Parsimonious potential energy surface expansions using dictionary learning with multi-pass greedy selection

Daniel R. Moberg, Ahren W. Jasper, and Michael J. Davis**

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

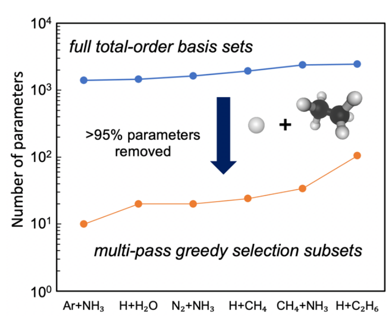
AUTHOR INFORMATION

Corresponding Authors

*E-mail: ajasper@anl.gov, davis@tcg.anl.gov

ABSTRACT. Potential energy surfaces fit with basis set expansions have been shown to provide accurate representations of electronic energies and have enabled a variety of high-accuracy dynamics, kinetics, and spectroscopy applications. The number of terms in these expansions scales poorly with system size, a drawback that challenges their use for systems with more than ~10 atoms. A solution is presented here using dictionary learning. Subsets of the full set of conventional basis functions are optimized using a newly developed multi-pass greedy regression method inspired by forward and backward selection methods from the statistics, signal processing, and machine learning literatures. The optimized representations have accuracies comparable to the full set but are one or more orders of magnitude smaller, and, notably, the number of terms in the optimized multi-pass greedy expansions scales approximately linearly with the number of atoms.

TOC GRAPHICS



KEYWORDS. Permutationally invariant polynomials, potential energy surfaces, machine learning, feature selection

Potential energy surfaces (PESs) describe the electronic energy dependence on nuclear geometry, and analytic representations are sought to enable efficient dynamics and kinetics calculations. Early strategies for constructing analytic PESs for molecular systems referenced the underlying electronic Hamiltonian, defining systems using natural coordinates such as bond lengths, bond angles, pairwise van der Waals interactions, etc. Recently it has become popular to instead represent PESs as mathematical expansions containing large numbers of parameters with no direct reference to molecular systems. The key advantages of this newer approach are that it is both broadly applicable and systematically improvable. Following the further improvement through enforcement of permutational invariance demonstrated by Braams and Bowman,[1] many permutationally invariant polynomial (PIP) expansions for molecular and chemical systems have been produced. For recent representative work from several groups see Refs. 2–16.

Nevertheless, a major limitation of PES expansions is their poor scaling with dimensionality. In fact, right away pioneers of this strategy addressed the challenge of a "ten atom limit," referring to the challenge of generating usable PIPs for larger systems due to the associated exponential increase in the number of terms.[8,17] This limit has been surpassed, sometimes using a fragmented approach to basis set generation and sometimes using high performance computing.[8,15–17,20] Here, we address this fundamental scaling problem using machine learning, noting that exponential scaling is clearly at odds with the low-order or even linear scaling of more traditional PES fitting approaches such as molecular mechanics, which likely better reflects the true complexity of molecular PESs.

We demonstrate here that dictionary learning[21–23] approaches provide optimized PIP expansions with improved scaling and without significant loss of accuracy. A new multi-pass greedy selection procedure is described based on methods developed in the statistics, signal processing, and machine learning literatures.[24–27] A key component of dictionary learning algorithms is a selection procedure wherein elements from the dictionary are chosen to best fit a set of training data. Here, the dictionary is a basis of permutationally invariant collections of terms of a PIP expansion, and the data are sampled ab initio energies describing a molecular potential energy surface. As PIPs have explicit functional forms and systematically improvable basis sets, they are better suited for dictionary learning than other popular PES generation methods such as Gaussian approximation potentials and neural networks.[28]

The dictionaries used in this Letter are total-order PIP expansion basis sets with the form

$$V = \sum_{k=1}^{N_t} c^{(k)} \prod_{j=1}^{p} y_{ji}^{\xi_j^{(k)}}, \tag{1}$$

where each term k is a product of the Morse variables[29] y_{ji} of all possible $p = n(n-1)/2$ atom–atom pairs, n is the number of atoms, $c^{(k)}$ are the linear expansion coefficients, and each of the N_t terms in the expansion is distinguished by a unique set of exponents $\xi_j^{(k)}$. A common strategy is to truncate Eq. (1) according to a total-order constraint $o \leq \sum_j \xi_j^{(k)}$, which results in expansions that scale as $N_t = (p+o)!/p!/o!$. The number of terms is somewhat smaller than this maximum value due to the removal of unphysical (so-called unconnected) terms to enforce the correct asymptotic limit, as described elsewhere.[8,16,30] Consider the limit $p \gg o$, generally valid for systems larger than about 5 atoms, where N_t scales as $\sim p^o$ or $\sim n^{2o}$. Typical values of o are $o = 4$ to 8, and this severe scaling limits the size of systems that can be treated using total-order PIP expansions.[8,17] The number of free parameters N_p is smaller than N_t when permutational invariance is enforced,[1] and the dictionaries used here are of size N_p and consist of the permutationally invariant collections of terms.

Figure 1 compares the values of N_t and N_p for six systems: three collisional energy transfer (CET) cases NH_3(+M), M = Ar, N_2, and CH_4, designated by dashed lines, and three reactive cases

$H + H_2O \rightleftarrows H_2 + O_2$, $H + CH_4 \rightleftarrows H_2 + CH_2$, and $H + C_2H_6 \rightleftarrows H_2 + C_2H_5$, designated by solid lines. Detailed descriptions of the construction of total-order PIP expansions for these six systems, including details of the ab initio methods and point selection for the training data sets, the permutational symmetrization algorithm, and trajectory applications for CETs and thermal kinetics, have been given recently.[16] The exponential increases in N_t and N_p with PIP order o is immediately evident in Figure 1.

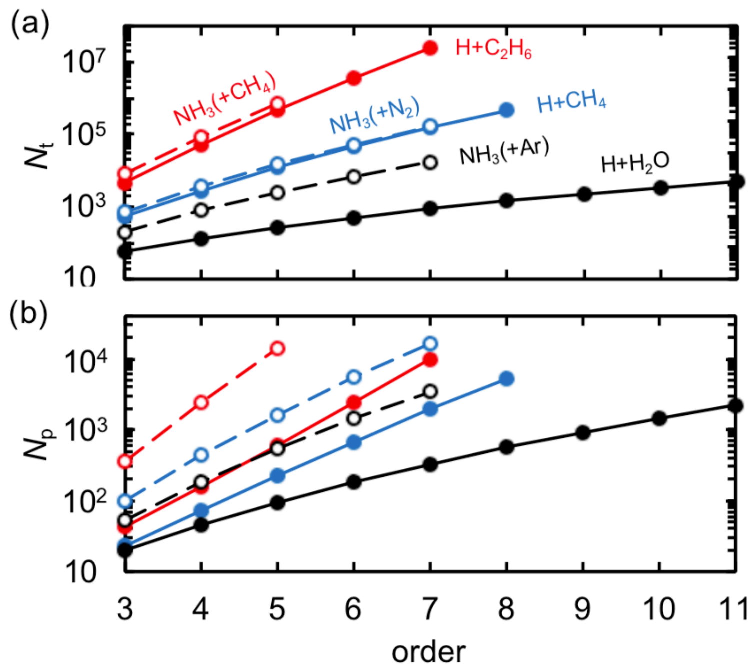


Figure 1. The number of (a) terms N_t and (b) free parameters N_p in total-order PIP expansions for six systems. Dashed lines indicate collisional energy transfer (CET) systems and solid lines indicate reactive systems.

A method of analyzing the scaling of PIP expansions and of generating much smaller optimized expansions with improved scaling is now demonstrated. The expansion coefficients in Eq. (1) are determined by minimizing the weighted sum

$$E_{\mathrm{rms}} = \sqrt{\frac{\sum_{i=1}^{M} w_i^2 (V(\mathbf{R}_i)\text{-}\hat{V})^2}{\sum_{i=1}^{M} w_i^2}}, \qquad (2)$$

where i labels values of a training data set consisting of M ab initio potential energies $V(\mathbf{R}_i)$ computed at the geometries $\mathbf{R}_i$, and $\widehat{V}$ is the fitted expansion value. The training data sets were limited to energies below 50 kcal/mol and 4000 cm^{-1} (~11.4 kcal/mol) relative to the reactants for the reactive and CET cases, respectively. The definition of the weights, w_i, is given in Ref. 31 and is designed to weight lower energies more than higher energies and to distribute the relative error approximately uniformly throughout the expansion. The values of the coefficients are obtained from numerical linear algebra,[32]

$$\mathbf{c} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y},\ \mathbf{y} = \mathbf{WV},\ \mathbf{X} = \mathbf{WB} \tag{3}$$

where the vector $\mathbf{V}$ is defined above as the values of the PES and the elements of $\mathbf{B}$, B_{ij}, describe the j^{th} basis function evaluated at the i^{th} geometry in the training set. Each PIP basis function is the collection of all permutationally equivalent terms, such that the size of $\mathbf{B}$ is N_{p} x M. The matrix $\mathbf{W}$ is the diagonal weight matrix from Eq. (2). Reference 33 and the supporting information for this paper have further information, including the calculation of a diagnostic leave-one-out cross-validation error, E_{loocv}.

To tame the exponential scaling in Figure 1, we show that we can make the following replacement

$$\mathbf{X}\left(N_{\mathrm{p}} \times M\right) \rightarrow \left(\text{dictionary learning}\right) \rightarrow \mathbf{X}_{\mathrm{d}}\left(N_{\mathrm{d}} \text{ x } M_{\mathrm{d}}\right),\ N_{\mathrm{d}} \ll N_{\mathrm{p}},\ M_{\mathrm{d}} \ll M, \tag{4}$$

and still have acceptable errors in E_{rms} and E_{loocv}. Furthermore, the sizes of the optimized expansions N_{d} scale favorably with system size and the number of training data required M_{d} can be reduced relative to the full training set of size M. We use a fixed dictionary rather than a learned dictionary,[22] and optimized subsets of the dictionary are "learned" via selection in a systematic manner to control the fitting error E_{rms} and the cross-validation error E_{loocv}.

Dictionary learning targeting small subsets of the dictionary works well if the problem of interest is sparse or nearly sparse. Sparsity has become a key concept in signal and image processing, as well as machine learning,[26,34,35] and Figure 2 shows a calculation with the idea of sparsity in mind. The results of 10^9 randomly chosen sets of size $N_d = 56$ from the PIP10 dictionary for H + H_2O were evaluated, and the resulting probability density function of E_{rms} is shown to peak around ~13 kcal/mol and no error is smaller than 3.1 kcal/mol, much larger than desirable. The lack of good fits from one billion random draws is not surprising as there are 10^{103} unique subsets of size 56 in the full PIP10 set with $N_p = 1461$.

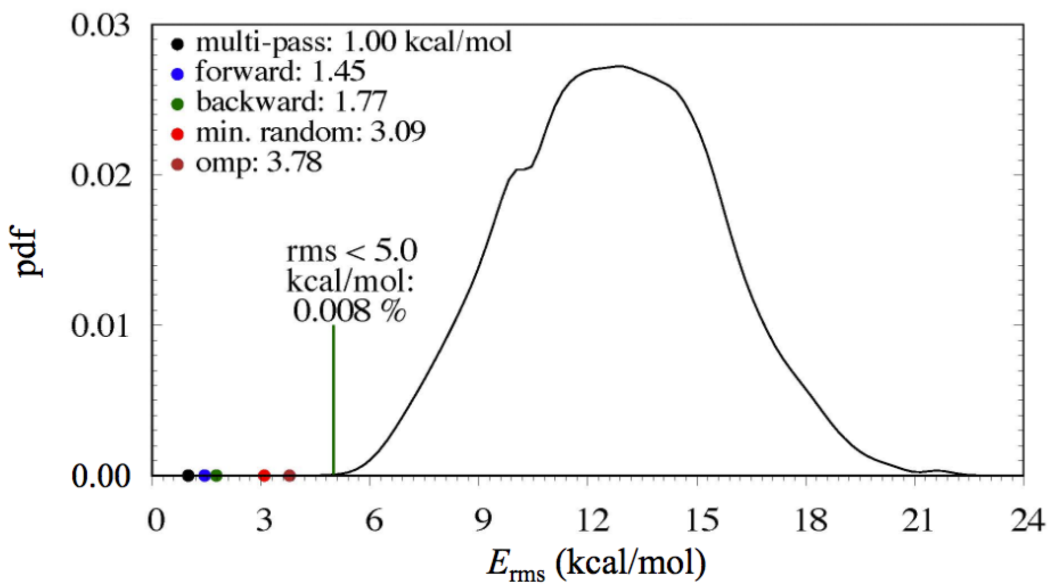


Figure 2. Probability distribution function (pdf) of E_{rms} for 10^9 randomly generated subsets of size 56 sampled from the PIP10 dictionary of H + H_2O. Errors for subsets of size 56 determined using four dictionary learning methods, as described in detail in the text, are also indicated.

Because a random draw is not sufficient for taking advantage of the presumed sparsity of our problem, we turned to greedy selection methods, otherwise known as stepwise regression methods. These methods are much faster than other sparse methods such as penalized regression techniques.[34,36–39] Three standard greedy methods were tested here: *forward greedy regression*[24] where basis functions are added sequentially and one-at-a-time from the dictionary such that the reduction in E_{rms} is maximized at each step; *backward greedy regression*[39] where basis functions are removed sequentially and one-at-a-time such that the increase in E_{rms} is minimized; and *orthogonal matching pursuit*[25,26] where basis functions are added sequentially based on their projection on the current residual vector. The cost of the backward greedy approach scales poorly with o as many large-N_d calculations must be done to reach small subset sizes, and it is the slowest of these methods by far. The selection procedure for the forward and backward methods is made more efficient by using update and downdate algorithms for the matrix inversion in Eq. (3).[40,41] Orthogonal matching pursuit (omp) is even faster than forward greedy regression, as the selection step is faster, but yields higher E_{rms} as shown next. For the applications considered here, the greedy selection calculations were much less computationally demanding than other aspects of PIP generation, such as symmetrizing the basis set for the largest expansions as well as generating the ab initio training and test data sets.

Figure 3a shows the results of the three standard greedy subset selection methods for H + H_2O using the PIP10 dictionary (N_p = 1461) and trained using M_d = 5449 ab initio energies. For all three standard approaches, E_{rms} is shown to rapidly decrease with N_d for small values of N_d, followed by a more gradual reduction in E_{rms} with increasing N_d. These methods achieve fitting errors of 1 kcal/mol with modest subset sizes of N_d = 88, 122, and 188 for the forward, backward, and omp methods, respectively. The subsets compare favorably with the results in Figure 2 for N_d = 56, with the forward and backward greedy methods outperforming one billion random samples. Regardless of the selection procedure used, the curves in Figure 3a demonstrate that only a small fraction of basis functions in total-order PIP expansions contribute significantly to reducing the overall error.

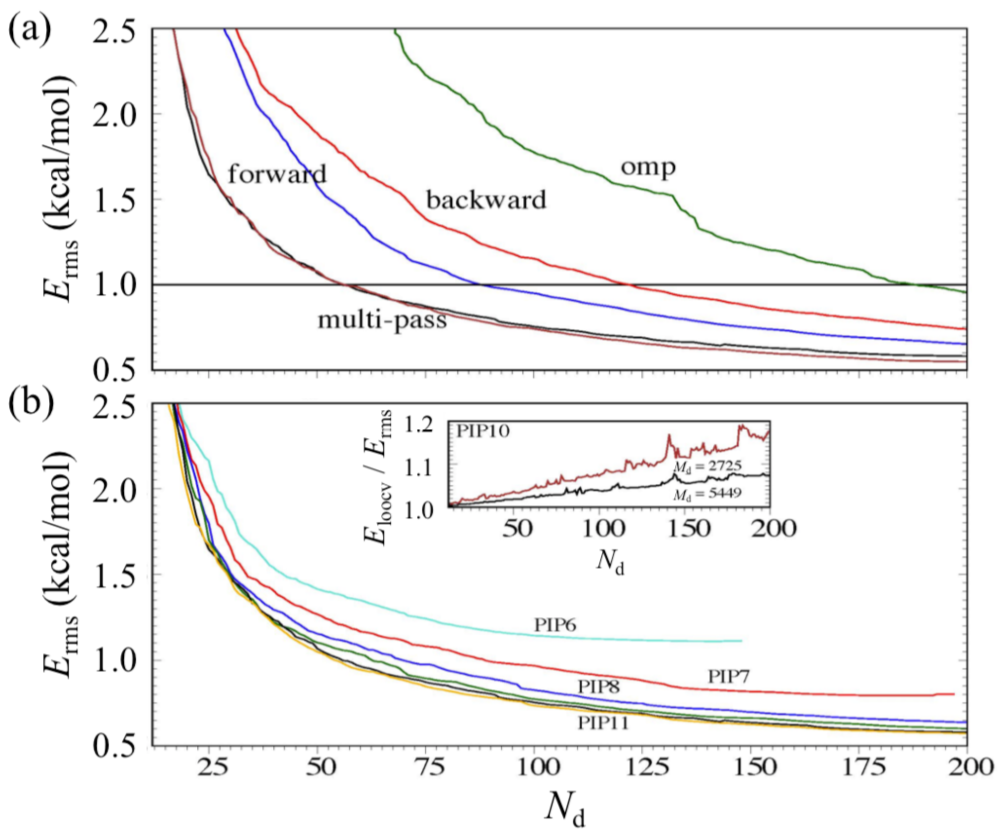


Figure 3. (a) Fitting errors E_{rms} for subsets of size N_d determined using the PIP10 dictionary of H + H_2O and four dictionary learning methods: multi-pass greedy, forward greedy, backward greedy, and orthogonal matching pursuit (omp). Also shown are the result of the multi-pass greedy method using half as many training data (brown). (b) Fitting errors E_{rms} for subsets of size N_d optimized using the multi-pass greedy method and several dictionaries, PIP6 through PIP11, of H + H_2O. The inset shows the ratio of the cross-validation error and rms error E_{loocv}/E_{rms} vs. N_d for PIP10 at two values of M_d.

Based on both its accuracy and relative cost, the forward greedy method is preferred among the standard approaches. One can do even better if a combination of forward and backward steps is used. Although it is not common, some applications have used a combination of forward steps with a few backwards steps,[24,27] often based on whether the backward step degrades the fit within a tolerance limit. We have used these methods as a starting point for the multi-pass algorithm, which uses many forward and backward steps.

The multi-pass greedy algorithm uses a series of forward and backward steps with the same rules as the forward and backward greedy methods described above, which are single-pass methods. A maximum subset size n_s and number of passes n_p are chosen. The first pass begins with the single basis function that minimizes E_{rms}, and the number of steps in the pass is chosen randomly with the constraint that the maximum subset size sampled on that pass is not greater than n_s. A series of backward steps is then taken from the final value of the forward pass, and this number of steps is again chosen randomly with the constraint that the number of steps is no greater than n_s and leads to subsets between 1 and the last value of the forward pass. The procedure is repeated n_p–1 additional times with each new pass starting from the last backward step of the previous pass. Errors for all chosen subsets of all passes are recorded. A series of calculations using a range of n_s values is done, each of which generates its own set of subsets. The final subsets are those with the smallest E_{rms} for each value of N_d from all the computations, which is shown in Figure 3a. The supporting information contains more details of the multi-pass greedy calculations. As shown in Figure 3a, the multi-pass greedy method produces optimized subsets requiring fewer basis functions than the standard greedy approaches, with E_{rms} < 1 kcal/mol for N_d = 56.

A variety of calculations were performed to examine the behavior of the multi-pass algorithm and its robustness. The dependence of E_{rms} for the optimized subsets on the order of the dictionary o is shown in Figure 3b for o = 6–11, again for H + H_2O. Figure 3a also compares the multi-pass greedy E_{rms} for M_d = 5449 and 2725, and the results are nearly indistinguishable. For all of the subsets presented here, we confirmed that E_{loocv} was within 20% of E_{rms} over the whole range of subsets considered, which is more than adequate from our experience.[33] Finally, we considered criteria other than E_{rms} alone for evaluating basis functions at each step, including enforcing a 10% or 20% cutoff in E_{loocv} relative to E_{rms} during the selection process, and that yielded nearly indistinguishable subsets.

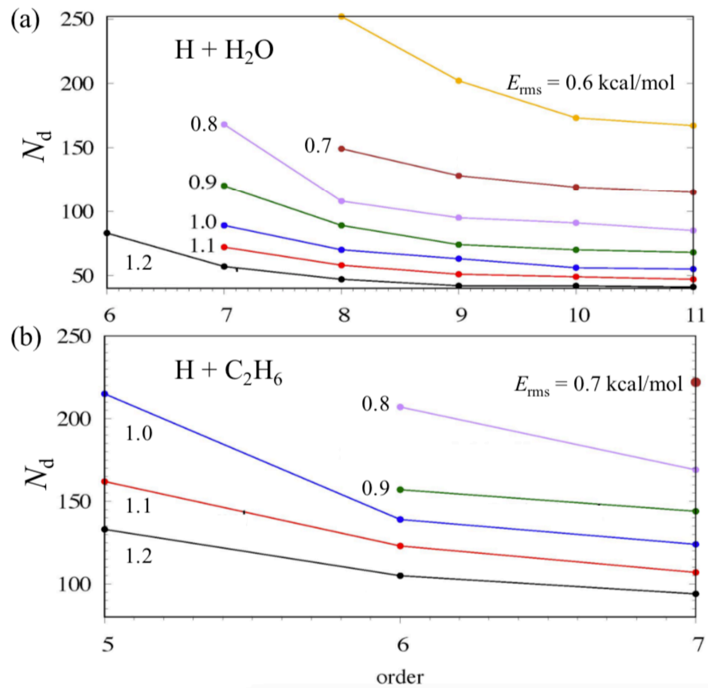


Figure 4. Subset sizes N_d with fitting errors of E_{rms} = 0.6 to 1.2 kcal/mol shown as functions of the order of the PIP dictionary for two systems: (a) $H + H_2O$ and (b) $H + C_2H_6$.

Perhaps the most important results of the dictionary learning methodology are presented in Figure 4 and Table 1. Figure 4 shows multi-pass greedy basis set sizes N_d for E_{rms} thresholds from 1.2 to 0.6 kcal/mol, for dictionaries of varying orders, and for two reactive systems: (a) $H + H_2O$ and (b) $H + C_2H_6$. The poor scaling with order of the total-order basis sets in Figure 1 has been mitigated and in fact reversed using dictionary learning. Higher-order PIP expansions include more flexible basis functions, inverting the trends seen in Figure 1. Furthermore, the number of basis functions N_d in the multi-pass greedy subsets required to accurately treat $H + C_2H_6$ is only moderately greater than for the smaller system $H + H_2O$, suggesting dramatic improvements in system size scaling, as well.

Table 1. Selected expansion information

System	V_{ref}[a]	n	d[b]	Dict.[c]	N_p[d]	M_d[e]	$N_d(0.15)$[f]	$N_d(0.10)$	$N_d(0.08)$	$\alpha(0.15)$[g]	$\alpha(0.10)$	$\alpha(0.08)$
$H+H_2O$	21.9	4	6	10	1461	5,449	12	20	25	1.01	1.01	1.01
$H+CH_4$	14.9	6	12	7	1933	5,339	14	24	32	1.01	1.01	1.02
$H+C_2H_6$	12.0	9	21	6	2455	5,000	59	105	146	1.02	1.04	1.06
$Ar+NH_3$	94.0	5	3	6	1408	14,000	5	10	15	1.01	1.01	1.01
N_2+NH_3	238	6	5	5	1638	14,000	7	20	27	1.01	1.13	1.08
CH_4+NH_3	244	9	6	4	2389	8,000	11	34	53	1.01	1.08	1.07

[a] V_{ref} is the saddle point energy (kcal/mol) for the reactive cases and the well depth (cm^{-1}) for the CET cases.
[b] The number of important degrees of freedom for each system. For reactive cases this is the number of internal coordinates and for the CET cases this is the number of internuclear coordinates.
[c] The order of the PIP dictionary used in the multi-pass greedy optimization.
[d] The number of basis functions in the dictionary.
[e] The number of training data.
[f] $N_d(f)$ is number of multi-pass greedy optimized basis functions required for an error equal to fV_{ref}.
[g] $\alpha = E_{loocv}/E_{rms}$

To quantify trends with respect to system size, we define a system-dependent target accuracy threshold for E_{rms} equal to a fixed fraction f of a relevant reference energy V_{ref}. For reactive cases V_{ref} is the saddle point energy relative to reactants, while for CET cases it is the van der Waals binding energy. We recently showed that these definitions of E_{rms}/V_{ref} are characteristic of relative errors in trajectory calculations of CET and thermal and nonthermal kinetics.[16] Results for three values of f are summarized in Table 1 for six systems where the number of atoms n varies from 4 to 9. We note that the PIP expansions for the CET cases are full dimensional, but the fitted interaction energies are only weakly dependent on the internal coordinates of the colliders. We therefore consider the scaling of $N_d(f)$ with the number of important degrees of freedom d, which is equal to the number of internal coordinates for the reactive cases but is set to the number of internuclear coordinates for the CET cases.

The optimized basis set sizes N_d scale approximately linearly with d, as shown in Table 1. Averaging over the six cases, $N_d(f)/d = 1.8 \pm 0.6$, 3.9 ± 1.3, and 5.5 ± 2.2 for $f = 0.15$, 0.10, and 0.08, respectively, reported to one standard deviation. This suggests that, to obtain a reduced PIP expansion with 10% error, just 4 parameters per degree of freedom are sufficient. These values are of course dependent on the choice of f, but for any useful value of f vast reductions in PIP expansion sizes are achieved along with improved scaling compared to the exponential scaling of total-order PIPs. Furthermore, the optimized sizes are close to the number of parameters per degree of freedom appearing in well-known analytic representations such as Morse potentials, Buckingham interactions, and molecular mechanics force fields.

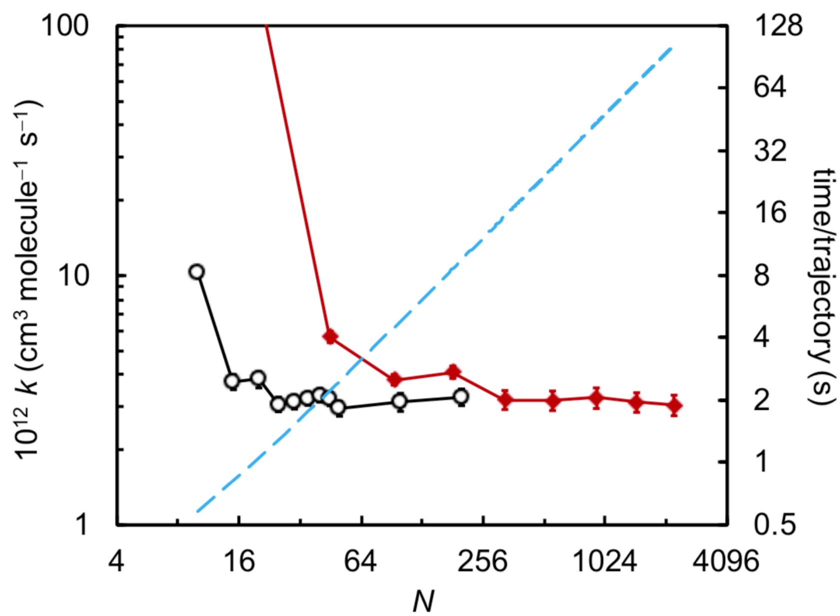


Figure 5. Thermal rate constant for H + H_2O at 2000 K calculated using quasiclassical trajectories shown as a function of the number of free parameters N_p or N_d for two series of PIP expansions: total order PIP expansions from PIP3 to PIP11 (red line with closed symbols) and optimized PIP expansions learned using multi-pass greedy selection and the PIP11 dictionary (black line with open symbols). The right axis and dashed line show the average computational cost of one trajectory, confirming that this cost scales linearly with N_p.

Finally, we tested the multi-pass optimized subsets against total-order PIP expansions in a trajectory-based application. Figure 5 shows calculated thermal rate constants for the hydrogen abstraction reaction H + H_2O at 2000 K, as studied in detail previously.[42] Results obtained using conventional total-order PIP expansions from PIP3 (N_p = 20) to PIP11 (N_p = 2235) are compared with multi-pass greedy PIP expansions with N_d = 10–200 using the PIP11 dictionary. For the total-order series, the rate constant is converged within 5% of the large-expansion limit (~3 x 10^{-12} cm^3 $molecule^{-1}$ s^{-1}) for the PIP7 expansion (with N_p = 327) and larger PIP expansions. The multi-pass greedy subsets predict rate constants within ~5% of the converged value with just $N_d = 25$, which is more than a factor of 10 reduction relative to the PIP7 total order expansion. Even larger reductions would be expected for larger systems, as discussed above. An additional test of the convergence in the vibrational frequencies for the expansions in Fig. 5 shows similarly good scaling with N_p and N_d and is presented as supporting information.

There are several ways in which the present procedure can be extended and improved. First, through the implementation of multiple basis sets, combining one which may be optimal for wells and others for steeper or flatter portions of the potentials, as examples. Furthermore, in this paper we did not establish any kind of model selection criterion or stopping point,[34] which will be a key part of our future efforts. A discussion of stopping behavior is presented in Ref. 43 which will guide our model-selection efforts. A drawback of the multi-pass greedy method is that the full total-order basis set must be computed before optimized subsets can be determined, and so larger systems still present a challenge. Although large basis sets can be generated using high performance computing,[16] a more efficient approach might be to use the fragmented basis set expansion[15] approach of Bowman and co-workers to generate the dictionary. Unsupervised methods[44–47] can also address this problem as they reduce the size of the **X** matrix *before* training data is generated. They also may provide a way to generate tractable PIP basis sets for much larger systems. Finally, applications other than trajectories such as quantum dynamics and spectroscopy should be explored, as these applications may feature different scaling than the near linear scaling observed here.

PIPs and other types of mathematical expansions are powerful tools for representing the potential energy surfaces of molecular systems. Their lack of reference to the underlying molecular problem can be limiting, particularly with respect to computational cost due to their exponential scaling. Dictionary learning, as applied here, revealed that many (more than 90%) of the terms in total-order PIP expansions are extraneous, with larger systems exhibiting increasingly higher fractions of wasted terms. Importantly, the near linear scaling with d of the optimized PIP expansions is likely representative of the true complexity of molecular potential energy surfaces, and dictionary learning should enable their application to larger systems.

Supporting Information

Details of the multi-pass greedy method and a convergence test of vibrational frequencies for the optimized PIP expansions

AUTHOR INFORMATION

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

This work was supported by the U. S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences operating under Contract Number DE-AC02-06CH11357 and by the Argonne–Sandia Consortium on High-Pressure Combustion Chemistry (ANL FWP 59044). An award of computer time was also provided by the ASCR Leadership Computing Challenge (ALCC) program using resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility. We gratefully acknowledge additional computing resources provided by Bebop, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory.

REFERENCES

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

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

3 Conte, R.; Houston, P. L.; Bowman, J. L. Communication: A Benchmark-Quality, Full-Dimensional Ab Initio Potential Energy Surface for Ar-HOCO, *J. Chem. Phys.* 2014, 140, 151101.

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

5 Jasper, A. W.; Harding, L. B.; Knight, C.; Georgievskii, Y. Anharmonic Rovibrational Partition Functions at High Temperatures: Tests of Reduced-Dimensional Models for Systems with up to Three Fluxional Modes, *J. Phys. Chem. A* **2019**, *123*, 6210–6228.

6 Nandi, A.; Qu, C.; Bowman, J. M. Full and Fragmented Permutationally Invariant Polynomial Potential Energy Surfaces for Trans and Cis N-methyl Acetamide and Isomerization Saddle Points. *J. Chem. Phys.* **2019**, *151*, 084306.

7 Jiang, B.; Li, J.; Guo, H. High-Fidelity Potential Energy Surfaces for Gas-Phase and Gas–Surface Scattering Processes from Machine Learning, *J. Phys. Chem. Lett.* **2020**, *11*, 5120–5131.

8 Conte, R.; Qu, C.; Houston, P. L.; Bowman, J. M. Efficient Generation of Permutationally Invariant Potential Energy Surfaces for Large Molecules. *J. Chem. Theory Comput.* **2020**, *16*, 3264–3272.

9 Koner, D.; Meuwly, M. Permutationally Invariant, Reproducing Kernel-Based Potential Energy Surfaces for Polyatomic Molecules: From Formaldehyde to Acetone. *J. Comp. Theory Comput.* **2020**, *16*, 5474–5484.

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

11 Babin, V.; Leforestier, C.; Paesani, F. Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient, *J. Chem. Theory Comput.* **2013**, *9*, 5395–5403.

[12] Tibor, G.; Czakó, G. Automating the Development of High-Dimensional Reactive Potential Energy Surfaces with the ROBOSURFER Program System, *J. Chem. Theory Comput.* **2020**, *16*, 51–66.

[13] 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.

[14] Riera, M.; Yeh, E. P.; Paesani, F. Data-Driven Many-Body Models for Molecular Fluids: CO_2/H_2O Mixtures as a Case Study, *J. Chem. Theory Comput.* **2020**, *16*, 2246–2257.

[15] Qu, C.; Conte, R.; Houston, P. L.; Bowman, J. M. Full-Dimensional Potential Energy Surface for Acetylacetone and Tunneling Splittings. *Phys. Chem. Chem. Phys.* **2021**, *23*, 7758.

[16] Moberg, D. R.; Jasper, A. W. Permutationally Invariant Polynomial Expansions with Unrestricted Complexity. *J. Chem. Theory Comput.* **2021**, Article ASAP. DOI:10.1021/acs.jctc.1c00352 (accessed 2021-09-08).

[17] Qu, C.; Bowman, J. M. A Fragmented, Permutationally Invariant Polynomial Approach for Potential Energy Surfaces of Large Molecules: Application to *N*-methyl Acetamide. J. Chem. Phys. **2019**, *150*, 141101.

[18] Tibor, G.; Czakó, G. Automating the Development of High-Dimensional Reactive Potential Energy Surfaces with the ROBOSURFER Program System, *J. Chem. Theory Comput.* **2020**, *16*, 51–66.

[19] Houston, P. L.; Conte, R.; Qu, C.; Li, J.; Bowman, J. M. Full-Dimensional, Ab Initio Potential Energy Surface for Glycine with Characterization of Stationary Points and Zero-Point Energy Calculations by Means of Diffusion Monte Carlo and Semiclassical Dynamics. *J. Chem. Phys.* **2020**, *153*, 024107.

[20] Qu, C.; Houston, P. L.; Conte, R.; Nandi, A.; Bowman, J. M. Breaking the Coupled Cluster Barrier for Machine-Learned Potentials of Large Molecules: The Case of 15-Atom Acetylacetone. *J. Chem. Phys. Lett.* **2021**, *12*, 4902–4909.

[21] Rubinstein R.; Bruckstein A. M.; Elad M. Dictionaries for Sparse Representation Modeling. *Proc. IEEE* **2010**, *98*, 1046-1057.

[22] Tosic I.; P. Frossard P. Dictionary Learning. *IEEE Signal Processing Magazine* **2011**, March, 27-38.

[23] Dumitrescu, B.; Irofti, P. *Dictionary Learning: Algorithms and Applications*; Springer (2018).

[24] Miller A. *Subset Selection in Regression*, *Second Edition*; Chapman & Hall/CRC, Boca Raton (2002).

[25] Tropp, J. A.; Gilbert, A. C. Signal Recovery from Random Measurements Via Orthogonal Matching Pursuit. *IEEE Trans. Information Theory* **2007**, *53*, 4655–4666.

[26] Elad, M. *Sparse and Redundant Representations*; Springer (2010).

[27] Zhang, T. Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. *IEEE Trans. Information Theory* **2011**, *57*, 4689–4708.

[28] Manzhos, S.; Carrington Jr., T. Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chem. Rev.* **2020**, *121*, 10187–10217.

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

[30] Paukku, Y.; Yang, K. R.; Varga, Z.; Truhlar, D. G. Global *ab initio* ground-state potential energy surface of N_4. *J. Chem. Phys.* **2013**, *139*, 044309, **2014**, 140, 019903(E).

[31] Jasper, A. W.; Davis M. J. Parameterization Strategies for Intermolecular Potentials for Predicting Trajectory-Based Collision Parameters, *J. Phys. Chem. A* **2019**, *123*, 3464–3480.
[32] Weisberg, S. *Applied Linear Regression, Third Edition*; John Wiley & Sons, Inc., Hoboken, New Jersey: 2005.
[33] 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.
[34] Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; Chapman & Hall/CRC: Boca Raton: 2015.
[35] Magnotti G. M.: Wang Z.; Liu W.; Sivaramakrishnan R.; Som S.; Davis M. J. Sparsity Facilitates Chemical-Reaction Selection for Engine Simulations. *J. Phys. Chem.* **2018**, *122*, 7227−7237
[36] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. B* **1996**, *58*, 267-288.
[37] Tibshirani, R. Regression Shrinkage and Selection via the Lasso: A Retrospective. *J. R. Stat. Soc. B* **2011**, *73*, 273-282.
[38] Hastie, T.; Tibshirani, R.; Friedman, J. *Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*; Springer: 2009.
[39] Couvreur C.; Bresler Y. On the Optimality of the Backward Greedy Algorithm for the Subset Selection Problem *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 797-808.
[40] Hager, W. G. Updating the Inverse of a Matrix, *SIAM Review* **1989**, *31*, 221-239.
[41] Khan E. Updating Inverse of a Matrix When a Column is Added/Removed, notes **2008**, https://emtiyaz.github.io/Writings/OneColInv.pdf, retrieved 6/28/2021.
[42] Tao, Y.; Jasper, A. W.; Georgievskii, Y.; Klippenstein, S. J.; Sivaramakrishnan, R. Termolecular Chemistry Facilitated by Radical-Radical Recombinations and its Impact on Flame Speed Predictions. *Proc. Combust. Inst.* **2021**, *31*, 515–522.
[43] Kenney, A.; Chiaromonte, F.; Felici, G. MIP-BOOST: Efficient and Effective L_0 Feature Selection for Linear Regression *Journal of Computational and Graphical Statistics* **2021**, online. DOI:10.1080/10618600.2020.1845184
[44] Boutsidis, C.; Mahoney, M. W.; Drineas, P. An Improved Approximation Algorithm for the Column Subset Selection Problem, SIAM, **2009**, 968-977.
[45] Mahoney, M. W.; Drineas, P. CUR matrix decompositions for improved data analysis *Proc. Nat. Acad. Sci.* **2009**, *106*, 697-702
[46] Boutsidis, C.; Drineas, P.; Magdon-Ismail M. Near-Optimal Column-Based Matrix Reconstruction, *SIAM J. Comput.* **2014**, *43*, 687–717.
[47] Hamm, K.; Huang, L. Perspectives on CUR decompositions. *Appl. Comput. Harmon. Anal.* **2020**, *48*, 1088-1099.