quantum calculations are currently insurmountable and so (quasi)classical molecular dynamics calculations are done instead using complex PESs. There are numerous examples of such calculations using robust MLPs, and some recent examples can be found in Refs. 36, 37, and 60–64.

In addition to quantum approaches and their demands on PESs, semi-classical (SC) methods, based on trajectories run on the PES, are able to reproduce quantum effects and also require an accurate description of the PES in the high energy region.[64–68] This is because SC methods respect the zero-point energy, and so dynamics are run at least at this energy.

Semi-quantum methods that sample high-energy regions of the PES are the path integral molecular dynamics (PIMD)[70] approach and methods related to it such as centroid molecular dynamics (CMD)[71,72] and ring polymer molecular dynamics (RPMD).[73,74]

To conclude this section, we comment on using direct-dynamics to generate datasets. This is a good approach as it can explore the configuration space based on the dynamics. However, one disadvantage is that amount of the data can be both enormous and also highly correlated. So, we generally keep data only for every tenth or so time step for the dataset. This is somewhat wasteful and, of course, is limited by classical sampling. Hence, we propose another approach that has the DMC zero-point energy specifically in mind as a target. The suggestion is inspired by Fig. 1, the distribution of potential energies from the DMC wavefunction (represented by thousands of "walkers"). This would be a good distribution to use for a dataset, if it could be obtained without the PES. Of course, that is not possible exactly. However, it could be done approximately using an approximate PES. Several possibilities for this range from using a separable harmonic-oscillator model (perhaps "morsified" to create a separable anharmonic PES) to a model from a force field. We plan to investigate this proposal in the near future.

VI. SUMMARY

This Perspective has focused on energy datasets for machine-learned potentials for "small" molecules. We have examined several from the MD17 dataset, which uses an approach that borrows heavily from the perspective of materials research. These have been contrasted with more extensive datasets for the same molecules from the perspective of isolated molecule, gas-phase chemical physics. MLPs for reactions of small molecules were also considered. In both cases, we have shown that datasets for small molecules often need to be much larger, both in energies and in configuration space, than those in the MD17 datasets. As such, these more extensive datasets, which are now available, could provide additional testing datasets for the numerous ML methods. These datasets constitute a new database, "QM-22", which currently consists of the datasets listed in Table III. The selection includes the datasets for the methane, ethanol, malonaldehyde, and glycine PESs discussed here. The ten additional datasets are recent ones from our group and include some that contain gradient data and and a variety of levels of electronic energies, as indicated. They all have been used in DMC calculations and, thus, are "DMC certified." The protocol to generate each dataset is molecule-specific and is aimed at specific scientific goals. The interested reader is referred to the original paper reporting the dataset for details of how the dataset was generated.

TABLE III. Datasets of indicated molecules in the QM-22 database, available for download.[a]

Molecule	Energies	Gradients
Hydronium H_3O^+	CCSD(T)	No
H_2CO, *cis* and *trans*-HCOH	MRCI	No
Methane CH_4	B3LYP	Yes
$OCHCO^+$	CCSD(T)	No
Malonaldehyde $C_3O_2H_4$	CCSD(T)	No
Acetaldehyde (singlet) CH_3CHO	CCSD(T)-MRCI	No
Acetaldehyde (triplet) CH_3CHO	CCSD(T)	No
syn-Criegee CH_3CHOO	CCSD(T)-MRCI	No
Ethanol CH_3CH_2OH	B3LYP	Yes
Formic acid dimer $(HCOOH)_2$	CCSD(T)	No
Glycine $C_2H_5NO_2$	B3LYP	Yes
N-methyl acetamide C_3H_7NO	B3LYP	Yes
Tropolone $C_7H_6O_2$	B3LYP	Yes
Acetylacetone $C_5H_8O_2$	MP2	Yes

[a]https://github.com/jmbowma/QM-22.

Perhaps this Perspective on "small" molecule-large datasets is the flip side of "large" molecules-small datasets, where hopefully the distinction between "large" and "small" datasets is clear.

Datasets for molecules that range from 4 to 15 atoms from our group (and one for acetylacetone from the Meuwly group) are publicly available, and we hope will be used in new tests of the many ML methods mentioned in the Introduction.

ACKNOWLEDGMENTS

J.M.B. thanks NASA (Grant No. 80NSSC20K0360) for financial support. R.C. thanks the Università degli Studi di Milano ("PSR, Azione A Linea 2 - Fondi Giovani Ricercatori") for support. Q.Y. thanks Professor Sharon Hammes-Schiffer and National Science Foundation (Grant No. CHE-1954348) for support. We also thank Gábor Czakó, Bina Fu, Markus Meuwly, and Meenu Upadhyay for

figures used here. Finally, we thank Markus Meuwly, Matthias Rupp, the handlng editor and the two reviewers for helpful comments.

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare that they have no conflicts of interest.

Author Contributions

Joel M. Bowman: Conceptualization (equal); Project administration (equal); Software (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Chen Qu:** Software (equal); Writing – review & editing (equal). **Riccardo Conte:** Writing – review & editing (equal). **Apurba Nandi:** Software (equal); Visualization (equal); Writing – review & editing (equal). **Paul L. Houston:** Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Qi Yu:** Software (equal); Visualization (equal).

DATA AVAILABILITY

Datasets for all molecules listed in Table III are available at https://github.com/jmbowma/QM-22.

REFERENCES

[1] J. Behler, J. Chem. Phys. **145**, 170901 (2016).
[2] J. Westermayr, M. Gastegger, K. T. Schütt, and R. J. Maurer, J. Chem. Phys. **154**, 230903 (2021).
[3] D. Koner, S. M. Salehi, P. Mondal, and M. Meuwly, J. Chem. Phys. **153**, 010901 (2020).
[4] T. Fröhlking, M. Bernetti, N. Calonaci, and G. Bussi, J. Chem. Phys. **152**, 230902 (2020).
[5] T. Mueller, A. Hernandez, and C. Wang, J. Chem. Phys. **152**, 050902 (2020).
[6] M. F. Langer, A. Goeßmann, and M. Rupp, npj Comput. Mater. **8**, 41 (2022).
[7] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).
[8] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, Comput. Phys. Commun. **240**, 38 (2019).
[9] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, J. Chem. Phys. **148**, 241725 (2018).
[10] C. Qu, Q. Yu, B. L. Van Hoozen, J. M. Bowman, and R. A. Vargas-Hernández, J. Chem. Theory Comput. **14**, 3381 (2018).
[11] H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, J. Chem. Phys. **150**, 114102 (2019).
[12] M. Pinheiro, F. Ge, N. Ferré, P. O. Dral, and M. Barbatti, Chem. Sci. **12**, 14396 (2021).
[13] P. L. Houston, C. Qu, A. Nandi, R. Conte, Q. Yu, and J. M. Bowman, J. Chem. Phys. **156**, 044120 (2022).
[14] A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015).
[15] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).
[16] L. Zhang, J. Han, H. Wang, R. Car, and W. E, Phys. Rev. Lett. **120**, 143001 (2018).
[17] O. T. Unke and M. Meuwly, J. Chem. Theory Comput. **15**, 3678 (2019).
[18] P. O. Dral, A. Owens, S. N. Yurchenko, and W. Thiel, J. Chem. Phys. **146**, 244108 (2017).
[19] P. O. Dral, J. Comput. Chem. **40**, 2339 (2019).
[20] V. Vassilev-Galindo, G. Fonseca, I. Poltavsky, and A. Tkatchenko, J. Chem. Phys. **154**, 094119 (2021).
[21] B. J. Braams and J. M. Bowman, Int. Rev. Phys. Chem. **28**, 577 (2009).
[22] See https://github.com/szquchen/MSA-2.0 for MSA software with gradients; accessed 20 January 2019.
[23] C. Qu, Q. Yu, and J. M. Bowman, Annu. Rev. Phys. Chem. **69**, 151 (2018).
[24] T. Győri and G. Czakó, J. Comput. Theory Chem. **16**, 51 (2020).
[25] G. Czakó, T. Győri, D. Papp, V. Tajti, and D. A. Tasi, J. Phys. Chem. A **125**, 2385 (2021).
[26] E. Lambros, S. Dasgupta, E. Palos, S. Swee, J. Hu, and F. Paesani, J. Chem. Theory Comput. **17**, 5635 (2021).
[27] D. R. Moberg and A. W. Jasper, J. Chem. Theory Comput. **17**, 5440 (2021).
[28] D. R. Moberg, A. W. Jasper, and M. J. Davis, J. Phys. Chem. Lett. **12**, 9169 (2021).
[29] C. Qu and J. M. Bowman, Phys. Chem. Chem. Phys. **18**, 24835 (2016).
[30] C. Qu and J. M. Bowman, Phys. Chem. Chem. Phys. **21**, 3397 (2019).
[31] R. Conte, C. Qu, P. L. Houston, and J. M. Bowman, J. Chem. Theory Comput. **16**, 3264 (2020).
[32] P. L. Houston, R. Conte, C. Qu, and J. M. Bowman, J. Chem. Phys. **153**, 024107 (2020).
[33] A. B. McCoy, B. J. Braams, A. Brown, X. Huang, Z. Jin, and J. M. Bowman, J. Phys. Chem. A **108**, 4991 (2004).
[34] A. B. McCoy, X. Huang, S. Carter, M. Y. Landeweer, and J. M. Bowman, J. Chem. Phys. **122**, 061101 (2005).
[35] A. B. McCoy, Int. Rev. Phys. Chem. **25**, 77 (2006).
[36] Y. M. Wang, V. Babin, J. M. Bowman, and F. Paesani, J. Am. Chem. Soc. **134**, 11116 (2012).
[37] J. M. Bowman, G. Czakó, and B. Fu, Phys. Chem. Chem. Phys. **13**, 8094 (2011).
[38] G. Czakó and J. M. Bowman, J. Phys. Chem. A **118**, 2839 (2014).
[39] A. Nandi, C. Qu, and J. M. Bowman, J. Chem. Theor. Comput. **15**, 2826 (2019).
[40] J. Li, C. Qu, and J. M. Bowman, Mol. Phys. **119**, e1976426 (2021).
[41] Y. Wang, B. J. Braams, J. M. Bowman, S. Carter, and D. P. Tew, J. Chem. Phys. **128**, 224314 (2008).
[42] T. Baba, T. Tanaka, I. Morino, K. M. T. Yamada, and K. Tanaka, J. Chem. Phys. **110**, 4131 (1999).
[43] W. Mizukami, S. Habershon, and D. P. Tew, J. Chem. Phys. **141**, 144310 (2014).
[44] T. Hammer and U. Manthe, J. Chem. Phys. **136**, 054105 (2012).
[45] M. Schröder and H.-D. Meyer, J. Chem. Phys. **141**, 034116 (2014).
[46] J. Zhu, V. Q. Vuong, B. G. Sumpter, and S. Irle, MRS Commun. **9**, 867–873 (2019).
[47] E. M. Orjan, A. B. Nacsa, and G. Czakó, J. Comput. Chem. **41**, 2001 (2020).
[48] R. Conte, P. L. Houston, C. Qu, J. Li, and J. M. Bowman, J. Chem. Phys. **153**, 244301 (2020).
[49] R. Conte, L. Parma, C. Aieta, A. Rognoni, and M. Ceotto, J. Chem. Phys. **151**, 214107 (2019).
[50] G. Botti, M. Ceotto, and R. Conte, J. Chem. Phys. **155**, 234102 (2021).
[51] G. Botti, C. Aieta, and R. Conte, J. Chem. Phys. **156**, 164303 (2022).
[52] B. Fu, Y.-C. Han, J. M. Bowman, L. Angelucci, N. Balucani, F. Leonori, and P. Casavecchia, Proc. Natl. Acad. Sci. U. S. A. **109**, 9733 (2012).
[53] K. Shao, B. Fu, and D. H. Zhang, Phys. Chem. Chem. Phys. **17**, 24098 (2015).
[54] N. M. Kidwell, H. Li, X. Wang, J. M. Bowman, and M. I. Lester, Nat. Chem. **8**, 509 (2016).
[55] M. Upadhyay and M. Meuwly, ACS Earth Space Chem. **5**, 3396 (2021).
[56] V. Tajti and G. Czako, Phys. Chem. Chem. Phys. **24**, 8166 (2022).
[57] J. Meyer, V. Tajti, E. Carrascosa, T. Győri, M. Stei, T. Michaelsen, B. Bastian, G. Czakó, and R. Wester, Nat. Chem. **13**, 977 (2021).
[58] T. E. Markland and M. Ceriotti, Nat. Rev. Chem. **2**, 0109 (2018).
[59] B. Fu and D. H. Zhang, J. Chem. Theory Comput. **14**, 2289 (2018).
[60] J. Li, B. Zhao, D. Xie, and H. Guo, J. Phys. Chem. Lett. **11**, 8844 (2020).
[61] G. Czakó and J. M. Bowman, Science **334**, 343 (2011).
[62] J. Espinosa-Garcia, J. Phys. Chem. A **120**, 5 (2016).
[63] A. Tasi and G. Czakó, J. Chem. Phys. **156**, 184306 (2022).

[64] M. Meuwly, J. Phys. Chem. B **126**, 2155 (2022).
[65] W. H. Miller, J. Phys. Chem. A **105**, 2942 (2001).
[66] E. J. Heller, Acc. Chem. Res. **14**, 368 (1981).
[67] M. F. Herman and E. Kluk, Chem. Phys. **91**, 27 (1984).
[68] A. L. Kaledin and W. H. Miller, J. Chem. Phys. **118**, 7174 (2003).
[69] C. Aieta, G. Bertaina, M. Micciarelli, and M. Ceotto, J. Chem. Phys. **153**, 214117 (2020).
[70] M. Ceriotti, M. Parrinello, T. E. Markland, and D. E. Manolopoulos, J. Chem. Phys. **133**, 124104 (2010).
[71] T. D. Hone, S. Izvekov, and G. A. Voth, J. Chem. Phys. **122**, 054105 (2005).
[72] T. J. H. Hele, M. J. Willatt, A. Muolo, and S. C. Althorpe, J. Chem. Phys. **142**, 191101 (2015).
[73] J. O. Richardson and S. C. Althorpe, J. Chem. Phys. **131**, 214106 (2009).
[74] S. Habershon, D. E. Manolopoulos, T. E. Markland, and T. F. Miller, Annu. Rev. Phys. Chem. **64**, 387 (2013).