Permutationally Invariant Polynomial Potential Energy Surfaces for Tropolone and H atom Tunneling Dynamics

Paul Houston,[1] Riccardo Conte,[2] Chen Qu,[3] and Joel M. Bowman[4]

[1] *Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, U.S.A. and Department of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332, U.S.A*[a]

[2] *Dipartimento di Chimica, Università Degli Studi di Milano, via Golgi 19, 20133 Milano, Italy*[b]

[3] *Department of Chemistry & Biochemistry, University of Maryland, College Park, Maryland 20742, U.S.A.*

[4] *Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, Georgia 30322, U.S.A.*[c]

(Dated: 19 June 2020)

We report permutationally invariant polynomial (PIP) fits to energies and gradients for 15-atom tropolone. These include standard, augmented, and fragmented PIP bases. Approximately 6600 energies and associated gradients are obtained from direct-dynamics calculations using DFT/B3LYP/6-31+G(d) supplemented by grid calculations spanning an energy range up to roughly 35 000 cm^{-1}. Three fragmentation schemes are investigated with respect to efficiency and fit precision. In addition several fits are done with reduced weight for gradient data relative to energies. These do result in more precision for the H-transfer barrier height. Properties of the fits such as stationary points, harmonic frequencies and the barrier to H-atom transfer are reported and compared to direct calculations. A 1-d model to obtain the tunneling splitting for the ground vibrational state and qualitative predictions for excited vibrational states is employed. The results are in encouraging agreement with experiment.

[a] Electronic mail: plh2@cornell.edu

[b] Electronic mail: riccardo.conte1@unimi.it

[c] Electronic mail: jmbowma@emory.edu

I. INTRODUCTION

In the past 15 years there has been dramatic progress in developing methods to represent *ab initio*-based potential energy surfaces (PESs) for complex chemical reactions with up to eight atoms, molecules with symmetric minima, and non-covalent complexes with up to 10 atoms. Much of this progress has been made in the development of non-parametric, machine learning approaches to fit large data sets of electronic energies for polyatomic molecules and clusters.[1–15] Among these there are several that have specifically been applied to molecules (as opposed to homogeneous materials). They are permutationally invariant polynomials (PIPs)[1,14], Neural Networks (NN),[3–6,16–19] NN with PIP inputs,[10–13,20,21] Gaussian Process regression (GPR),[8,15,22] and GPR with PIP inputs.[23] These methods are now basically routine for 4–8 atoms. There are a number of issues to go beyond 10 atoms. One that is common to all fitting methods is the highly non-linear increase in the computational effort in using the "gold standard" level of *ab initio* theory, i.e., CCSD(T), for larger molecules. Other issues include the number of variables, which is equivalent to the mathematical dimensionality of the space. For PIP methods, both direct PIP regression and using PIPs for inputs for NN and GPR, the number of polynomials grows rapidly with the number of variables. This particular bottleneck has been the motivation for a recent strategy from our group to break the 10-atom limit.[24–26] The basic idea (given in more detail below) is to fragment the PIP basis. We investigate this approach here for 15-atom tropolone. While this is a significant step beyond the 10-atom limit, we are able to generate a full PIP basis (of maximum polynomial order of 3) and this provides a "benchmark" PES.

Before we present the new work on tropolone, we make some general remarks about the plethora of methods mentioned above. They all have in common that they do not rely on a model for representing potentials, e.g., Lennard-Jones, LEPS, exp/6, force-fields, etc. In that sense they are all non-parametric in the language of machine-learning. Also, they use a universal set of inputs, also known as descriptors, for the fitting. For PIP, PIP-NN, and PIP-GPR these are all the internuclear distances or transformations of these, e.g., Morse variables. This is distinct from approaches using models, where the variables are molecule-specific. So, based on these basic aspects, the above are all machine-learning methods; however, there is a fundamental difference in the way the data are used in these methods. In PIP and NN the parameters contained in those approaches are optimized, generally using a

2

least squares minimization criterion. In PIP, where parameters are linear this is linear-least-squares. In NN fitting the parameters are non-linear and so the least squares optimization is non-linear. The number of parameters in both approaches can be thousands or even tens of thousands, depending on the dimensionality of the space. As an aside, it is worth noting that if Morse variables are used as the inputs for PIP or PIP-NN, then the range parameter(s) of the variables could be treated as additional non-linear parameters. The number of such parameters is very small, however, compared to the thousands of parameters just mentioned. In any case, in these methods the data are used in the optimizations but not for prediction. By contrast, other machine-learning methods use the data explicitly for prediction. GPR is a prominent example of this method. Modified Shepard interpolation[27] and kernel ridge regression[28] are other examples. All these methods have the same goals, which are a precise representation of the known data (GP can reproduce the data exactly but since this generally leads to a singular matrix for prediction "noise" is added to the data) and smooth and accurate predictions. It is of course of interest to compare the performance of these different ML approaches. We refer the interested reader to a recent assessment of GPR and PIP regression for four case studies, where both methods perform with about equal high quality.[23]

Of course all machine-learning approaches aim for ever increasing ranges of applicability, in particular to larger molecules. As noted above, the PIP approach has been applied for molecules with as many as 10 atoms and this value has been cited in the literature as the practical limit for the PIP, while the PIP-NN approach has been applied to the seven atom $OH+CH_4$ reaction.[29] The "ten-atom limit" using the PIP approach was broken for the 12-atom *trans-N*-methyl acetamide (*trans*-NMA).[24] The major point of that paper, which is preliminary to the present one, was to describe a fragmented PIP approach able to extend the PIP method to larger molecules. We also note very recent progress for larger molecules using other methods. These include Gaussian Process Regression (using a different fragmentation strategy) to obtain a PES for the 19-atom protonated imidazole dimer,[30] and a high-dimensional NN approach to develop PESs for H-atom transfers in three molecules with up to 15 atoms[31] using the PhysNet software.[32]

Returning to the PIP approach, we recently suggested that there is a regime of applicability for molecules with more than 10 atoms but probably less than hundreds of atoms where the PIP approach could be extended.[24,25] This suggestion arises from the fact that Morse

variables go asymptotically to zero at large inter-nuclear distances. Thus for large molecules, with necessarily a large number of large inter-nuclear distances, many Morse variables are essentially zero so that any PIP basis function containing these variables is zero and can be dropped from the total basis. This observation allows one to fragment the larger target molecule into smaller moieties for which PIP basis sets can be generated efficiently.[24,25] This approach is indeed successful and it was used with recent, extended MSA software that includes gradient data for fits.[33] However, several issues with the algorithm were noted. These included redundant terms in the basis and also the cost of evaluating gradients. New software, described in a recent paper,[26] solves these two problems. In addition, a novel approach to prune or augment a full PIP basis was also proposed in that paper; more details are now provided.

In this paper we apply a combination of full basis, fragmentation, augmentation, and pruning to develop potential energy surfaces for the 15-atom molecule tropolone. Like the previously studied malonaldehyde,[34,35] tropolone is of interest because of its hydrogen transfer dynamics.[36] The small mass of hydrogen allows it to tunnel through the barrier of a double well, resulting in a splitting of vibrational levels. Building on substantial FTIR work by Redington and coworkers,[37–44] as well as jet-cooled infrared fluorescence dip spectroscopy,[45] Vaccaro and coworkers have recently used two-color resonant four-wave mixing techniques to measure the splitting in tropolone as a function of vibrational level.[46] Although there have been several previous theoretical approaches to this problem,[47–50] these new measurements beg for a detailed explanation, one that no doubt will require an accurate potential energy surface as well as a quantum dynamics approach for determining the splitting and its dependence on mode. Because the double well in tropolone is symmetric, the surface needs to maintain permutational symmetry, posing yet another challenge.

In the next section we briefly describe the various new approaches to developing a PES for tropolone, then we present details of the database of energies and gradients followed by an analysis of the various PESs. In the penultimate section, by presenting an approximate calculation of the H-atom tunneling dynamics obtained from the PESs as well as qualitative predictions of mode-specific tunneling, we make close contact with the recent experiments performed by the Vaccaro group.[46] A short summary and conclusions are given in the final section.

II.   BRIEF RECAP OF THEORY

To begin, recall that the PIP approach represents the potential as follows, using compact notation:

$$V = \sum_{i=1}^{n_p} c_i p_i, \tag{1}$$

where c_i are coefficients, p_i are permutationally invariant polynomials, denoted as PIPs, (the basis set) and n_p is the total number of polynomials for a given maximum polynomial order. The p_i are typically functions of Morse variables, which themselves are functions of the inter-atomic distances, $r_{\alpha,\beta}$ (by the usual exponential relationship $\exp(-r_{\alpha,\beta}/\lambda)$, where λ is commonly chosen to be equal to 2 bohr). Following the practice of the MSA software,[51] Morse variables are in this work denoted by x_l. We further stipulate that the basis functions should be chosen to maintain permutational invariance among identical atoms, or at least the most important of them. The linear coefficients are obtained using standard least squares fits to large data sets of electronic energies at "scattered" geometries.

In standard PIP approaches, the computational issue arises when the basis set for the parent molecule becomes completely cumbersome – too large to be practically useful, either because the number of coefficients is so large that the least squares optimization becomes problematic or because calculating the proper basis set takes too long. The size of the basis varies in a complicated and generally non-linear way with respect to the maximum polynomial order, the number of Morse variables, and the order of the symmetric group.[1] This growth in the size of the PIP basis is the basic consideration in stating the 10-atom limit for the approach.

However, as noted above, the fragmented basis approach is an effective way to break this limit. Clearly, by fragmenting a parent molecule into groups of smaller molecular moieties the basis for each smaller moiety can be calculated rapidly and then combined with those of other fragments to provide a compact and hopefully still precise representation of the potential energy.[24] To be specific, consider a simple example of a 5-atom molecule with atoms labeled as 1–5 and a scheme in which the molecule is fragmented into three fragments, say $\{1,2,3\}$, $\{2,3,4\}$, $\{3,4,5\}$. In this 3-fragment scheme the potential is given compactly by

$$V = \sum_i c_i\, p_i(\boldsymbol{x}_1, \boldsymbol{m}_1) + \sum_j c_j'\, p_j'(\boldsymbol{x}_2, \boldsymbol{m}_2) + \sum_k c_k''\, p_k''(\boldsymbol{x}_3, \boldsymbol{m}_3), \tag{2}$$

where $\{p\}$, $\{p'\}$, and $\{p''\}$ are PIP bases for the nth fragment, $n = 1,2,3$, $\{c\}$, $\{c'\}$, $\{c''\}$ are

the corresponding linear coefficients, $\boldsymbol{x}_n$ represent the set of corresponding Morse variables, and $\boldsymbol{m}_n$ indicate a set of monomials built from the Morse variables. Morse variables between atoms 1 and 4, atoms 1 and 5, and atoms 2 and 5 are assumed to be zero and hence not in the fragmented bases. In this example there are some Morse variables in common among the fragments, and it should be clear that there are indeed some redundant basis functions in this expression in terms of common Morse variables.

These issues were pointed out previously;[24,25] however, they were not "fatal" ones, because the linear least squares method used was able to deal with a modest number of identical basis functions. Nevertheless, there is compelling motivation to eliminate these redundant basis functions and thereby reduce the size of the basis. We do note the redundant-term issue is similar to an issue two of us identified earlier for developing PIP representations of interaction potentials that should rigorously vanish in asymptotic regions where there is no interfragment interaction. In that case the issue concerned basis functions involving Morse variables of fragments that do not go to zero at large internuclear distances where rigorously there is no inter-fragment interaction. An effective "empirical" pruning procedure was then employed to eliminate such basis functions and applied to several systems.[52–54]

An "empirical" pruning or adding approach is also developed here as a post-processing task performed on the standard complete PIP bases of the fragments.

III. COMPUTATIONAL METHODS

Development of an accurate, permutationally invariant potential energy surface for tropolone is based on the choice and application of *ab initio* methods. Because the surface requires energies and gradients at many geometries, we chose for most of the work to use density functional theory at the B3LYP level with a 6-31+G(d) basis set, as implemented in Molpro.[55,56] Characterization of stationary points on the surface at the B3LYP/6-31+G(d) level was first carried out for the geometries of the global minimum (GM), for the transition state for hydrogen transfer between the two oxygen atoms, for conformer I, in which the OH is pointed away from the second oxygen, and for the transition state between conformer I and the global minimum. Optimum geometries and harmonic vibrational frequencies were determined at each of these stationary points.

An *ab initio* molecular dynamics program was used to calculate B3LYP/6-31+G(d) en-

ergies and gradients at geometries encountered on trajectories that started from specific locations. For example, 6000 steps were calculated starting from the optimized GM for energies of 4000, 10000, 20000, 30000, and 40000 cm^{-1}, while 3000 steps were calculated starting from the optimized geometry for conformer I at an energy of 6000 cm^{-1}. For each trajectory, a data set of geometries, energies and gradients was constructed by taking data from every tenth step of each trajectory, giving data at 3300 geometries. An additional 3301 data points were taken from calculations on grids centered on the geometries of the GM and of the H-transfer transition state. This data set of energies and gradients at 6601 geometries was used for fitting all PES surfaces. Based on the 6601 geometry data set, two smaller data sets were also investigated: a 6249 geometry set with an upper energy cutoff of 20000 cm^{-1} and a 5717 geometry set with an upper energy cutoff of 15000 cm^{-1}. The 6601 geometry set was augmented to 6768 later, to fix the "holes" found in the diffusion Monte Carlo calculation of the zero-point energy. This 6768 set was only used in two best fits. A smaller set of 3601 geometries, including all the *ab initio* molecular dynamics points and 301 grid points centered on the transition state to hydrogen transfer, was also investigated in some fits. Finally, a "test", data set was also taken from the trajectories, using approximately every 40th point and separated maximally from points used in the main data set; the test set contained energies and gradients at 825 geometries.

Information about the GM and the transition state for H-atom transfer was also obtained using CCSD(T)-F12 with the aVDZ basis sets. Optimized geometries were calculated and harmonic vibrational frequencies were obtained.

The fits are made to an *ab initio* set of, typically, 6601 energies and $(3N) \times 6601$ gradient components, where $N = 15$ is the number of atoms in the parent molecule, tropolone. A distribution of energies is shown in Fig. 1. The fit is compared to the original *ab initio* data by calculating the root-mean squared errors (RMSE) for comparison of the potential energies and the gradient components between the fit and the *ab initio* data as well as by determining the time it takes to evaluate the potential energy and gradient components for a particular geometry. The times we will report are those for evaluating 3000 geometries to obtain an energy and to obtain 45 gradient components for each. There is usually a trade-off between the speed and accuracy of the calculation, and typically the larger the basis set, the slower the calculation but the higher the accuracy. Thus, choosing the basis set presents a challenge: we need to keep the basis set small enough so that the calculation is efficient,

while keeping it big enough to ensure an adequate fit to the *ab initio* data.
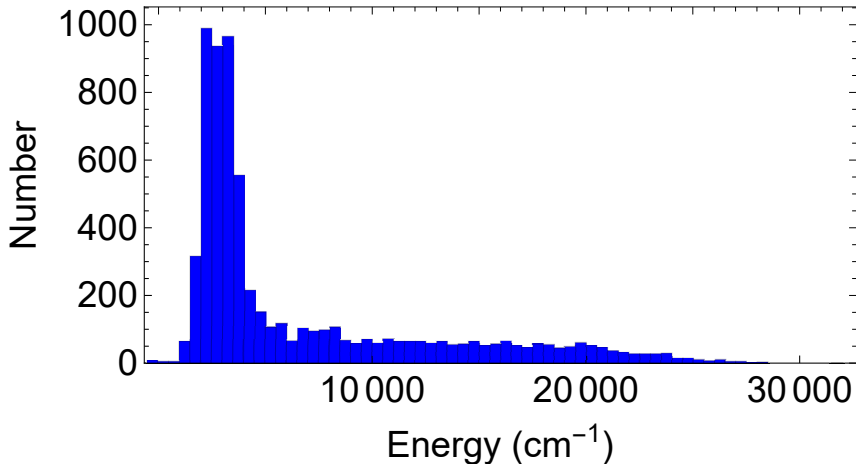

FIG. 1. Histogram of energies for geometries used as the data set for tropolone. The bin size for the abscissa is 500 cm^{-1}. Most energies are for configurations close to that of the global minimum. The data base has 6601 energies and 45×6601 gradient components.

Extensive use of fragmentation[25,26] and permutationally invariant polynomial basis sets allowed us to investigate several possible fits to the potential energy surface. First, software described and made available elsewhere[26] was used to investigate fits based on fragmentation. It allows one to specify various "fragments" of the parent molecule so as to reduce the size of the basis set of polynomials. The software ensured the elimination of duplicate polynomials, and it also provides for more rapid evaluation of gradient components, if desired. Second, we used software, outlined previously,[26] that allows one either to "prune" from or to add to an existing basis set in a manner that eliminates the least important existing polynomials or adds the most important new ones.

Both the fragmentation software and the pruning/adding software are based on the same idea; that is, to emphasize polynomials that involve Morse variables of high value. The Morse functions are transformed interatomic distances, $x(r_{\alpha,\beta}) = \exp(-r_{\alpha,\beta}/\lambda)$, where λ is commonly chosen to be equal to 2 bohr. Thus, the Morse variables range from 0 to 1. Because the interactions of atoms at large distance do not create much attractive or repulsive contribution to the potential, it makes sense not to include their Morse variables, whose values are nearly zero. The first step in determining which polynomials are most or least important is to evaluate all the Morse variables for all the geometries included in the

ab initio data set and to record the maximum value of each Morse variable. An example of this evaluation, for the 6768 data set, is given in Table SI-3. One then examines each of the polynomials of interest to determine its value when evaluated with the group of maximal Morse variable values. An ordered list of polynomial values is thus developed. The next step for pruning is to use the list in order to eliminate a number of polynomials, starting from that with the lowest value and moving upward in value until the desired number of polynomials has been eliminated. For adding polynomials, the current software looks at all pair-wise products of existing polynomials. If the product is a polynomial already in use, it skips to the next pair. If not, it evaluates the new polynomial using the maximal Morse values. Using an ordered list of existing polynomial values, the program examines pairs in a sequence such that the largest-valued possible products are examined first and then lesser-valued ones are evaluated subsequently. The program stops when the desired number of new polynomials has been identified. Further steps in the software involve deleting duplicates, renumbering the polynomials, adding derivatives, etc., but these are all steps that have already been developed and included in the fragmentation software.[26] Note that either for adding polynomials or for pruning polynomials, we maintain the permutational symmetry of the original basis set of polynomials. This preservation is due to the fact that each polynomial added or deleted has the correct permutational symmetry.

As described elsewhere,[26] it is convenient to establish an arbitrary but consistent numbering scheme for the atoms of the parent molecule. One numbering was used for a basis set of full permutational invariance and is shown in Fig. 2a, which depicts the minimum energy structure of tropolone. A second numbering was used in the fragmented basis sets and is shown in Fig. 2b. The equilibrium geometry is of planar symmetry and the primary molecular interest to chemical physics interest is the transfer of H atom at the top of each diagram between the two oxygen atoms over a symmetrical barrier.

Tables I-III summarize several of the fragmentation schemes that we have investigated for the quality of fit and the time for execution. The basis sets are denoted by the number of fragments and the atoms in each fragment: 1 fragment of 15 atoms in Table I, 2 fragments of 11 and 12 atoms in Table II, and 4 fragments of 9, 9, 8, and 5 atoms in Table III. In each basis we investigated fits to the full data set of 6601 geometries (whose histogram is shown in Fig. 1), as well as to subsets of this data set with 6249, 5717, and 2717 geometries, corresponding to upper energy cutoffs of 20000, 15000, and 15000 cm^{-1}, respectively, where the 2717 set
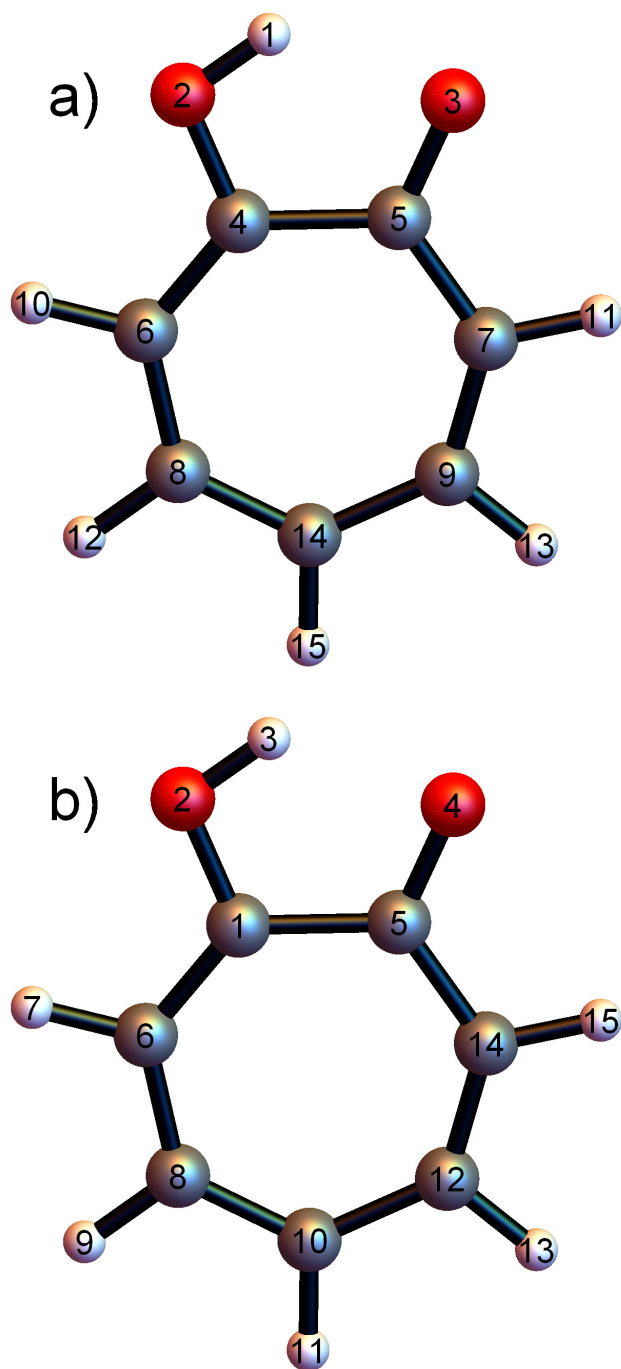
FIG. 2. Numbering schemes used in the text for tropolone. H, C, and O atoms are white, gray, and red, respectively. a) scheme used for full-symmetry calculations. b) scheme used for fragmented basis sets.

is based only on the 3601 geometries obtained from the *ab initio* molecular dynamics and 301 of the grid points. The quality of the fit may be judged from the RMSE values for the potential and gradient components as measured by the difference between the fit and the relevant data set, using inverse energy weighting (wRMSE(pot) and wRMSE(grad)), as described in Section SI-1. For each fit, the height of the barrier to H transfer with respect to the global minimum was determined; these values are shown in cm^{-1}. Finally, the times for evaluating 3000 potential energies and 45×3000 gradient components are shown; these are relative to a value of 100 for the 1-(15)/25103 basis. We next describe each basis and its results.

We start with the full-symmetry basis description of tropolone, for which the MSA software can provide a permutationally invariant basis set in 3rd order. We designate this basis as 1-(15), indicating that there is one fragment (the parent) of 15 atoms. All 105 Morse variables are included, and there are 11522 monomials that must be evaluated before the polynomials can be evaluated. The number of polynomials is the number of coefficients in the fit. The permutational symmetry is designated as {1,2,2,2,2,2,2,1,1}, the atom numbering is {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}, and the numbering scheme is that of Fig. 2a. Thus, the following pairs of atoms permute: {2,3}, {4,5}, {6,7}, {8,9}, {10,11}, and {12,13}. The full permutational symmetry ensures that as the H atom at position 1 moves between the two oxygen atoms, the barrier will be symmetrical. As can be seen from Table I, the use of a large basis set (25103 polynomials) and a complete set of 105 Morse variables gives an excellent fit to the full 6601 geometry data set, with weighted RMS errors for the potential and the gradient components of 25 cm^{-1} and 23 cm^{-1}/bohr, respectively. (We defer discussion of the 6768 data sets for the moment.) The pruning of the basis set to 14843 coefficients still provides a good fit, with weighted RMS values of 30 cm^{-1} and 29 cm^{-1}/bohr, and the speed is improved by more than a factor of 2. In general, for this 1-(15) example as well as for the others, the weighted RMSE values decrease as the number of points is decreased (i.e., they decrease as the energy cut-off is decreased). The calculated barrier heights increase as the energy cut-off decreases. For comparison, the barrier height obtained by comparing the optimized H transition state to the optimized GM at the B3LYP 6-31+G(d) level used for all data set point calculations is 2131 cm^{-1}. We note that the fitting errors are small compared to the intrinsic inaccuracy of the DFT level of theory and the basis set adopted, so that fit approximations to the barrier height that are within 50–100 cm^{-1} are not too

11

bad. At the CCSD(T)-F12/aVDZ level, the barrier is 2374 cm^{-1}, about 1 kcal/mol larger than that of the best of our fits.

TABLE I. Tropolone 3rd-order fitting results for 1-(15), full symmetry. Weighted RMS error values (wRMSE) are in cm^{-1} and cm^{-1}/bohr for potentials and gradient components, respectively. Times (see text) are relative to that for 1-(15) full sym/25103 gradient components, taken to be 100.

basis / ncoef	1-(15) full sym / 25103					1-(15) full sym / 14843					
N (points)	2717	5717	6249	6601	6768*	2717	5717	6249	6601	6768	6768*
wRMSE (pot)	16	21	24	25	10	27	26	29	30	38	16
wRMSE (grad)	14	18	21	23	8	23	24	27	29	34	12
Barrier	2064	2010	1983	1971	2083	2002	1974	1952	1940	1931	2055
Splitting	3.89	4.32	4.50	4.58	3.78	4.23	4.56	4.69	4.77	4.85	3.90
Time (pot / grad)	1.45 / 100.0					1.03 / 46.5					

* The weight of gradient components is reduced to 1/3 of the corresponding energy

We now turn to the fragmentation results. Designing a fragmentation scheme is partly art and partly science. While there is no requirement that the fragments be connected in the way they are in the parent molecule, most successful schemes put in one fragment mostly atoms that are likely to be close to one another; these have the largest values of the Morse variables. One can intentionally omit Morse variables involving pairs of distant atoms by ensuring that these atoms never appear in the same fragment. However, even if all the Morse variables were included, the number of monomials and polynomials would still be fewer for the fragments than for the parent molecule as a whole. This is because the fragmentation also determines the cross-terms between the Morse variables that appear in their products up to the polynomial order used in the fit. Only cross terms between Morse variables of pairs atoms that appear together in a fragment can appear. Fragments will often overlap, that is, they will have atoms in common. The Delete Duplicates program ensures that Morse variables, monomials, and polynomials based on the same pairs of atoms are not included more than once. For example, in the 2-(1112) fragmentation (described below), the tropolone is considered as having an upper "half" (atoms 1–8, 12, 14, and 15)

and a lower "half" (atoms 1, 5–15); they overlap since atoms 1, 5–8, 12, 14, and 15 appear in both fragments.

Two fragmentation schemes are included here, among many that were studied. The first, designated 2-(1112), has two fragments, one of 11 atoms and one of 12. The basis set uses 93 of the 105 Morse variables and there are 6643 monomials in addition to the 10271 polynomials. The permutation symmetries of the two fragments are {2,2,2,2,2,1} and {2,2,2,2,2,1,1}, with atom identification numbers or {1,5,2,4,6,14,7,15,8,12,3} and {1,5,6,14,7,15,8,12,9,13,10,11}, respectively. The atoms that permute, now using the numbering system of Fig. 2b, are the same ones that permute in the full symmetry case. Consequently, this fragmentation scheme likewise maintains the symmetry of the barrier to H transfer. A problem with this fragmentation scheme, however, is that in order to permute, atoms must be contained in the same fragment. For example, consider the distant atom pairs {7,15} and {9,13} in Fig. 2b. These will have small Morse variables, and from a fragmentation point of view, it would be better to have each member of the pair in a separate fragment so that we could eliminate these small-valued Morse variables. However, they would then not permute, and the transition barrier would no longer be symmetrical. The price we pay for the fragmentation is that the fit to the data sets is not as good. For the 6601 geometry data set, the RMSE values are 55 cm^{-1} for the potential and 134 cm^{-1}/bohr for the gradient components. We see from this example that the goals of fragmentation and those of permutational symmetry can often be at conflict with one another. An attempt to increase the accuracy by adding new polynomials to a total of 11001 was not particularly effective, as can be seen from the right three columns.

The second fragmentation scheme, designated 4-(9985), has four fragments of 9, 9, 8 and 5 atoms. The basis set uses 89 of the 105 Morse variables, and there are 317 monomials and 16138 polynomials. The fragment permutational symmetries are {2,1,1,1,1,1,1,1}, {2,1,1,1,1,1,1,1}, {1,1,1,1,1,1,1,1} and {1,1,1,1,1}, with atom identifications of {2,4,1,3,5,6,7,8,9}, {2,4,1,3,5,12,13,14,15}, {6,8,9,10,11,12,13,14}, and {1,5,7,10,15}, respectively. Only the two oxygen atoms (numbered 2 and 4 in Fig. 2b) are assumed to permute, which means that without further effort this scheme cannot guarantee a symmetrical transition barrier. We circumvent this problem by duplicating the basis set in such a way that whenever a geometry

TABLE II. Tropolone 3rd-order fitting results for 2-(1112) fragmentation. Weighted RMS error values (wRMSE) are in cm^{-1} and cm^{-1}/bohr for potentials and gradient components, respectively. Times (see text) are relative to that for 1-(15) full sym/25103 gradient components, taken as 100 (Table I).

basis / ncoef	2-(1112) / 10271				2-(1112) / 11001			
N (points)	2717	5717	6249	6601	2717	5717	6249	6601
wRMSE (pot)	55	50	54	55	-	49	53	54
wRMSE (grad)	110	129	133	134	-	129	133	133
Barrier	1903	1911	1886	1873	-	1916	1890	1877
Splitting	5.06	5.24	5.48	5.61	-	5.15	5.40	5.53
Time (pot / grad)	0.680 / 29.4				0.740 / 32.4			

appears, the same geometry with permuted atom identities also appears carrying the same energy and gradient components, as described in Section SI-2. This fragmentation scheme provides less accurate fits to the *ab initio* data sets.

TABLE III. Tropolone 3rd-order fitting results for 4-(9985) fragmentation. Weighted RMS error values (wRMSE) are in cm^{-1} and cm^{-1}/bohr for potentials and gradient components, respectively. Times (see text) are relative to that for 1-(15) full sym/25103 gradient components, taken as 100 (Table I).

basis / ncoef	4-(9985) / 16138			
N (points)	2717	5717	6249	6601
wRMSE (pot)	56	54	58	61
wRMSE (grad)	136	169	171	171
Barrier	1943	1929	1905	1889
Splitting	4.78	5.00	5.24	5.37
Time (pot / grad)	0.641 / 35.6			

The 4-(9985) scheme was also used to see if the fit to the 3601 geometry data set also

fit the "test" set (825 geometries). These test-set geometries were not included in the determination of the coefficients, so this comparison offers a good estimate on how the fit works for unknown geometries. The test gave evenly weighted RMSE values for the potential and gradient components of 74 and 118 cm^{-1}, whereas the fit to the 3601 geometry set gave 73 and 181 cm^{-1}. Thus, it is expected that the fit should be as good for unknown geometries as for the known ones used in the large data set.

IV. RESULTS AND DISCUSSION

A. Tropolone Stationary Points

For each of the tropolone fitting schemes listed in Tables I-III and using the 5717 data set, we determined the harmonic vibrational frequencies for the global minimum (GM); the transition state TS(H) to H atom transfer between the oxygen atoms; Conf I, in which the O-H bond is pointed away from rather than toward the remaining O; and TS(I), the transition state to Conf I from the GM. A comparison of each of these frequency sets to those for the optimized structures (at the B3LYP/6-31+G(d) level) gave a mean absolute error, MAE. These values (in cm^{-1}) are shown in Table IV. The 1-(15) full symmetry, and the 2-(1112) fitting schemes provide excellent agreement with the *ab initio* data, while the 4-(9985) scheme has very good, but somewhat less accurate, agreement.

Based on the B3LYP/6-31+G(d) optimization results for the stationary points, the energy landscape for tropolone is shown in Fig. 3. The barrier to hydrogen transfer between the two oxygen atoms is about 2131 cm^{-1}. The energy of conformer I lies at about 4523 cm^{-1}, and the barrier between that conformer and the ground state is about 6345 cm^{-1} above the latter, although this latter barrier may have significant (10-20 %) error.

Benchmark CCSD(T)-F12/aVDZ calculations were performed using Molpro.[56] The energy of the optimized GM was found to be -420.15609372 hartree, while that of the optimized transition state to H transfer was found to be -420.14527707. Thus, the height of the transition state above the GM is 2374 cm^{-1}. The gradient components for the GM and transition state were 10^{-7} or less. Geometries for the optimized GM and TS(H) are listed in Table SI-1, while harmonic frequencies are listed in Table SI-2.

TABLE IV. Mean absolute errors (in cm^{-1}) for the harmonic frequencies calculated from tropolone PES fits as compared to those for optimized structures, where GM is the global minimum, TS(H) is the transition state to H transfer, Conf I is the conformer with the H pointing away from the oxygen to which it is not bound, and TS(I) is the transition state to Conf I from the GM. In all cases the number of geometries used was 5717, corresponding to an energy cut-off of 15,000 cm^{-1}.

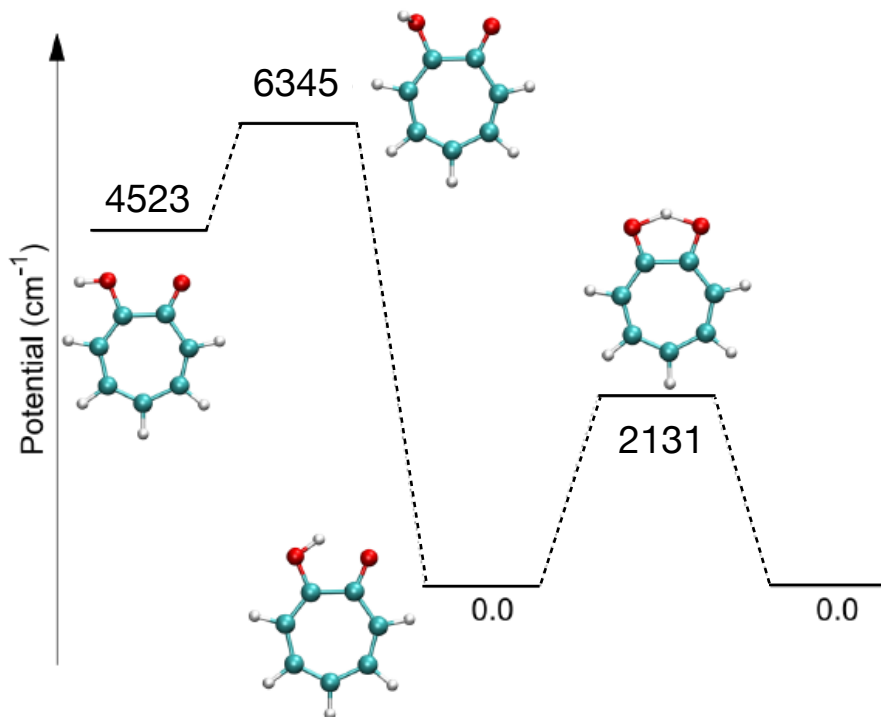| basis/structure	GM	TS(H)	Conf I	TS(I)
1-(15)/25103	1.7	8.4	22.4	30.4
1-(15)/14843	1.8	8.7	17.6	28.9
2-(1112)/10271	10.5	14.4	20.2	26.2
2-(1112)/11001	10.5	14.7	21.2	27.1
4-(9985)/16138	12.7	18.6	32.4	39.3

FIG. 3. Stationary states in tropolone, including the global minimum (0 cm^{-1}), conformer I (4523 cm^{-1}), the transition state to hydrogen transfer (2131 cm^{-1}) and the transition state to conformer I were determined from optimized structures at the B3LYP/6-31G(d) level.

16

B. Diffusion Monte Carlo Calculations

Diffusion Monte Carlo (DMC) simulation is an approach to compute quantum zero-point energy (ZPE) of a molecule.[57,58] In this work, the simple unbiased algorithm[59] was applied. In brief, we use an ensemble of random walkers to represent the nuclear wavefunction of the molecule. At each step, a random displacement is assigned to each walker, and a walker may remain alive or be killed by comparing its potential energy with a reference energy. The reference energy, $E_r(\tau)$, is updated after each step using the equation

$$E_r(\tau) = V(\tau) - \alpha \frac{N(\tau) - N(0)}{N(0)}, \tag{3}$$

where τ is the imaginary time; $V(\tau)$ is the average potential over all the walkers that are alive; $N(\tau)$ is the number of live walkers at time τ; α is a parameter that can control the fluctuations of the number of walkers and the reference energy. Finally, the average of the reference energy over the imaginary time gives an estimate of ZPE.

Five DMC simulations were performed, using the fit that is based on 1-(15)/14843 basis and 6768 geometry data set, and in each simulation, 20000 walkers were equilibrated for 10000 steps, and then were propagated for another 50000 steps to compute the ZPE, with step size of 5.0 au.

Our estimate of tropolone ZPE is 25085 ± 20 cm^{-1}. This is the average of the five simulations, and the uncertainty is the standard deviation of these five simulations. The harmonic ZPE using the same PES is 25315 cm^{-1}; therefore the anharmonicity is more than 200 cm^{-1}.

C. Tunneling Dynamics

As noted above, the ground vibrational state tunneling splitting has been measured accurately to be 0.96 cm^{-1}.[43,46] Given that this is quite small, say by comparison to the 21 cm^{-1} splitting in malonaldehyde, it would clearly be a challenge to calculate it starting with a fit to electronic energies and gradient components. We do note that that challenge was met in the case of malonaldehyde using a PIP PES that was a fit to highly accurate focal point energies.[34] The present PES is based on fitting to DFT electronic energies which as noted above underestimate the H-atom transfer barrier by almost 1 kcal/mol. Thus, even an "exact" calculation of the tunneling splitting would almost certainly be larger than the

experimental value just due to the error in the barrier height. Such calculations of small splittings are possible using the ring-polymer instanton method, as illustrated for the formic acid dimer, where the even smaller splitting of 0.013 cm^{-1} was reproduced with remarkable accuracy[60] using an accurate PIP PES for formic acid dimer developed by two of us.[61]

With the above caveat in mind we did calculate the ground-state tunneling splitting for several fits using a fast 1-d Q_{im}-path tunneling method.[35,62] In this approach the potential along the imaginary frequency normal mode of the H-atom TS is fully relaxed with respect to all other modes. In this way a rectilinear reaction path is established. This path is shown in Figure 4, and as seen the potential is a symmetric double-well. The details of this potential vary a bit depending on the fit. However, in all cases, numerically exact eigenvalues of this potential were obtained, and the ground state splitting is then the prediction from this theory. The results, shown in tables above, vary from around 3 to 5 cm^{-1}, depending on the fit. As expected these are larger than experiment, due at least to the error in the barrier height. Of course the 1-d method also has errors. For example, for malonaldehyde this method produced a tunneling splitting of 24 cm^{-1}, which is in error by around 3 cm^{-1} both from experiment and essentially exact calculations. For FAD the 1-d method produced a splitting of 0.44 cm^{-1} which again is larger than experiment and the result from the ring-polymer instanton calculation using the same PES. Thus it is likely that a calculation of the ground state splitting using that method would produce a result less than the few cm^{-1} reported here. With the new PESs available, hopefully, such a calculation can be done in the near future.

Next, we briefly address the interesting question of how the splitting varies with vibrational excitation. This has been examined experimentally for excitations up to around 1500 cm^{-1}.[46] Even with a realistic full dimensional PES, as developed here, the calculation of the effect of excitation of tropolone vibrational modes on the splitting is a major challenge. We developed a simple projection model to predict mode-specific effects in tunneling splittings and applied it successfully to malonaldehyde.[35] A similar model termed sudden vector projection (SVP) was developed independently by Guo and co-workers and aimed at predicting mode-specific effects in chemical reactions.[63] In these models normal modes of the molecule are projected onto the imaginary frequency normal mode of the transition state and those modes with large projections are predicted to have large positive effects to enhance tunneling in the former case and in reactivity in the latter cases. We performed this projection
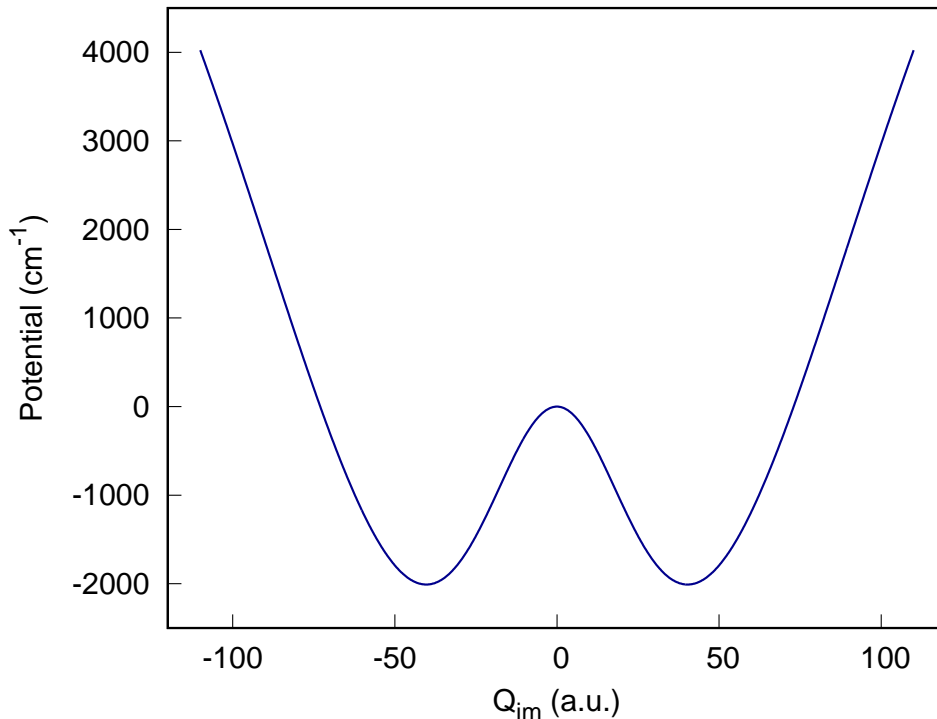
FIG. 4. The potential along the imaginary coordinate corresponding to hydrogen transfer between the two oxygen molecules. The potential is in cm^{-1}, while the imaginary coordinate is in mass-weighted atomic units.

analysis using the present PES for all the normal modes of tropolone. The results are given, along with the depiction and frequency of these modes, in the SI. Among these we single out modes 3 and 4. The projections of theses two modes are roughly 0.19 and 0.18, respectively whereas for other modes in this spectral range the projections are zero or near zero, except for mode 7 which has a projection of 0.18 magnitude. Modes 3, 4 and 7 in our ordering scheme (simply by increasing harmonic frequencies) correspond to modes 36, 35 and 33, respectively in the experimental scheme, which is in terms of decreasing experimental frequency.[46] The measured tunneling splittings for the fundamental of modes 36 and 35 are 5.1 and 1.3 cm^{-1}, respectively. The experimental splitting for mode 33 was not reported; however, by visual examination of normal modes it was predicted to enhance tunneling.

Normal modes 3 and 4 are depicted in Fig. 5; as seen they do involve O atom and the transferring H atom motions. Thus, pictorially at least, it is clear that these are H-transfer promoting modes.
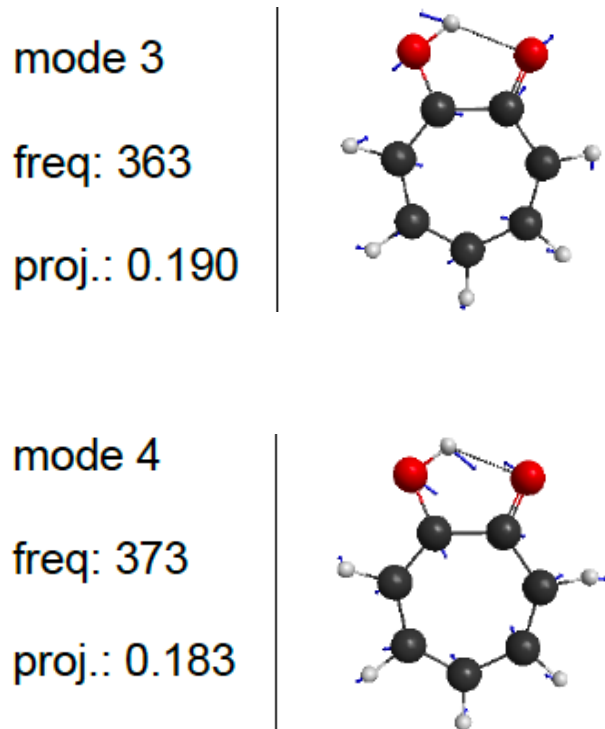
FIG. 5. Normal modes of tropolone predicted to increase the tunneling splitting

The Q_{im}-path model (unlike the SVP one) can make quantitative predictions of the effects of mode excitation and that was done successfully for malonaldehdye.[35] Doing so here is beyond the scope of this paper; however,we plan to do this in the future for the complete and extensive set of experimental data.[46]

D.   Recommended PES

We now return to Tables I-III to determine which of the fits might be best, with emphasis on the H-transfer dynamics. The fits using fragmentation are less accurate but only slightly faster than the pruned 1-(15) basis with 14843 polynomials. The most accurate surface, not surprisingly, is the full-symmetry surface with 25103 coefficients. The fit using the smallest data set (2717 geometries, 15000 cm^{-1} energy cut-off) gives very good wRMSE values and a close fit to the optimized barrier (at 2131 cm^{-1}), but the low cut-off energy and sparse number of geometries suggest that a fit to a larger data set might have fewer problems and

be of more general use. In examining the 6601 geometry data set with DMC, we noted that there were some "holes" in the surface that were fixed by adding potentials and gradient components at a modest number of new geometries; this new data set has 6768 geometries. Finally, the calculation of the barrier was closest to the optimized value when we underweighted the gradient components in the fit by a factor of three relative to the potential energies. Thus, the most accurate fit is the one with 25103 coefficients using the 6768* data set.

Unfortunately, this accurate surface is also the slowest to calculate, particularly for molecular dynamics simulations, where the calculation of the gradient components dominates the time. The corresponding 14843 coefficient version using the 6768* data set is twice as fast for this purpose with nearly the same accuracy with regard to the TS(H) barrier, the splitting, and the wRMSE values. The most accurate surface and the 6768 geometry data set are provided in the SI, and both surfaces as well as the data set are available from the authors.[64]

V. SUMMARY AND CONCLUSIONS

We reported permutationally invariant polynomial (PIP) fits to energies and gradient components for 15-atom tropolone. These include standard, pruned, augmented, and fragmented PIP bases. Approximately 6600 energies and associated gradient components are obtained from direct-dynamics calculations using DFT/B3LYP/6-31+G(d) supplemented by grid calculations spanning an energy range up to roughly 35000 cm^{-1}. Three fragmentation schemes were investigated with respect to efficiency and fit precision. Properties such as stationary points, harmonic frequencies and the barrier to H-atom transfer are reported and compared to direct calculations. A 1-d model to obtain the tunneling splitting for the ground vibrational state was applied to each fit and the splittings in the range of 3–5 cm^{-1} are in encouraging agreement with experiment (0.96 cm^{-1}). Based on new CCSD(T)-F12/aVDZ calculations, the barrier for H transfer for the DFT-based PESs is probably low by about 1 kcal/mol and this fact is at least partially responsible for the overestimate of the experimental splitting.

The fits used five basis sets and five data sets and employed inverse energy weighting and an adjustment of the weights between the *ab initio* potentials and gradient components. The most accurate fit was the 1-(15) full symmetry basis with 25103 coefficients, but it

was also the slowest. Pruning this basis to 14843 coefficients improved the speed by a bit more than a factor of two while maintaining most of the accuracy. Fits using the two fragmentation schemes, involving either two or four fragments, were not as accurate as the two full-symmetry fits, but slightly faster. The largest data set with 6601 geometries (later augmented to 6768) provides the most stable and accurate surface. It is clear that most of the data required for fitting the large number of coefficients is the result of the gradient components. However, this strong reliance on the gradients adversely affects the calculation of the energies, in particular the energy of the barrier to H-atom transfer. A significant improvement in the fit for this barrier was obtained by de-emphasizing the gradients in the fit by dividing their weight by a factor of three. Although we explored this option only for the most accurate two fits, it is likely to be effective in other systems where most of the *ab initio* data is given by the gradient components. The prospects for extending these methods to somewhat larger molecules are good, but because at least 10^4 coefficients will be needed, it is clear that gradient components as well as potential energies must be used in the fit. While these are readily available from DFT methods, they are much more costly for CCSD(T) calculations. One 15-atom molecule, acetyl acetone, is being studied at this level by our group and that of Meuwly,[31] with promising results.

VI. ACKNOWLEDGEMENTS

VII. SUPPORTING INFORMATION

The Supporting Information contains information on:

I. Weighting of *ab initio* data

II. Duplication of *ab initio* data: Fig. SI-1

III. CCSD(T)-F12/AVDZ results: Tables SI-1 and SI-2

IV. Morse variable values for the 6768 data set: Table SI-3

V. Normal modes and projections of these onto the imaginary frequency normal mode: Figs. SI-2 and SI-3

In addition, the SI includes a zip file containing the fitting program using the 6768*
data set and a program for calculating potentials and gradient components using the full-
symmetry/25103 basis set.

The data that support the findings of this study are available from the corresponding
author upon reasonable request.

REFERENCES

[1] B. J. Braams and J. M. Bowman, Int. Rev. Phys. Chem. **28**, 577 (2009).

[2] C. M. Handley and P. L. A. Popelier, J. Phys. Chem. A **114**, 3371 (2010).

[3] J. Behler, Phys. Chem. Chem. Phys. **13**, 17930 (2011).

[4] J. Behler, J. Chem. Phys. **134**, 074106 (2011).

[5] M. Sergei, D. Richard, and T. Carrington, Int. J. Quantum Chem. **115**, 1012 (2014).

[6] J. Behler, Int. J. Quantum Chem. **115**, 1032 (2015).

[7] A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015).

[8] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[9] E. Uteva, R. S. Graham, R. D. Wilkinson, and R. J. Wheatley, J. Chem. Phys. **147**, 161706 (2017).

[10] B. Jiang and H. Guo, J. Chem. Phys. **139**, 054112 (2013).

[11] K. Shao, J. Chen, Z. Zhao, and D. H. Zhang, J. Chem. Phys. **145**, 071101 (2016).

[12] B. Jiang, J. Li, and H. Guo, Int. Rev. Phys. Chem. **35**, 479 (2016).

[13] B. Fu and D. H. Zhang, J. Chem. Theory . **14**, 2289 (2018).

[14] C. Qu, Q. Yu, and J. M. Bowman, Annu. Rev. Phys. Chem. **69**, 6.1 (2018).

[15] A. Kamath, R. A. Vargas-Hernández, R. Krems, T. Carrington, and S. Manzhos, J. Chem. Phys. **148**, 241702 (2018).

[16] D. I. Doughan, L. M. Raff, M. G. Rockley, M. Hagan, P. M. Agrawal, and R. Komanduri, J. Chem. Phys. **124**, 054321 (2006).

[17] S. Manzhos and T. Carrington, J. Chem. Phys. **125**, 084109 (2006).

[18] A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, R. Narulkar, S. Bukkapatnum, and R. Komanduri, J. Chem. Phys. **130**, 134101 (2009).

[19] A. Nongnuch and J. Behler, Phys. Rev. B **85**, 045439 (2012).

[20] J. Li, B. Jiang, and H. Guo, J. Chem. Phys. **139**, 204103 (2013).

[21] Y. Guan, S. Yang, and D. H. Zhang, J. Phys. Chem. A **122**, 3140 (2018).

[22] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[23] C. Qu, Q. Yu, B. L. Van Hoozen, J. M. Bowman, and R. A. Vargas-Hernández, J. Chem. Theory Comput. **14**, 3381 (2018).

[24] C. Qu and J. M. Bowman, J. Chem. Phys. **150**, 141101 (2019).

[25] A. Nandi, C. Qu, and J. M. Bowman, J. Chem. Phys. **151**, 084306 (2019).

[26] R. Conte, C. Qu, P. L. Houston, and J. M. Bowman, J. Chem. Theory Comput. **16**, 3264 (2020).

[27] M. A. Collins, Theor. Chem. Acc. **108**, 313 (2002).

[28] P. O. Dral, A. Owens, S. N. Yurchenko, and W. Thiel, J. Chem. Phys. **146**, 244108 (2017).

[29] J. Li and H. Guo, J. Chem. Phys. **143**, 221103 (2015).

[30] H. Sugisawa, T. Ida, and R. V. Krems, "Interpolation and extrapolation in a 51-dimensional variable space: system-agnostic construction of high-dimensional pes," (2020), arXiv:2001.07271 [physics.comp-ph].

[31] S. Käser, O. Unke, and M. Meuwly, New Journal of Physics **xx**, xxxx (2020).

[32] O. T. Unke and M. Meuwly, J. Chem. Theory Comput. **15**, 3678 (2019).

[33] "Msa software with gradients," `https://github.com/szquchen/MSA-2.0` (2019), accessed: 2019-01-20.

[34] Y. Wang, B. Braams, J. M. Bowman, S. Carter, and D. P. Tew, J. Chem. Phys. **128**, 224314 (2008).

[35] Y. Wang and J. M. Bowman, J. Chem. Phys. **139**, 154303 (2013).

[36] J. T. Hynes, J. P. Klilnman, H.-H. Limbach, and R. L. Schowen, *Hydrogen-Transfer Reactions* (Wiley-VCH Verlag GmbH & Co., Weinheim, Germany, 2007).

[37] R. L. Redington, J. Chem. Phys. **113**, 2319 (2000).

[38] R. L. Redington, T. E. Redington, and J. M. Montgomery, J. Chem. Phys. **113**, 2304 (2000).

[39] R. L. Redington and R. L. Sams, Chem. Phys. **283**, 135 (2002).

[40] R. L. Redington and R. L. Sams, J. Phys. Chem. A **106**, 7494 (2002).

[41] R. L. Redington, T. E. Redington, T. A. Blake, R. L. Sams, and T. J. Johnson, J. Chem. Phys. **122**, 224311/1 (2005).

[42] R. L. Redington, T. E. Redington, and R. L. Sams, J. Phys. Chem. A **110**, 9633 (2006).

[43] R. L. Redington, T. E. Redington, and R. L. Sams, Z. Phys. Chem. **222**, 1197 (2008).

[44] R. L. Redington, T. E. Redington, and R. L. Sams, J. Phys. Chem. A **112**, 1480 (2008).

[45] R. K. Frost, F. C. Hagemeister, C. A. Arrington, T. S. Zwier, and K. D. Jordan, J. Chem. Phys. **105**, 2595 (1996).

[46] D. Murdock, L. A. Burns, and P. H. Vaccaro, Phys. Chem. Chem. Phys. **12**, 8285 (2010).

[47] M. V. Vener, S. Scheiner, and N. D. Sokolov, J. Chem. Phys. **101**, 9755 (1994).

[48] Y. Guo, T. D. Sewell, and D. L. Thompson, J. Phys. Chem. A **102**, 5040 (1998).

[49] K. Giese and O. Kühn, J. Chem. Phys. **123**, 054315 (2005).

[50] K. Giese, M. Petkovic, H. Naundorf, and O. Kühn, Phys. Rep. **430**, 211 (2006).

[51] Z. Xie and J. M. Bowman, J. Chem. Theory Comput. **6**, 26 (2010).

[52] R. Conte, P. L. Houston, and J. M. Bowman, J. Chem. Phys. **140**, 151101 (2014).

[53] R. Conte, C. Qu, and J. M. Bowman, J. Chem. Theory Comput. **11**, 1631 (2015).

[54] C. Qu, R. Conte, P. L. Houston, and J. M. Bowman, Phys. Chem. Chem. Phys. **17**, 8172 (2015).

[55] H. J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Scchutz, WIREs Comput. Mol. Sci. **2**, 242 (2012).

[56] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, *et al.*, https://www.molpro.net (2019).

[57] J. B. Anderson, J. Chem. Phys. **63**, 1499 (1975).

[58] J. B. Anderson, J. Chem. Phys. **65**, 4121 (1976).

[59] I. Kosztin, B. Faber, and K. Schulten, Am. J. Phys. **64**, 633 (1996).

[60] J. O. Richardson, Phys. Chem. Chem. Phys. **19**, 966 (2017).

[61] C. Qu and J. M. Bowman, Phys. Chem. Chem. Phys. **18**, 24835 (2016).

[62] Y. Wang and J. M. Bowman, J. Chem. Phys. **129**, 121103 (2008).

[63] H. Guo and B. Jiang, Accnts of Chem, Res. **47**, 3679 (2014).

[64] https://scholarblogs.emory.edu/bowman/potential-energy-surfaces/.