Prepared for *JPCA*
April 3, 2019

Parameterization Strategies for Intermolecular Potentials for Predicting Trajectory-Based Collision Parameters

Ahren W. Jasper* and Michael J. Davis

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

Abstract. The accuracy of separable strategies for constructing full-dimensional potential energy surfaces for collisional energy transfer and collision rate calculations is studied systematically for three alcohols (A = methanol, ethanol, and butanol) and three bath gases (M = Ar, N_2, and H_2O). The fitting efficiency (defined as the number of *ab initio* data required to achieve parameterizations of a desired accuracy) is quantified for both pairwise (Buckingham or "exp6") and nonpairwise (permutationally invariant polynomials, PIPs, of Morse variables) functional forms, for four sampling strategies, and as a function of the complexity and anisotropy of the interaction potential. We find that convergence with respect to the number of sampled *ab initio* data is largely independent of the choice of functional form but instead varies nearly linearly with the number of adjustable parameters and depends strongly on the sampling strategy. Specifically, the use of biased Sobol quasirandom sampling is ~7x more efficient than using unbiased pseudorandom sampling, on average, requiring just ~3 computed *ab initio* energies per adjustable fitting parameter. The pairwise exp6 functional form is shown to provide accurate and transferable parameterizations for M = Ar but is unable to accurately describe alcohol interactions with M = N_2 and H_2O. The nonpairwise PIP functional form, which is systematically improvable, can produce separable parameterizations with arbitrarily small fitting errors. However, these can suffer from overfitting, which is demonstrated using dynamics calculations of collision parameters for a large number of exp6 and PIP parametrizations. The tests described here validate a robust strategy for automatically generating A + M potential energy surfaces with minimal human intervention, including a quantifiable out of sample metric for judging the accuracy of the fitted surface. We further analyze this set of automatically generated potential energy surfaces to identify areas where more sophisticated fitting strategies may be desired, including pruning of the PIP expansions for large systems and improved sampling strategies more closely coupled with the description of the functional form.

*E-mail: ajasper@anl.gov

1. INTRODUCTION

Collision rates and energy-transfer efficiencies of so-called "third bodies" are fundamental aspects of gas phase reactivity, controlling pressure dependence in elementary kinetics[1–4] and molecular transport.[5–7] Computing collision information using classical trajectories has a long history (for representative work from several groups see Refs. 8–16), including some of the first polyatomic trajectory calculations.[9,17,18] Trajectory studies have provided important mechanistic details of collisional energy transfer and have been useful for interpreting experimental results.[16,19–22]

Recently, quantitative predictions of pressure-dependent kinetics have appeared using trajectory-based collision parameters,[23,24] and quantum[25,26] and classical[27] scattering calculations have been used to accurately predict diffusion coefficients and other transport properties. The quantitative accuracy of these recent *a priori* calculations can be attributed in part to the use of *ab initio*–based fitted potential energy surfaces and in particular to accurate descriptions of the intermolecular (interaction) potential,[28] which largely controls collision outcomes.

A great number of strategies have been devised for constructing potential energy surfaces.[29–42] Using a relatively simple approach, we previously built a family of full-dimensional potential energy surfaces for nonreactive C_xH_y + M collisions, where M is one of several common atomic and diatomic baths.[43,44] The parameterizations assumed separable and pairwise intermolecular atom–atom interactions that depended only on the identities of the interacting atoms (with small adjustments sometimes made for unsaturated C atoms[45]). The parameters in these fits did not depend on the number of

atoms or on the internal structure of C_xH_y, and parameters obtained for CH_4 + M were assumed to be transferable to larger hydrocarbons.

The good accuracy of these assumptions for constructing potential energy surfaces for computing collision properties was confirmed in several subsequent dynamics and kinetics studies. For example, pressure-dependent rate constants based on trajectory results using these potential energy surfaces agreed with experimental results to better than a factor of two when used with simple energy transfer and state counting models[46,47] and to better than 25% when used with more detailed models.[24] In other work, diffusion coefficients calculated using trajectories and these same potential energy surfaces agreed with measured values to better than 1%.[27]

A major goal of the present work is to systematically test this potential energy surface approach for more challenging systems, specifically for oxygenated systems (here, alcohols) and the polyatomic bath gas M = H_2O. These systems feature larger anisotropy than the hydrocarbon species and the atomic and diatomic baths considered previously, and we therefore include consideration of flexible nonpairwise functional forms as well as the simpler pairwise functional forms.

Five systems are studied: CH_3OH + Ar, CH_3OH + N_2, CH_3OH + H_2O, C_2H_5OH + Ar, and C_4H_9OH + Ar. For each system, several pairwise and nonpairwise functional forms and parameterization strategies are applied, with the goal of quantifying the "fitting efficiency" of the different approaches. Here, fitting efficiency is defined to emphasize minimizing the number of *ab initio* data required to generate analytic representations of a desired accuracy, but a more comprehensive and useful metric might include consideration of the costs to generate and to evaluate the fitted potential energy surface as

well. In addition to fitting efficiency, we consider the accuracy of the resulting parameterizations, which we characterize both in terms of fitting errors using an "out of sample" test set of *ab initio* energies and dynamical errors in the desired collision parameters.

As part of this work, we introduce systematic strategies for potential energy surface parameterization, for demonstrating convergence with respect to sampled energies, and for defining interpretable and quantifiable error functions for the resulting analytic representations. One goal of these efforts is to produce strategies that may be applied without human judgment and are therefore automated. The consideration of A + M intermolecular systems allows for robust tests of these strategies on relatively simple systems, as well as systematic studies of the scaling of fitting efficiency and accuracy with respect to both the size of the alcohol and the size of the bath gas, which in turn determines the dimensionality of the interaction potential.

This paper is organized as follows. In Sec. 2, we describe a separable approximation for constructing A + M potential energy surfaces, and we describe the pairwise and nonpairwise functional forms used to represent the interaction potential. This section also includes descriptions of other details of the parameterizations, including: the *ab initio* methods used to calculate energies, descriptions of the four sampling strategies tested here, the definition of the out of sample fitting error, and details of the collision parameter calculations. In Sec. 3.1, we provide a validation study of the *ab initio* method used to generate the training and out of sample test sets. In Sec. 3.2, the specific functional forms considered are introduced and discussed. In Sec. 3.3, these functional forms are used to test the four sampling strategies for the CH_3OH + Ar

system. We define and demonstrate the utility of the converged prediction error f_∞ for identifying accurate functional forms and the fitting efficiency M^* for quantifying the effectiveness of the sampling strategies. In Sec. 3.4, the tests of Sec. 3.3 are repeated for larger alcohols and for larger bath gases (and thus for higher-dimensional interaction potentials), and trends in these two series are discussed. In Sec. 3.5, results of dynamical collision parameter calculations are presented for a large number of the parameterizations developed in Secs. 3.3 and 3.4, and the relationship between the fitting errors f_∞ and dynamical errors is discussed. Section 4 summarizes our major conclusions and identifies areas for future work.

2. THEORY

Background. The potential energy surface (PES) describing an A + M collision can be written

$$V(\mathbf{R}) = V_A(\mathbf{R}_A) + V_M(\mathbf{R}_M) + V_I(\mathbf{R}), \quad (1)$$

where here the coordinates are chosen to be all of the atom–atom distances $\mathbf{R}$. The collections of intramolecular atom–atom distances for the unimolecular species A and for the bath gas M are labeled $\mathbf{R}_A$ and $\mathbf{R}_M$, respectively, and the remaining intermolecular atom–atom distances are labeled $\mathbf{R}_I$ such that $\mathbf{R} = (\mathbf{R}_A, \mathbf{R}_M, \mathbf{R}_I)$. The term "coordinates" is used to refer to the (typically redundant) set of $n(n–1)/2$ atom–atom distances, where n is the number of atoms, and "dimensions" is used to refer to the (typically) $3n–6$ internal independent geometric parameters upon which V depends. In eq 1, V_A and V_M are lower-dimensional PESs describing the isolated species A and M, and V_I is their interaction potential, which has the same dimensionality as the full system.

At long range and more generally whenever A and M are only weakly interacting, V_I is a more sensitive function of intermolecular distances than intramolecular ones. In such a situation, the terms $\frac{d}{d\mathbf{R}_A}V_I$ and $\frac{d}{d\mathbf{R}_M}V_I$ are small, and the total force can be approximated

$$-\frac{d}{d\mathbf{R}}V \approx -\frac{d}{d\mathbf{R}_A}V_A - \frac{d}{d\mathbf{R}_M}V_M - \frac{d}{d\mathbf{R}_I}V_I . \quad (2)$$

This assumption defines a "separable" approximation for V_I and greatly simplifies the construction of PESs. This is a natural approximation for describing intermolecular potentials for collisional energy transfer; it has been employed repeatedly over the decades, from early studies of triatomics and rare gas colliders[17,18,48] to recent studies of gas-surface collisions involving complex molecules and ions.[49,50] The challenging problem of fitting V in its full dimensionality is replaced with the more manageable one of fitting V_A, V_M, and V_I separately, with the fit of V_I requiring the consideration of no more than six dimensions: the center-of-mass separation and the up to five angles defining the relative orientations of A and M. Here, A is always a nonlinear polyatomic alcohol, and the dimensionality of V_I is 3, 5, and 6 for M = Ar, N_2, and H_2O, respectively.

The separable approximation of eq 1 adopted here is

$$V(\mathbf{R}) = V_A(\mathbf{R}_A) + V_M(\mathbf{R}_M) + V_I(\mathbf{R}_I, \mathbf{R}_A=\mathbf{R}_A^{(e)}, \mathbf{R}_M=\mathbf{R}_M^{(e)}), \quad (3)$$

where $\mathbf{R}_A^{(e)}$ and $\mathbf{R}_M^{(e)}$ are fixed reference geometries upon which V_I depends parametrically, and "e" refers to the equilibrium geometries of A and M. By assumption, V_I is weakly dependent on the reference geometry. Because V_I is written as a function of atom–atom distances, the resulting intermolecular potential depends indirectly on $\mathbf{R}_A$ and $\mathbf{R}_M$ inasmuch as physical motions that change $\mathbf{R}_M$ and $\mathbf{R}_A$ will typically also result in changes in $\mathbf{R}_I$. The good accuracy of V_I when evaluated at geometries other than $\mathbf{R}_A^{(e)}$ and

$\mathbf{R}_{\mathrm{M}}^{(\mathrm{e})}$ is required for the accurate use of separable PESs in trajectory calculations, where A and M are not rigid. This type of transferability will be tested in Sec. 3.5. Both pairwise and nonpairwise descriptions of V_{I} are considered and discussed next.

The Separable Pairwise Approximation for V_I. The separable pairwise intermolecular potentials tested here have the form

$$V_{\mathrm{I}} = \sum_i A_i \exp(-R_i / B) - C_i^6 / (R_i^6 + D_i^6), \tag{4}$$

where R_i is the i^{th} intermolecular atom–atom distance, and each atom–atom interaction is described using a four-parameter cutoff Buckingham (exp6) expression. The number of unique atom–atom pairs included in the fit is labeled N_{u}, and the number of adjustable parameters in eq 4 is therefore $4N_{\mathrm{u}}$. Symmetry is enforced in eq 4 by setting the parameters of chemically equivalent intermolecular atom pairs to be equal. Additional constraints are sometimes included to further reduce the number of parameters, where chemically-non-equivalent H atoms are treated as equivalent. For example, there are four chemically distinct atom types in CH_3OH and therefore the maximum value of N_{u} is 4 for CH_3OH + Ar, and we considered fits for CH_3OH + Ar using eq 4 with 16 adjustable parameters, four for each unique intermolecular atom–atom interaction. We also considered parameterizations where the hydroxyl H atom was artificially constrained to have the same interaction parameters with Ar as the three chemically equivalent methyl ones, and in such a fit N_{u} = 3 and there were 12 adjustable parameters. In this paper, pairwise functional forms are labeled exp6(N_{u}).

The Separable Nonpairwise Approximation for V_I. Equation 4 is not flexible enough to accurately describe some complex interaction potentials, and fits obtained using eq 4 are not readily systematically improvable. We therefore also considered a

more flexible, nonpairwise approach where the interaction potential is expanded using permutationally invariant polynomials[31] (PIPs) of Morse variables,[51]

$$V_{\mathrm{I}} = \sum_{\beta} c_{\beta} \prod_{i} \exp(-R_i / 1\,\text{Å})^{\xi_i^{(\beta)}}, \tag{5}$$

each basis function, β, is distinguished by its unique set of exponents $\xi_i^{(\beta)}$, and symmetry-equivalent terms in the expansion were forced to have the same coefficients c_β. The size of the basis was controlled by restricting the range of the exponent on each factor from 0 to ξ_{factor} as well as their sum, which is the total order of the product, from 0 to ξ_{order}. The separable nonpairwise functional forms are labeled PIP$\xi_{\text{factor}}\xi_{\text{order}}(N_{\text{u}})$, where N_{u} again is the number of symmetry-unique pairs of intermolecular atoms. In some parameterizations N_{u} was determined by the true symmetry of the interaction potential, and in other parameterizations it was made artificially smaller by forcing similar, but not chemically-equivalent, atoms (e.g., all H atoms) to share interaction parameters. The use of artificially reduced values of N_{u} can significantly reduce the number of independent parameters in the expansion without significant loss of fitting accuracy, as demonstrated below.

A significant advantage of the nonpairwise PIP functional form in eq 5 over the pairwise exp6 functional form in eq 4 is that it is systematically improvable via increases in ξ_{max} and ξ_{order}, and it is therefore able to accurately treat interaction potentials of arbitrary complexity. Furthermore, the linear PIP parameters in eq 5 are readily determined via least-squares optimization, whereas the nonlinear exp6 parameters in eq 4 require more complex nonlinear fitting approaches, as described below. When accurate exp6 parameterizations can be obtained, however, they may be preferred over PIP

parameterizations, as: (1) exp6 parameterizations likely require relatively fewer *ab initio* calculations of the interaction potential to accurately constrain its (typically fewer) parameters, and (2) one can anticipate that the physically-motivated exp6 functional form produces transferable parameters, both with respect to the choice of reference geometries in eq 3 as well as the identity of the alcohol. That is, exp6 parameterizations for CH_3OH + M may be accurate for larger alcohols as well, thus greatly facilitating the study of collision parameters for large numbers of alcohols. This type of transferability was tested previously for hydrocarbons[24,43–47] and is tested here for alcohols in Sec. 3.5. The PIP functional forms, in contrast, generally feature different terms in their expansions for different alcohols, and this strategy is therefore less amenable to transferable parameterizations. Section 3 has a further examination of these tradeoffs.

Intramolecular PESs for A and M. In the present applications, V_A does not need to be able to describe bond breaking. Molecular mechanics parameterizations are therefore suitable, and the parameterizations distributed with Tinker[52] were tested. The effect of making small system-specific adjustments to equilibrium distances and force constants to improve rotational constants and vibrational frequencies was examined, and, as expected for common organic molecules such as alcohols, the unaltered Amber force field[53] gave good descriptions of the species considered here. Collision parameters have been shown to be relatively insensitive to the accuracy of V_A,[10,12] and so the sensitivity to this part of the PES was not included in the present tests. For V_M, we used our previous fit[44] for N_2 and Partridge and Schwenke's fit[54] for H_2O.

Electronic Structure Theory. The quantum chemistry method used to compute interaction energies was validated as follows. Convergence with respect to basis set and

level of electron correlation was tested by comparing energies obtained using CCSD(T) and MP2, with and without the counterpoise correction (cc), and for four different complete basis set (CBS) extrapolation schemes. We employed a two point CBS formula[55] and extrapolated based on cc-pVDZ and cc-pVTZ energies (CBS(d,t)), aug'-cc-pVDZ and aug'-cc-pVTZ energies (CBS(a'd,a't)), or aug-cc-pVDZ and aug-cc-pVTZ energies (CBS(ad,at)). The above basis set notation indicates Dunning's basis sets with (aug-) and without (no prefix) additional diffuse functions, as usual, while the "aug'-" notation indicates the use of augmented basis sets for non-H atoms and un-augmented basis sets for H atoms.[56] For a few geometries, we also considered CBS extrapolations using the larger aug-cc-pVTZ and aug-cc-pVQZ basis sets (CBS(at,aq)). As reported in Sec. 3.1 and in more detail as supporting information, good basis set convergence was achieved using two-point double-ζ/triple-ζ extrapolations, and therefore more complicated extrapolation schemes were not tested.

Fitting Error and Out of Sample Error. For all of the PESs generated here, parameters were determined by minimizing the weighted least-squares error

$$f^2 = \sum_k^M w_k^2 \left(V(\mathbf{R}_k) - \widehat{V}_k \right)^2 / \sum_k^M w_k^2 \ , \tag{6}$$

where M is the number of *ab initio* energies included in the training set, k labels the computed *ab initio* energies $\widehat{V}_k$ and the associated geometries $\mathbf{R}_k$, and w_k is a weight. The choice of w_k is one way to control the distribution of errors in the fitted PES. When w_k = 1 for all k, for example, the optimizations would tend to give fits with similar *absolute* errors throughout the potential. This would likely be a poor choice in the current applications, as in such a fit a small number of high energies deep in the repulsive region

could "spoil" the fit in the weakly bound van der Waals region. In fact, errors in the van der Waals region could be made arbitrarily large by including more and more high energy data in the training set. Here we would prefer the fits to have more or less an even distribution of *relative* errors throughout the PES, as, for example, a 50 cm^{-1} error deep in the repulsive wall is likely to have little impact on the accuracy of computed collision parameters, while an error of that same magnitude could alter the anisotropy in the van der Waals region. We therefore set

$$w_k = v / (\widehat{V}_k - \min\{\widehat{V}_k\} + v), \tag{7}$$

where v is a parameter and $\min\{\widehat{V}_k\}$ is the minimum sampled *ab initio* energy. We further chose v to be 200 cm^{-1} for CH_3OH + Ar, 1600 cm^{-1} for CH_3OH + H_2O, and 350 cm^{-1} otherwise, which are close to the van der Waals well depths for these systems. This choice gives f the following useful interpretation. At low energies near the van der Waals minimum, f is representative of the *absolute* error in the fit. In this same region, f/v is approximately the relative error, and this same relative error persists throughout the PES and up to high energies. PESs produced using this approach will tolerate larger and larger absolute errors deeper and deeper into the repulsive wall and will therefore be relatively insensitive to very high energies that may be sampled.

Equation 6 defines the "fitting error" for the training data set and is one measure of the quality of the fitted surfaces. We also report an out of sample error f_{oos}, which is the result of eq 6 evaluated for a separate out of sample test set of 165 *ab initio* energies calculated along six cuts through the interaction potential. Out of sample errors, which may be called prediction errors, can provide a better representation of the expected

accuracy of the fit and provide some measure of overfitting.[57,58] Both errors are considered in detail below.

The six cuts comprising the out of sample test set were generated automatically as follows. A small set of geometries was sampled in the van der Waals region, and the lowest-energy geometry was identified. The first cut was obtained by varying the center-of-mass distance so as to include this geometry. A second cut was generated with the opposite sides of both A and M interacting and again varying the center of mass distance. The remaining four cuts were constructed to be perpendicular to the first two. For a given A + M system, the same six cuts were always used as the test set to allow for consistent comparisons among competing functional forms and sampling strategies.

Sampling Strategies for the Training Data. One of the main goals of this paper is to quantify the relative efficiencies of competing strategies for sampling the interaction potential to generate training data. For each of the five A + M systems and for several pairwise and nonpairwise functional forms, a large number of parameterizations were obtained using training data sets consisting of from M = 20 to 10,000 *ab initio* energies sampled using either pseudorandom (P) or Sobol quasirandom (S) distributions. Sobol's quasirandom sampling is one of many schemes for avoiding clustering in the sampled geometries[59] and is commonly used for calculating multidimensional integrals.[60,61] Here, we use Sobol's sequence as implemented in Numerical Recipes.[62]

In both types of sampling, the angular coordinates were sampled over all allowed values and from unbiased spatial distributions, while the remaining center-of-mass distance r was sampled from r_{min} to r_{max}. We did not attempt to optimize the minimum and maximum values of the sampled r. Instead, we used an empirical rule where r was

sampled from r_{min} = 2/3 σ_{LJ} to r_{max} = 5/2 σ_{LJ}, where σ_{LJ} is the effective (i.e., spherically averaged) Lennard–Jones diameter for the A + M interaction. Direct evaluations of V_I and the "one-dimensional optimizations" method described previously [63] were used to calculate σ_{LJ} for each of the five systems considered here, and we obtained σ_{LJ} = 3.59, 3.73, 3.61, 3.91, and 4.60 Å for CH_3OH + Ar, CH_3OH + N_2, CH_3OH + H_2O, C_2H_5OH + Ar, and C_4H_9OH + Ar, respectively.

We also tested the use of a sampling bias, where small values of r were sampled more often than large values of r. The unbiased pseudorandom and Sobol quasirandom training data sets are labeled P and S, respectively, and the biased training data sets are labeled bP and bS. The bias is described next.

Sampling Bias in the bP and bS Training Data. We first consider CH_3OH + Ar. A set of unbiased geometries for this system was generated by pseudorandomly sampling over the two angles defining relative orientation and sampling r from r_{min} to r_{max}. cc MP2/CBS(a'd,a't) interaction energies were computed at the sampled geometries. The randomly generated values of r were binned using bin sizes of 0.1 Å, and the average, maximum, and minimum energies in each bin were calculated and are shown in Fig. 1(a). Even for this relatively simple and low-dimensional interaction potential, the anisotropy in V_I is large at small r, as indicated in Fig. 1(a) by the large discrepancy in the minimum and maximum energies for $r \lesssim 4$ Å. We note that the Lennard–Jones estimate of the minimum in the interaction potential for this system is $r_{LJ} = 2^{1/6}\ \sigma_{LJ}$ = 4.0 Å, which empirically serves as a useful indicator for the approximate onset of significant anisotropic dispersion.

At r = 3.1 Å, for example, $\langle V_I \rangle$ is 2700 cm^{-1}, which is near the upper end of the range of energies relevant to high temperature collisions, but the maximum and minimum V_I vary widely and are 20000 and 100 cm^{-1}, respectively. One may anticipate that relatively more sampling is needed to sufficiently explore dynamically-relevant low-energy configurations when anisotropy is large, and this is the reason that a bias is introduced.

Figure 1(b) shows the weights defined by eq 7 for the same set of sampled energies and geometries shown in Fig. 1(a), again binned with respect to r. Center-of-mass distances where the weights are the largest are again located around r_{LJ} = 4.0 Å with the largest maximum weights appearing at somewhat shorter r and closer to σ_{LJ} = 3.6 Å. For very short r, the *average* weights are small due to the repulsive wall, but relatively more sampling is needed to find the infrequent but important low-energy orientations.

We designed the bias to sample most often in regions where the weights are largest *and* where additional sampling is needed due to significant anisotropy. Together, these considerations are represented by the ratio of the maximum weight to the average weight, which is plotted in Fig. 1(c). Also shown is a "coarse grained" version of this ratio, where σ_{LJ} and r_{LJ} were chosen as the locations of the two steps. The relative heights of the three regions in the coarse-grained bias were set to 8:4:1 to approximate the more detailed curve.

Similar plots (not shown) were made for $CH_3OH + N_2$, $CH_3OH + H_2O$, C_2H_5OH + Ar, and C_4H_9OH + Ar, and we found that coarse grained biases with the ratios 8:4:1 worked well for all three CH_3OH + M systems, while 12:6:1 and 16:8:1 were found to be better matches to the detailed curves for the larger C_2H_5OH + Ar and C_4H_9OH + Ar

systems. The increased relative weighting at short separations for the larger alcohols reflects the increased anisotropy for these systems. In all five case, the location of the two steps were well approximated by r_{LJ} and σ_{LJ}, and the coarse-grained biases were used in all subsequent calculations. The coarse-grained biasing function has the significant advantage over the detailed curves in that it can be determined without needing to first generate a set of unbiased sampled energies and geometries and is therefore readily automated.

Several of the choices made throughout this procedure could be optimized, such as the minimum and maximum values of r, the range parameter in the weighting function v, and the design of the out of sample test set. As one of the goals here is to develop automated strategies, we have chosen to apply simple but broadly applicable rules for these choices instead of optimizing them. The effects of these choices on fitting efficiency and accuracy are likely less important than the effects of sampling strategy and functional form, which are analyzed in detail here.

Parameterization Procedure. The linear PIP coefficients in eq 5 were obtained using a standard least-squares solver, providing a unique solution for a given fitting basis (ξ_{max}, ξ_{order}), sampling strategy, and set of training data. For S and bS sampling, the geometries in the data sets are ordered, and the first M data were always used. For both P and bP sampling, a large set of 10,000 *ab initio* energies were calculated, and we studied the effect of using different portions of this set of 10,000 by generating 100 sub-samples of size M and calculating the maximum, minimum, average, and standard deviation of f and f_{oos} for the resulting 100 parameterizations.

A genetic algorithm[64] (GA) was used to optimize the nonlinear exp6 parameters in eq 4. In the present implementation, the GA was not used to find the globally best set of parameters. Instead the GA found many similar-quality parameterizations with (sometimes very) different parameters, and despite the different parameters we refer to the nonlinear GA parameterizations as converged when the fitting error f and out of sample error f_{oos} are converged with respect to GA iteration.

The GA was implemented as follows: The initial ranges for each parameter were set to standard values ($\log_{10}A_i = 2 - 10$, $B_i = 0.1 - 0.4$, $C_i = 0.5 - 10$, and $D_i = 0.5 - 10$ in cm^{-1} and Å), and a first GA run consisting of 5000 "generations" was carried out. The GA was then restarted with parameter ranges set at ±10% of the optimized values from the previous run. This process was repeated for a total of 120 GA iterations. Results from the first 20 iterations were discarded to eliminate the dependence of the statistics on our choices for the initial parameter ranges, and the maximum, minimum, average, and variance of the fitting and test set errors for the 100 remaining iterations were calculated.

This procedure, which includes ~20 million evaluations of eq 6, could certainly be made more efficient. Here, though, we do not wish to test aspects of the GA, and so the present strategy was designed so that the fitting errors obtained were largely insensitive to internal choices of the GA. When using P and bP sampling, M data were selected randomly from the full set of 10,000 data for each GA iteration, while for the ordered S and bS data sets, the first M data were always used. Each parameterization differed in the random number seeds used to initialize the GA algorithm, and so even for a fixed training data set (as in S and bS sampling), the resulting nonlinear parameterizations differ from one another.

Trajectory-Based Collision Parameters. In addition to the fitting error f and the out of sample prediction error f_{oos}, the accuracies of the parameterized A + M potential energy surfaces were tested for two dynamical quantities: the Lennard-Jones collision rate Z and the average energy in deactivating collisions α.

The "one-dimensional minimization"[63] method was used to calculate Z. In this approach, random relative orientations of A + M are chosen, and, for each orientation, the center-of-mass separation is varied to determine the local well depth and the inner turning point at the energy of the asymptote. These two quantitates are averaged over several orientations to obtain estimates of the Lennard–Jones well depth and diameter, σ_{LJ}, from which Z is calculated.

Ensembles of classical trajectories were used to calculate the average energy in deactivating collisions $\alpha = \langle \Delta E_d \rangle$, which is a key parameter in some models for predicting pressure dependent kinetics. As discussed elsewhere,[14,24] α is one moment of the detailed state-to-state energy and angular momentum transfer function, and quantitative models of energy transfer require consideration of more than just α. For simplicity, we restrict attention here to α. In general, α is a function of the initial internal state of A as well as the temperature and identity of the bath gas M. As is appropriate for predicting energy transfer parameters relevant to unimolecular decomposition,[14,24] we set the initial internal energy of A to be close to the dissociation threshold (here we used 90 kcal/mol), and we sampled the initial rotational state of A and the bimolecular A + M collision parameters from thermal distributions at T. We report α rescaled to the Lennard–Jones collision rate Z, and we used a large enough value of the maximum impact parameter ($b_{max} = 8 + N_{heavy}/2$ Å, where N_{heavy} is the number of O and C atoms in

the alcohol) in the trajectory calculations to ensure that the calculated value of α did not depend on this choice. Results for T = 300–2000 K were obtained, with the most detailed results presented for 1000 K. Additional details of the trajectory calculations have been given previously. [24,44]

Comparing the performance of the fitted PESs for these two observables provides a test of a key assumption in the separable approximation defined by eq 3, namely the transferability of V_I parametrizations from the reference geometries $\mathbf{R}_A^{(e)}$ and $\mathbf{R}_M^{(e)}$ to geometries where A and/or M are distorted from these values. The calculated Lennard–Jones collision rate Z depends only on the rigid-body intermolecular PES, and so the accuracy of Z is independent of this assumption. In contrast, the trajectory calculations of α sample the full-dimensional PES, including geometries outside the domain of rigid-body geometries from which the training data sets were drawn, and therefore accurate predictions of α require small errors associated with this type of transferability and the separability assumption. We will show in Sec. 3.5 that while this assumption is often an accurate one, it can break down when large PIP basis sets are used.

3. RESULTS AND DISCUSSION

3.1. Quantum Chemistry

Interaction energies for the five systems considered here (CH_3OH + Ar, C_2H_5OH + Ar, C_4H_9OH + Ar, CH_3OH + N_2, and CH_3OH + H_2O) were computed along the cuts through V_I comprising the out of sample test sets described in Sec. 2. Figures comparing energies computed along these cuts at several levels of theory are given as supporting information, with the main conclusions from these comparisons summarized next.

For these five systems, MP2 and CCSD(T) energies are typically in good agreement with one another for a given basis set and counterpoise correction (cc) strategy. This is a useful practical result, as, for CH_3OH + Ar for example, an MP2 energy calculation requires just ~1/20th of the computational time of a CCSD(T) energy calculation using the basis sets considered here.

As may be expected for weak nonbonding interactions, the cc strategy has a larger effect on the energy than the choice of the electron correlation method. Generally, the MP2/CBS(at,aq) results agree with the smaller-basis-set MP2/CBS(a'd,a't) and MP2/CBS(ad,at) energies (differing by just a few cm^{-1} in the interaction energy) but only when the cc is applied. This suggests basis set convergence in the cc series already at the CBS(a'd,a't) level. Somewhat slower convergence with respect to basis set was found when the cc correction was not applied. An earlier study of the van der Waals interactions of CF_4 + Ar came to very similar conclusions, including a demonstration of the good performance of MP2 relative to CCSD(T).[65]

Based on these comparisons, we chose to sample the interaction potentials using the relatively modest cc MP2/CBS(a'd,a't) level of theory for all five systems. For CH_3OH + Ar, the cc MP2/CBS(a'd,a't) energies differ from the highest level of theory employed here (cc CCSD(T)/CBS(at,aq)) by 6 cm^{-1} in the van der Waals well and by less than 3% up the repulsive wall (at energies around 2500 cm^{-1} relative to the isolated fragments). Similar relative differences between the cc MP2/CBS(a'd,a't) method and the highest-level calculations performed here were found for the four other systems tested, with larger absolute differences for the stronger van der Waals interactions. Specifically, we computed the mean unsigned difference of the cc MP2/CBS(a'd,a't) energies and the

cc CCSD(T) energies with the largest basis sets for geometries in the van der Waals well to be: 12, 12, 15, and 33 cm^{-1} for C_2H_5OH + Ar, C_4H_9OH + Ar, CH_3OH + N_2, and CH_3OH + H_2O, respectively. These values provide a simple estimate of the error in the cc MP2/CBS(a'd,a't) quantum chemistry approach against which the magnitudes of the PES fitting errors discussed extensively below can be calibrated.

More definitive quantum chemistry calculations could be carried out, but this is not the goal of the present study. One motivation for the present study is to identify parameterization strategies that are efficient with respect to the number of quantum chemistry evaluations required so as to enable future studies of more challenging systems and/or using higher-level and more expensive quantum chemistry methods.

3.2. Functional Form Flexibility

In this section, we motivate the specific pairwise exp6(N_u) and nonpairwise PIP$\xi_{factor}\xi_{order}$(N_u) functional forms to be systematically tested in Secs. 3.3–3.5. We do so by temporarily using the six cuts comprising the "out of sample" test sets as training sets, such that in this section $f = f_{oos}$. The values of f_{oos} obtained in this section are therefore the minimum possible values of f_{oos} for each functional form and system. These tests aid in the identification of appropriate choices of N_u, ξ_{factor}, and ξ_{order} by eliminating strategies with large minimum values of f_{oos}. We also determined the number of adjustable parameters, N_p, for each functional form considered, and we eliminate from further testing strategies with excessive numbers of parameters. A table showing the full results of this study (f_{oos} and N_p for 47 exp6 and PIP functional forms including all five systems) is given as supporting information (Table S1). The principal results are summarized next.

Figure 2 compares *ab initio* and fitted energies along the six cuts in the out of sample test set for CH_3OH + Ar and for the exp6(3) functional form (a 12-parameter, N_u = 3 implementation of eq 4 where all H atoms were forced to have the same Ar–H interaction parameters). For the parameterization in Fig. 2, the fitting error is only f_{oos} = 17 cm^{-1}, where f_{oos} is the fitting error f defined by eq 6 evaluated for the data comprising the six cuts. This value of f_{oos} can be interpreted as representative of the absolute fitting error near the bottom of the well, with larger absolute—but similar relative—fitting errors at energies up the repulsive wall. Specifically, for six cuts shown in Fig. 2, the fitting errors near the local minimum along each cut are 4, 36, 1, 25, 2, and 11 cm^{-1}, respectively, with an average of 13 cm^{-1}, which is close to the value of f_{oos} = 17 cm^{-1} for this system. For these same six local minima, the average relative fitting error is 9%, which is similar to the average relative fitting error for energies up the repulsive wall (~7% for energies along the six cuts close to 2500 cm^{-1} above the asymptote). Note that errors this size are already ~2x larger than the error assigned in Sec. 3.1 to the cc MP2/CBS(a'd,a't) method for this system.

For CH_3OH + Ar, we tested two implementations of the exp6 functional form, depending on whether the interaction of Ar with hydrogen was the same for all H atoms [exp6(N_u = 3)] or differed for only the hydroxyl H atom [exp6(N_u = 4)]. The choice of N_u had only a small effect on f_{oos}.

For C_2H_5OH and C_4H_9OH + Ar, the values of N_u that reflect the true C-atom and H-atom symmetries are 6 and 10, respectively. We also considered an N_u = 3 implementation where all H atoms and all C atoms were treated as equivalent, and an N_u = 4 implementation which differed from the N_u = 3 implementation in that the hydroxyl

H atom was distinguished from the other H atoms. Allowing the hydroxyl H atom to have unique interaction parameters with Ar [exp6(4)] does improve the accuracy of the fits relative to the exp6(3) functional form for the larger alcohols (reducing f_{oos} by ~25%, as shown in Table S1), whereas the additional flexibility from allowing the symmetry-nonequivalent C and H atoms along the hydrocarbon backbone to have unique parameters [as in the exp6(6) and exp6(10) fits] does not. We therefore limit attention to the exp6(3) and exp6(4) strategies throughout the rest of this work. This restriction represents a considerable reduction in complexity of the functional forms for V_I and in the number of adjustable parameters. The validation of the use of "universal" C–Ar and H–Ar exp6 parameters for the backbone C and H atoms, as in the exp6(3) and exp6(4) fits, enables these fits to perform well for larger alcohols, and this transferability is demonstrated in Sec. 3.5.

For the larger baths (M = N_2 and H_2O), the exp6 function form cannot accurately reproduce the anisotropy in the interaction potential, leading to errors of f_{oos} = ~100 cm^{-1}. These errors are large enough to lead to qualitative errors in the fitted interaction potentials, such as incorrectly ordering the interaction strengths along the different cuts in the test set. An example of this is shown in Fig. 3(a) for the exp6(3) fit to $CH_3OH + N_2$. We therefore restrict the use of the exp6 functional form to M = Ar in the rest of this work.

Seven nonpairwise PIP functional forms were tested for each system, with detailed results again summarized in Table S1. For M = Ar, the smallest PIP basis sets considered here, PIP22(N_u), have similar numbers of independent linear parameters, N_p, as the exp6(N_u) functional forms, and, for N_u = 4, the two approaches result in similar

values of f_{oos} despite the different underlying functional forms. For $N_u = 3$, the PIP22(3) functional form gives significantly larger values of f_{oos} than the exp6(3) functional form, despite the similar numbers of adjustable parameters; the source of this result is not clear.

As shown in Table S1 and as expected, the accuracy of the PIP functional forms can be made arbitrarily small by increasing ξ_{factor}, ξ_{order}, and/or N_u and therefore increasing the number of adjustable parameters, N_p. Whereas the exp6 approach was unable to represent the anisotropy in the $CH_3OH + N_2$ interaction potential, as shown in Fig. 3(a), the relatively small PIP22(4) functional form ($N_p = 27$) is flexible enough ($f_{oos} <$ 1 cm^{-1}), as shown in Fig. 3(b).

For M = H_2O, the H and O atoms in the bath have distinct interactions with the alcohol atoms, doubling the value of unique atom–atom interactions, N_u, from 3 or 4 to 6 or 8 and increasing the number of terms in the PIP expansions. For example, $N_p = 38$ for the smallest PIP fit considered for M = H_2O (PIP22(6)), and this functional form is flexible enough to accurately fit the six cuts through the $CH_3OH + H_2O$ surface ($f_{oos} = 6$ cm^{-1}).

Several additional PIP functional forms were generated for each system by increasing ξ_{factor} up to 3 and ξ_{order} up to 4. Even these small increases generate large numbers of adjustable parameters, N_p, particularly for the larger baths. For example, $N_p =$ 1262 for PIP24(8) for $CH_3OH + H_2O$. As a result, many parameterizations considered in this section would be underdetermined for the 156 energies comprising the present parameterization sets, as indicated in Table S1. Because of the severe scaling of N_p with ξ_{factor} and ξ_{order} for these systems, we do not consider larger values of these constraints. While the additional flexibility of the PIP strategy relative to the exp6 strategy allows for

more accurate fits of complex interaction potentials, it also requires more robust fitting strategies and larger training sets, such as those considered next.

3.3. Detailed Fitting Efficiency Tests for CH_3OH + Ar

A detailed study is presented for the simplest system considered here, CH_3OH + Ar. We first define and quantify the fitting efficiencies of the four sampling strategies described in Sec. 2: unbiased pseudorandom (P) sampling, unbiased Sobol (S) quasirandom sampling, and their biased-in-*r* versions (bP and bS). For each sampling strategy and functional form, 100 parameterizations were generated for several different training set sizes M. The resulting 100 values of f and f_{oos} were averaged, and the convergence of $\langle f \rangle$ and $\langle f_{oos} \rangle$ as a function of M was studied. Figure 4 shows this convergence for the exp6(3) functional form. We define f_∞ as the $M \to \infty$ limit of $\langle f_{oos} \rangle$, and this value is tabulated in Table 1 and shown as a horizontal line in Fig. 4. This quantity represents the error inherent in the functional form and is necessarily independent of sampling strategy; it may be distinguished from the error due to incomplete and inefficient sampling for finite M. The rate of convergence of $\langle f_{oos} \rangle$ to f_∞ as a function of M is a measure of fitting efficiency, as detailed next.

For the unbiased sampling strategies, both the average fitting error $\langle f \rangle$ and the average out of sample prediction error $\langle f_{oos} \rangle$ converge to similar values (~34 cm^{-1}) at large M, as seen in Figs. 4(a) and 4(c). As here we did not generate our out of sample test set using the same prescription as our training set, it is not guaranteed that $\langle f \rangle$ and $\langle f_{oos} \rangle$ will converge to similar values. This agreement does suggest, however, that the six cuts we are using as a test set are representative of the full-dimensional interaction potential.

For the biased data sets, $\langle f \rangle$ converges to a larger value due to the relatively higher number of sampled points at short distances. We are concerned here with the rate of convergence of $\langle f \rangle$ with M and not its numerical value, and so the $\langle f \rangle$ curve was offset to tend to f_∞ when plotted in Fig. 4(b) and 4(d) for easier comparisons among the four panels.

In each panel of Fig. 4, the average fitting error $\langle f \rangle$ and the average out of sample prediction error $\langle f_{oos} \rangle$ are seen to converge to their $M \to \infty$ values at similar rates, with $\langle f \rangle$ converging from below and $\langle f_{oos} \rangle$ converging from above. This behavior was previously described in a different context.[58] For this simple system (which we describe as simple because it is relatively isotropic and just three-dimensional), both $\langle f \rangle$ and $\langle f_{oos} \rangle$ converge quickly, requiring just ~100s of *ab initio* energies for convergence. The maximum errors (and, similarly, the standard deviations in the fitting errors for the 100 fits) converge more slowly with M. Qualitatively, it is evident that both biasing in r (as in Figs. 4(b) and 4(d)) and employing the low-discrepancy Sobol sequence (as Figs. 4(c) and 4(d)) result in faster convergence with respect to M relative to the more typical unbiased pseudorandom sampling strategy (Fig. 4(a)). Notably, the use of Sobol quasirandom sampling results in much less dispersion in the fitting errors for a given M, as evidenced by the smaller standard deviations in the fitted errors, and therefore much faster convergence of the maximum fitting errors relative to pseudorandom sampling.

The qualitative observations in the previous paragraph were quantified as follows. Instead of considering the maximum value of f_{oos}, which depends arbitrarily on the number of parameterizations generated, we define

$$f_{oos}^{95} = \langle f_{oos} \rangle + 2 s_{oos}, \tag{8}$$

where s^2_{oos} is the variance in f_{oos} over 100 parameterizations, and the superscript "95" indicates the use of two standard deviations in eq 8. We monitored convergence in f^{95}_{oos} with respect to M, and we chose our convergence threshold to be 30% of f_∞ (e.g., 1.3 f_∞ = 44 cm^{-1} for the fits shown in Fig. 4). The value of M where this threshold is met is labeled M* and is one measure of fitting efficiency. M* was calculated by fitting log$M$ = m (f^{95}_{oos} – 1.3 f_∞) + log$M$* to the three values of M closest to the convergence threshold.

In Fig. 5(a), f^{95}_{oos} computed using the data from Fig. 4 is shown as a function of M for the P, bP, S, and bS data sets, with the convergence threshold (1.3 f_∞) and M* indicated and tabulated in Table 1. Using M* as a measure of fitting efficiency, the bS sampling strategy is 16x more efficient than the unbiased P sampling strategy, converging f^{95}_{oos} to better than 30% with just M* = 43 sampled *ab initio* energies. The P sampling strategy is the least efficient requiring M* = 699 *ab initio* energies for convergence.

The use of biased sampling in r lowers M* by a factor of 2.6 for both the pseudorandom and Sobol sampling strategies. This arises from nearly equal contributions from reductions in both the average and variance of the fitting errors, as shown in Figs. 5(b) and 5(c).

The low-discrepancy Sobol sampling strategy improves the fitting efficiency (i.e., lowers M*) by a factor of ~6 for both the unbiased and biased data sets. These significant efficiency improvements arise for the most part from reductions in the variance (see Fig. 5(c)), suggesting that, for small M, clustering of geometries in the sampled P and bP data sets can leave some regions of the intermolecular potential unexplored, leading to large prediction errors f_{oos}.

The analysis in Fig. 5 was repeated for several functional forms, and the resulting converged prediction errors f_∞ and fitting efficiencies M^* are given in Table 1. Similar trends in the sampling efficiency M^* with respect to the four sampling strategies are observed here as detailed above for exp6(3). Specifically, the reduction in M^* due to biased sampling in r relative to unbiased sampling is a factor of ~2.5 when averaged over all nine functional forms in Table 1, while the use of Sobol quasirandom sampling instead of pseudorandom sampling has a larger effect, reducing M^* by a factor of ~6, on average. When both strategies are employed (as in the bS strategy), the fitting efficiency improves by a factor of ~8.5 relative to the P sampling strategy. The consistency of the observed efficiency improvements suggests these improvements are related to the efficient exploration of the underlying interaction potential and its inherent anisotropy, independent of the functional forms being parameterized.

The exp6 parameterizations have a more physically-motivated functional form than the PIP parameterizations and are indeed more accurate than the smallest PIP22 parameterizations, which have similar numbers of adjustable parameters. Otherwise, the converged prediction errors f_∞ are found to exponentially decrease with the number of adjustable parameters, N_p. As an example, the PIP23(N_u) functional form has 78 terms in its expansion, with many of these terms constrained to have the same coefficients by symmetry. Increasing N_u from 3 to 4, i.e., allowing the H atom on the hydroxyl group to interact differently with Ar than the methyl H atoms, nearly doubles the number of free adjustable parameters, N_p, from 22 to 37. The doubling of N_p reduces the converged prediction error f_∞ by more than a factor of four from 37 to 8 cm^{-1}.

While the prediction errors f_∞ are observed to scale quadratically with the number of parameters in the fit N_p, the sampling efficiencies M^* instead scale nearly linearly with N_p. This results in fitting efficiencies per parameter (M^*/N_p) that are fairly independent of the functional form, at least for the smaller basis sets, as shown in Table 1. The least efficient sampling strategy (P) requires 18–78 training set energies per parameter for convergence, whereas the most efficient sampling strategy (bS) requires just 3–8 training set energies per parameter. This latter ratio of *ab initio* data to adjustable parameters is already quite small (one would not expect it to be less than 1), despite the relatively simple bias used here and the consideration of an unoptimized low-discrepancy (Sobol) sampling sequence.

3.4. Scaling with Respect to System Size and Dimensionality

Results analogous to those in Table 1 are given in Tables 2 and 3 for C_2H_5OH + Ar and C_4H_9OH + Ar, respectively, and in Table 4 for $CH_3OH + N_2$ and $CH_3OH + H_2O$. In a few cases for $CH_3OH + H_2O$ when using the larger PIP basis sets, the out of sample prediction error f_∞ was not converged for $M < 10{,}000$ and M^* could not be quantified. Figure 6(a) plots f_∞ from Tables 1–4 as a function of the number of fitted parameters N_p for all 31 cases considered. In Fig. 6(b) these same data are converted to relative errors by scaling them by the approximate well depths v used in eq 7. Recall that f_∞ is a measure of the inherent error of the functional form in representing the sampled *ab initio* interaction potential independent from incomplete sampling errors associated with finite M.

Several conclusions may be drawn from Fig. 6, many of which were already noted in Sec. 3.3 for CH_3OH + Ar. Briefly, when applicable and when the number of

adjustable parameters N_p is small (≤ 20), the physically-motivated exp6 functional forms are more accurate than the PIP functional forms, with relative representation errors f_∞ of just 10–15% for the exp6 parameterizations compared to ~30% for the PIP parameterizations. PIP parameterizations are systematically improvable, however, and f_∞ is shown to decrease with increasing N_p regardless of how N_p is increased, i.e., regardless of whether ξ_{factor}, ξ_{order}, or N_u is increased. There is no clear preference for choosing N_u = 3 or 4 (or 6 or 8 for M = H_2O) independent from the overall trend of f_∞ decreasing with increasing N_p.

For M = Ar, the relative representation error, f_∞/v, is shown to be largely independent of the size of the alcohol and to scale as N_p^{-2}, i.e., doubling N_p reduces f_∞ by a factor of four. For the larger bath gases (and consequently higher-dimensional interaction potentials), larger relative errors and weaker scaling with N_p is observed, although the converged representation errors are in fact quite small for all the parameterizations considered (< 10%). The weaker scaling of f_∞ with N_p for the higher-dimensional systems may reflect inherent inefficiencies in the way the PIP basis sets are generated here via the constraints ξ_{factor} and ξ_{order}. One may anticipate that these higher-dimensional systems with large numbers of intermolecular bond distances (and therefore with large numbers of higher-order terms in the PIP expansions) may benefit from alternate strategies for generating and/or pruning these terms, and this is being explored in ongoing work.

Next, we consider fitting efficiencies, M^*. Plots analogous to Fig. 5(a) were generated for all 31 cases in Tables 1–4 and for the four sampling types P, bP, S, and bS. The determination of M^* for each of these 124 combinations required evaluations of f_{oos}^{95}

(eq 8) for several values of M (we used 18 values varying uniformly in $\log_{10}M$ from M = 10 to 1000), and 100 fits were generated for each value of M with different training sets and random number seeds. The total number of parameterizations performed as part of this work is therefore over 200,000. We mention this to highlight the utility of automation in potential energy surface construction and analysis and, in particular, in demonstrating convergence and quantifying fitting efficiency.

The values of M^* from Tables 1–4 for the most efficient (bS) and least efficient (P) sampling types are plotted Fig. 7 as a function of the number of adjustable parameters N_p. Although there is some scatter, M^* is shown to vary linearly with N_p, on average, for all five systems and for both sampling types. For the bS sampling strategy, the number of *ab initio* data required per parameter (M^*/N_p) is as low as just 2.8 and is 5.6 ± 2.9, on average, where the uncertainty indicated is one standard deviation of the results in Fig. 7(c). For the P sampling strategy (which may be thought of as a "standard" approach to sampling), M^*/N_p is never smaller than 16 and is 39 ± 16, on average. There is no apparent trend in M^*/N_p with respect to alcohol size, the dimensionality of V_I, or the choice of functional form for either sampling type. The linear dependence of M^* on N_p further supports the search for improved basis set selection strategies (e.g., by pruning the PIP expansions) to both improve the scaling of f_∞ with N_p, as noted above, while also lowering N_p and thus requiring fewer computed *ab initio* data.

The fitting efficiencies M^* were used to quantify the effectiveness of our biased sampling strategy as well as the use of Sobol's low-discrepancy sampling as a function of dimensionality and system size. For each of the 31 cases in Tables 1–4, we computed averaged relative efficiency improvements by comparing M^* for unbiased pseudorandom

sampling (P) with M^* for the other sampling types (bP, S, and bS). The overall efficiency improvement of using *both* the bias in r and quasirandom sampling is defined as $M^*(\mathrm{P})/M^*(\mathrm{bS})$.

This ratio was computed and averaged over the different functional forms considered in Tables 1–4 and plotted as a function of the dimensionality of V_1 in Fig. 8(a) and as a function of alcohol size in Fig. 8(b). For CH_3OH + Ar, this ratio is 9.3, indicating that the bS sampling strategy requires nearly an order of magnitude fewer *ab initio* energies than the unbiased P sampling strategy to achieve convergence with respect to M. For the more complex systems, this ratio remains greater than one but, as shown in Fig. 8(b), the overall efficiency improvement of the bS strategy relative to the P strategy decreases monotonically with system size (down to 5.2 for butanol + Ar) and, as shown in Fig. 8(a), decreases and then increases with respect to dimensionality (and is 8.2 for N_2 and 17 for H_2O).

These trends can be rationalized by separating the efficiency improvements due to the bias in r from the improvements due to quasirandom Sobol sampling. To do so, we computed the ratios $M^*(\mathrm{P})/M^*(\mathrm{bP})$ and $M^*(\mathrm{S})/M^*(\mathrm{bS})$ and again averaged them over several functional forms to isolate the improvement due to the use of the bias in r, and similarly we used the average of $M^*(\mathrm{P})/M^*(\mathrm{S})$ and $M^*(\mathrm{bP})/M^*(\mathrm{bS})$ to isolate the improvements due to the use of Sobol quasirandom sampling. These averages are also show in Fig. 8.

Considering these isolated contributions in Fig. 8(a), we see that the Sobol sequence becomes less effective as the dimensionality increases, whereas the bias in r becomes more effective. The behavior of these two effects, and the fact that the two

curves cross one another, suggests that addressing the increased anisotropy of the M = N_2 and H_2O systems relative to M = Ar (as the bias in *r* is designed to do) can be more important than addressing sampling uniformity (as the Sobol sequence is designed to do) for strongly anisotropic systems, whereas the trend is reversed for weakly anisotropic systems. In general, both effects are important and combining the two strategies is more effective than using either approach individually.

For the series of increasingly large alcohols + Ar in Fig. 8(b), we see that the bias in *r* is less effective than low discrepancy sampling in improving sampling efficiency and again that combining the two strategies can be particularly effective. The bias in *r* is not very effective for butanol, however, and the pursuit of more sophisticated biasing strategies may be of particular utility when considering large systems where the center of mass coordinate *r* is strongly coupled to the angular coordinates.

Results for both series in Fig. 8 support the conclusion drawn in Sec. 3.3 that fitting strategies for more complex systems are likely to benefit from more complex sampling biases, such as those using optimal experimental design or leverage based sampling.[66,67] For example, it may be more natural to sample uniformly from the bond distances directly (or using Morse variables of these distances), instead of sampling coordinates for the rigid body separation and orientation as was done here.

3.5. Dynamical Tests

In this section, we quantify the convergence of dynamical predictions with respect to the various fitting strategies considered here. Collision parameters at several temperatures were calculated using ensembles of classical trajectories (as described in

Sec. 2) for every 10th parameterized surface of the 100 generated for each value of M and discussed above. This was done for several systems, functional forms, and sampling types, and the results were used to construct convergence plots of collision parameters with respect to M. An example is shown in Fig. 9 for the exp6(3) functional form, CH_3OH + Ar, and 1000 K comparing convergence in the average energy transferred in deactivating collisions α for two sampling strategies: P and bS. At each value of M, the average of the ten calculated values of α is shown, along with bounds indicating two standard deviations of the ten results. The horizontal lines indicate the large-M-limit (i.e., the converged) value of α along with the two-sigma *statistical* error bounds arising from the use of a finite-sized ensemble of trajectories. Each ensemble consisted of 6000 trajectories, and the converged value of α for this system and functional form is 389 ± 26 cm^{-1}.

As shown in Fig. 9, the results of trajectory calculations converge to the same large M limit and with the same statistical distribution for both sampling types. Convergence is faster with respect to M when using bS sampling than when using P sampling. The relevant values of M^* taken from Table 1 for these two sampling types are indicated in Fig. 9, where it can be seen that M^* (which is based on the convergence of the out of sample fitting error) correlates well with convergence in this dynamical result.

Figure 10 extends this analysis to include the collision rate Z along with α, now shown for all four sampling strategies: P, bP, S, and bS. For each, we show the convergence of expressions analogous to f_{oos}^{95}, where δ_{α}^{95} and δ_{Z}^{95} have been defined as relative error in the mean values of α and Z, respectively, plus two standard deviations of the results for each value of M. By construction, these quantities converge to their relative

2-sigma statistical errors in the limit of large M. At smaller values of M, the errors include the error due to the underlying, un-converged PES. The rates of convergence in the curves shown in Fig. 10 are then measures of the convergence in the *a priori* error in the predicted collision parameters due to the PES.

Figure 10(a) shows the values of M^* taken from Table 1, and it can be seen that at these values of M^*, α is converged to ~15% on average. Recall that M^* was defined as the value of M where the fitting error was converged to 30% of its large-M limit. Figure 10 thus demonstrates the useful result that the dynamical predictions that are based on highly averaged properties, such α, are likely to converge more rapidly than errors in the underlying energetics and forces. Relative errors in the collision rates Z are even smaller (by nearly a factor of 10) than those for α. Similar convergence plots were made for the other systems considered here and for other temperatures, and it was observed generally that dynamical errors were converged to 10–25% at the critical values of $M = M^*$.

Finally, we discuss two types of transferability of practical relevance. We first consider the transferability of parameterizations with respect to the identity of the alcohol. In past work, we found that exp6 parameterizations for CH_4 + M were just as accurate for larger hydrocarbons + M as system-specific parameterizations,[68] and this transferability was tested here for alcohols. Several converged (large-M-limit) PESs were used to predict the collision parameters Z and α for an entire series of normal alcohols as large as octanol and for M = Ar. A total of twelve PESs were tested, including both the exp6(3) or exp6(4) functional forms, those generated using the P or bS sampling types, and all three alcohols from Tables 1–3. Each of the twelve parameterizations was used to

compute Z and α for an entire series of alcohols: n-$C_xH_{2x+1}OH$ + Ar with x =1, 2, 4, 6, and 8 at 1000 K.

We found very little difference in the twelve predictions generated for each alcohol. Specifically, the deviations of the individual results from their average for a given alcohol were typically less than 10%, which is equal to the 2-sigma statistical sampling of the finite trajectories ensembles. Just four of the sixty cases had larger deviations, which is consistent with the usual interpretation of 2-sigma error bounds as 95% confidence limits. Similar results were found at 300 and 2000 K. We conclude that there is no significant error associated with the transferability of exp6 parameters generated for one alcohol when used for other alcohols. These results further confirm the utility of generating "universal" exp6 atom–atom interaction parameters associated with common functional groups. Analogous tests could not be performed for the PIP functional forms, as the PIP basis set expansions contain different terms for the different alcohols making the application of a fitted CH_3OH + M PES to a larger alcohol ambiguous.

Next, we consider transferability with respect to the reference geometries for A and M ($\mathbf{R}_A^{(e)}$ and $\mathbf{R}_M^{(e)}$) used to generate the *ab initio* training and out of sample test sets. Comparing the performance of the fitted PESs for two observables α and Z provides insight into this type of transferability as the collision rate Z depends only on the rigid-body interaction potential, whereas the energy transfer per collision α is the result of full-dimensional trajectory calculations. The calculation of α therefore includes PES evaluations with A and M displaced from their reference (equilibrium) geometries, while the calculation of Z does not.

Figure 11 shows the computed values of Z and α for CH_3OH + Ar using converged (large-M-limit) parameterizations for all of the rows shown in Table 1, i.e., for all 36 combinations of sampling strategy and basis set choice. The same sets of computed observables are shown twice, once in panels (a) and (c) as functions of the number of parameters N_p and again in panels (b) and (d) as functions of the fitting error f_∞.

Convergence in the collision rate Z with respect to the number of parameters N_p is apparent in Fig. 11(a), where the results using the exp6 functional forms are converged already at N_p = 12 while the PIP functional forms are not converged until a somewhat larger value, N_p = 22. This trend is consistent with the discussions provided above regarding f_∞, and indeed convergence in Z is readily apparent when plotted against f_∞ as in Fig. 11(b). In fact, Z is converged to better than 2% when f_∞ is fairly large (~40 cm^{-1}). Such systematic convergence with respect to increasing N_p and decreasing f_∞ is what one would expect for a rigid-body observable, like Z, which depends only on A + M interaction energies from the same domain as was sampled to generate the training set and test set data.

Trends in α are less clear, and they demonstrate a failure of the transferability assumption central to the separable approximation when large PIP functional forms are used. As seen in Fig. 11(c), 26 of the 36 parameterizations tested predict α within 10% of ~400 cm^{-1}, including results obtained using the exp6 PESs, which in past work have been shown to be very accurate for similarly simple systems.[24] We therefore take 400 cm^{-1} to be the accurate dynamical result, and we note that deviations from 400 cm^{-1} occur for both larger *and* smaller values of N_p.

For small N_p, errors in the PIP parameterizations may be straightforwardly attributed to poorly fitted PESs, just as was observed for Z. Comparing Figs. 11(b) and 11(d), we see that both Z and α are inaccurate for $f_\infty > 40$ cm^{-1}. Unlike Z, however, errors in α appear at small f_∞. These errors may be attributed to poor transferability of the fitted PES when evaluated at geometries for A and M different from the reference geometries. Such evaluations do not appear in the calculation of Z, which depends only on the rigid-body PES, and so this type of error is expected only for α, just as is seen in Fig. 11. These errors represent a breakdown of the separable approximation central to this work, which relies on the transferability of the fitted PES to geometries of A and M other than those used as reference geometries. This breakdown appears for the largest PIPs considered here, where the values of f_∞ are very small (< 10 cm^{-1}), and this breakdown can therefore be considered as a type of overfitting.

Plots analogous to Fig. 11 were generated for C_2H_5OH + Ar and CH_3OH + N_2 and H_2O. Similar trends were observed for these systems, where again Z was found to systematically converge with respect to both N_p and f_∞. Convergence in α exhibited overfitting for large N_p and small f_∞, including trajectories that found "holes" in the fitted potentials for these larger systems. The cutoff for overfitting was found to be $f_\infty \approx 20$ cm^{-1} for these larger systems. While this particular example of overfitting may be unique to the use of the separable approximation for fitting nonreactive many body potentials, these results demonstrate an inherent difficultly in observing the consequences of overfitting in general. If just a few of the results in Fig. 11(d) had been available, instead of the whole set, one might assume that the largest PIP expansion with smallest value of f_∞ was the

most accurate. Instead overfitting resulted in additional dynamical errors as large as a factor of two in α.

4. CONCLUSIONS

Fitting a potential energy surface (PES) is a multiobjective optimization problem,[69] involving at least three competing objectives: (1) maximizing the accuracy of the fit, (2) minimizing the computational cost and human effort required to obtain the fit, and (3) minimizing the cost of evaluating the fitted PES during the dynamics calculation. The first two have been characterized in detail in this paper, while the third objective, which is related to the number of terms in the expansion and to their complexity, has only been briefly addressed.

Consider how the objectives compete. The fitting error inherent in a given functional form can be lowered either by adding adjustable parameters, N_p, (e.g., by increasing the number of terms in the PIP expansions) or by using more physically-motivated functional forms (e.g., using exp6 Buckingham interactions instead of polynomials of Morse variables). The improvement of objective (1) by increasing N_p necessarily competes with objectives (2) and (3), as more training data are required for larger N_p and more terms in expansion result in more expensive PES evaluations.

To quantify this competition, we defined an out of sample fitting error, f_{oos}, such that f_{oos} represents the absolute error in the PES at low energies and f_{oos}/v, where v is the well depth, represents the relative error throughout the PES. By computing f_{oos} as a function of the number of training data, we determined the number of data required for

convergence M^* for several PIP and exp6 functional forms and for five different A + M systems.

We found that the ratio M^*/N_p was largely independent of functional form and system but did depend sensitively on sampling strategy. An empirical sampling bias toward short separations lowered M^*/N_p, as did regularizing the distribution of sampled geometries using Sobol's sequence of quasirandom numbers. Combining these two approaches lead to overall sampling efficiency improvements of close to a factor of 10 and parameterizations requiring just M^*/N_p = 3–8 *ab initio* energies per parameter. This ratio is close to the ideal value of 1, which is perhaps a surprising result considering the relatively simple bias and sampling strategies being used here.

Improved sampling strategies for further reducing M^*/N_p and motivated by the current set of tests are being studied in ongoing work. For example, we identified limitations in the one-dimensional bias used here, particularly for the more complex systems studied. We are studying the effectiveness of tying sampling more closely to the fitting basis, such as sampling directly in Morse variable coordinates and using optimal experimental design strategies and leverage-based sampling as discussed in Refs. 66 and 67. Such approaches provide a multidimensional bias that is tuned to the functional forms being parameterized, which are likely more powerful than the simple one-dimensional bias used here.

We quantified the reduction of the converged fitting error f_∞ with respect to N_p for the PIP functional forms, and we found very efficient scaling ($f_\infty \alpha N_p^{-2}$) for the three-dimensional PESs for M = Ar. For the five- and six-dimensional PESs, with M = N_2 and H_2O, respectively, scaling was found to be worse, with f_∞ scaling approximately as $N_p^{-1/2}$.

This suggests inefficiencies in the PIP expansions for these larger systems, which have large numbers of intermolecular atom–atom distances and therefore have many similarly-flexible higher-dimensional terms. In ongoing work, we are studying the effectiveness of pruning the PIP basis sets. There are many ways that this can be accomplished, including model selection, variable selection, feature selection, and subset selection.[57,70,71]

Finally, the present study outlined and tested a strategy for automating PES construction. Key to this was the development of a meaningful (i.e., interpretable) out of sample error function, f_{oos}, as well as a strategy for demonstrating convergence. We furthermore correlated these fitting errors with errors in the dynamical quantities of interest, namely the collision rate Z and collision efficiency α. We tested the transferability of the exp6 parameterizations constructed for one alcohol to a series of alcohols, and again validated the good accuracy of generating "universal" parametrizations for different classes of molecules, at least for M = Ar. While not transferable in the same way, PIP functional forms can be just as accurate for M = Ar and are much more accurate for M = N_2 and H_2O. We showed, however, that large PIP expansions may suffer from overfitting and a breakdown of the separability assumption employed here.

ASSOCIATED CONTENT

Supporting Information

Figures of the quantum chemistry tests described in Sec. 3.1.

Tables of the results of the functional form flexibility tests described in Sec. 3.2.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ajasper@anl.gov

ACKNOWLEDGMENTS

This work was supported by the U. S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences through Argonne National Laboratory. Argonne is a U. S. Department of Energy laboratory managed by UChicago Argonne, LLC, under Contract Number DE-AC02-06CH11357. We gratefully acknowledge computing resources provided by Blues and Bebop, two high-performance computing clusters operated by the Laboratory Computing Resource Center at Argonne National Laboratory.

Table 1. Converged prediction errors f_∞ and critical sample sizes M^* for CH_3OH + Ar

Functional form	N_p	Sampling	f_∞, cm^{-1}	M^*	$M^*(P)/M^*$	M^*/N_p
exp6(3)	12	P	34	699	≡ 1	58
		bP	34	262	2.6	22
		S	35	111	6.1	9.3
		bS	34	43	16	3.6
exp6(4)	16	P	27	437	≡ 1	27
		bP	28	159	2.7	10
		S	27	143	3.1	8.9
		bS	27	63	6.9	3.9
PIP22(3)	11	P	82	311	≡ 1	28
		bP	91	153	2.0	14
		S	81	87	3.6	7.9
		bS	87	41	7.6	3.7
PIP22(4)	16	P	57	441	≡ 1	28
		bP	68	214	2.1	13
		S	59	116	3.8	7.3
		bS	68	82	5.4	5.1
PIP23(3)	22	P	36	1725	≡ 1	78
		bP	38	1169	1.5	53
		S	35	155	11	7.1
		bS	37	107	16	4.9
PIP33(3)	25	P	19	1298	≡ 1	52
		bP	19	601	2.2	24
		S	19	174	7.5	6.9
		bS	20	80	16	3.2
PIP23(4)	37	P	7	648	≡ 1	18
		bP	8	623	1.0	17
		S	7	350	1.9	9.5
		bS	8	179	3.6	4.8
PIP33(4)	41	P	5	1459	≡ 1	36
		bP	7	778	1.9	19
		S	5	583	2.5	14
		bS	7	225	6.5	5.5
PIP24(4)	70	P	2	2984	≡ 1	43
		bP	2	1129	2.6	16
		S	2	930	3.2	13
		bS	2	541	5.5	7.7

Table 2. Converged prediction errors f_∞ and critical sample sizes M^* for C_2H_5OH + Ar

Functional form	N_p	Sampling	f_∞, cm^{-1}	M^*	M^*(P)/M^*	M^*/N_p
exp6(3)	12	P	59	552	≡ 1	46
		bP	58	232	2.4	19
		S	59	110	5.0	9.2
		bS	58	80	6.9	6.7
exp6(4)	16	P	29	757	≡ 1	47
		bP	27	262	2.9	16
		S	27	411	1.8	26
		bS	27	177	4.3	11
PIP23(3)	26	P	33	900	≡ 1	35
		bP	31	674	1.3	26
		S	32	169	5.3	6.5
		bS	31	123	7.3	4.7
PIP33(3)	29	P	18	1280	≡ 1	44
		bP	18	533	2.4	18
		S	16	474	2.7	16
		bS	17	167	7.7	5.8
PIP23(4)	42	P	15	1776	≡ 1	42
		bP	15	418	4.2	10
		S	15	580	3.1	14
		bS	16	199	8.9	4.7
PIP33(4)	46	P	12	1668	≡ 1	36
		bP	12	578	2.9	13
		S	12	546	3.1	12
		bS	12	163	10	3.5
PIP24(4)	87	P	4	6896	≡ 1	79
		bP	4	2105	3.3	24
		S	4	805	8.6	9.3
		bS	4	643	11	7.4

Table 3. Converged prediction errors f_∞ and critical sample sizes M^* for C_4H_9OH + Ar

Functional form	N_p	Sampling	f_∞, cm^{-1}	M^*	M^*(P)/M^*	M^*/N_p
exp6(3)	12	P	65	407	≡ 1	34
		bP	68	405	1.0	34
		S	64	83	4.9	6.9
		bS	68	65	6.3	5.4
exp6(4)	16	P	36	421	≡ 1	26
		bP	37	297	1.4	19
		S	36	301	2.1	13
		bS	36	137	3.1	8.6
PIP23(3)	27	P	51	693	≡ 1	26
		bP	49	586	1.2	22
		S	45	162	4.3	5.8
		bS	47	119	5.8	4.4
PIP33(3)	30	P	22	1064	≡ 1	35
		bP	22	841	1.3	28
		S	20	173	6.2	4.1
		bS	22	133	8.0	3.2
PIP23(4)	43	P	22	1179	≡ 1	27
		bP	24	1032	1.1	24
		S	22	469	2.5	11
		bS	22	379	3.1	8.8
PIP33(4)	47	P	17	2013	≡ 1	43
		bP	18	1126	1.8	24
		S	17	416	4.9	8.9
		bS	18	284	7.2	6.0
PIP24(4)	93	P	5	4959	≡ 1	53
		bP	5	4323	1.1	46
		S	5	1562	3.2	17
		bS	5	1612	3.1	17

Table 4. Converged prediction errors f_∞ and critical sample sizes M^* for CH_3OH+N_2 and $CH_3OH + H_2O$

Functional form	N_p	Sampling	f_∞, cm^{-1}	M^*	$M^*(P)/M^*$	M^*/N_p
CH_3OH+N_2						
PIP23(3)	54	P	30	1414	≡ 1	26
		bP	31	1080	1.3	20
		S	28	779	1.8	14
		bS	32	159	8.9	2.9
PIP33(3)	57	P	32	2158	≡ 1	38
		bP	33	1003	2.2	18
		S	32	592	3.6	10
		bS	35	161	13	2.8
PIP23(4)	97	P	25	2150	≡ 1	22
		bP	29	1021	2.1	11
		S	24	886	2.4	9.1
		bS	27	413	5.2	4.3
PIP24(4)	309	P	15	5030	≡ 1	16
		bP	16	2113	2.4	6.8
		S	14	2856	1.8	9.2
		bS	15	892	5.6	2.9
CH_3OH+H_2O						
PIP23(3)	160	P	75	9989	≡ 1	63
		bP	74	1609	6.2	10
		S	76	2363	4.2	15
		bS	74	476	21	3.0
PIP33(3)	166	P	73	10200	≡ 1	63
		bP	72	1878	5.3	12
		S	74	4404	2.3	28
		bS	74	777	13	4.9
PIP23(4)	297	P		a		
		bP	48	3597		12
		S		a		
		bS	47	1873		6.3
PIP24(4)	1262	P		a		
		bP	44	7838		6.3
		S		a		
		bS	42	5475		4.3

[a] $M^* >> 10000$

REFERENCES

[1] Hase, W. L. Dynamics of Unimolecular Reactions. In *Dynamics of Molecular Collisions. Modern Theoretical Chemistry, vol. 2*; Miller, W. H., Ed.; Springer: Boston, MA, 1976.

[2] Barker, J. R.; Golden, D. M. Master Equation Analysis of Pressure-Dependent Atmospheric Reactions. *Chem. Rev.* **2003**, *103*, 4577–4592.

[3] Pilling, M. J. Reactions of Hydrocarbon Radicals and Biradicals. *J. Phys. Chem. A* **2013**, *117*, 3697–3717.

[4] Klippenstein, S. J. From Theoretical Reaction Dynamics to Chemical Modeling of Combustion. *Proc. Combust. Inst.* **2017**, *36*, 77–111.

[5] Mason, E. A. Transport in Neutral Gases, In *Kinetic Processes in Gases and Plasmas*, Hochstim, A. R.; Academic Press: London, 1969; pp. 57–97.

[6] Brown, N. J.; Bastian, L. A. J.; Price, P. N. Transport Properties for Combustion Modeling. *Prog. Energy Combust. Sci.* **2011**, *37*, 565–582.

[7] Dagdigian, P. J. Combustion Simulations with Accurate Transport Properties for Reactive Intermediates. *Combust. Flame* **2015**, *162*, 2480–2486.

[8] Taxman, N. Classical Theory of Transport Phenomena in Dilute Polyatomic Gases. *Phys. Rev.* **1958**, *110*, 1235–1239.

[9] Brown, N. J.; Miller, J. A. Collisional Energy Transfer in the Low-Pressure-Limit Unimolecular Dissociation of HO_2. *J. Chem. Phys.* **1984**, *80*, 5568–5580.

[10] Lendvay, G.; Schatz, G. C. Choice of Gas Kinetic Rate Coefficients in the Vibrational Relaxation of Highly Excited Polyatomic Molecules. *J. Phys. Chem.* **1992**, *96*, 3752–3756.

[11] Lenzer, T.; Luther, K.; Troe, J.; Gilbert, R. G.; Lim, K. F. Trajectory Simulations of Collisional Energy Transfer in Highly Excited Benzene and Hexafluorobenzene. *J. Chem. Phys.* **1995**, *103*, 626–641.

[12] Meroueh, O.; Hase, W. L. Collisional Activation of Small Peptides. *J. Phys. Chem.* A **1999**, *103*, 3981–3990.

[13] Oref, I. Collisional Energy Transfer in Polyatomic Molecules in the Gas Phase. *Isr. J. Chem.* **2007***, 47*, 205–214.

[14] 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.

[15] Conte, R.; Houston, P. L.; Bowman, J. M. Classical Trajectory Study of Energy Transfer in Collisions of Highly Excited Allyl Radical with Argon. *J. Phys. Chem. A* **2014**, *118*, 7742–7752.

[16] Paul, A. K.; West, N. A.; Winner, J. D.; Bowersox, R. D. W.; North, S. W.; Hase, W. L. Non-statistical Intermolecular Energy Transfer from Vibrationally Excited Benzene in a Mixed Nitrogen-Benzene Bath. *J. Chem. Phys.* **2018**, *149*, 134101.

[17] Gallucci, C. R.; Schatz, G. C. Energy Transfer, Stabilization and Dissociation in Collisions of He with Highly Excited HO_2. *J. Phys. Chem.* **1982**, *86*, 2352–2358.

[18] Gelb, A. Classical Trajectory Study of Energy Transfer Between Argon Atoms and Vibrationally-Rotationally Excited Ozone Molecules. *J. Phys. Chem.* **1985**, *89*, 4189–4194.

[19] Barker, J. R.; Yoder, L. M.; King, K. D. Vibrational Energy Transfer Modeling of Nonequilibrium Polyatomic Reaction Systems. *J. Phys. Chem. A* **2001**, *105*, 796–809.

[20] Gilbert, R. G. Theory of Collisional Energy Transfer of Highly Excited Molecules. *Int. Rev. Phys. Chem.* **1991**, *10*, 319–347.

[21] Bernshtein, V.; Oref, I. Energy Transfer Between Azulene and Krypton: Comparison Between Experiment and Computation. *J. Chem. Phys.* **2006**, *125*, 133105.

[22] Steill, J. D.; Jasper, A. W.; Chandler, D. W. Determination of the Collisional Energy Transfer Distribution Responsible for the Collision-Induced Dissociation of NO_2 with Ar. *Chem. Phys. Lett.* **2005**, *636*, 1–14.

[23] Troe, J.; Ushakov, V. G. The Dissociation/Recombination Reaction CH_4 (+M) ⇄ CH_3 + H (+M): A Case Study for Unimolecular Rate Theory. *J. Chem. Phys.* **2012**, *136*, 214309.

[24] Jasper, A. W.; Pelzer, K. M.; Kamarchik, E.; Harding, L. B.; Klippenstein, S. J. Predictive A Priori Pressure-Dependent Kinetics. *Science* **2014**, *346*, 1212–1215.

[25] Ma, L.; Alexander, M. H.; Dagdigian, P. J. Theoretical Investigation of Rotationally Inelastic Collisions of $CH_2(X)$ with Helium. *J. Chem. Phys.* **2011**, *134*, 154307.

[26] Dagdigian, P. J.; Klos, J.; Warehime, M.; Alexander, M. H. Accurate Transport Properties for O(^3P)–H and O(^3P)–H_2. *J. Chem. Phys.* **2016**, *145*, 164309.

[27] 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.

[28] Maitland, G. C.; Rigby, M.; Smith, E. B.; Wakeham, W. A. *Intermolecular Forces: Their Origin and Determination*; Clarendon: Oxford, 1987.

[29] Murrell, J. N.; Carter, S.; Farantos, S. C.; Huxley, P.; Varandas, A. J. C. *Molecular Potential Energy Functions*; John Wiley: Chichester, 1984.

[30] Collins, M. A. Molecular Potential-Energy Surfaces for Chemical Reaction Dynamics. *Theor. Chem. Acc.* **2002**, *108*, 313–324.

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

[32] Ho, T. S.; Rabitz, H. A General Method for Constructing Multidimensional Molecular Potential Energy Surfaces from Ab Initio Calculations. *J. Chem. Phys.* **1996**, *104*, 2584–2597.

[33] Bender, J. D.; Doraiswamy, S.; Truhlar, D. G.; Candler, G. V. An Improved Potential Energy Surface and Multi-Temperature Quasiclassical Trajectory Calculations of N_2 + N_2 Dissociation Reactions. *J. Chem. Phys.* **2014**, *140*, 054302.

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

[35] Dawes, R.; Ndengue, S. A. Single- and Multireference Electronic Structure Calculations for Constructing Potential Energy Surfaces. *Int. Rev. Phys. Chem.* **2016**, *35*, 441–478.

[36] Metz, M. P.; Piszczatowski, K.; Szalewicz, K. Automatic Generation of Intermolecular Potential Energy Surfaces. *J. Chem. Theory Comput.* **2016**, *12*, 5895–5919.

[37] 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.

[38] Kolb, B.; Zhao, B.; Li, J.; Jiang, B.; Guo, H. Permutation Invariant Potential Energy Surfaces for Polyatomic Reactions Using Atomistic Neural Networks. *J. Chem. Phys.* **2016**, *144*, 224103.

[39] Lui, Y.; Huang, Y.; Ma, J.; Li, J. Classical Trajectory Study of Collision Energy Transfer between Ne and C_2H_2 on a Full Dimensional Accurate Potential Energy Surface. *J. Phys. Chem. A* **2018**, *122*, 1521–1530.

[40] Bhandari, H. N.; Ma, X.; Paul, A. K.; Smith, P.; Hase, W. L. PSO Method for Fitting Analytic Potential Energy Functions. Application to $I^-(H_2O)$. *J. Chem. Theory Comput.* **2018**, *14*, 1321–1332.

[41] 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.

[42] Qu, C.; Yu, Q.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces. *Annu. Rev. Phys. Chem.* **2018**, *69*, 151–175.

[43] Jasper, A. W.; Miller, J. A. Collisional Energy Transfer in Unimolecular Reactions: Direct Classical Trajectories for $CH_4 \rightleftarrows CH_3 + H$ in Helium. *J. Phys. Chem. A* **2009**, *113*, 5612–5619.

[44] Jasper, A. W.; Miller, J. A. Theoretical Unimolecular Kinetics for $CH_4 + M \rightleftarrows CH_3 + H + M$ in Eight Baths, M = He, Ne, Ar, Kr, H_2, N_2, CO, and CH_4. *J. Phys. Chem. A* **2011**, *115*, 6438–6455.

[45] Mebel, A. M.; Georgievskii, Y.; Jasper, A. W.; Klippenstein, S. J. Temperature- and Pressure-Dependent Rate Coefficients for the HACA Pathways from Benzene to Naphthalene. *Proc. Combust. Inst.* **2017**, *39*, 919–926.

[46] Klippenstein, S. J.; Miller, J. A.; Jasper, A. W. Kinetics of Propargyl Radical Dissociation. *J. Phys. Chem. A* **2015**, *119*, 7780–7791.

[47] Tranter, R. S.; Jasper, A. W.; Randazzo, J. B.; Lockhart, J. P. A.; Porterfield, J. P. Recombination and Dissociation of 2-methyl Allyl Radicals: Experiment and Theory. *Proc. Combust. Inst.* **2017**, *36*, 211–218.

[48] Stace, A. J.; Murrell, J. N. A Classical Trajectory Study of Collisional Energy Transfer in Thermal Unimolecular Reactions. *J. Chem. Phys.* **1978**, *68*, 3028–3039.

[49] Yan, T.; Hase, W. L. Comparisons of Models for Simulating Energy Transfer in Ne-Atom Collisions with an Alkyl Thiolate Self-Assembled Monolayer. *J. Phys. Chem. B* **2002**, *106*, 8029–8037.

[50] Majumder, M.; Gibson, K. D.; Sibener, S. J.; Hase, W. L. Chemical Dynamics Simulations and Scattering Experiments for O_2 Collisions with Graphite. *J. Phys. Chem. C* **2018**, *122*, 16048–16059.

[51] Špirko, V.; Jensen, P.; Bunker, P. R.; Čejchan, A. The Development of a New Morse-Oscillator Based Rotation–Vibration Hamiltonian for H_3^+. *J. Molec. Spectroscopy* **1985**, *112*, 183–202.

[52] Ponder, J. W. TINKER – Software Tools for Molecular Design, Version 6; Washington University Medical School, 2011.

[53] Case, D. A.; Betz, R. M.; Botello-Smith, W.; Cerutti, D. S.; Cheatham, III, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; et al. AMBER 2016; University of California, San Francisco, 2016.

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

[55] Martin, J. M. L.; Uzan, O. Basis Set Convergence in Second-Row Compounds. The Importance of Core Polarization Functions. *Chem. Phys. Lett.* **1998**, *282*, 16–24.

[56] Del Bene, J. E. Proton Affinities of NH_3, H_2O, and HF and Their Anions: A Quest for the Basis-Set Limit Using the Dunning Augmented Correlation-Consistent Basis Sets. *J. Phys. Chem.* **1993**, *97*, 107–110.

[57] Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer, 2006.

[58] 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.

[59] Fang, K.-T.; Li, R.; Sudjianto, A. *Design and Modeling for Computer Experiments*; Chapman & Hall/CRC.

[60] Caflisch, R. E. Monte Carlo and Quasi-Monte Carlo Methods, *Acta Numerica* **1998**, *7*, 1–49.

[61] Lemieux, C. *Monte Carlo and Quasi-Monte Carlo Sampling*; Springer, 2009.

[62] Press, W. H. *Numerical Recipes in Fortran: The Art of Scientific Computing*; Cambridge University Press: New York, 1992; p. 309.

[63] 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.

[64] Carroll, D. L. In *Developments in Theoretical and Applied Mechanics*, *Vol. XVII*; Wilson, H.; Batara, R.; Bert, C.; Davis, A.; Schapery, R.; Stewart, D.; Swinson, F., Eds.; School of Engineering, The University of Alabama: Tuscaloosa, AL, 1996; p. 411.

[65] Vayner, G.; Alexeev, Y.; Wang, J.; Windus, T. L.; Hase, W. L. Ab Initio and Analytic Intermolecular Potentials for Ar–CF_4. *J. Phys. Chem. A* **2016**, *110*, 3174–3178.

[66] Ma, P.; Mahoney, M. W.; Yu, B. A Statistical Perspective on Algorithmic Leveraging, *J. Machine Learning Res.* **2015**, *16*, 861–911.

[67] Wang, Y.; Yu, A. W.; Singh, A. On Computationally Tractable Selection of Experiments in Measurement-Constrained Regression Models, *J. Machine Learning Res.* **2017**, *18*, 5238–5278.

[68] Jasper, A. W.; Oana, C. M.; Miller, J. A. "Third-Body" Collision Efficiencies for Combustion Modeling: Hydrocarbons in Atomic and Diatomic Baths *Proc. Combust. Inst*. **2015**, *35*, 197–204.

[69] Diwekar, U. *Introduction to Applied Optimization*; Springer, 2008.

[70] Miller, A. J. *Subset Selection in Regression, Second Edition*; Chapman & Hall/CRC, 2002.

[71] Hastie, T.; Tibshirani, R.; Friedman, J. *Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*; Springer, 2009.

Figure Captions

Fig. 1. Average, minimum, and maximum (a) V_I calculated using cc MP2/CBS(a'd,a't) and (b) w calculated using eq 7 for CH_3OH + Ar as a function of the center of mass distance r. Panel (c) shows an empirical biasing function calculated as max$w$/<w> (solid line) and a coarse grained approximation to it (dotted line).

Fig. 2. CH_3OH + Ar interaction energies for the six cuts through V_I used as the out of sample test set. The filled symbols are the calculated cc MP2/CBS(a'd,a't) energies (connected with dotted splines), and the solid lines show the exp6(4) fit to these energies. The two panels contain the same data and highlight (a) the repulsive wall and (b) the van der Waals region.

Fig. 3. CH_3OH + N_2 interaction energies for the six cuts through V_I used as the out of sample test set. The filled symbols are the calculated cc MP2/CBS(a'd,a't) energies (connected with dotted splines), and the solid lines show the (a) exp6(4) and (b) PIP22(4) fits to these energies.

Fig. 4. Convergence of the in sample f (black solid lines) and out of sample f_{oos} (blue dashed lines) fitting errors for CH_3OH + Ar as a function of the number of *ab initio* data M in the training set for the (a) P, (b) bP, (c) S, and (d) bS sampling strategies and for the exp6(3) functional form. For each M, 100 parameterizations were generated, and the average fitting errors are shown as thick lines, with the minimum and maximum fitting errors shown as surrounding thin lines. The error bars indicate one standard deviation in f_{oos}. The large-M limit of f_{oos} (f_∞) is shown as a green dotted horizontal line.

Fig. 5. Convergence of (a) $f_{oos}^{0.95}$, (b) <f_{oos}>, and (c) s_{oos} for CH_3OH + Ar and the exp6(3) functional form as a function of the number of fitted *ab initio* data M for the P

(black solid lines), bP (black dashed lines), S (blue solid lines), and bS (blue dotted lines) data sets. In (a), the horizontal line and circles indicate 1.3 f_∞ and M^*, respectively.

Fig. 6. Log-log plots of the fitting-basis representation error f_∞ as a function of N_p using the data from Tables 1–4 and with the following notation: exp6 (open symbols and dotted lines), PIP (closed symbols and solid lines), $N_u = 3$ (triangles), and $N_u = 4$ (squares). The colors indicate different systems as noted in the legend. In (b), the absolute errors in (a) are shown scaled by approximate van der Waals well depths, v, to give relative errors.

Fig. 7. Log-log plots of the fitting efficiencies M^* for the (a) bS and (b) P data from Tables 1–3, using the same notation as in Fig. 6. In (c) and (d), the results in (a) and (b) are shown scaled by N_p.

Fig. 8. Average efficiency improvements as a function of (a) dimensionality of the interaction potential and (b) alcohol size.

Fig. 9. Convergence of the average energy in deactivating collisions, α, as a function of the number of fitted *ab initio* data M for the P (black solid lines) and bS (blue dashed lines) training sets for the exp6(4) functional form and CH_3OH + Ar at 1000 K. Trajectory simulations were carried out for 10 parameterizations for each value of M. The averages of the resulting α are shown as the thick lines, with the thin lines indicating two standard deviations. The converged value of α is shown as a horizontal solid green line, with the dashed green lines indicating two standard deviations of the statistical error. The vertical dotted lines indicate the relevant values of M^* from Table 1.

Fig. 10. Convergence of the computed trajectory-based collision parameters (a) α and (b) Z with respect to M for four sampling strategies, the exp6(3) functional form, and CH_3OH + Ar at 1000 K. The vertical lines show the values of M* from Table 1.

Fig. 11. Trajectory-based collision parameters (a,b) Z and (c,d) α for CH_3OH + Ar at 1000 K shown as functions of N_p or f_∞ for four sampling strategies and nine functional forms (see Table 1). Error bars on the exp6 results (open symbols) indicate the 8% 2-sigma statistical uncertainties in these predictions. Similar 2-sigma error bars are present in the PIP results (closed symbols) but are not shown to reduce clutter.

Figure 1

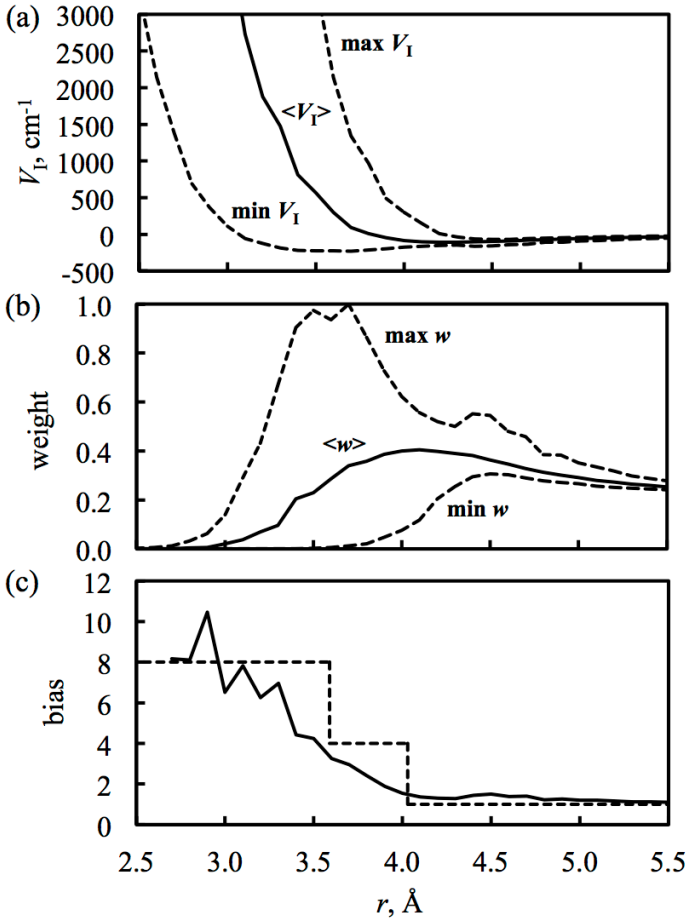


Fig. 1. Average, minimum, and maximum (a) V_I calculated using cc MP2/CBS(a'd,a't) and (b) w calculated using eq 7 for CH_3OH + Ar as a function of the center of mass distance r. Panel (c) shows an empirical biasing function calculated as max$w$/<w> (solid line) and a coarse grained approximation to it (dotted line).

Figure 2

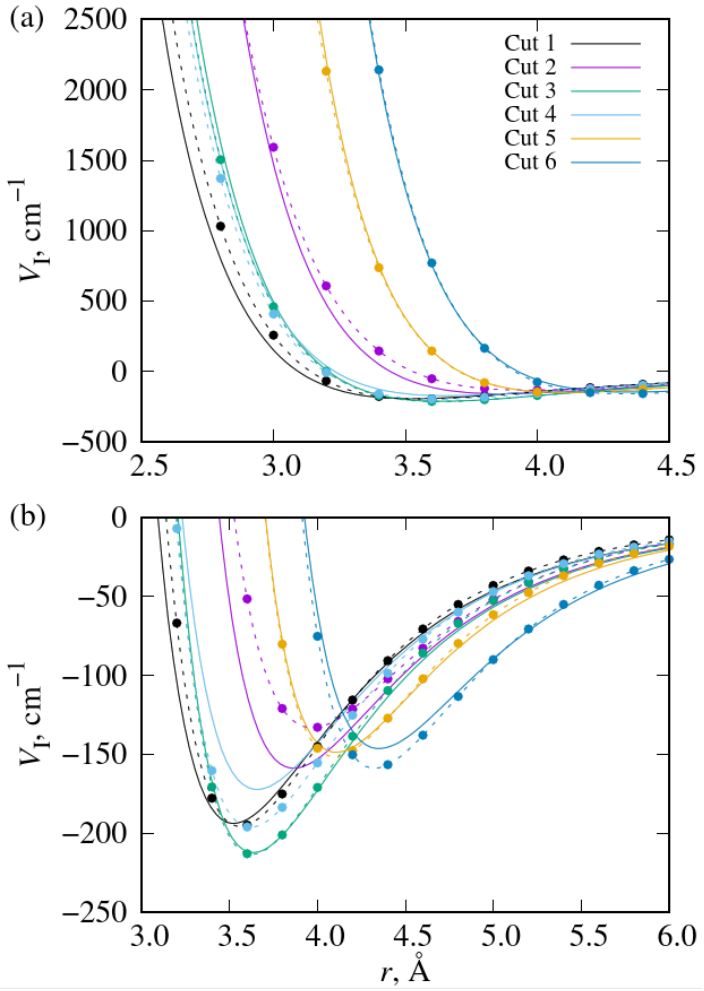


Fig. 2. CH_3OH + Ar interaction energies for the six cuts through V_I used as the out of sample test set. The filled symbols are the calculated cc MP2/CBS(a'd,a't) energies (connected with dotted splines), and the solid lines show the exp6(4) fit to these energies. The two panels contain the same data and highlight (a) the repulsive wall and (b) the van der Waals region.

Figure 3

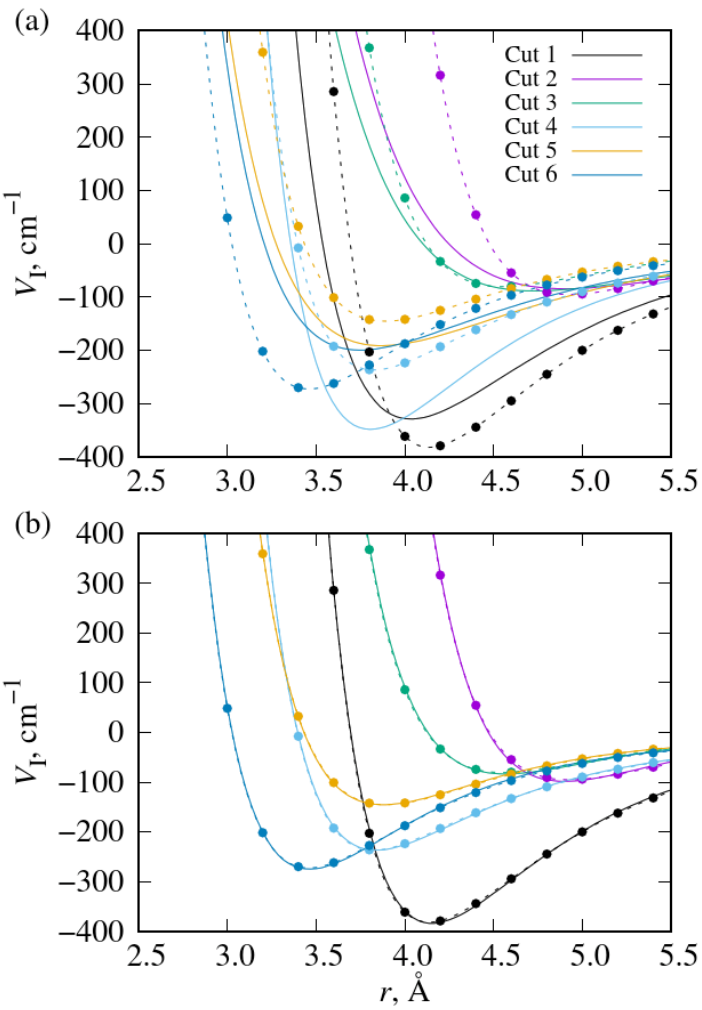


Fig. 3. $CH_3OH + N_2$ interaction energies for the six cuts through V_1 used as the out of sample test set. The filled symbols are the calculated cc MP2/CBS(a'd,a't) energies (connected with dotted splines), and the solid lines show the (a) exp6(4) and (b) PIP22(4) fits to these energies.

Figure 4

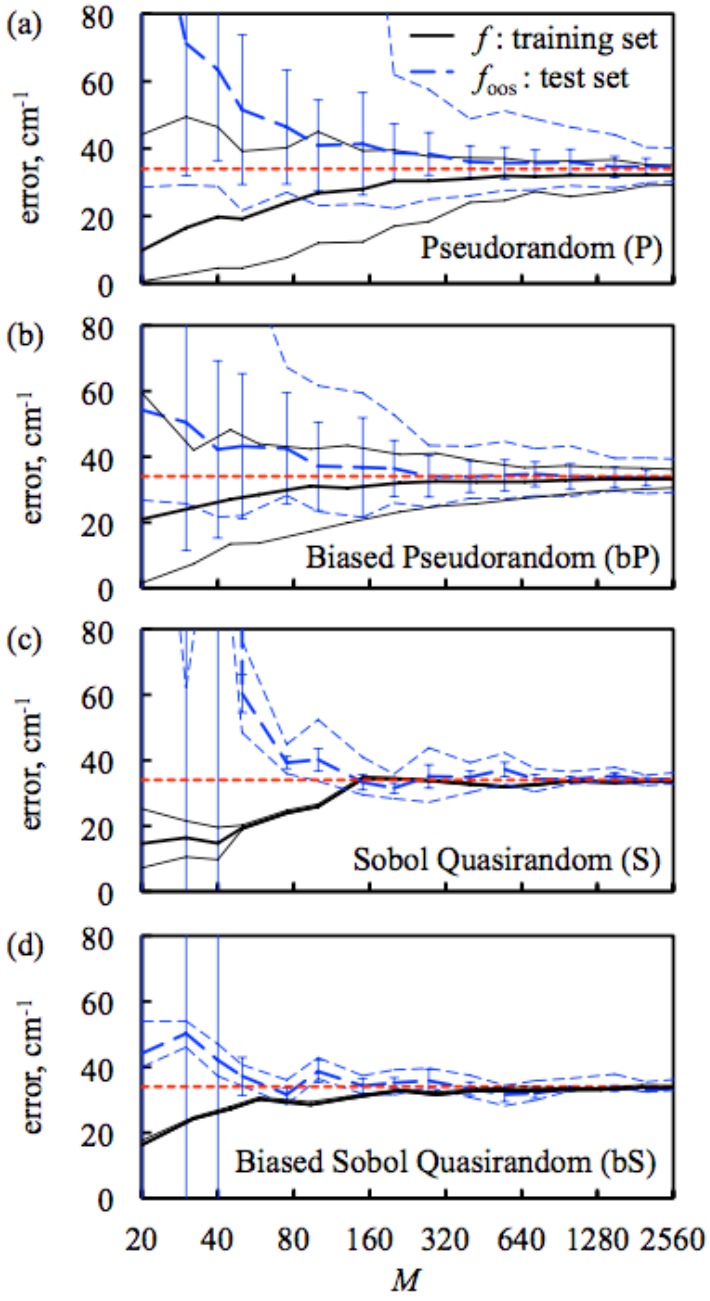


Fig. 4. Convergence of the in sample f (black solid lines) and out of sample f_{oos} (blue dashed lines) fitting errors for CH_3OH + Ar as a function of the number of *ab initio* data M in the training set for the (a) P, (b) bP, (c) S, and (d) bS sampling strategies and for the exp6(3) functional form. For each M, 100 parameterizations were generated, and the average fitting errors are shown as thick lines, with the minimum and maximum fitting errors shown as surrounding thin lines. The error bars indicate one standard deviation in f_{oos}. The large-M limit of f_{oos} (f_∞) is shown as a green dotted horizontal line.

Figure 5

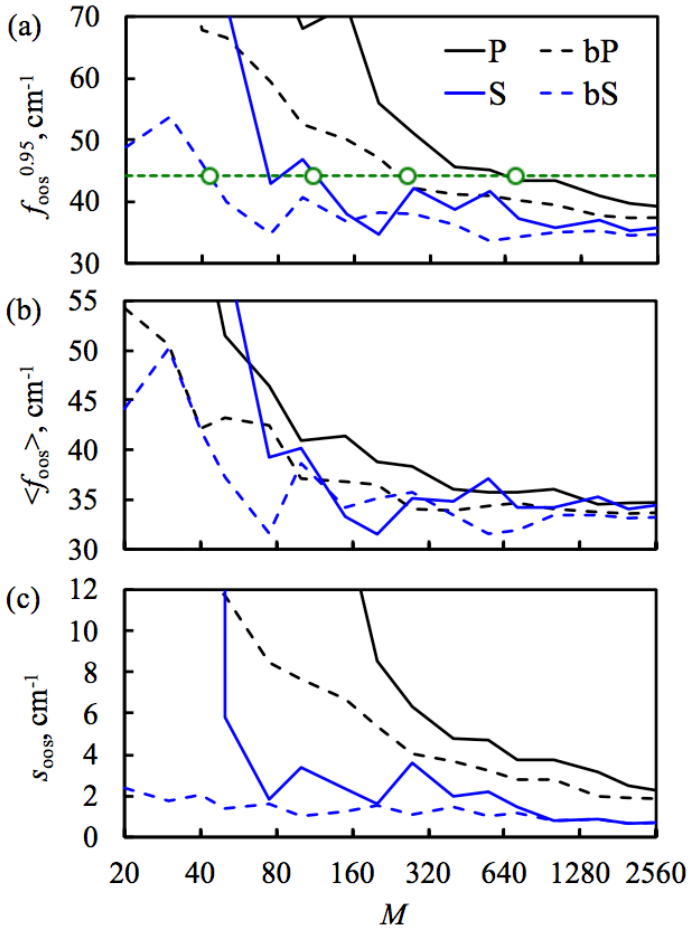


Fig. 5. Convergence of (a) $f_{oos}^{0.95}$, (b) <f_{oos}>, and (c) s_{oos} for CH_3OH + Ar and the exp6(3) functional form as a function of the number of fitted *ab initio* data M for the P (black solid lines), bP (black dashed lines), S (blue solid lines), and bS (blue dotted lines) data sets. In (a), the horizontal line and circles indicate 1.3 f_∞ and M*, respectively.

Figure 6

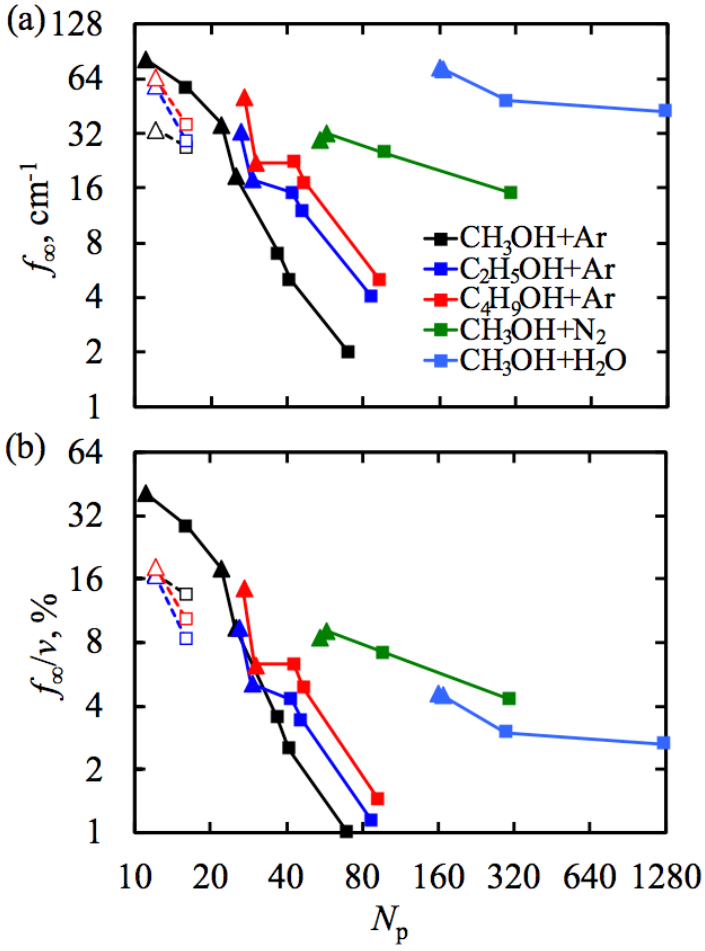


Fig. 6. Log-log plots of the fitting-basis representation error f_∞ as a function of N_p using the data from Tables 1–4 and with the following notation: exp6 (open symbols and dotted lines), PIP (closed symbols and solid lines), $N_u = 3$ (triangles), and $N_u = 4$ (squares). The colors indicate different systems as noted in the legend. In (b), the absolute errors in (a) are shown scaled by approximate van der Waals well depths, v, to give relative errors.

Figure 7

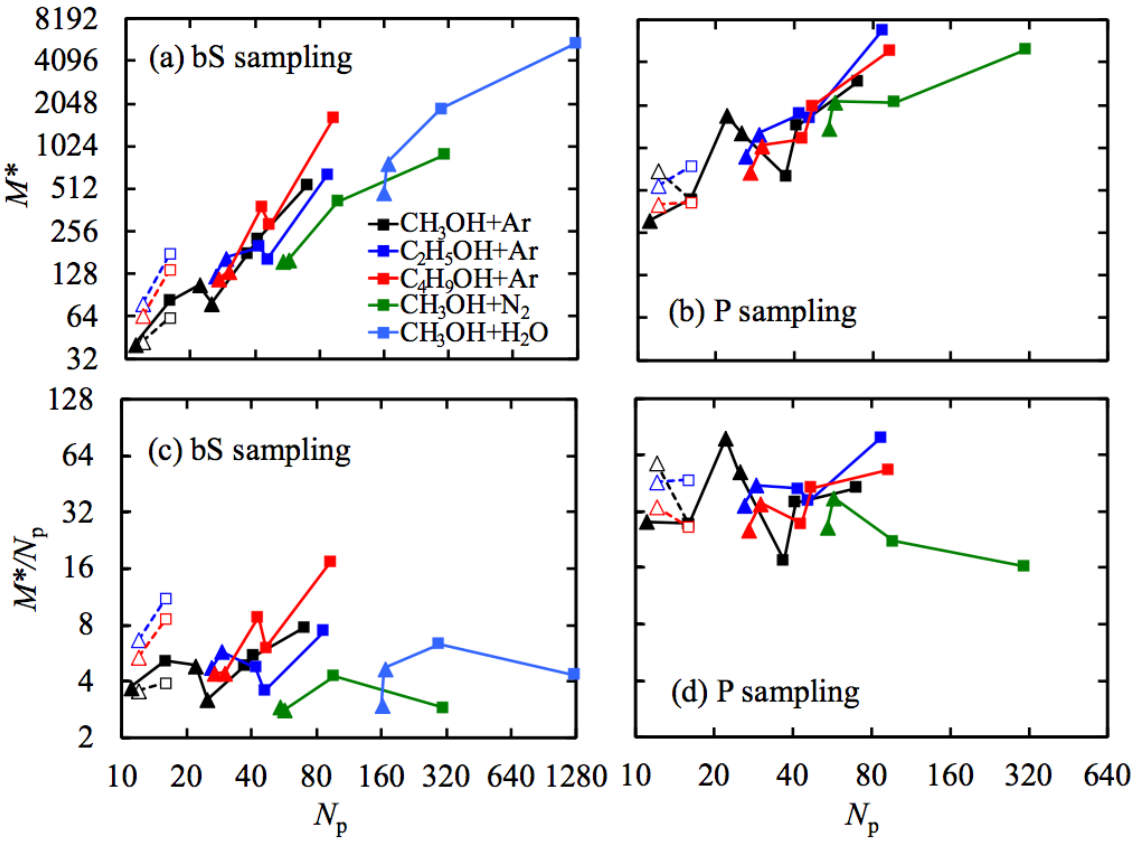


Fig. 7. Log-log plots of the fitting efficiencies M^* for the (a) bS and (b) P data from Tables 1–3, using the same notation as in Fig. 6. In (c) and (d), the results in (a) and (b) are shown scaled by N_p.

Figure 8

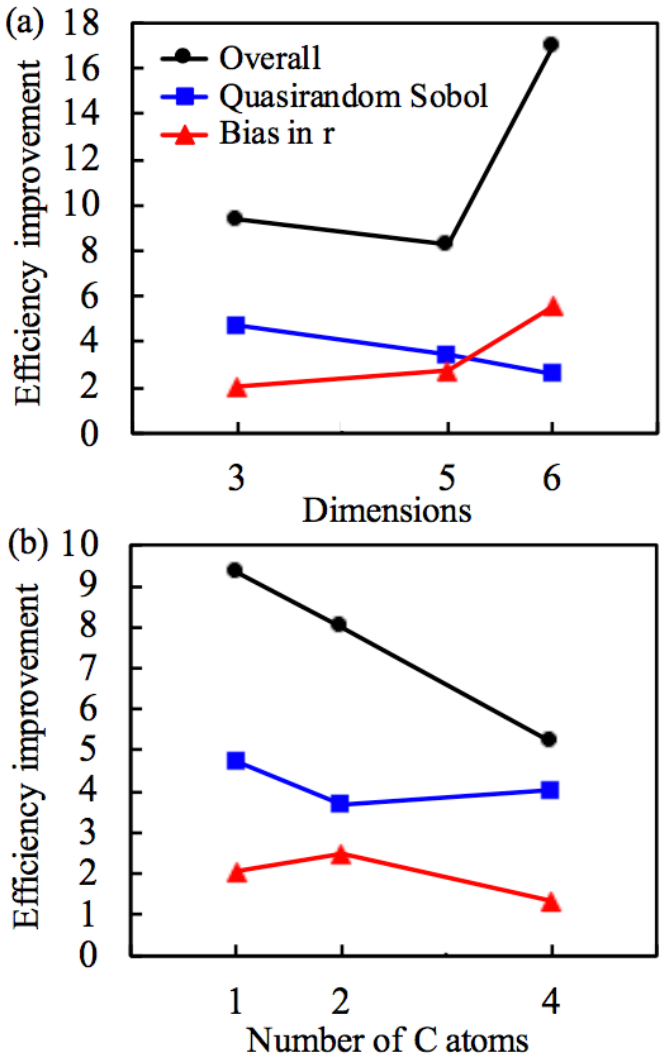


Fig. 8. Average efficiency improvements as a function of (a) dimensionality of the interaction potential and (b) alcohol size.

Figure 9

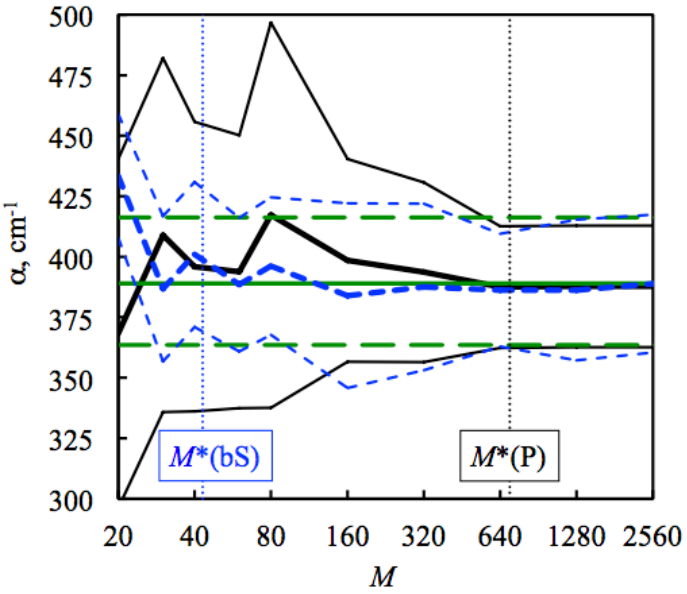


Fig. 9. Convergence of the average energy in deactivating collisions, α, as a function of the number of fitted *ab initio* data M for the P (black solid lines) and bS (blue dashed lines) training sets for the exp6(4) functional form and CH_3OH + Ar at 1000 K. Trajectory simulations were carried out for 10 parameterizations for each value of M. The averages of the resulting α are shown as the thick lines, with the thin lines indicating two standard deviations. The converged value of α is shown as a horizontal solid green line, with the dashed green lines indicating two standard deviations of the statistical error. The vertical dotted lines indicate the relevant values of M* from Table 1.

Figure 10

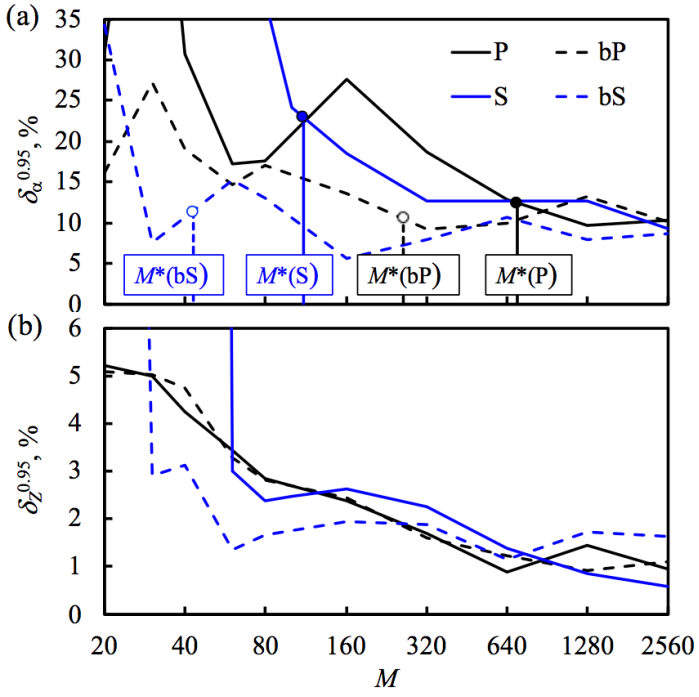


Fig. 10. Convergence of the computed trajectory-based collision parameters (a) α and (b) Z with respect to M for four sampling strategies, the exp6(3) functional form, and CH_3OH + Ar at 1000 K. The vertical lines show the values of M* from Table 1.

Figure 11

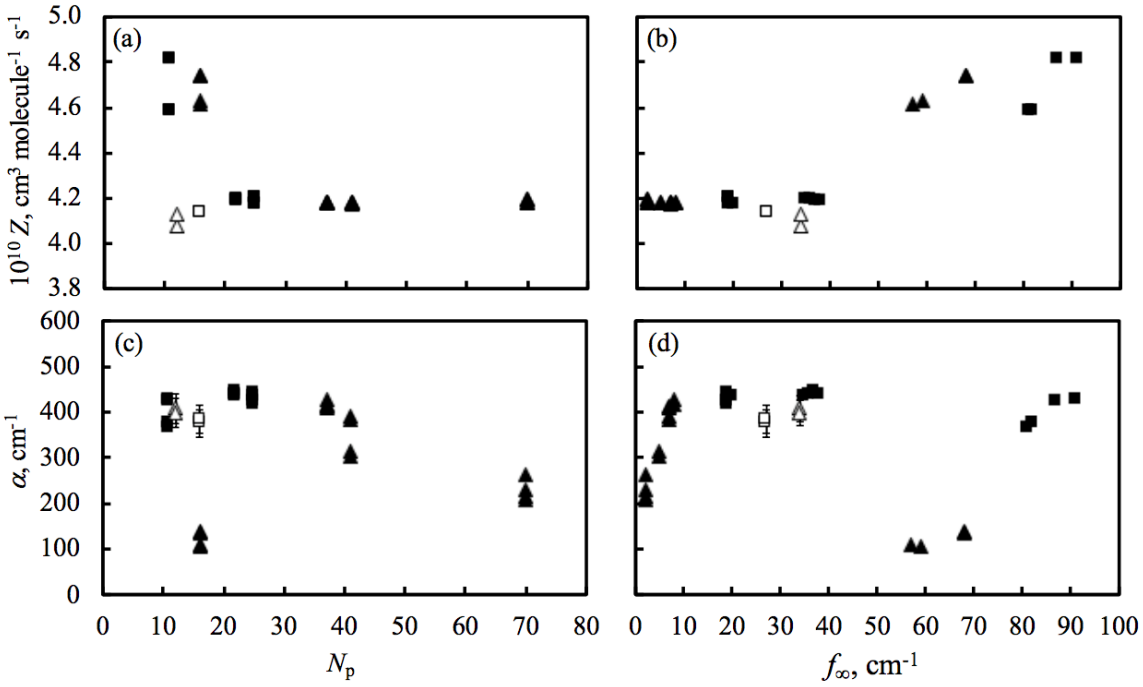
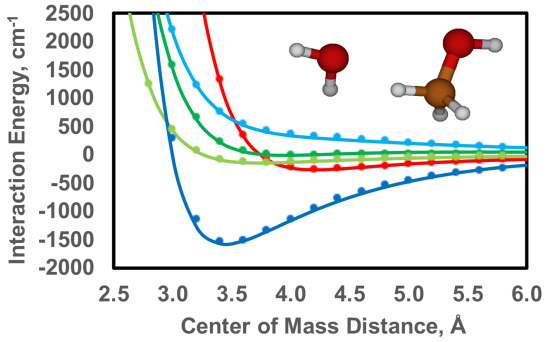


Fig. 11. Trajectory-based collision parameters (a,b) Z and (c,d) α for CH_3OH + Ar at 1000 K shown as functions of N_p or f_∞ for four sampling strategies and nine functional forms (see Table 1). Error bars on the exp6 results (open symbols) indicate the 8% 2-sigma statistical uncertainties in these predictions. Similar 2-sigma error bars are present in the PIP results (closed symbols) but are not shown to reduce clutter.

TOC Graphic


Interaction Energy, cm-1
2500
2000
1500
1000
500
0
-500
-1000
-1500
-2000
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Center of Mass Distance, Å