July 28, 2021

Permutationally invariant polynomial expansions with unrestricted complexity

Daniel R. Moberg and Ahren W. Jasper*

Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, IL 60439

Abstract. A general strategy is presented for constructing and validating permutationally invariant polynomial (PIP) expansions for chemical systems of any stoichiometry. Demonstrations are made for three categories of gas phase dynamics and kinetics: collisional energy transfer trajectories for predicting pressure dependent kinetics, three-body collisions for describing transient van der Waals adducts relevant to atmospheric chemistry, and nonthermal reactivity via quasiclassical trajectories. In total, 30 systems are considered with up to 15 atoms and 39 degrees of freedom. Permutational invariance is enforced in PIP expansions with as many as 13 million terms and 13 permutationally-distinct atom types by taking advantage of petascale computational resources. The quality of the PIP expansions is demonstrated through the systematic convergence of in-sample and out-of-sample errors with respect to both the number of training data and the order of the expansion, and these errors are shown to predict errors in the dynamics for both reactive and nonreactive applications. The parallelized code distributed as part of this work enables the automation of PIP generation for complex systems with multiple channels and flexible user-defined symmetry constraints and for automatically removing unphysical unconnected terms from the basis set expansions, all of which are required for simulating complex reactive systems.

*E-mail: ajasper@anl.gov

1. INTRODUCTION

Analytic potential energy surfaces (PESs) that describe the dependence of the electronic energy on the nuclear geometry are prerequisites for efficient dynamics and kinetics calculations. Early strategies for PES construction featured physically motivated functional forms,[1–6] in part to limit the number of adjustable parameters and to associate them with molecular properties. Popular modern approaches, in contrast, employ mathematical expansions with large numbers of uninterpretable parameters, such as permutationally invariant polynomials (PIPs),[7] neural networks,[8–10] Gaussian approximation potentials,[11–14] and interpolated moving least squares,[15,16] to name a few. Regardless of the number of parameters in these expansions, it is often possible to calculate many more electronic energies, and so, since the influential work of Braams and Bowman who demonstrated the enforcement of permutational invariance,[7] a great number of PES expansions for molecular and chemical systems have been produced. For recent representative work from several groups, see Refs. 17–29.

Because of the large numbers of parameters and the often ad hoc strategies used in their construction, it can be difficult to assess the accuracy of PES expansions for high-dimensional systems and for systems with multiple reaction channels, even when dynamics applications appear successful. Fitting errors are almost always reported to be small, but small fitting errors alone can be artifacts of incomplete or insufficient training data and so can be complemented by other measures of accuracy.[21,27–31] For example, we recently demonstrated the use of an out-of-sample error for characterizing the quality of training data, including its completeness.[32] This analysis further determines the number of training data required for producing a PES of a desired accuracy, and this number was shown to scale in predictable ways with the training data sampling strategy, the size and complexity of the expansion basis, and the dimensionality of the system.

Another practical challenge complicating PES expansion construction is the identification and elimination of unphysical artifacts, including so-called "unconnected terms" and "holes." Unconnected terms, described previously[22,33] and removed in some applications,[34–37,42] are terms in the expansion that are functions of the internal coordinates of two or more separated fragments but are not functions of any relative coordinates connecting the fragments. If unconnected terms are not removed, motion in these coordinates leads to unphysical energy exchange between fragments even when they are widely separated. Holes,[37] as recently discussed,[38] are regions where the expansion energy decreases to large negative values. Holes are a dramatic example of a more general of type of defect that arises whenever an important region lacks sufficient training data resulting in unconstrained extrapolations from neighboring regions.

The main goals of the present study are to describe and demonstrate general strategies for preparing permutationally-invariant polynomial (PIP) expansions for systems with unrestricted complexity, for validating the quality of the PIP expansions including the completeness of the training data, and for automatically identifying and removing unphysical artifacts. Several versatile codes for PIP generation are currently distributed by other groups (for example[22,23,39–44]). These implementations sometimes include restrictions on allowed stoichiometries, numbers of atom types, or the number of permutational operations, or are limited to non-reactive systems. The implementation described here is general, and the associated software distributed as part of this work[45] is parallelized to take advantage of petascale resources.

This paper is organized as follows. In Sec. 2, we review terminology for characterizing PIP expansions, describe general algorithms for PIP construction, summarize our parameterization strategy including tests of training data completeness, and introduce three categories of dynamics applications: two-body collisional energy transfer trajectories for predicting pressure dependent

elementary kinetics, three-body collisions for characterizing transient van der Waals adducts relevant to atmospheric chemistry, and nonthermal quasiclassical trajectories for calculating the reactivity of energized reactants in plasmas and flames. In Sec. 3.1, collisional energy transfer results for ten HO_2(+M) and H_2O_2(+M) systems are presented. These systems are used to demonstrate methods for verifying convergence with respect to both the training data and the completeness of the PIP expansion, the relationship between the intrinsic fitting error of the PIP expansion with dynamical errors, and the overall accuracy of the converged PIP expansions via comparisons with experimental kinetics. In Sec. 3.2, eleven NH_3(+M) and N_2H_4(+M) systems are used to discuss the performance of our scheme for enforcing permutational invariance as well as to characterize errors associated with unphysical unconnected terms in bimolecular collisions. In Sec. 3.3, six two- and three-body van der Waals complexes involving the Criegee intermediate CH_2OO are considered, and the usefulness of artificial permutational symmetry is demonstrated. In Sec. 3.4, we consider three reactive cases, H + H_2O, H + CH_4, and H + C_2H_6, where we characterize the performance of the PIP expansions for nonthermal and thermal applications. The reactive cases again demonstrate the usefulness of the intrinsic fitting error as a measure of the quality of the PIP expansion. Section 4 is a summary.

2. THEORY

2.1. Background and notation. The potential energy surface V for a molecular system may be expanded as

$$V = \sum_{k=1}^{N_t} c_{p^{(k)}} \prod_i y_i^{\xi_i^{(k)}}, \tag{1}$$

where each term k is a product of the Morse variables,[46]

$$y_i = \exp(-R_i/1\ \text{Å}), \tag{2}$$

i labels the atom–atom distances $\mathbf{R} = \{R_i\}$, and each of the N_t terms in the expansion is distinguished by a unique set of exponents $\boldsymbol{\xi}^{(k)} = \{\xi_i^{(k)}\}$. Permutational invariance is enforced by collecting permutationally-equivalent terms into groups labeled $p^{(k)}$, which share expansion coefficients $c_{p^{(k)}}$. The number of adjustable parameters in eq 1, N_p, is equal to the number of permutationally-unique groups and is often much smaller than the number of terms N_t, as emphasized by Braams and Bowman.[7]

A common strategy is to generate terms in eq 1 according to two constraints on the allowed values of $\boldsymbol{\xi}^{(k)}$: ξ_f, which sets the maximum value of the exponent for each factor $\xi_i^{(k)}$, and ξ_o, which sets the total order of the fit, i.e., the maximum value of $\sum_i \xi_i^{(k)}$. Expansions generated subject to these constraints are sometimes labeled PIP$\xi_f\xi_o$. Here we only consider cases where $\xi_f = \xi_o$ and label the expansions PIPξ_o, although the algorithm detailed below will work for cases where they are not equal. We will comment on the inefficiency of this strategy below. Improved basis set selection strategies are under development.[47]

2.2. Permutationally invariant polynomial expansions with unrestricted complexity. A general strategy for constructing permutationally invariant PES expansions is described next. First, the full list of terms (or, equivalently, their exponents $\boldsymbol{\xi}^{(k)}$) is generated subject to the user-specified constraints ξ_f and ξ_o as follows:

1. The set of exponents $\xi_i = 0$ is initialized, where i labels bond distance pairs and runs from 1 to N_{pair}. This initial set of exponents represents the constant term in the expansion, and it is retained as $\boldsymbol{\xi}^{(1)}$, i.e., as the first term in the expansion.
2. A running index j is set to N_{pair}.
3. The exponent ξ_j is increased by one to generate a new candidate term for the expansion.
4. If $\xi_j > \xi_f$ or if $\Sigma_{i=1,j}\xi_i > \xi_o$, then the candidate term is not valid. To proceed, we check the value of j.
    a. If $j = 1$, the term generation procedure is ended, as no more valid terms are possible.

b. If $j > 1$, ξ_j is reset to 0 and ξ_{j-1} is increased by one, while the running index j is decreased by one. The term generation procedure is resumed at Step 4.

5. If we have found a valid term, the index k that counts terms is increased by one, and the newly discovered valid term is retained as $\boldsymbol{\xi}^{(k)}$.

6. The procedure is repeated starting from Step 2.

Permutational invariance is enforced according to group labels assigned by the user to each atom. Atom group labels may be chosen to reflect the true permutational symmetry of the system, or they can be chosen to artificially enforce extra symmetry to reduce the number of adjustable parameters. The utility of flexible symmetry assignment is demonstrated in Sec. 3.3. The total number of symmetry permutations, $N_!$, is equal to the product of the factorials of the number of atoms assigned to each group, and $N_!$ can be quite large.

Atom group assignments are converted to a list of $N_!$ atom–atom pair permutations, and the full set of permutation operations is denoted $\boldsymbol{\mathcal{P}}$. Terms are collected into permutation groups $p^{(k)}$ in the order they were generated above. Specifically, to assign the k^{th} term to the appropriate group, its indices $\boldsymbol{\xi}^{(k)}$ are permuted up to $N_!$ times according to $\boldsymbol{\mathcal{P}}$, and each permuted set of indices is compared with $\boldsymbol{\xi}^{(k')}$ for all $k' < k$ working backwards from $k' = k - 1$. If a match is found, i.e., if $\boldsymbol{\mathcal{P}}\boldsymbol{\xi}^{(k)} = \boldsymbol{\xi}^{(k')}$, $p^{(k)}$ is assigned to the permutation group of the matching term, $p^{(k)} = p^{(k')}$. If no match is found, the k^{th} term is assigned to a newly created permutation group. Although we did not implement their method, we note that Xie and Bowman previously described an efficient scheme for assigning permutation groups.[41]

2.3. Removing unphysical unconnected terms. Next, we discuss the removal of unphysical unconnected terms. We first consider a case with one bimolecular fragment channel. Generalizing from the description of Truhlar and coworkers,[33] we label a term $\boldsymbol{\xi}^{(k)}$ "unconnected" when it features nonzero exponents for intramolecular distances appearing in both bimolecular fragments

but no nonzero exponents for intermolecular distances connecting them. Using this definition, the following algorithm is readily implemented:

1. Each atom a is assigned by the user to the fragment group f_a. For a bimolecular channel, for example, f_a equals either 1 or 2.
2. The index j labels the atom pair a_j and a'_j where the atoms are ordered such that $f_{a_j} \leq f_{a'_j}$.
3. For each term in the expansion k, we calculate

$$u_{ff'}^{(k)} = \sum_j \xi_j^{(k)} \, \delta(f - f_{a_j}) \, \delta(f' - f_{a'_j}), \tag{3}$$

   where $u_{ff}^{(k)}$ for $f' = f$ is the total order of the factors describing fragment f in term k, while $u_{ff'}^{(k)}$ for $f \neq f'$ is the total order of the factors connecting fragments f and f' in term k.
4. If $u_{ff}^{(k)} > 0$ for more than one value of f, we check for a nonzero cross term $u_{ff'}^{(k)}$.
   a. If $u_{ff'}^{(k)} = 0$ the k^{th} term is an unconnected term and is thrown out.
   b. If $u_{ff'}^{(k)} > 0$ the k^{th} term is a connected term and is retained.

As written, the algorithm could be applied to a channel with three or more fragments, but additional checks at step 4 would be required. For example, a term may feature nonzero exponents for just one fragment $u_{ff}^{(k)}$, nonzero exponents for cross terms connecting two *other* fragments $u_{f'f''}^{(k)}$, and no other nonzero exponents. This term is an unconnected term as it allows for energy transfer from modes associated with fragments f' and f'' into and out of fragment f even if fragment f is infinitely separated from the other fragments. One could modify the above approach to identify and remove such terms, but we have not done so. Instead, we allow for the specification of multiple bimolecular fragment channels, and terms are removed when they are unconnected in *any* channel. This strategy allows for the treatment of a termolecular channel, for example, by checking for unconnected terms appearing in *any* of the possible permutations of bimolecular channel assignments (e.g., unconnected terms for a termolecular case 1 + 2 + 3 can be removed by

considering the three bimolecular cases 1 + 23, 2 + 13, and 3 + 12). This procedure removes the type of termolecular unconnected terms mentioned above.

2.4. Applications. *Collisional energy transfer for pressure-dependent unimolecular kinetics. A priori* calculations of collisional energy transfer have been shown to enable quantitative fully theoretical treatments of pressure-dependent kinetics[48] and molecular transport.[34] For these applications, it is convenient to generate PIP expansions describing the two-body A(+M) *intermolecular* potential energy surface V_I, where

$$V_I(\mathbf{R}) = V(\mathbf{R}) - V_A(\mathbf{R}_A) - V_M(\mathbf{R}_M). \quad (4)$$

The required total potential energy surface V is then obtained by adding (typically) molecular mechanics force fields describing V_A and V_M to V_I, as described previously.[32] Note that the dimensionality of V_I is formally the same as that of the full system as it depends on the internal coordinates of both the unimolecular reactant A ($\mathbf{R}_A$) and the bath gas M ($\mathbf{R}_M$) as well as the intermolecular coordinates $\mathbf{R}_I$, but only inasmuch as they couple with intermolecular modes.

PIP expansions were generated here for the following 21 two-body systems: HO_2(+M), M = He, Ar, N_2, H_2O, and CO_2; H_2O_2(+M), M = Ar, N_2, CO_2, H_2O, and H_2O_2; NH_3(+M), M = Ar, O_2, N_2, CH_4, C_2H_6, CO_2, and NH_3; and N_2H_4(+M), M = Ar, O_2, N_2, and NH_3. This varied test set includes species with internal-coordinate dimensionalities that vary from 6 to 30 and features systems that range from nearly isotropic (e.g., M = Ar) to strongly anisotropic (e.g., HO_2 ($+H_2O$)). These specific systems were chosen to allow for comparisons with the relatively few available experimental collision efficiencies, as detailed in Secs. 3.1. and 3.2.

Based on the definition in eq 4, terms that describe a single fragment (i.e., that are functions of only $\mathbf{R}_A$ or only $\mathbf{R}_M$) should not be included in these applications, and they were therefore removed. We sometimes discuss PIP expansions of V_I where intramolecular-only terms are not

removed to demonstrate their effect on collisional energy transfer, and these expansions are indicated with "+i," e.g., PIP3+i. Similarly, to demonstrate the effect of unconnected terms on collisional energy transfer dynamics, we sometimes produce unphysical PIP expansions where unconnected terms have not been removed. These expansions are indicated by "+u," e.g., PIP3+u.

It can be a useful approximation to neglect the dependence of $V_I(\mathbf{R})$ on $\mathbf{R}_A$ and $\mathbf{R}_M$ when parametrizing V_I.[32] Expansions of this kind may be called "separable" expansions and are labeled s-PIPξ_o. The separable approximation, when suitable, greatly simplifies PES construction, including enabling so-called transferable PES parameterizations that allow for the treatment of large numbers of similar systems from a single PES parameterization.[49] Here we quantify the error associated with the separable approximation by comparing results obtained using s-PIP expansions with those obtained using full-dimensional PIP expansions. We note that an alternative and common method for dealing with the non-separability of PIPs is switching functions which allow for a smooth transition between fragment regions of the PES.[25,50–52]

Collisional energy transfer for the A(+M) systems was computed as follows and as described in detail elsewhere.[49] Briefly, orientationally-averaged rigid body optimizations[53] using the PIP expansions of the A(+M) interaction energies V_I were used to compute effective Lennard–Jones parameters[54] σ and ε from which the total A(+M) collision rate Z was calculated. Full-dimensional classical trajectories,[55–61] which are also quite sensitive to the accuracy of V_I, were then used to compute the collision efficiency range parameter $\alpha = <\Delta E_d>$, which by convention is set equal to the average energy in deactivating collisions,[62] along with several other low-order moments describing collisional energy and angular momentum transfer. As appropriate for describing unimolecular kinetics, the reactant A was initially prepared with an internal energy close to its dissociation energy, while the bath gas and relative coordinates were prepared at the

temperature of interest. Notably, the PIP expansions produced for these applications must therefore be able to accurately describe bath interactions with the unimolecular reactants A at high internal energies.

The experimental observable of interest here is the relative collision efficiency $\boldsymbol{\mathcal{M}}_{\mathrm{M}}$, which may be formally defined as the ratio of the low-pressure-limit rate constant k_0 for the decomposition of A in some bath M relative to k_0 for a reference bath (here we use either Ar or N_2 as the reference bath). $\boldsymbol{\mathcal{M}}_{\mathrm{M}}$ is readily computed from σ, ε, and α using Troe's β_c model,[62,63] which has been shown to predict $\boldsymbol{\mathcal{M}}_{\mathrm{M}}$ in close agreement[64,65] with the results of detailed *a priori* two-dimensional master equation calculations for weak colliders ($\beta_c << 1$).[48] For the strongest colliders considered here $\beta_c \approx 1$, and when this is the case the results of the β_c model depend sensitively on the necessarily arbitrary choice of the collision rate Z. For some strong colliders, we therefore computed collision efficiencies by solving a two-dimensional master equation[48,66] (2DME) for k_0. The 2DME is a detailed *a priori* approach that requires more trajectory-based collisional information than the β_c method (specifically, more low-order moments of the collisional energy and angular momentum transfer function), and it does not depend in an arbitrary way on the choice of Z.

Three-body van der Waals complexes in the atmosphere. The next set of applications involves the low-temperature dynamics of transient van der Waals complexes of the Criegee intermediate, CH_2OO. Namely, we considered the nonreactive interaction of CH_2OO with two hydroperoxides: hydrogen peroxide (H_2O_2) and hydroperoxymethylformate ($HOOCH_2OCHO$ or HPMF). The lifetimes and reactivities of the transient adduct complexes $CH_2OO\cdots H_2O_2$ and $CH_2OO\cdots HOOCH_2OCHO$ are of importance to the modeling of aerosol formation in the

atmosphere; when these adducts are sufficiently long-lived, they can alter product branching in CH_2OO insertion reactions and thus alter the rate of molecular growth of aerosols.[67]

For simplicity, we again consider collisional energy transfer and characterize $\alpha = <\Delta E_d>$ for Criegee complexes, noting that increased collisional stabilization leads to longer lived van der Waals complexes. In contrast to the two-body applications described above, however, here we demonstrate the use of PIP expansions in three-body collisions describing the energy transfer dynamics of the van der Waals complexes interacting with the baths M = He and H_2O.

Training data for the nonreactive systems. Training data sets for the two- and three-body collisional energy transfer applications were generated using A(+M) sampling strategies described previously.[32] Briefly, the relative orientations of A and M were sampled pseudorandomly and uniformly, while the center of mass separation distance was sampled pseudorandomly from a biased distribution out to 20 Å. We computed two batches of training energies for each system: one with the two fragments held fixed at their equilibrium geometries, and one with fragment geometries sampled uniformly in time from relevant isolated single-fragment microcanonical or thermal trajectory ensembles. Each batch typically consisted of 36,000 sampled energies for a total training data set size of 72,000. For some systems we computed larger training sets, as noted throughout Sec. 3. For the three-body A+B(+M) applications, this approach was applied twice: first for A with respect to B to generate a distribution of A···B geometries, and then for A···B with respect to M to generate the training set.

Energies at the sampled geometries were calculated using counterpoise corrected density fitted MP2 extrapolated to the complete basis set limit (cc df-MP2/CBS(cc-pVDZ,cc-pVTZ)). The error in this level of theory is typically small,[32] but higher-level methods may be merited for some

systems. As we are focused here on details of the fitting procedure more than the accuracy of the training data, we did not pursue higher-level calculations.

Reactive nonthermal dynamics. In a final set of applications, we considered full-dimensional reactive PIP expansions for use in high-energy quasiclassical trajectory calculations for the bimolecular reactions $H_2O^* + H \rightleftharpoons OH + H_2$, $CH_4^* + H \rightleftharpoons CH_3 + H_2$, and $C_2H_6^* + H \rightleftharpoons C_2H_5 + H_2$ with the molecular reactant initially excited close to its dissociation threshold. Reactions of this type can be very fast, with rate constants that approach collision rate constants.[68] Nonthermal reactions can impact global combustion properties, where radical–radical chemistry produces high energy molecules like H_2O^*, CH_4^*, and $C_2H_6^*$, and, before they can be collisionally cooled, they rapidly react with radicals like H atoms to alter the available radical pool.[69]

The nonthermal kinetics applications involve a straightforward implementation of quasiclassical trajectories,[70,71] and their main challenge is the generation of suitable PIP expansions for global trajectories that are accurate at high internal energies. The $H + H_2O$ and $H + CH_4$ reactions are in fact both frequently studied as prototypical chemical systems,[72–82] but the accuracy of the many parameterized PESs that have been developed is not generally known for high-energy (>110 kcal/mol) applications. Nonthermal rate constants are labeled $k^*(E_{vib},T)$ and are sensitive functions of both the initial vibrational energy of the molecular reactant, E_{vib}, and local temperature, T, while thermal rate constants, $k(T)$, are functions only of T.

Training data for the reactive systems. Training and testing data sets for the two smaller reactive cases were calculated using multireference perturbation theory (CASPT2), as implemented in Molpro,[83] and the aug-cc-pVTZ basis set. For $H + CH_4$, a full valence active space was used. For $H + H_2O$, an active space consisting of the full valence space plus the O(1s) was used, along with dynamic weighting for the two lowest electronic states.[84] Several levels of theory

for H + H_2O were recently tested against high level benchmarks,[69] and the level of theory chosen here provides a useful compromise of accuracy and computational cost. Briefly, the classical (zero-point-inclusive) barrier heights at these levels of theory are 21.9 (20.4 kcal/mol) and 14.9 (13.4 kcal/mol) for H + H_2O and H + CH_4, respectively, which may be compared with benchmark values of 21.5 (20.1 kcal/mol)[69] and 15.1 (13.6 kcal/mol).[85] For simple abstractions like the ones considered here, multireference approaches have the benefit of providing a robust strategy for handling geometries with high energies, even those with energies far too high to be accessed during the trajectories but that might be sampled. This reduces the amount of human intervention required to identify and remove unconverged or otherwise erroneous energies that might appear in the data sets.

H + C_2H_6 energies were calculated using CCSD(T)/aug'-cc-pVTZ, where the prime denotes that the diffuse basis functions were removed for the H atoms. This level of theory predicts an electronic barrier height, electronic reaction energy, and 0 K reaction enthalpy of 12.3, –0.59, and –3.91 kcal/mol, respectively, in good agreement with recent larger basis set calculations[86] (12.0, –0.61, and 3.58 kcal/mol).

To generate the training data for the reactive systems, we first applied the nonreactive two-body sampling strategy described above for both the reactant and product channels. These data sets were augmented with 72,000 additional geometries sampled from a harmonic expansion about the saddle point for a total of 216,000 training set energies for each system.

2.5. Parameterization details. The PIP expansions were parameterized by minimizing the weighted least-squares error

$$\varepsilon^2 = \sum_m^M w_m^2 \left(V(\mathbf{R}_m) - \widehat{V}_m \right)^2 / \sum_m^M w_m^2 \ , \tag{5}$$

where M is the number of *ab initio* energies in the training set, and m labels the computed *ab initio* energies $\widehat{V}_m$ and their associated geometries $\mathbf{R}_m$. The weight in eq 5 is given by

$$w_m = v/|\widehat{V}_m - V_{\text{ref}} + v|, \tag{6}$$

where v is a parameter that controls the range of the weighting function, and V_{ref} is a reference energy. For the nonreactive cases, we chose V_{ref} to be the minimum sampled *ab initio* energy, $\widehat{V}_{\text{min}}$, and, as before,[32] we also set $v \approx \widehat{V}_{\text{min}}$. For the reactive cases, we set V_{ref} equal to the saddle point energy and v equal to the initial internal energy of the nonthermal molecular collider. The zero of energy is always set to the classical equilibrium geometry of the reactant asymptote. The resulting fitting errors have the following useful interpretation: ε is a measure of the absolute fitting error at energies near V_{ref}, and $|\varepsilon/V_{\text{ref}}|$ is a measure of the relative error throughout the sampled region. Notably, the use of eq 6 ensures a uniform distribution of relative errors throughout the broad range of internal energies required for the present nonthermal and near-threshold collisional energy transfer applications.

In addition to the fitting error in eq 5, we considered an out-of-sample error, sometimes called a prediction error, labeled ε_{oos}, which is the result of eq 5 evaluated for a large test set of sampled energies that are not included in the training set. Both ε and ε_{oos} converge to the same large-M limit ε_∞, where ε_∞ can be labeled the "intrinsic fitting error" and is distinguished from the finite-M error associated with incomplete training data.[32] Importantly, the fitting error ε approaches ε_∞ from below while ε_{oos} approaches it from above. If one only considers ε (as is often the case), one might erroneously interpret a small unconverged value of ε as indicating a high accuracy fit when instead it should be interpreted as a sign of insufficient training data. The consideration of both ε and ε_{oos} allows for a clearer appraisal of the quality of the PES expansions, as demonstrated throughout Sec. 3.

3. RESULTS AND DISCUSSION

3.1. Two-body systems: HO_2(+M) and H_2O_2(+M). In this Section, we demonstrate convergence with respect to the completeness of both the training data and the expansion basis set, and we test the quality of the converged PIP expansions via comparisons of computed trajectory-based collisional energy transfer kinetics with experimental values. PIP expansions for several choices of ξ_f and ξ_o were generated for ten nonreactive two-body systems: HO_2(+M), M = He, Ar, N_2, CO_2, and H_2O, and H_2O_2(+M), M = Ar, N_2, CO_2, H_2O, and H_2O_2. Fitting statistics for $\xi_f = \xi_o$ = 3, 4, and 5 are shown in Table 1, including the converged fitting errors ε_∞, and the numbers of adjustable parameters N_p and terms N_t for each expansion.

Convergence with respect to the training data set was determined as described previously.[32,87] Briefly, weighted least-squares fitting errors ε and out-of-sample test errors ε_{oos} were computed as functions of the number of the training data M. For each value of M, 100 parameterizations were generated using M training data randomly sampled from the full training set. The average values of ε and ε_{oos} converge with M to the same fitting error ε_∞ but from below and above, respectively, and their difference is a measure of convergence. This analysis is demonstrated for HO_2(+N_2) in Fig. 1 for the PIP3, PIP4, and PIP5 expansions. Importantly, when plotted as a function of M/N_p, as in Fig. 1, the curves are nearly independent of the choice of ξ_o, despite the variation in N_p from 79 to 825 for the three expansions. This result shows that the number of required training data for PIP expansions scales in predictable ways with the number of parameters N_p.

Table 1. PIP expansion fitting statistics for several HO_2(+M) and H_2O_2(+M) systems

System (V_{ref}/cm^{-1})	PIP ξ_o	ε_∞/cm^{-1}	N_p	N_t	N_t/N_p
HO_2 + He (56)	PIP3	20	38	65	1.7
	PIP4	13	99	176	1.8
	PIP5	7.8	223	407	1.8
HO_2 + Ar (248)	PIP3	40	38	65	1.7
	PIP4	34	99	176	1.8
	PIP5	24	223	407	1.8
HO_2 + N_2 (671)	PIP3	75	79	252	3.2
	PIP4	59	278	932	3.4
	PIP5	39	825	2,878	3.5
HO_2 + CO_2 (1188)	PIP3	117	247	733	3.0
	PIP4	84	1,138	3,667	3.2
	PIP5	48	4,407	15,043	3.4
HO_2 + H_2O (2639)	PIP3	163	247	733	3.0
	PIP4	121	1,138	3,667	3.2
	PIP5	78	4,407	15,043	3.4
H_2O_2 + Ar (216)	PIP3	27	70	203	2.9
	PIP4	12	249	792	3.2
	PIP5	6.1	752	2,542	3.4
H_2O_2 + N_2 (609)	PIP3	53	131	697	5.3
	PIP4	26	599	3,547	5.9
	PIP5	15	2,297	14,713	6.4
H_2O_2 + CO_2 (1001)	PIP3	73	378	1,805	4.8
	PIP4	25	2,165	11,936	5.5
	PIP5	*a*	10,446	63,779	6.1
H_2O_2 + H_2O (1968)	PIP3	104	378	1,805	4.8
	PIP4	55	2,165	11,936	5.5
	PIP5	*a*	10,446	63,779	6.1
H_2O_2 + H_2O_2 (1995)	PIP3	136	547	4,041	7.4
	PIP4	74	3,759	34,141	9.1
	PIP5	*a*	21,863	231,149	10.6

[a]The training data sets are too small to reliably parameterize these expansions.

These observations are quantified by considering the average prediction error ε_{oos} plus twice the standard deviation of the prediction errors as a 95% confidence limit measure of convergence with respect to M. Using this metric and a convergence tolerance of 30% of ε_{∞}, we can read from Fig. 1 (looking for where the thick dashes lines cross the horizontal dashed line) that $M/N_p \approx 35$ sampled energies are required per adjustable parameter N_p to converge the fitting error to 30%. This value of M/N_p agrees with values obtained previously (M/N_p = 20–30) when using biased pseudorandom sampling to train interaction potentials for other systems.[32] Empirically, then, we draw the useful conclusion that M/N_p is relatively insensitive to system as well as functional form, at least when parameterizing nonreactive two-body interaction potentials. This rule incorporates the idea that higher order fits, which necessarily have more parameters, require more training data, and it allows for ready estimates of the required number. Using this rule, training sets with M = 72,000 cannot be reliably used to parameterize PIP expansions for N_p > ~2,000. For the entries in Table 1 where N_p > 2,000, we therefore employed larger training sets of 144,000 energies.

Figure 1 demonstrates the danger of considering the training set's fitting error ε exclusively when characterizing the quality of parameterized PIP expansions. If, for example, we had trained our fits using $M/N_p \approx 10$ *ab initio* energies per adjustable parameter, as might be assumed to be suitable for parametrizing the interaction potential of a relatively simple system like $HO_2(+N_2)$, we would erroneously conclude based on the fitting error alone that the resulting parameterizations are quite accurate. In fact, the resulting fitting errors would be 50% *smaller* than the converged fitting errors ε_{∞} given in Table 1. The consideration of ε_{oos} alongside ε clearly reveals that fits with $M/N_p \approx 10$ are unreliable, as ε and ε_{oos} are seen to differ by factors of 2–3.

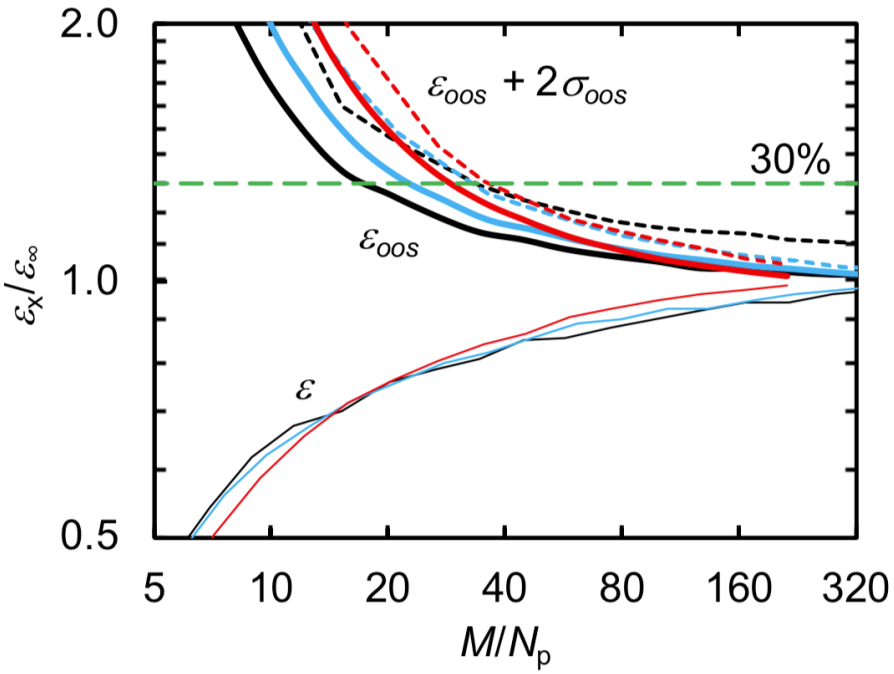

Fig. 1. Convergence of the average out-of-sample error ε_{oos} (thick solid lines) plus two standard deviations σ_{oos} (thick dashed lines) with respect to the number of training data M per adjustable parameter N_p for $HO_2(+N_2)$ for three expansions: PIP3 (black), PIP4 (blue), and PIP5 (red). Also shown is the average in-sample training error ε (thin solid lines). The green dashed horizontal line indicates a 30% threshold for convergence.

We note that in Ref. 32 low discrepancy (quasirandom) sampling was found to reduce bunching of the sampled geometries and in turn reduce M/N_p by nearly a factor of eight to as low as ~4. Low discrepancy sampling is most useful when tailored to the application of interest, and it was not pursued here to provide more general insights into PIP expansion construction. Improved sampling methods such as low discrepancy sampling are likely promising strategies for reducing the number of required training data.[32,88]

Next, we consider the relationship between the converged fitting error ε_∞ and the size and complexity of the expansion. The fitting errors ε_∞ in Table 1 are fairly small for all of the PIP expansions shown, typically between 5–15% of the van der Waals binding energies (indicated by V_{ref}), with the exception of larger relative errors for He due to its small binding energy. Figure 2(a) shows the dependence of $|\varepsilon_\infty/V_{ref}|$ on the number of adjustable parameters N_p, again for $HO_2(+N_2)$.

The figure shows the expansions listed in Table 1 as well as a series of separable expansions (s-PIPξ_o) alongside a more complete set of full-dimensional expansions (PIPξ_o). For both series, the relative error decreases as ξ_f and ξ_o (and therefore N_p) are increased and as flexibility is added to the expansions. Notably, the relative errors scale roughly as $\sim N_p^{-½}$ with exponents varying from –0.3 to –0.6 and closer to –0.5 for the larger expansions. This "standard error" scaling of ε_∞ with N_p suggests that while the larger basis set expansions indeed increase the available parameter space and reduce the fitting error, they do so in an inefficient way that is not ideally tailored to chemical applications.

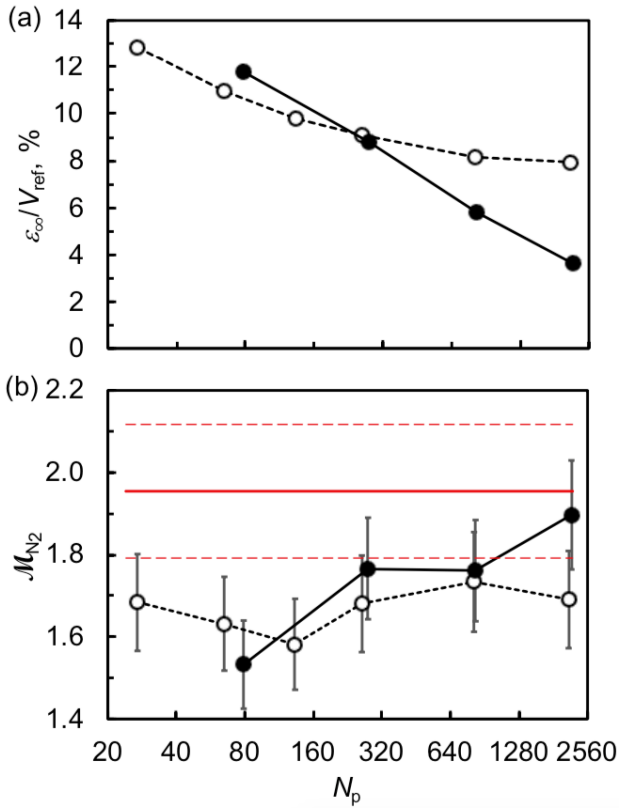

Fig. 2. (a) Relative converged fitting errors $\varepsilon_\infty/V_{ref}$ for several $HO_2(+N_2)$ PIP expansions as a function of the number of adjustable parameters N_p. The open symbols indicate expansions employing the separable approximation (s-PIPξ_o for ξ_o = 3–6,8,10), and the closed symbols indicate full-dimensional expansions (PIPξ_o for ξ_o = 3–6). (b) Bath gas efficiency and two-sigma statistical uncertainties for $HO_2(+N_2)$ at 300 K relative to a reference value for $HO_2(+Ar)$ for the same set of PIP expansions and again shown as a function of N_p. Michael and co-workers' experimental value[89] and their assigned 2-sigma error bars are indicated by red horizontal lines.

Figure 2(a) is representative of similar plots made for the other systems in Table 1, and we do not reproduce them here. In general, the separable s-PIPξ_o expansions outperform full-dimensional PIPξ_o expansions for small N_p, but the s-PIPξ_o series asymptotically approach a finite error as $N_p \rightarrow \infty$ (of about 8% or 53 cm^{-1} in Fig. 2(a)), while the error in the PIPξ_o series can be made arbitrarily small. Physically, the residual error in the s-PIPξ_o series represents the error due to potential energy coupling between the intermolecular and intramolecular degrees of freedom that is neglected by the separable assumption.

Figure 2(b) shows an observable of interest (the collision efficiency of N_2 relative to Ar, i.e., $\mathcal{M}_{N_2}$, at 300 K) and its dependence on the complexity of the PIP expansion for the same expansions shown in Fig. 2(a). For all the PIP expansions shown, fairly good agreement is found between the computed ($\mathcal{M}_{N_2}$ = 1.5 – 1.9) and experimental[89] ($\mathcal{M}_{N_2}$ = 1.95 ± 0.16) values, with the results differing by close to the combined experimental (8%) and theoretical statistical (6%) 2-sigma uncertainties. Encouragingly, the results improve as N_p increases, and these improvements are correlated with reductions in the fitting errors shown in Fig. 2(a). The absolute accuracy of this result should not be overinterpreted, however, due to the approximations involved in converting the trajectory-based information into $\mathcal{M}_{N_2}$. Instead, Fig. 2(b) is presented to emphasize that, in agreement with our earlier study,[32] relative deviations in highly averaged dynamical observables like $\mathcal{M}_{N_2}$ are of the same order as (and often somewhat smaller than) relative fitting errors when defined as in eq 6. We will confirm this important result in the context of reactive PIP expansions in Sec. 3.4.

Similar comparisons were made with available experimental results for several HO_2(+M)[89] and H_2O_2(+M)[90] systems. Briefly, for HO_2(+He) the computed value of $\mathcal{M}_{He}$ at 300 K ranges from 0.74 to 0.78 for the PIP3–PIP6 expansions, in good agreement with the experimental value of 0.82

± 0.07. For M = H_2O at 300 K, the weak collider correction $\beta_c \approx 0.95$. When β_c is close to unity, collisions are not "weak," and predictions made using this model are sensitive to the choice of total collision rate constant *Z*. Using the Lennard–Jones collision rate constant described in Sec. 2, we calculated $\mathcal{M}_{H_2O} \approx 6$, in poor agreement with experiment ($\mathcal{M}_{H_2O} = 23 \pm 2$). A proposed[91] implementation of the β_c model that sets *Z* to a larger, capture rate constant improves the calculated value somewhat to $\mathcal{M}_{H_2O} = 12$. To avoid difficulties in motivating further adjustments to *Z*, we instead solved an *a priori* two-dimensional master equation (2DME) to compute k_0 for M = H_2O and Ar, and this first principles approach does not depend on the choice of *Z*. Using trajectory-based 2DMEs (as described in Ref. 48), we calculated $\mathcal{M}_{H_2O} = 23.3$ and 24.0 for the PIP3 and PIP4 expansions, respectively, in quantitative agreement with experiment.

For H_2O_2(+M), we highlight a recent theoretical study[92] from Matsugi who employed trajectory and kinetics strategies similar to our own. Matsugi used direct (scaled MP2) force evaluations, however, and so avoided detailed characterizations of low temperatures as well as detailed calculations for the larger bath gases due to computation cost considerations. As would be expected due to the very similar trajectory and kinetics approaches involved, the present results (e.g., $\mathcal{M}_{N_2} = 2.1 \pm 0.2$ at 1000 K for H_2O_2(+M), again relative to Ar and with an error indicating the 2-sigma bootstrapped sampling error of 6% for our 14,400 trajectory ensembles) are in close agreement with Matsugi's results ($\mathcal{M}_{N_2} = 1.7 \pm 0.6$, which has a larger 2-sigma statistical sampling error due to his use of smaller 2000 trajectory ensembles). Using the present PIP expansions, we extended these results to include larger baths and lower temperatures. For example, at 1000 K we calculated $\mathcal{M}_{H_2O} = 8.7$, $\mathcal{M}_{CO_2} = 4.8$, and $\mathcal{M}_{H_2O_2} = 8.1$ which may be compared with experimental values close to 1000 K of 6–9, 1.8, and 8–10, respectively.

To conclude this section, we emphasize that the largest source of error in pressure-dependent chemical kinetics is very likely the kinetic model used to describe collisional energy transfer (here, β_c vs the 2DME) and *not* the computation of trajectory-based collision information using *a priori* PIP expansions, even when relatively low-order PES expansions like PIP3 are used. We demonstrated above that the quality of the PIP expansion can be predicted by considering the dependence of the out-of-sample error on the number of training data M, as in Fig. 1. Converged training errors in turn serve as useful predictors of the accuracy of dynamical results, as validated in Fig. 2. These observations, along with increasing access to large scale computing, opens up the possibility for further automating "recipes" for PIP expansion construction and validation enabling further automated kinetics calculations without the need for human intervention.[93]

3.2. Two-body systems: NH_3(+M) and N_2H_4(+M). Next, we consider collisional energy transfer for NH_3(+M), M = Ar, O_2, N_2, CH_4, C_2H_6, CO_2, NH_3, and N_2H_4(+M), M = Ar, O_2, N_2, NH_3. These applications share similarities with those considered in Sec. 3.1 but feature more degrees of freedom, larger bath gases, larger PIP expansions, and more permutational symmetry. We do not repeat the above analyses for NH_3(+M) and N_2H_4(+M), but we note instead a recent modeling study including a comparison of computed collision efficiencies with experimental values.[94] Systems in this Section are used to comment on the symmetrization algorithm and its parallelization efficiency and to demonstrate the effects of unphysical intramolecular and unconnected terms on collisional energy transfer.

Fitting statistics for NH_3(+M) and N_2H_4(+M) are summarized in Table 2, where the number of atomic permutations $N_!$ is indicated and is as large as 3!2!6! = 8640 for NH_3(+C_2H_6). The reduction in the number of fitting parameters N_p relative to the number of terms N_t in the expansion is significant, with N_t/N_p as large as ~450 for the PIP5 expansion of NH_3(+C_2H_6). Considering all

the systems in Table 2, we note that empirically N_t/N_p goes as $\sim N_!^{1/2}$, although there is significant scatter in this relationship.

Table 2. PIP expansion fitting statistics for several NH_3(+M), N_2H_4(+M), and H_2O_2(+M) systems

System (V_{ref}/cm^{-1})	$N_!$	PIP ξ_o	ε_∞, cm^{-1}	N_p	N_t	N_t/N_p	N_u	N_i
NH_3 + Ar (94)	6	PIP3	9.2	53	203	3.8	0	83
		PIP4	4.9	181	792	4.4	0	209
		PIP5	2.9	534	2,542	4.8	0	461
NH_3 + O_2 (133)	12	PIP3	16	100	697	7.0	33	86
		PIP4	11	438	3,547	8.1	116	213
		PIP5	7.5	1,638	14,713	9.0	325	466
NH_3 + N_2 (238)	12	PIP3	24	100	697	7.0	33	86
		PIP4	15	438	3,547	8.1	116	213
		PIP5	10	1,638	14,713	9.0	325	466
NH_3 + CH_4 (244)	144	PIP3	22	359	8,171	23	600	368
		PIP4	10	2,389	86,546	36	3,635	1,209
		PIP5	*a*	14,222	729,050	51	16,885	3,463
NH_3 + C_2H_6 (348)	8640	PIP3	34	507	44,625	88	3,192	4,577
		PIP4	*a*	3,939	843,081	214	37,646	36,168
		PIP5	*a*	27,929	12,444,153	446	337,960	237,796
NH_3 + CO_2 (877)	12	PIP3	38	296	1,805	6.1	117	102
		PIP4	20	1,615	11,936	7.4	471	243
		PIP5	*a*	7,557	63,779	8.4	1,485	516
NH_3 + NH_3 (1039)	36	PIP3	62	347	4,041	12	288	166
		PIP4	33	2,161	34,141	16	1,401	418
		PIP5	*a*	11,759	231,149	20	5,265	922
N_2H_4 + Ar (170)	48	PIP3	16	93	1,209	13	0	815
		PIP4	8.4	457	8,775	19	0	3,875
		PIP5	4.1	2,017	50,277	25	0	15,503
N_2H_4 + O_2 (210)	96	PIP3	25	161	3,527	22	150	818
		PIP4	16	944	31,116	33	965	3,879
		PIP5	*a*	4,921	216,988	44	4,840	15,508
N_2H_4 + N_2 (375)	96	PIP3	35	161	3,527	22	150	818
		PIP4	20	944	31,116	33	965	3,879
		PIP5	*a*	4,921	216,988	44	4,840	15,508
N_2H_4 + NH_3 (1494)	288	PIP3	52	505	15,273	30	1,125	898
		PIP4	*a*	3,858	199,227	52	8,565	4,084
		PIP5	*a*	26,357	2,052,981	78	49,815	15,964

[a]The training data sets are too small to reliably parameterize these expansions.

Enforcing permutational symmetry for the largest systems considered here required significant computational effort and was enabled by parallelization. For example, symmetrizing the PIP5 expansion for NH_3(+C_2H_6) required, in principle, $N_!N_t(N_t-1)/2 = \sim 10^{18}$ comparisons of

the exponents to identify symmetry matches. (Fewer comparisons were made in practice, as comparisons for a given term are no longer made once a match is found.) With the symmetrization step parallelized over 144 cores, the computation time for the PIP5 expansion for $NH_3(+C_2H_6)$ took just 1 day, while on a single processor this step would have taken ~3.5 months.

Scaling efficiency in the symmetrization step was quantified on Argonne's 1.7 PFlop/s Bebop (Intel Broadwell) cluster for the PIP5 expansion of the $NH_3(+CH_4)$ interaction energy, where we found a strong scaling efficiency of 75% on 36 cores (1 node), 70% on 144 cores (4 nodes), and 50% on 1152 cores (32 nodes). Larger scaling efficiency tests were performed using Argonne's 10 PFlop/s Theta (Intel Xenon Phi) cluster. For example, for the PIP4 expansion for the three-body system $CH_2OO+H_2O_2+H_2O$ (discussed in more detail in Sec. 3.3), we quantified a strong scaling efficiency of 56% using 10,240 cores (160 nodes). Scaling efficiency was found to be better for the larger PIP expansions where petascale resources are most needed.

Table 2 shows the number of unphysical unconnected terms N_u and intramolecular-only terms N_i removed from the expansions. For these systems, just 4% of terms are unconnected, on average, with somewhat larger fractions (9%) of intramolecular-only terms. The corresponding ~13% reduction in N_p due to the removal of these terms leads to a correspondingly modest reduction in the number of required training data and in the cost of evaluating the expansion.

More importantly, however, the values assigned to these unphysical parameters depend in uncontrollable ways on the particulars of the fitting routine and training data. In fact, the intramolecular-only terms are largely unconstrained by the present training data sets consisting of intermolecular energies. The coefficients that end up assigned to these terms are artifacts of the fitting routine and are not generally set to the correct value of zero, although they may be small. The unconnected terms, in contrast, are in principle constrained by long-range data. Just how close

the coefficients converge to zero, however, depends on the relative amount of long-range data to short-range data as well as the overall fitting error. Errors in dynamics associated with retaining these unconstrained and uncontrolled terms are characterized next.

We generated several unphysical PIP expansions where unconnected terms were not removed (labeled by "+u") and where both unconnected terms and intramolecular-only terms were not removed (labeled by "+ui"). The singular value decomposition (svdfit) routine from Numerical Recipes[95] was used to perform the least-squares optimizations, and the values this routine assigned to the unconstrained coefficients were used unaltered.

Collision parameters were calculated using several PIPξ_o+u and PIPξ_o+ui expansions, and we found that including unconnected and intramolecular-only terms increased the calculated value of α by as much as 25% relative to PIPξ_o results for the systems tested from Table 2. An error this size would typically lead to a correspondingly large error in the predicted low-pressure-limit reaction rate constant. The magnitude of this error rivals the overall accuracy of unimolecular kinetics calculations,[48,66] and furthermore its size depends in an uncontrolled way on details of the fitting procedure. It also depends on details of the trajectory calculations, as described next.

Figure 3 shows the deactivating wing of collision outcomes for NH_3(+NH_3) at 300 K represented as a pseudo-opacity function. The energy transferred in deactivating collisions $\langle \Delta E_d \rangle$ for an ensemble of 5000 trajectories was binned and plotted as a function of the impact parameter b (the initial offset of the center-of-mass momentum vectors for the two colliders). Energy transfer is seen to be most significant for small b, i.e., for "head on" collisions, where the average energy transferred is around 1500 cm^{-1} and is very similar for the three PIP expansions, PIP3, PIP3+u, and PIP3+ui. In PIP applications where long-range descriptions are not necessary, such as in the

many body water potential MB-pol for describing condensed phases,[27] Fig. 3 suggests that unconnected terms do not introduce substantive errors.

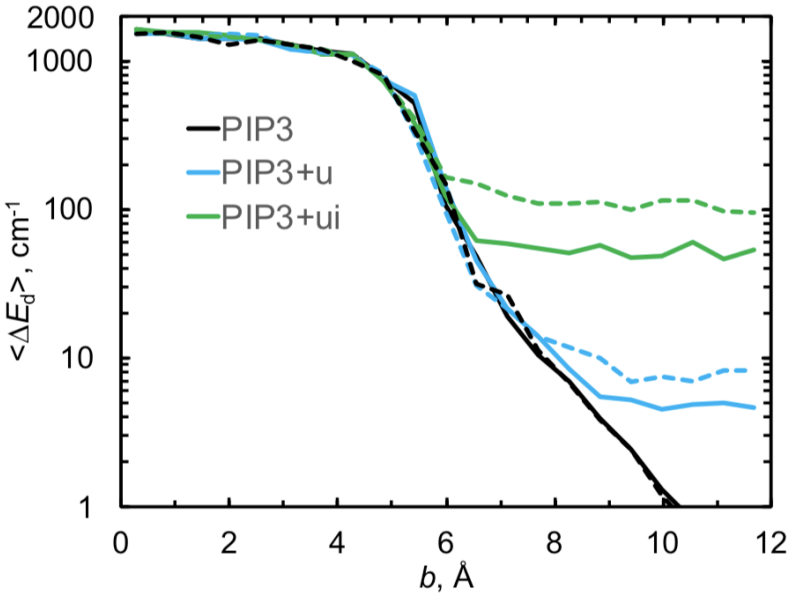

Fig. 3 Pseudo-opacity function for the deactivating wing of NH_3(+NH_3) collision outcomes at 300 K. The average energy transferred in deactivating collisions, $\langle\Delta E_d\rangle$, is shown binned as a function of impact parameter *b* for the PIP3 (solid black line), PIP3+u (solid blue line), and PIP3+ui (solid green line) expansions. The dotted lines show the effects of removing training data with center of mass separations greater than 8 Å from the training data set.

At large *b*, however, the unphysical terms in the PIP33+u and PIP33+ui expansions introduce noticeable errors. Using the correctly pruned PIP3 expansion, the energy transfer exponentially decays to zero as *b* increases and as the two colliders increasingly "miss" each other. For the PIP3+u and PIP3+ui expansions, on the other hand, energy transfer erroneously tends toward constant values of ~5 and ~50 cm^{-1}, respectively. The magnitudes of these large-*b* errors are small relative to the magnitude of energy transfer in head-on collisions for the NH_3(+NH_3) system, but this need not be the case. Furthermore, the magnitude of the coefficients for the unconnected terms depends on the relative amount of training data at long range. To demonstrate this, we refit the PIP3+u and PIP3+ui expansions with training data sampled out to center-of-mass

distances of 8 Å instead of 20 Å. Results for these new expansions are shown as dashed lines in Fig. 3 where it can be seen that the associated long-range errors are doubled.

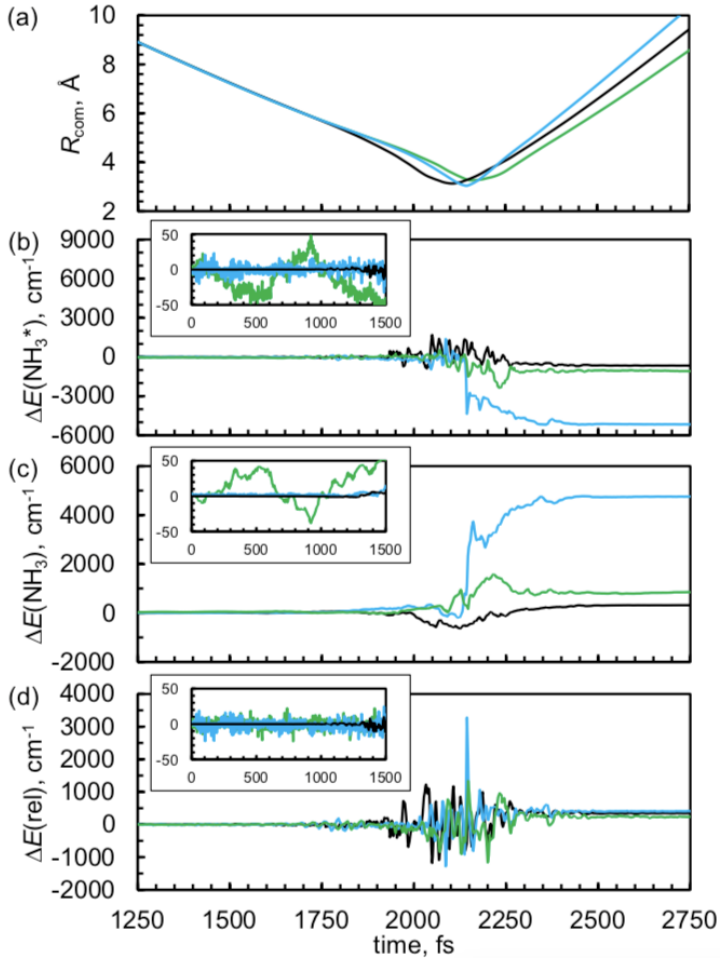

Fig. 4. Representative NH_3(+NH_3) trajectory at 300 K, showing (a) the center of mass separation R_{com} and the change in the energy of the (b) excited unimolecular reactant NH_3*, (c) the bath gas NH_3, and (d) their relative coordinates as functions of time. Results are compared for the PIP3 (black), PIP4+u (blue), and PIP3+ui (green) expansions for identical initial conditions.

Unphysical long-range energy transfer is presented in more detail in Fig. 4 for a single NH_3(+NH_3) collision. Figure 4 shows the center of mass distance of the colliders, their internal energies, and their relative energy as functions of time along a representative trajectory at 300 K. The same initial conditions were used to launch trajectories governed by each of the three PIP3,

PIP3+u, and PIP3+ui expansions. As seen in Fig. 4(a), the inclusion of additional unphysical PIP expansion terms causes the trajectories to diverge from one another at a center-of-mass distance of ~5 Å, leading to very different trajectory outcomes.

The insets in Figs. 4(b)–(d) explicitly show the artificial long-range energy transfer for the early portions of the trajectories before the colliders interact appreciably. As would be expected, using the correctly prepared PIP3 expansion results in negligible exchange between the three types of energy over this region of the trajectory. Using the PIP3+u expansion, in contrast, we see high-frequency energy exchange between the highly excited collider NH_3* and the relative coordinates with less energy exchange into the colder, thermally populated NH_3. This is presumably due to high-order unconnected terms affecting regions more likely to be accessed at the high energies in NH_3* than in thermal NH_3.

For the PIP3+ui expansion, we again see the effects of the unconnected terms leading to high-frequency energy exchange between NH_3* and the relative coordinates along with a new type of artificial energy exchange due to the intramolecular-only terms. These latter terms exchange energy between the fragments with a relatively large amplitude and lower frequency than the artificial energy transfer associated with unconnected terms only. The magnitudes of these two effects correspond with the errors at large-*b* quantified in Fig. 3.

To conclude this Section, we emphasize that the errors associated with unphysical unconnected terms are small enough to be difficult to notice in many applications, unlike, say, the catastrophic failures associated unphysical holes. The errors are large enough to spoil quantitative assessments of collisional energy transfer, however, and furthermore introduce unphysical dependences on numerical parameters of the calculation such as the maximum impact parameter.

3.3. Three-body systems: CH_2OO···HPMF(+M) and CH_2OO···H_2O_2(+M). Next we consider nonreactive collisions for two systems relevant to atmospheric chemistry,[67] namely the transient complexes formed between the Criegee intermediate (CH_2OO) and hydrogen peroxide (H_2O_2) or hydroperoxymethylformate ($HOOCH_2OCHO$, HPMF). The results in this Section are used to demonstrate the severe scaling of PIP expansions for three-body systems and the utility of artificial permutational symmetry for dealing with this complexity.

For each of the two hydroperoxides, two-body PIP expansions were generated for describing its van der Waals interaction with CH_2OO and three-body PIP expansions were generated for describing the interaction of the van der Waals complexes with the bath gases M = He and H_2O. The six expansions are characterized in Table 3. For these systems, V_{ref} = 3000 to 6000 cm^{-1} such that the intrinsic fitting errors ε_∞ in Table 3 are less than 10% of V_{ref}, even for the relatively low-order PIP2 expansions.

Table 3. PIP expansion fitting statistics for six van der Waals complexes of the Criegee intermediate CH_2OO

System	$N_!$	PIP ξ_o	ε_∞, cm^{-1}	N_p	N_t	N_u	N_i
$CH_2OO + H_2O_2$	8	PIP2	256	164	551	60	92
		PIP3	199	1,933	8,171	600	368
		PIP4	*a*	17,336	86,546	3,635	1,209
$CH_2OO + H_2O_2 + He$	8	PIP2	300	291	859	130	92
		PIP3	134	3,952	15,203	1,725	368
		PIP4	*a*	42,323	197,112	13,555	1,209
$CH_2OO + H_2O_2 + H_2O$	16	PIP2	400	511	1,799	378	101
		PIP3	*a*	8,809	44,150	9,189	387
		PIP4	*a*	123,814	817,081	108,342	1,243
CH_2OO + HPMF	4	PIP2	280	2,402	4,076	450	1,145
		PIP3	*a*	86,638	173,301	13,275	17,580
		PIP4	*a*	2,232,923	5,108,126	242,250	212,875
CH_2OO + HPMF + He	4	PIP2	121	3,361	5,461	775	1,145
		PIP3	*a*	160,251	302,621	27,300	17,580
		PIP4	*a*	3,904,110	8,582,626	585,750	212,875
CH_2OO + HPMF + H_2O	8	PIP2	167	4,649	9,041	1,740	1,154
		PIP3	*a*	213,217	520,731	82,290	17,599

[a]The training data sets are too small to reliably parameterize these expansions.

Notable in Table 3 is the severe scaling of N_p and N_t with both the order of the expansion and the increasing complexity of the third body. In fact, parametrizing PIP3 expansions was possible using training sets of 144,000 energies for just two of the six cases due to the severe scaling of N_p for these systems. PIP3 expansions indeed improve upon the fitting error relative to PIP2 but at the cost of significantly increasing the size of the expansion by factors varying from 15x to 58x. Adding an atomic third body (M = He) approximately doubles the number of terms and parameters, while adding a molecular third body (M = H_2O) requires 2–9 times as many terms, with increased scaling for the larger basis sets.

Generating the millions of training data required to parametrize improved fits for the three-body systems in Table 3 is clearly undesirable if not prohibitive. Furthermore, this severe scaling would seem to be at odds with the relatively simple nature of many-body interaction potentials, which are often accurately described with cutoffs and/or pairwise expressions with much weaker scaling.

The generality of the permutational invariance scheme described in Sec. 2 provides a way to reduce the complexity of the PIP expansions and to achieve less severe scaling. Table 4 summarizes fitting statistics for the two-body CH_2OO + HPMF system for six choices of permutation group assignments. Scheme A reflects the true permutational symmetry of this system with its 13 distinct atom types for this 15 atom system, giving $N_! = 2!2! = 4$. In Scheme F, by contrast, atoms in each molecule are assigned to just three groups, one group each for H, C, and O, for a total of six groups giving $N_! = 2!2!4!4!2! = 4608$. Schemes B–E are intermediate cases motivated by chemical considerations. For example, in Scheme B the two peroxy O atoms on each molecule are treated as equivalent to one another, reducing the number of groups by 2 and

increasing $N_!$ by 4 relative to Scheme A, while in Scheme C the ether O atom on HPMF is included in this group as well, resulting in a factor of three more permutations and giving $N_! = 48$.

Table 4. Fitting statistics for six permutation group assignment strategies for CH_2OO + $HOOCH_2OCHO$ ($V_{ref} = 3697$ cm^{-1})

																			PIP2 (N_t = 4,076)		PIP3 (N_t = 173,301)	
Scheme	C	H	H	O	O	+	H	O	O	C	H	H	O	C	H	O	$N_!$	ε_∞, cm^{-1}	N_p	ε_∞, cm^{-1}	N_p
A	1	2	2	3	4		5	6	7	8	9	9	10	11	12	13	4	280	2,402	*a*	86,638
B	1	2	2	3	3		4	5	5	6	7	7	8	9	10	11	16	317	1,327	*a*	41,192
C	1	2	2	3	3		4	5	5	6	7	7	5	8	9	10	48	335	980	*a*	26,383
D	1	2	2	3	3		4	5	5	6	7	7	5	6	7	8	288	356	522	*a*	11,206
E	1	2	2	3	3		4	5	5	6	7	7	5	6	7	5	1152	368	352	*a*	6,492
F	1	2	2	3	3		4	5	5	6	4	4	5	6	4	5	4608	385	226	141	3,600

[a]The training data set is too small to parameterize these expansions.

In some contexts, adding artificial symmetry might be a severe approximation, but here we see that the increased fitting error ε_∞ introduced by this approximation does not significantly increase the overall error in the fit. Meanwhile, the benefit of employing extra symmetry is significant, as Scheme F reduces the number of adjustable parameters N_p by factors of 10 and 25 relative to Scheme A for the PIP2 and PIP3 expansions, respectively.

Reductions in N_p lead to corresponding reductions in the required number of training data M and thus considerably simplify the fitting procedure, as shown in Fig. 5 where the six permutation group assignment Schemes were tested for convergence with respect to the number of training data M using the same out-of-sample error 95% confidence criterion used in Fig. 1. For all six Schemes, we find that $M/N_p \approx 20$ training data are needed to converge the fits within 30% of the converged intrinsic fitting error despite the variation in N_p from 226 to 2402 for the six

Schemes. The reduction in N_p by a factor of ~10 when using the artificial symmetry of Scheme F thus allows for the use of an order of magnitude fewer training data.

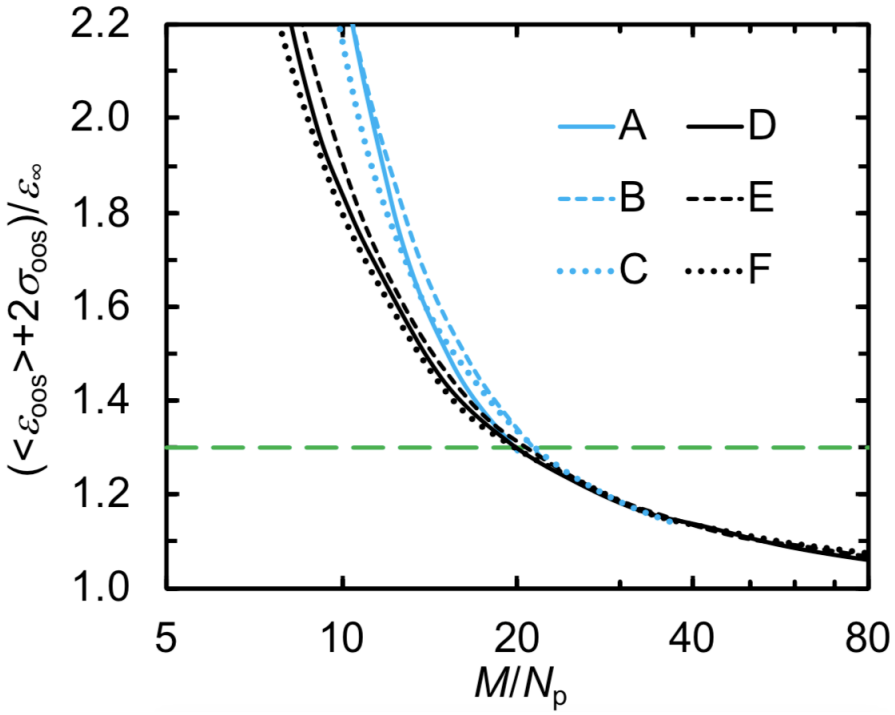

Fig. 5 The convergence of the average out-of-sample error plus two standard deviations for six permutation Schemes A–F for CH_2OO + HPMF, showing its relative independence of M/N_p. The green dashed horizontal line indicates a 30% threshold for convergence.

Artificial symmetry enables the use of higher-order PIP expansions as well. For example, without using artificial permutational symmetry, improving upon the PIP2 expansion with the PIP3 expansion leads to an increase in N_p of 36x to over 86,000 parameters and would require a training data set of nearly 2 million energies. Using Scheme F's artificial symmetry assignments, the larger PIP3 expansion can be parameterized using a manageable training data set, and we see from Table 4 that its fitting error is reduced by ½ compared to Scheme A's PIP2 expansion.

The Schemes A–F in Table 4 were used to parametrize three-body fits for M = He and H_2O, as shown in Table 5. The number of adjustable parameters N_p is again reduced by almost an

order of magnitude using Scheme F relative to the true permutational symmetry in Scheme A, with the fitting errors remaining small (~100–200 cm^{-1}) relative to the well depths (2500–6000 cm^{-1}).

Table 5. PIP2 expansion fitting statistics for two three-body system CH_2OO + $HOOCH_2OCHO$ + M using the six permutation group assignment strategies defined in Table 4

	M = He (N_t = 5,461)			M = H_2O (N_t = 9,041)		
Scheme	$N_!$	ε_∞, cm^{-1}	N_p	$N_!$	ε_∞, cm^{-1}	N_p
A	4	121	3361	8	167	4,649
B	16	129	1967	32	186	2,857
C	48	144	1472	96	197	2,177
D	288	154	808	576	222	1,244
E	1152	166	559	2,304	227	888
F	4608	179	372	9,216	234	615

The performance of the three-body PIP expansions was tested using collisional energy transfer trajectory ensembles describing changes in the internal energies of the CH_2OO···hydroperoxide van der Waals complexes. The use of artificial symmetry, even as extreme as in Scheme F, was found to introduce no appreciable error in the collisional energy transfer dynamics. Specifically, for CH_2OO···HPMF(+He) at 300 K, we computed values of $\alpha = \langle \Delta E_d \rangle$ that varied from 444 to 512 cm^{-1} for the six Schemes, and the small deviations showed no clear trend with ε_∞. The scatter in the six results is in fact close to the 2-sigma statistical uncertainties of 10%.

Comparing results for the two hydroperoxides, our computed values of $\alpha = \langle \Delta E_d \rangle$ for CH_2OO···HPMF (+ He) at 300 K were found to be just ~20% larger, on average, than the computed values of α for CH_2OO···H_2O_2 (+ He). This trend agrees with past work showing weak scaling of collisional energy transfer strengths with system size for systems with more than a few heavy atoms.[49]

As expected, M = H_2O was found to be a significantly more efficient collider than He for both hydroperoxides, with $\alpha = \langle \Delta E_d \rangle$ close to ~2700 cm^{-1} at 300 K. Notably, for M = H_2O a large number of long-lived trajectories were identified, leading to a significant fraction of van der Waals exchange reactions. In fact, the three possible van der Waals products (e.g., $CH_2OO\cdots H_2O_2$, $CH_2OO\cdots H_2O$, and $H_2O_2\cdots H_2O$) were found to be formed in nearly equal amounts. A detailed description of these dynamics is beyond the scope of the present work. Instead, these results are intended to highlight the general utility of using artificial permutational symmetry to efficiently generate PIP expansions for complex many body systems.

3.4. Reactive systems: $H + H_2O \rightleftharpoons H_2 + O_2$, $H + CH_4 \rightleftharpoons H_2 + CH_3$, and $H + C_2H_6 \rightleftharpoons H_2 + C_2H_5$. In a final set of applications, we consider three reactive systems relevant to nonthermal kinetics. The two smaller systems were studied previously, and their nonthermal reaction kinetics were characterized[68,69] with the molecular reactant initially vibrationally excited close to its dissociation threshold and the other degrees of freedom sampled thermally. For our earlier H + CH_4 study, we used the analytic PES of Corchado and co-workers,[85] which they demonstrated gives quantitative thermal kinetics up to 2000 K using transition state theory. They also showed,[85] and we reproduced,[68] that high-temperature thermal kinetics calculated using quasiclassical trajectories and their PES agreed well with the TST results, so long as nuclear quantum effects were accounted for. The quality of this surface for nonthermal kinetics, which features high-energy distorted CH_4 geometries not accessed under thermal conditions, was unable to be assessed, however, and is explored next via comparisons to converged PIP expansions parametrized against training data generated to describe a wide range of energies.

Reactive PIP expansions for H + CH_4 ⇌ H_2 + CH_3 were generated for several values of ξ_f and ξ_o, and fitting statistics are shown in Table 6 alongside results for H + H_2O and H + C_2H_6. For

these reactive cases, $N_!$ can be quite large due to the required equivalence of all of the H atoms. Specifically, $N_!$ increases from 6 to 120 to 10,080 for H + H_2O, CH_4, and C_2H_6, respectively. Because of this, the number of adjustable parameters N_p increases ~10x more slowly than the number of terms N_t across the series. The PIP6 expansion for H + C_2H_6, for example, has more than three million terms but just 3447 parameters, a reduction of more than 1000x. Trends in the fitting statistics for the reactive cases in Table 6 are similar to those identified for the nonreactive cases discussed in Tables 1–3. We again find that the intrinsic fitting errors ε_∞ scale as $\sim N_p^{-½}$ and furthermore that these errors are predictive of the dynamical errors, as shown next.

Table 6. PIP expansion fitting statistics for two reactive systems

PIP ξ_o	ε_∞, kcal/mol	N_p	N_t	N_u
H + H₂O (V_{ref} = 21.9 kcal/mol; $N_!$ = 6)				
PIP3	15.7	20	75	9
PIP4	6.5	45	192	18
PIP5	4.3	93	432	30
PIP6	3.1	181	879	45
PIP7	2.4	327	1,653	63
PIP8	1.9	562	2,919	84
PIP9	1.7	924	4,897	108
PIP10	1.5	1461	7,873	135
H + CH₄ (V_{ref} = 14.9 kcal/mol; $N_!$ = 120)				
PIP3	9.5	23	546	270
PIP4	4.6	72	2,866	1,010
PIP5	2.6	224	12,554	2,950
PIP6	1.7	678	46,929	7,335
PIP7	1.2	1933	154,294	16,250
PIP8	0.9	5209	457,334	32,980
H + C₂H₆ (V_{ref} = 12.3 kcal/mol; $N_!$ = 10,080)				
PIP3	7.6	43	4,666	4,473
PIP4	3.1	159	51,994	39,396
PIP5	1.7	610	480,388	269,010
PIP6	0.9	2455	3,732,862	1,512,924

We used the PIP expansions in Table 6 to calculate several thermal and nonthermal rate constants, and a subset of results are shown in Fig. 6(a) for the PIP4, PIP5, PIP6, and PIP7

expansions describing $H + CH_4 \rightleftharpoons H_2 + CH_3$. The three rate constants shown in Fig. 6(a) are: the thermal rate constant at 1000 K, k(1000 K), and two nonthermal rate constants at 1000 K with different initial internal energies: k^*(56 kcal/mol,1000 K) and k^*(91 kcal/mol, 1000 K). The lower-energy nonthermal case has a similar average total energy as the thermal case, whereas the higher-energy nonthermal case is significantly hotter and approaches methane's dissociation threshold.

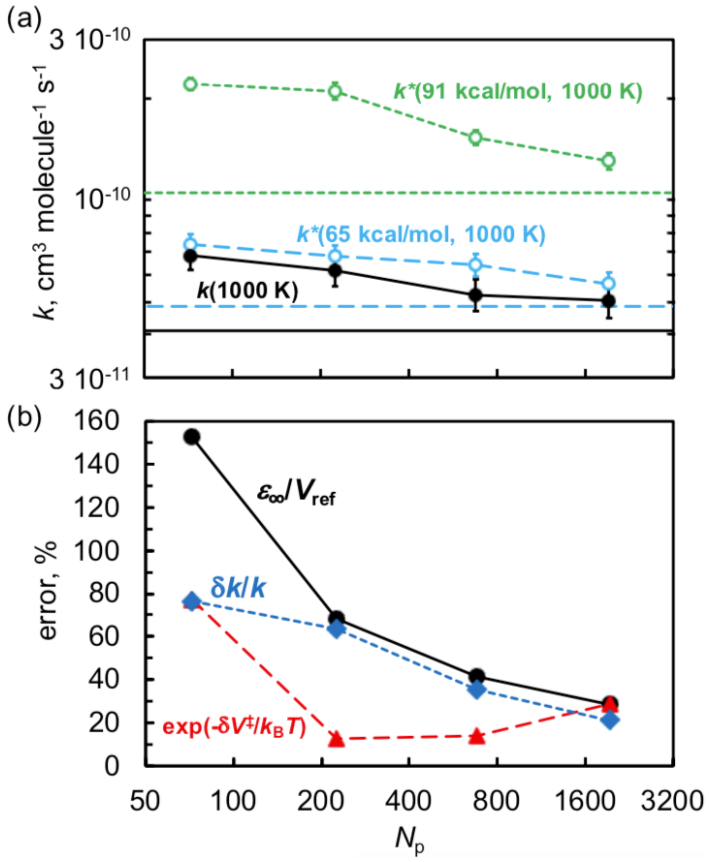

Fig. 6. (a) Thermal and nonthermal rate constants and two-sigma statistical uncertainties for four CH_5 expansions (PIPξ_0 for ξ_0 = 4–7) as a function of N_p. The horizontal lines indicate the results obtained using the analytic PES of Ref. 85. (b) Relative fitting errors (black circles), Boltzmann-weighted saddle point energy errors (red triangles), and the average relative error in the three rate constants (blue diamonds) for the same four PIP expansions.

The calculated rate constants are compared in Fig. 6(a) with results obtained previously using the analytic surface of Corchado et al.[85] (shown as horizontal lines). Trends in the predicted

rate constants are seen to be similar for all three initial conditions, with somewhat larger deviations from the reference rate constant for the higher-energy nonthermal case than for the lower-energy cases. In every case, the results converge with respect to N_p to our previous results[68] obtained using the analytic PES of Corchado et al.[85]

Figure 6(b) quantifies this observation, where the average of the relative deviation in k and k^* from Corchado's analytic PES reference value for the three cases ($\delta k/k$) is shown to decrease from ~80% for the PIP4 expansion down to just 20% for the PIP7 expansion. Notably, this deviation is well predicted by the relative intrinsic fitting errors $\varepsilon_\infty/V_{ref}$ (recall that V_{ref} is set to the saddle point energy when parametrizing the reactive PIP expansions). This result generalizes the conclusion drawn from Fig. 2 and previously in Ref. 32 of the utility of $\varepsilon_\infty/V_{ref}$ in predicting the quality of the PIP expansion for kinetics and dynamics applications.

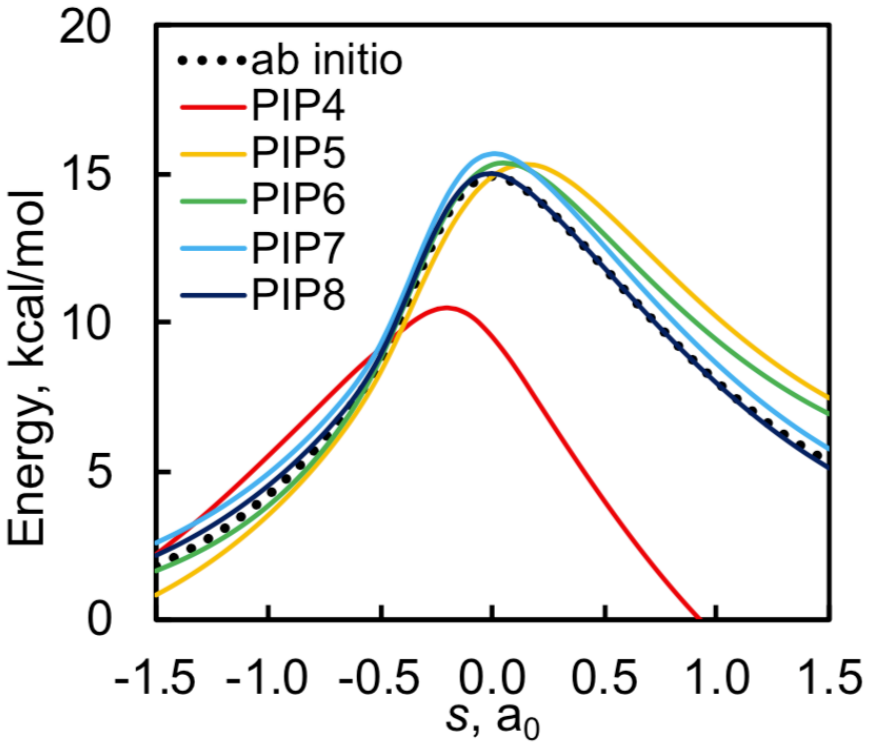

Fig. 7. H + CH_4 energies along the *ab initio* (full valence CASPT2/aug-cc-pVTZ) minimum energy path for five PIP expansions, where *s* is the distance along the mass-scaled reaction coordinate.

Minimum energy paths (MEPs) for five PIP expansions for H + CH_4 are shown in Fig. 7. It is common practice to show the quality of the fitted MEP as an indication of the expected accuracy of the fitted PIP expansion along the reaction coordinate (here shown as the mass-scaled distance s). While it is certainly a good sign to fit the MEP well, this alone cannot provide a guarantee of the accuracy of the dynamical results. For example, in Fig. 6(b) $\exp(-\delta V^{\ddagger}/k_{\mathrm{B}}T)$ is used to estimate the error in the rate constant due to the error in the barrier height $\delta V^{\ddagger}$ and is shown to poorly correlate with the error in the predicted kinetics $\delta k/k$. Furthermore, figures like Fig. 7 can be made to look arbitrarily good by including MEP geometries as training data and increasing their weights. Again, we emphasize that $\varepsilon_{\infty}/V_{\mathrm{ref}}$ provides an accurate measure of the quality of the PIP expansion (although the near perfect agreement for three of the four cases in Fig. 6(b) is likely fortuitous).

We note that the utility of $\varepsilon_{\infty}/V_{\mathrm{ref}}$ is due in part to the use of the weight function defined in eq 6. For energies close to the reference energy, such as those along the MEP, ε_{∞} is a measure of the absolute fitting error. For example, the mean fitting errors for the MEPs shown in Fig. 7 are 4.9, 2.9, 1.7, 0.7, and 0.5 kcal/mol for PIP4–8, respectively, which agree well with the fitting errors ε_{∞} reported in Table 6 (4.6, 2.6, 1.7, 1.2, and 0.9 kcal/mol). Furthermore, eq 6 ensures that the relative errors are distributed nearly uniformly throughout the energy range of interest. For example, when binned as a function of energy the average relative fitting error for the PIP7 expansion of H + CH_4 is 2.5% for energies within 5 kcal/mol of V_{ref}. This relative error varies from 1.7 to 2.4% for binned energies up to 120 kcal/mol, which is the highest energy relevant to the nonthermal quasiclassical trajectory calculations.

The training data sets for the reactive cases included reactant geometries representative of high-energy H_2O, CH_4, and C_2H_6 configurations. As a test, we refit the H + CH_4 PIP expansions

using a training set modified to include CH_4 geometries sampled thermally at 2000 K, while retaining the same training data for the product and saddle point geometries, i.e., the training sets was modified to exclude high-energy CH_4 geometries that are not accessed thermally. Rate constants predicted using these PIP expansions agreed well with those shown in Fig. 6 for the thermal and low energy nonthermal rate constants, although noticeable but small (<2%) fractions of trajectories did find "holes" and had to be thrown out. For high-energy nonthermal cases, however, a significant fraction (>50%) of trajectories found holes, and rate constants computed for the remaining trajectories deviated significantly from the converged results.

Sometimes an iterative procedure is used to correct holes, where a provisional PIP expansion is produced, trajectories are run, and the geometries of holes are collected. These geometries are then included as training data and the process is repeated. We attempted such a procedure, and, after a few iterations, we were indeed able to produce PIP expansions that allowed us to run large ensembles of trajectories without finding holes. However, the rate constants obtained in this way deviated from the ones in Fig. 6 by factors of 2 to 10! This perhaps surprising result is likely due to the fact that simply eliminating holes is not enough to ensure that poorly sampled regions of the PES are being represented accurately. Instead of the catastrophic infinities associated with holes, poorly sampled regions might instead feature inaccurate valleys or hills that are more difficult to notice. This error highlights the utility of the continuing advance of systematic approaches to PIP generation and validation.

4. CONCLUSIONS

A general strategy for constructing physical PIP expansions was described and applied to 30 chemical systems of varying complexity. The demonstrations included cases with large internal

energies requiring semiglobal descriptions, cases with as many as 15 atoms and 39 internal degrees of freedom, two- and three-body van der Waals complexes, and both nonreactive and reactive systems. For each system, several PIP expansions were generated and symmetrized, and unphysical unconnected terms were automatically identified and removed. The PIP expansions included as many as 13 million terms and symmetry reductions of nearly a factor of 450, and the construction of the largest expansions was enabled via parallelization and access to petascale resources. The software written for this work, PIPPy, is freely available.[45]

The quality of the PIP expansions was tested in a variety of ways. Convergence with respect to the number and completeness of training data was demonstrated by considering an out-of-sample error alongside the usual training error. Likewise, by considering series of PIP expansions, convergence with the respect to the complexity of the expansion was quantified. Importantly, these metrics were shown to correlate with the convergence of trajectory-based dynamics and kinetics results for both nonreactive and reactive cases. These convergence checks help reveal whether sampled regions have been undersampled and whether as a result the PIP expansions are overfitting the data to produce unphysical wiggles and "holes." Notably, however, these convergence checks *cannot* reveal whether the training data set neglects entire swaths of configurations, such as those representing one of several product channels. This latter type of sampling error is likely to produce holes as well, and the identification and removal of holes remain important areas of study.[20,26,38]

For several nonreactive two-body systems, converged PIP expansions were used in collisional energy trajectory calculations and were shown to predict pressure dependent kinetics and relative collision efficiencies in quantitative agreement with experiment. By comparing the results of two trajectory-based kinetics models (a one-parameter weak collider β_c model and more

detailed master equation model), we showed that the largest source of error in typical a priori pressure dependent kinetics calculations is likely the choice of kinetic model and not the quality of the PIP expansions, as is sometimes assumed.

In these applications, attention was given to the removal of unphysical unconnected terms and the consequences of erroneously retaining these terms. The errors associated with these terms were shown to be small enough to be difficult to notice in highly averaged collisional energy transfer moments but nonetheless large enough to spoil quantitative assessments of collisional energy transfer. Furthermore, the effects of artificial long-range energy transfer associated with these unphysical terms was identified and shown to greatly alter trajectory outcomes as well as lead to uncontrolled dependencies on numerical details of the parametrization and trajectory calculations.

The flexibility of the algorithm and the use of artificial permutational symmetry was shown to enable the parametrization of accurate PIP expansions for complex many-body systems like $CH_2OO + HOOCH_2OCHO + H_2O$. Flexible symmetry assignments allowed for parameterizations using training data sets of thousands of sampled energies instead of the millions that would have been otherwise required.

The consideration of several reactive cases allowed us to generalize the conclusions drawn from the nonreactive cases. Most importantly, the converged fitting errors ε_∞, when defined as in eqs 5 and 6, were shown to be useful predictors of the accuracy of the nonthermal rate constants. In fact, we found the near quantitative agreement between $|\varepsilon_\infty/V_{\text{ref}}|$, where here V_{ref} is the saddle point energy, and the relative deviation in predicted rate constants from their converged values. This is in notable contrast to the utility of assessing energies at the saddle point or along the reaction path to demonstrate the quality of PIP expansions.

Finally, although we demonstrated the construction of PIP expansions for complex systems with up to 15 atoms, the factorial scaling of N_p with dimensionality severely limits the utility of this approach for much larger systems. The observation quantified here that the fitting error ε_∞ scales as $N_p^{-1/2}$ (i.e., the standard error scaling of ε_∞ scales with N_p) suggests that higher-order terms in a PIP expansion seemingly "randomly" contribute to its flexibility and are not optimized for chemical applications. Such a Monte Carlo style approach to improving PIPs is inherently inefficient and is especially undesirable considering the underlying structure of molecular systems, which is known to be well described, at least qualitatively, by a small number of linearly-scaling stretches, bends, and torsions.

Looking ahead, automated PES generation via the PIP expansion parametrization methods described here is an important component toward the full automation of characterizing complex chemical systems from first principles. Automation requires fast and general techniques for handling a wide variety of stoichiometries, numbers of reaction channels, and many body interactions, as well as strategies for assessing the accuracy of the PIP expansions that do not require human judgement. The historical approach to PES fitting involving ad hoc system-specific strategies is clearly not readily applied at scale.

Acknowledgments

This work was supported by the U. S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences operating under Contract Number DE-AC02-06CH11357 and by the Argonne–Sandia Consortium on High-Pressure Combustion Chemistry (ANL FWP 59044). We gratefully acknowledge computing resources provided by Bebop, a high-performance computing cluster operated

by the Laboratory Computing Resource Center at Argonne National Laboratory. Software development was supported in part by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations, the Office of Science and the National Nuclear Security Administration. An award of computer time was also provided by the ASCR Leadership Computing Challenge (ALCC) program using resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility.

References

[1] London, F. Quantenmechanische Deutung des Vorangs der Aktivierung, Z. Elektrochem. *Angew. Phys. Chem.* **1929**, *35*, 552–555.

[2] Eyring, H.; Polanyi, M. On Simple Gas Reactions, *Z. Phys. Chem.* **1931**, *12*, 279–311.

[3] Rydberg, R. Graphische Darstellung einiger bandenspektroskopischer Ergebnisse, *Z. Phys.* **1932**, *73*, 376–385.

[4] Axilrod, B. M.; Teller, E. J. Interaction of the van der Waals Type Between Three Atoms, *Chem. Phys.* **1943**, *11*, 299–300.

[5] Sato, S. Potential Energy Surface of the System of Three Atoms, *J. Chem. Phys.* **1955**, *23*, 2465–2466.

[6] Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39*, 1397–1412.

[7] Braams, B. J.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces in High Dimensionality, *Int. Rev. Phys. Chem.* **2009**, *28*, 577–606.

[8] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[9] Le, H. M.; Huynh, S.; Raff, L. M. Molecular Dissociation of Hydrogen Peroxide (HOOH) on a Neural Network Ab Initio Potential Surface with a New Configuration Sampling Method Involving Gradient Fitting, *J. Chem. Phys.* **2009**, *131*, 014107.

[10] Behler, J. Constructing High-Dimensional Neural Network Potentials: A Tutorial Review, *Int. J. Quantum Chem.* **2015**, *115*, 1032.

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

[12] Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments, *Phys. Rev. B.* **2013**, *87*, 184115.

[13] Qu, C.; Yu, Q.; Van Hoozen, Jr., B. L.; Bowman, J. M.; Vargas-Hernández, R. A. Assessing Gaussian Process Regression and Permutationally Invariant Polynomial Approaches to Represent High-Dimensional Potential Energy Surfaces. *J. Chem. Theory Comput.* **2018**, *14*, 3381–3396.

[14] Laude, G.; Calderini, D.; Tew, D. P.; Richardson, J. O. Ab Initio Instanton Rate Theory Made Efficient using Gaussian Process Regression, *Faraday Discuss.* **2018**, *212*, 237–258.

[15] Dawes, R.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: A Strategy for Efficient Automatic Data Point Placement in High Dimensions, *J. Chem. Phys.* **2008**, *128*, 084107.

[16] Majumder, M.; Ndengue, S. A.; Dawes, R. Automated Construction of Potential Energy Surfaces, *Mol. Phys.* **2016**, *114*, 1–18.

[17] Jiang, B.; Li, J.; Guo, H. Potential Energy Surfaces from High Fidelity Fitting of Ab Initio Points: The Permutation Invariant Polynomial-Neural Network Approach. *Int. Rev. Phys. Chem.* **2016**, *35*, 479−506.

[18] Kidwell, N. M.; Li, H.; Wang, X.; Bowman, J. M.; Lester, M. I. Unimolecular Dissociation Dynamics of Vibrationally Activated CH_3CHOO Criegee Intermediates to OH Radical Products. *Nat. Chem.* **2016**, *8*, 509−504.

[19] Jasper, A. W.; Harding, L. B.; Knight, C.; Georgievskii, Y. Anharmonic Rovibrational Partition Functions at High Temperatures: Tests of Reduced-Dimensional Models for Systems with up to Three Fluxional Modes, *J. Phys. Chem. A* **2019**, *123*, 6210–6228.
[20] Nandi, A.; Qu, C.; Bowman, J. M. Full and Fragmented Permutationally Invariant Polynomial Potential Energy Surfaces for Trans and Cis N-methyl Acetamide and Isomerization Saddle Points. *J. Chem. Phys.* **2019**, *151*, 084306.
[21] Jiang, B.; Li, J.; Guo, H. High-Fidelity Potential Energy Surfaces for Gas-Phase and Gas–Surface Scattering Processes from Machine Learning, *J. Phys. Chem. Lett.* **2020**, *11*, 5120–5131.
[22] Conte, R.; Qu, C.; Houston, P. L.; Bowman, J. M. Efficient Generation of Permutationally Invariant Potential Energy Surfaces for Large Molecules. *J. Chem. Theory Comput.* **2020**, *16*, 3264–3272.
[23] Koner, D.; Meuwly, M. Permutationally Invariant, Reproducing Kernel-Based Potential Energy Surfaces for Polyatomic Molecules: From Formaldehyde to Acetone. *J. Comp. Theory Comput.* **2020**, *16*, 5474–5484.
[24] Dawes, R.; Ndengue, S. A. Single- and Multireference Electronic Structure Calculations for Constructing Potential Energy Surfaces. *Int. Rev. Phys. Chem.* **2016**, *35*, 441–478.
[25] Babin, V.; Leforestier, C.; Paesani, F. Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient, *J. Chem. Theory Comput.* **2013**, *9*, 5395–5403.
[26] Tibor, G.; Czakó, G. Automating the Development of High-Dimensional Reactive Potential Energy Surfaces with the ROBOSURFER Program System, *J. Chem. Theory Comput.* **2020**, *16*, 51–66.
[27] Reddy, S. K.; Straight, S. C.; Bajaj, P.; Pham, C. H.; Riera, M.; Moberg, D. R.; Morales, M. A.; Knight, C.; Götz, A. W.; Paesani, F. On the Accuracy of the MB-pol Many-Body Potential for Water: Interaction Energies, Vibrational Frequencies, and Classical Thermodynamic and Dynamical Properties from Clusters to Liquid Water and Ice, *J. Chem. Phys.* **2016**, *145*, 194504.
[28] Riera, M.; Yeh, E. P.; Paesani, F. Data-Driven Many-Body Models for Molecular Fluids: CO_2/H_2O Mixtures as a Case Study, *J. Chem. Theory Comput.* **2020**, *16*, 2246–2257.
[29] Qu, C.; Conte, R.; Houston, P. L.; Bowman, J. M. Full-Dimensional Potential Energy Surface for Acetylacetone and Tunneling Splittings. *Phys. Chem. Chem. Phys.* **2021**, *23*, 7758.
[30] Dral, P. O.; Owens, A.; Yurchenko, S. N.; Thiel, W. Structure-Based Sampling and Self-Correcting Machine Learning for Accurate Calculations of Potential Energy Surfaces and Vibrational Levels. *J. Chem. Phys.* **2017**, *146*, 244108.
[31] Qu, C.; Bowman, J. M. High-Dimensional Fitting of Sparse Datasets of CCSD(T) Electronic Energies and MP2 Dipole Moments, Illustrated for the Formic Acid Dimer and Its Complex IR Spectrum. *J. Chem. Phys.* **2018**, *148*, 241713.
[32] Jasper, A. W.; Davis, M. J. Parameterization Strategies for Intermolecular Potentials for Predicting Trajectory-Based Collision Parameters. *J. Phys. Chem. A* **2019**, *123*, 3464–3480.
[33] Paukku, Y.; Yang, K. R.; Varga, Z.; Truhlar, D. G. *J. Chem. Phys.* **2013**, *139*, 044309, **2014**, 140, 019903(E).
[34] Jasper, A. W.; Kamarchik, E.; Miller, J. A.; Klippenstein, S. J. First-Principles Binary Diffusion Coefficients for H, H_2, and Four Normal Alkanes + N_2. *J. Chem. Phys.* **2014**, *141*, 124313 ().

[35] Conte, R.; Qu, C.; Bowman, J. M. Permutationally Invariant Fitting of Many-Body, Non-Covalent Interactions with Application to Three-Body Methane–Water–Water. *J. Chem. Theory Comput.* **2015**, *11*, 1631.
[36] Li, J.; Varga, Z.; Truhlar, D. G.; Guo, H. Many-Body Permutationally Invariant Polynomial Neural Network Potential Energy Surface for N_4. *J. Chem. Theory Comput.* **2020**, *16*, 4822–4832.
[37] Rheinecker, J.; Bowman, J. M. The calculated infrared spectrum of Cl^-H_2O using a new full dimensional *ab initio* potential energy surface and dipole moment surface, *J. Chem. Phys.* **2006**, *125*, 133206.
[38] Pandey, A.; Poirier, B. An Algorithm to Find (and Plug) "Holes" in Multi-Dimensional Surfaces, *J. Chem. Phys.* **2020**, *152*, 214102.
[39] Yang, K. R.; Varga, Z.; Truhlar, D. G. PIPFit, version 2015, University of Minnesota, Minneapolis, 2015. comp.chem.umn.edu/pipfit
[40] Kamarchik, E.; Braams, B.; Krylov, A. I.; Bowman, J. M. ezPES library supported by iOpenShell Center. iopenshell.usc.edu/downloads/ezpes/
[41] Xie, Z.; Bowman, J. M. Permutationally Invariant Polynomial Basis for Molecular Energy Surface Fitting via Monomial Symmetrization. *J. Chem. Theory Comput.* **2010**, *6*, 26–34.
[42] Abbott, A. S.; Turney, J. M.; Zhang, B.; Smith, D. G. A.; Altarawy, D.; Schaefer, III, H. F. PES-Learn: An Open-Source Software Package for the Automated Generation of Machine Learning Models of Molecular Potential Energy Surfaces, *J. Chem. Theory Comput.* **2019**, *15*, 4386–4398.
[43] Rocha, C. M. R.; Varandas, A. J. C. A General Code for Fitting Global Potential Energy Surfaces via CHIPR method: Direct-Fit Diatomic and Tetratomic Molecules, *Comput. Phys. Comm.* **2021**, *258*, 107556.
[44] MSA Software with Gradients. https://github.com/szquchen/MSA-2.0
[45] Moberg, D. R.; Jasper, A. W. PIPPy, Argonne National Laboratories, 2021. github.com/auto-mech/pippy
[46] Špirko, V.; Jensen, P.; Bunker, P. R.; Čejchan, A. The Development of a New Morse-Oscillator Based Rotation−Vibration Hamiltonian for H_3^+. *J. Mol. Spectrosc.* **1985**, *112*, 183−202.
[47] Moberg, D. R.; Jasper, A. W.; Davis, M. J., "Potential Energy Surface Expansions with Improved Scaling Using Dictionary Learning Based on Multipass Greedy Selection," in preparation.
[48] Jasper, A. W.; Pelzer, K. M.; Miller, J. A.; Kamarchik, E.; Harding, L. B.; Klippenstein, S. J. Predictive A Priori Pressure Dependent Kinetics, *Science* **2014**, *346*, 1212–1215.
[49] Jasper, A. W. "Third-body" Collision Parameters for Hydrocarbons, Alcohols, and Peroxides and an Effective Internal Rotor Approach for Estimating Them, *Int. J. Chem. Kinet.* **2020**, *52*, 387–402.
[50] Wang, Y.; Shepler, B.; Braams, B.; Bowman, J. M. Flexible, ab initio potential, and dipole moment surfaces for water. I. Tests and applications for clusters up to the 22-mer. *J. Chem. Phys.* **2011**, *134*, 094509.
[51] Jin, Z.; Braams, B. J.; Bowman, J. M. An ab Initio Based Global Potential Energy Surface Describing $CH_5^+ \rightarrow CH_3^+ + H_2$. *J. Phys. Chem. A* **2006**, *110*, 1569-1574.
[52] Xie, Z.; Braams, B. J.; Bowman, J. M. Ab initio global potential-energy surface for $CH_5^+ \rightarrow CH_3^+ + H_2$. J. *Chem. Phys.* **2005**, *122*, 224307.

[53] Jasper, A. W.; Miller, J. A. Lennard-Jones Parameters for Combustion and Chemical Kinetics Modeling from Full-Dimensional Intermolecular Potentials, *Combust. Flame* **2014**, *161*, 101–110.
[54] Brown, N. J.; Bastian, L. A. J.; Price, P. N. Transport Properties for Combustion Modeling, *Prog. Energy Combust. Sci.* **2011**, *37*, 565–582.
[55] Lenzer, T.; Luther, K.; Troe, J.; Gilbert, R. G.; Lim, K. F. Trajectory Smulations of Collisional Energy Transfer in Highly Excited Benzene and Hexafluorobenzene. *J. Chem. Phys.* **1995**, *103*, 626.
[56] Bernshtein, V.; Oref, I. Collisional Energy Transfer in Polyatomic Molecules in the Gas Phase. *Isr. J. Chem.* **2007**, *47*, 205.
[57] Jasper, A. W.; Miller, J. A. Collisional Energy Transfer in Unimolecular Reactions: Direct Classical Trajectories for $CH_4 \rightarrow CH_3 + H$ in Helium. *J. Phys. Chem. A* **2009**, *113*, 5612–5619.
[58] Barker, J. R.; Weston, R. E. Collisional Energy Transfer Probability Densities $P(E, J; E', J')$ for Monatomics Colliding with Large Molecules. *J. Phys. Chem. A* **2010**, *114*, 10619–10633, **2012**, 116, 799(E).
[59] Conte, R.; Houston, P. L.; Bowman, J. M. Trajectory Study of Energy Transfer and Unimolecular Dissociation of Highly Excited Allyl with Argon. *J. Phys. Chem. A* **2014**, *118*, 7742.
[60] Matsugi, A. Collision Frequency for Energy Transfer in Unimolecular Reactions. J. *Phys. Chem. A* **2018**, *8*, 1972–1985.
[61] Lendvay, G. In: Roberton, S. H. ed. *Unimolecular Kinetics: Parts 2 and 3: Collisional Energy Transfer and the Master Equation, Vol. 43.* Elsevier: Amsterdam, the Netherlands, 2019, pp. 110–272.
[62] Troe, J. Theory of Thermal Unimolecular Reactions at Low Pressures. 1. Solutions of the Master Equation. *J. Chem. Phys.* **1977**, *66*, 4745−4757.
[63] Tardy, D. C.; Rabinovitch, B. S. Intermolecular Vibrational Energy Transfer in Thermal Unimolecular Systems. *Chem. Rev.* **1977**, *77*, 369–408.
[64] Jasper, A. W.; Miller, J. A.; Klippenstein, S. J. The Collision Efficiency of Water in the Unimolecular Reaction CH_4 $(+H_2O) \rightarrow CH_3 + H$ $(+H_2O)$: One-Dimensional and Two-Dimensional Solutions of the Low-Pressure-Limit Master Equation. *J. Phys. Chem. A* **2013**, *117*, 12243–1255.
[65] (a) Labbe, N. J.; Jasper, A. W.; Miller, J. A.; Klippenstein, S. J.; Ruscic, B.; Sivaramakrishnan, R., in preparation. See also (b) N. J. Labbe, A. W. Jasper, J. A. Miller, S. J. Klippenstein, B. Ruscic, and R. Sivaramakrishnan, High Accuracy Thermochemical Kinetics for $H + CH_3$ (+M) ⇄ CH_4 (+M), Paper 1A01, 10th US National Combustion Meeting, College Park, MD, April 23−26, 2017.
[66] Jasper, A. W. Microcanonical Rate Constants for Unimolecular Reactions in the Low-Pressure Limit. *J. Phys. Chem. A* **2020**, *124*, 1205–1226.
[67] Rousso, A. C.; Hansen, N.; Jasper, A. W.; Ju, Y. Identification of the Criegee Intermediate Reaction Network in Ethylene Ozonolysis: Impact on Energy Conversion Strategies and Atmospheric Chemistry. *Phys. Chem. Chem. Phys.* **2019**, *21*, 7341–7357.
[68] Jasper, A. W.; Sivaramakrishnan, R.; Klippenstein, S. J. Nonthermal Rate Constants for CH_4* $+ X \rightarrow CH_3 + HX$, X = H, O, OH, and O_2. *J. Chem. Phys.* **2019**, *150*, 114112.

[69] Tao, Y.; Jasper, A. W.; Georgievskii, Y.; Klippenstein, S. J.; Sivaramakrishnan, R. Termolecular Chemistry Facilitated by Radical-Radical Recombinations and its Impact on Flame Speed Predictions. *Proc. Combust. Inst*., online. DOI: https://doi.org/10.1016/j.proci.2020.06.342

[70] Truhlar, D. G.; Muckerman, J. T. In: Bernstein, R. D., ed., *Atom-Molecule Collision Theory: A Guide for the Experimentalist*, Plenum: New York, 1979, pp. 505–566.

[71] Raff, L. M.; Thompson, D. L. In: Baer, M., ed. *Theory of Chemical Reaction Dynamics, Vol. 3*, CRC Press: Boca Raton, 1985, pp. 1–121.

[72] Steckler, R.; Dykema, K. J.; Brown, F. B.; Hancock, G. C.; Truhlar, D. G.; Valencich, T. A Comparative Study of Potential Energy Surfaces for $CH_3 + H_2 = CH_4 + H$. J. *Chem. Phys.* **1987**, *87*, 7024–7035.

[73] Wu, G.-S.; Schatz, G. C.; Lendvay, G.; Fang, D.-C.; Harding, L. B. A New Potential Energy Surface and Quasiclassical Trajectory Study of $H + H_2O$. *J. Chem. Phys.* **2000**, *113*, 3150.

[74] Zhang, D. H.; Collins, M. A.; Lee, S.-Y. First-Principles Theory for the $H + H_2O$, D_2O Reactions. *Science* **2000**, *290*, 961–963.

[75] Fu, B.; Kamarchik, E.; Bowman, J. M. Quasiclassical Trajectory Study of the Postquenching Dynamics of OH by H_2/D_2 on a Global Potential Energy Surface. *J. Chem. Phys.* **2010**, *113*, 164306.

[76] Li, J.; Jiang, B.; Guo, H. Permutation Invariant Polynomial Neural Network Approach to Fitting Potential Energy Surfaces. II. Four-Atom Systems. *J. Chem. Phys.* **2013**, *139*, 204103

[77] Li, J.; Chen, J.; Zhao, Z.; Xie, D.; Zhang, D. H.; Guo. H. A permutationally invariant full-dimensional *ab initio* potential energy surface for the abstraction and exchange channels of the $H + CH_4$ system. *J. Chem. Phys.* **2015**, *142*, 204302.

[78] Zhao, B.; Sun, Z.; Guo, H. A Reactant-Coordinate-Based Wave Packet Method for Full-Dimensional State-to-State Quantum Dynamics of Tetra-Atomic Reactions: Application to Both the Abstraction and Exchange Channels in the $H + H_2O$ reaction. *J. Chem. Phys.* **2017**, *144*, 064104.

[79] Fu, B.; Shan, X.; Zhang, D. H.; Clary, D. C. Recent Advances in Quantum Scattering Calculations on Polyatomic Bimolecular Reactions. *Chem. Soc. Rev.* **2017**, *46*, 7625–7649.

[80] Ellerbrock, R.; Manthe, U. $H + CH_4 \rightarrow H_2 + CH_3$ Initial State-Selected Reaction Probabilities on Different Potential Energy Surfaces, *Chem. Phys.* **2017**, *482*, 106–112.

[81] Chen, J.; Xu, X.; Xu, X.; Zhang, D. H. A global potential energy surface for the $H_2 + OH \leftrightarrow H_2O + H$ reaction using neural networks. *J. Chem. Phys.* **2013**, *138*, 154301.

[82] Xu, X.; Chen, J.; Zhang, D. H. Global Potential Energy Surface for the $H+CH_4 \leftrightarrow H_2+CH_3$ Reaction using Neural Networks, *Chin. J. Chem. Phys.* **2014**, *373*, 373.

[83] Celani, P.; Werner, H.-J. Multireference Perturbation Theory for Large Restricted and Selected Active Space Reference Wave Functions. *J. Chem. Phys.* **2000**, *112*, 5546–5557.

[84] Dawes, R.; Jasper, A. W.; Tao, C.; Richmond, C.; Mukarakate, C.; Kable, S. H.; Reid, S. A. Theoretical and Experimental Spectroscopy of the S2 State of CHF and CDF: Dynamically Weighted Multireference Configuration Interaction Calculations for High-Lying Electronic States. *J. Phys. Chem. Lett.* **2010**, *1*, 641–646.

[85] Corchado, J. C.; Bravo, J. L.; Espinosa-Garcia, J. The Hydrogen Abstraction Reaction $H + CH_4$. I. New Analytical Potential Energy Surface Based on Fitting to *Ab Initio* Calculations. *J. Chem. Phys.* **2009**, *130*, 184314.

[86] Espinosa-Garcia, J.; Garcia-Chamorro, M.; Corchado, J. C. The Hydrogen Abstraction Reaction $H + C_2H_6 \rightarrow H_2(v,j) + C_2H_5$. Part I. A Full-Dimensional Analytical Potential Energy Surface Based on Ab Initio Calculations, *Phys. Chem. Chem. Phys.* **2019**, *21*, 13347–13355.
[87] Davis, M. J.; Liu, W.; Sivaramakrishnan, R. Global Sensitivity Analysis with Small Sample Sizes: Ordinary Least Squares Approach, *J. Phys. Chem A* **2017**, *121*, 553−570.
[88] Chen, X.; Goldsmith, C. F. Accelerating Variational Transition State Theory via Artificial Neural Networks, *J. Phys. Chem. A* **2020**, *124*, 1038–1046.
[89] Michael, J. V.; Sun, M.-C.; Sutherland, J. W.; Carroll, J. J.; Wagner, A. F. Rate Constants for $H + O_2 + M \rightarrow HO_2$ in Seven Bath Gases, *J. Phys. Chem. A* **2002**, *106*, 5297–5131.
[90] Hong, Z.; Farooq, A.; Barbour, E. A.; Davidson, D. F.; Hanson, R. K. Hydrogen Peroxide Decomposition Rate: A Shock Tube Study Using Tunable Laser Absorption of H_2O near 2.5 μm. *J. Phys. Chem. A* **2009**, *113*, 12919–12925.
[91] Maergoiz, A. I.; Nikitin, E. E.; Troe, J.; Ushakov, V. G. Trajectory and Adiabatic Channel Study of the Transition From Adiabatic to Sudden Capture Dynamics. III. Dipole-Dipole Capture. *J. Chem. Phys.* **1996**, *105*, 6277.
[92] Matsugi, A. Modeling Third-Body Effects in the Thermal Decomposition of H_2O_2, *Combust. Flame* **2021**, *225*, 444–452.
[93] Keçeli, M.; Elliott, S.; Li, Y.-P.; Johnson, M. S.; Cavallotti, C.; Georgievskii, Y.; Green, W. H.; Pelucchi, M.; Wozkiak, J. M.; Jasper, A. W.; Klippenstein, S. J. Automated Computational Thermochemistry for Butane Oxidation: A Prelude to Predictive Automated Combustion Kinetics, *Proc. Combust. Inst.* **2019**, *37*, 363–371.
[94] Glarborg, P.; Hashemi, H.; Cheskis, S.; Jasper, A. W. On the Rate Constant for $NH_2 + HO_2$ and Third Body Collision Efficiencies for NH_2+H(+M) and NH_2+NH_2(+M), *J. Phys. Chem. A* **2021**, *125*, 1505–1516.
[95] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran: The Art of Scientific Computing, 2nd ed.* Cambridge University Press: USA, 1993.
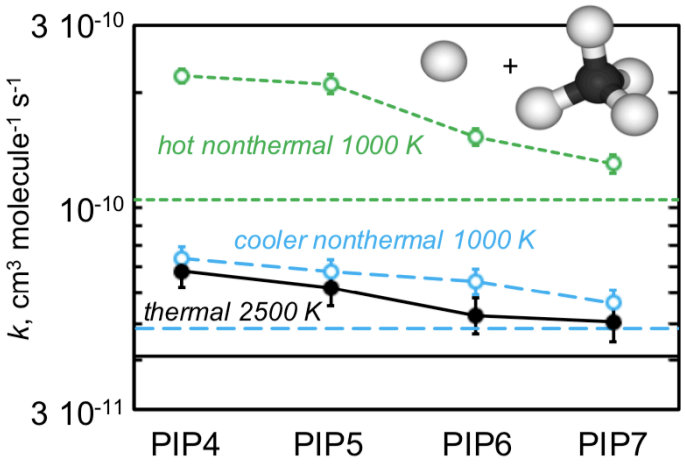
Table of Contents Figure