PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2019, **21**, 4513

Modelling potential energy surfaces for small clusters using Shepard interpolation with Gaussian-form nodal functions

Haina Wang *[a] and Ryan P. A. Bettens *[b]

Received 14th December 2018,
Accepted 1st February 2019

DOI: 10.1039/c8cp07640e

rsc.li/pccp

The potential energy surface (PES) of a chemical system is an analytical function that outputs the potential energy of the system when a nuclear configuration is given as input. The PESs of small atmospheric clusters have theoretical as well as environmental significance. A common method used to generate analytical PESs is the Shepard interpolation, where the PES is a weighed sum of Taylor series expansions (nodal functions) at *ab initio* sample points. Based on this, in this study we present a new method based on the Shepard interpolation, where the nodal functions are composed of a symmetric Gaussian term and an asymmetric exponential term in each dimension. Corresponding sampling methods were also developed. We tested the method on several atmospheric bimolecular clusters and achieved root mean square errors (RMSE) below 0.13 kJ mol^{-1} in 150 samples for Ar-rigid H_2O and Ne-rigid CO_2, and below 0.39 kJ mol^{-1} in 1800 samples for rigid N_2-rigid CO_2.

Introduction

The potential energy surface (PES) of a chemical system is an important concept in theoretical chemistry that has wide applications including spectral analysis,[1] protein folding,[2–4] and reaction dynamics.[5] Based on the Born–Oppenheimer approximation, the PES refers to a function that outputs the electronic potential energy of the system when a nuclear configuration is given as input. For a nonlinear molecule with N atoms, the PES is a function of $3N - 6$ independent degrees of freedom.[6,7] However, "redundant dimensions" can be present depending on the choice of the coordinate system that describes the nuclear configuration. In the 1990s and early 2000s, there has been controversy or reservation about using all internuclear distances as coordinates because of this "redundancy", and there were suggestions of using different sets of $3N - 6$ internal coordinates in different regions of the configuration space, *e.g.* in ref. 8. However, starting with Bowman's group, recent decades have seen successful uses of all internuclear distances, or "Morse variables" derived therefrom.[9,10]

Recently, due to the increasing awareness of human impact on the atmosphere, there has been extensive theoretical interest in the PESs of atmospheric clusters.[11–13] These are small systems of common atmospheric molecules, *e.g.* N_2, O_2, O_3, H_2O, Ar and CO_2, interacting through van der Waals forces or hydrogen bonds. Careful investigations of these clusters can shed new light on current environmental issues concerning atmospheric chemistry. For example, it has been predicted that as CO_2 forms clusters, its usually symmetric stretch mode ν_1 can be significantly asymmetric and absorb infrared radiation, contributing to the greenhouse effect together with its usually asymmetric vibrational modes, ν_2 and ν_3.[11]

Due to the small sizes of atmospheric clusters, *ab initio* single point energy (SPE) evaluations with very high accuracy (1.0 kJ mol^{-1}) can be achieved.[14] However, a thorough understanding of their possible interactions demands the generation of accurate analytical PESs over a global range of chemically and physically interesting configurations using a finite set of *ab initio* data, which remains a challenging task. The advantages of an analytical PES over pointwise evaluations include faster SPE retrieval, easy mathematical treatment and convenience of visualization. Typically, generating an analytical PES involves three aspects: the basic form of the function, the sampling of the configuration space (for *ab initio* calculations), and finally, the fitting or interpolation method that gives the global PES.[6]

There are many milestones for all these aspects. The requirement for the PES to be symmetric with respect to permutations of like atoms has been explicitly stated by Murrell and colleagues in their classic monograph.[15] To incorporate this permutational symmetry, theories of symmetrized coordinates[16] and invariant fitting bases[17] have been developed for triatomic and homonuclear tetra-atomic systems. More recently, the least squares fitting of permutationally invariant polynomials (PIP) was developed for high dimensionality

[a] *Department of Chemistry, Princeton University, Princeton, USA. E-mail: hainaw@princeton.edu*
[b] *Department of Chemistry, National University of Singapore, Singapore. E-mail: chmbrpa@nus.edu.sg*

by Braams and Bowman.[18] The many-body expansion (MBE), first applied by Varandas and Murrell to PESs of H_n systems,[16] has become a widely used approach to model asymptotic regions of configuration spaces.[6,15,19–21] Common sampling methods include random sampling from a grid of bond lengths and angles. There are, however, more sophisticated sampling methods such as Mutually Orthogonal Latin Squares that aim to cover a wide range of a high dimensional space.[4] The Bowman group has used scattered sampling methods, where samples are generated from molecular dynamics simulation data with widely different trajectories, effectively sampling a large range of configurations and energies.[22]

One of the most successful interpolation methods is the modified Shepard interpolation proposed by Collins and colleagues in 1994.[8,23] It represents the PES as a weighed sum of Taylor expansions in inverse internuclear distances up to the second order on each sample. The weight of a sample at a configuration is inversely proportional to the distance between the configuration and the sample. In addition, the group proposes an iterative sampling scheme that outperforms random sampling, adding in each round new samples that are distant from existing samples. Problems of the modified Shepard method include bump- or step-shaped artifacts where distinct samples compete for dominance. Various studies have been focusing on reducing such effect.[20,24] Another problem with the modified Shepard interpolation is that the number of *ab initio* points needed to calculate the second derivatives grows as the square of the number of dimensions, making the method computationally expensive for large systems.

More recently, some groups have studied the application of Gaussian Process (GP) in PES generation.[25–30] As another method of interpolation, it uses Gaussians instead of Taylor expansion as nodal functions. Being a less dramatically varying function than the polynomial, the Gaussian has promises to reduce artifacts. It also provides a natural estimation of error at not-yet sampled configurations, thereby providing an iterative sampling scheme where points with large estimated errors are added to the sample set. GP has been used successfully in modeling three-dimensional PESs of proton transfer in crystals.[25] Very recently, PIP-GP approaches have been developed to incorporate permutation invariance into GP.[30] The application of GP to higher dimensional gaseous systems, however, has been largely limited by the inflexible width of the Gaussian nodal functions and the fact that Gaussians are inevitably symmetric at samples, making it hard to account for first derivatives.

In this study, we developed a new form of nodal functions that aims to tackle the limitations of GP. It composes of a symmetric Gaussian part and an asymmetric inverse-exponential part for each dimension, both having flexible widths and amplitudes. We also proposed an iterative sampling scheme compatible with these nodal functions. We tested our method on three atmospheric clusters: Ar-rigid H_2O,[31] Ne-rigid CO_2,[28] and rigid CO_2-rigid N_2.[11] The focus for this present paper is on non-covalent interactions, with expectations of possible application to atmospheric science. For example, the PES for the clusters can serve as a perturbation term for the rovibration spectrum of interacting atmospheric molecules, with the monomers as the "base case", in a similar spirit to ref. 32. Nevertheless, we believe that the method itself is general enough to also deal with reacting molecules.

We focus only on interpolation of *ab initio* data in this current study, while noting that the dissociation regime can be well described by MBE methods, and that potential functions accurate for different regimes can be combined to give a global PES, *e.g.* with the energy switching (ES) approach developed by Varandas.[33–35]

We will first explain the general methodology and then discuss the specific implementations and results for the different clusters.

Methodology

Nodal functions

We use internuclear distances to describe configurations. As all clusters we study are rigid, only intermolecular internuclear distances are included. For example, in the cluster Ar-rigid H_2O, configurations are defined by the internuclear distances (Ar–O, Ar–H^1, Ar–H^2). Given the configuration vector $\mathbf{r}$ of a sample with *ab initio* energy $E_{\mathbf{r}}$, the nodal function $V_{\mathbf{r}}$ is expressed by a Gaussian symmetric term (resembling the s orbital) and an asymmetric term (resembling the p orbital) in each dimension, plus the constant $E_{\mathbf{r}}$, ensuring that the nodal function passes through the sample, *i.e.* $V_{\mathbf{r}}(\mathbf{r}) = E_{\mathbf{r}}$, as shown in (1).

$$V_{\mathbf{r}}(\tilde{\mathbf{r}}) = \sum_{i=1}^{\dim(\mathbf{r})} \left[A_i(\tilde{r}_i - r_i)\mathrm{e}^{-\alpha_i(\tilde{r}_i - r_i)^2} + B_i\left(\mathrm{e}^{-\beta_i(\tilde{r}_i - r_i)^2} - 1\right)\right] + E_{\mathbf{r}} \tag{1}$$

where $\tilde{\mathbf{r}}$ is any arbitrary configuration in the vicinity of $\mathbf{r}$. For each dimension, the parameters A_i and α_i are amplitude and width of the asymmetric term. B_i and β_i are amplitude and width of the symmetric term. α_i and β_i must be positive.

All parameters A_i, B_i, α_i, β_i need to be determined. The most obvious way is to fit the parameters based on *ab initio* calculations on a set of configurations surrounding each sample, which we terms as co-samples. We then optimize the width parameters with a rather "brute force" line search, first letting α_1 increase from 0.01 to 50 with increment 0.01, while all other widths are initially fixed at 0.01. For each value of α_1, the amplitude parameters A_i and B_i are obtained by linear regression; a least-squares regression error is thus generated. The value of α_1 that gives the smallest regression error is taken to be optimal and fixed, while the next width parameter, *e.g.* α_2, undergoes optimization, and so on for all α_i, β_i. After all widths undergo optimization, the whole process starts over from α_1 again and repeats for totally three rounds. This regression scheme apparently does not search for the global optimum of α_i and β_i, but as we will see in the next section, it turns out to be a rather practical scheme.

The permutational invariance, *i.e.* the requirement that when two atoms of the identical element switch position, the potential energy should remain the same, can be incorporated easily. Whenever a sample is chosen, we also add its "permutational twins" into the sample set. The nodal functions of the twin samples are obtained by permuting the terms of the nodal function of

the original sample. For example, if a sample of Ar-rigid H_2O is given by the internuclear distances (Ar–O, Ar–H^1, Ar–H^2) = (r_1, r_2, r_3) = $\mathbf{r}$ with nodal function (1), then it has a twin sample $\mathbf{r}'$ = (r_1', r_2', r_3') = (r_1, r_2, r_3) with the nodal function (2).

$$\begin{aligned} V_{\mathbf{r}'}(\tilde{\mathbf{r}}) = {} & A_1\left(\tilde{r}_1 - r_1'\right)\mathrm{e}^{-\alpha_1\left(\tilde{r}_1 - r_1'\right)^2} + B_1\left(\mathrm{e}^{-\beta_1\left(\tilde{r}_1 - r_1'\right)^2} - 1\right) \\ & + A_3\left(\tilde{r}_2 - r_2'\right)\mathrm{e}^{-\alpha_3\left(\tilde{r}_2 - r_2'\right)^2} + B_3\left(\mathrm{e}^{-\beta_3\left(\tilde{r}_2 - r_2'\right)^2} - 1\right) \\ & + A_2\left(\tilde{r}_3 - r_3'\right)\mathrm{e}^{-\alpha_2\left(\tilde{r}_3 - r_3'\right)^2} + B_2\left(\mathrm{e}^{-\beta_2\left(\tilde{r}_3 - r_3'\right)^2} - 1\right) + E_{\mathbf{r}} \end{aligned} \tag{2}$$

After the nodal functions of all samples are obtained, the global PES $V(\tilde{\mathbf{r}})$ is expressed as a weighed sum of nodal functions using inverse distance weighing, as given by (3) and (4), identical to the modified Shepard interpolation.[23]

$$V(\tilde{\mathbf{r}}) = \sum_{\text{sample } \mathbf{r}} V_{\mathbf{r}}(\tilde{\mathbf{r}}) w_{\mathbf{r}}(\tilde{\mathbf{r}}) \tag{3}$$

where the weights $w_{\mathbf{r}}(\tilde{\mathbf{r}})$ are such that for any arbitrary configuration $\tilde{\mathbf{r}}$

$$w_{\mathbf{r}}(\tilde{\mathbf{r}}) \propto \frac{1}{\| \tilde{\mathbf{r}} - \mathbf{r} \|^{p}} \quad \text{and} \quad \sum_{\text{sample } \mathbf{r}} w_{\mathbf{r}}(\tilde{\mathbf{r}}) = 1 \tag{4}$$

for a constant $p > 0$.

We remark here that we have used a rather "brute force" way to incorporate permutatioanal symmetry. In the systems studied in the paper, the symmetry is relatively low and the limiting step for generating the PES is the generation of the nodal functions, *i.e.* the nonlinear and linear regressions to optimize A_i, B_i, α_i, β_i. These steps do not need to be replicated for the replicated data. However, for systems with higher permutational symmetry like CH_5, the "inference" step (3), *i.e.* the generation of the final PES by weighing all the nodal functions may well be the limiting step. In such cases, fast search algorithms exist in computer science to search for the closest samples to an arbitrary configuration (range queries),[36,37] because only those contribute significantly to the energy estimation. Alternatively, instead of replicating data, PIP methods could be incorporated into our approach to make the PES permutationally invariant in an efficient way, where "primary invariant polynomials" described in ref. 18 of the internal coordinates replace the expressions $(\tilde{r}_i - r_i)$ in (1).[18,30] For non-covalent interactions, reduced permutational symmetry can be used to further reduce computational cost.[38]

Sampling

Simple random sampling over a grid of *ab initio* points in Jacobian coordinates (defined in the next section) was tested for all clusters studied here. The Jacobian coordinates must be converted to internal coordinates to generate the nodal functions. On top of random sampling, as we have mentioned in the introduction, there are iterative sampling algorithms that gradually improve the PES estimation as more samples are added, and they generally make significant improvement over random sampling. Inspired by these works, we came up with two ideas that could enhance sampling.

One of them is "error probing". Since we use linear regression to obtain the parameters and of the nodal functions, a regression error is generated automatically at each sample. We reason that if this regression error is large, the Gaussian nodal function does not describe the energy in the vicinity of the sample very well. Therefore, more samples are needed in that area to reduce the PES estimation error effectively.

The second idea is "distance probing", similar to that in Collins' work.[23] The estimation error is expected to be large at regions where samples are sparse. We want the range of the samples to be as broad as possible, so that unfavorable extrapolation can be prevented. Therefore, points far from existing samples are chosen as new samples.

In practice, error probing and distance probing should alternate with random sampling. Exactly how this can be done will be described for individual systems in the next section.

For the more complex cluster rigid CO_2-rigid N_2, we also developed a "Divide and Conquer" method, which probes for samples separately in the wall, well and dissociation regions before pooling samples together.

Application to atmospheric clusters

General aspects: co-samples, regression and the decay exponent

The selection of co-samples around each sample is a crucial step in generating the nodal functions. We choose co-samples in a random fashion such that the sample–co-sample distances (in the Euclidean metric on internal coordinates) are between $0.1a_0$ and $1.0a_0$. The number of co-samples L around each sample needs to be equal to or slightly larger than the number of unknown parameters in (1), or $L \geq 4\cdot\dim(\mathbf{r})$. Therefore, for Ar-rigid H_2O and Ne-rigid CO_2, L = 13. For rigid CO_2-rigid N_2, L = 25. Taking fewer co-samples than $4\cdot\dim(\mathbf{r})$ can lead to very inaccurate nodal functions.

As we will explain for each cluster in the next subsections, while the samples are truly *ab initio* points, the co-samples are pseudo *ab initio*, generated from very accurate analytical functions. We remark that this is a simplification, since some topological features, *e.g.* nearby crossings that may also happen for non-covalent interactions, do not appear in analytical functions used to generate the pseudo *ab initio* points.[39]

For all clusters studied, the "brute force" nonlinear regression for α_i and β_i, together with linear regression for A_i and B_i, usually gives least-squares errors of less than $0.05E_{\mathbf{r}}$ for samples in the well region (*i.e.* 5% of the sample energy) and 0.15–$0.50E_{\mathbf{r}}$ for samples in the wall region. As expected, the asymmetrical amplitudes A_i are larger at samples with large first derivatives, especially for samples in the wall region, where local asymmetry of the PES is significant. Therefore, we believe that we have found a rather practical method for parameter determination. We also found that the regression accuracy is much more sensitive to α_i than to β_i. One could even save computational resources by fixing all β_i at 0.5 without

changing the order of magnitude of the regression error. The accuracy is insensitive to α_i only within an order of magnitude.

The "decay exponent" p in (4) is an important parameter in all interpolations using inverse distance weighing. It controls the "blurryness" of the generated PES. Previous studies on the modified Shepard interpolation shows that one must choose $p \gg 3N - 3$ to ensure that samples far from $\tilde{\mathbf{r}}$ makes no contribution to the energy estimation at $\tilde{\mathbf{r}}$.[24] It turns out that our investigations of the Gaussian nodal functions yield similar results. p between 10 to 15 works well for Ar-rigid H_2O and Ne-rigid CO_2. For rigid CO_2-rigid N_2, we set p between 20 to 30. It is observed that, as long as p lies in the ranges specified above, the RMSE of the PES estimation is not sensitive to the specific value that p takes.

Ar-rigid H_2O

Coordinate system and *ab initio* details. The Jacobian coordinates on which the *ab initio* data grid is based is shown in Fig. 1(a). The water molecule is rigid with $r(OH) = 1.810a_0$ and $\angle HOH = 104.51^\circ$. The H_2O is placed in the xy-plane with its center of mass at the origin. The y-axis bisects $\angle HOH$ and O is on the positive half. The configuration can now be described by the polar coordinates (R, θ, ϕ) of Ar, where $0 < \theta < 90^\circ$, $0 < \phi < 180^\circ$.

The *ab initio* data consists of 1584 symmetry-unique points on a grid of (R, θ, ϕ), described in ref. 31 with R ranging from $4.5–20a_0$. The grid points are denser near the equilibrium values of R ($6.2–7.2a_0$). The theory level and basis set is CCSD(T)-F12(a)/AVQZ with standard counterpoise correction of basis set superposition effect (BSSE).

In our algorithm of PES interpolation, the Jacobian coordinates are converted to the internal coordinates (Ar–O, Ar–H^1, Ar–H^2). A test set of 100 symmetry-unique points is chosen from the grid. All samples are also chosen from the grid, but points in the test set are not allowed to be samples. The energies of the randomly chosen co-samples are obtained from a fitted analytical form given in ref. 31 which has been shown to give discrepancy within 10^{-3} cm^{-1} (1.2×10^{-5} kJ mol^{-1}) compared to direct *ab initio* calculations.

Analysis of sampling. The sampling algorithm, described by three integers (S, D, E), is an iterative process, in which each round consists of three actions that add new samples, described as following.

1. Randomly select S symmetry-unique samples.

2. ("Distance probing") randomly select 100 symmetry-unique "probe points" that are not yet samples. For each probe point, obtain its distance to its nearest existing sample (using the Euclidean metric on internal coordinates). The D probes points with largest distances to existing samples are added as new samples.

3. ("Error probing") randomly select 100 symmetry unique "probe points" that are not yet samples. For each probe point, we obtain an estimation of error by averaging the linear regression errors of its three nearest samples. The E probes with largest estimated error are added to the sample set.

Steps 1 through 3 are repeated while the RMSE of the test set is monitored, until 250 samples are obtained. Fig. 2(a) shows the variation of RMSE with the number of samples for four settings of (S, D, E). Note that the setting (50, 0, 0) is identical to simple random sampling. Each curve is the average result of 5 runs.

It can be observed from Fig. 2(a) that the error probing step is very helpful in reducing the error from early on. RMSE of around 100 μE_h (0.26 kJ mol^{-1}) can be achieved within 50 samples using the setting (35, 0, 15), and the final error converges at 20–50 μE_h (0.05–0.13 kJ mol^{-1}). On the other hand, using distance probing without error probing seems to be detrimental to the PES estimation in this system, probably because in the first rounds the distance probing step finds distant points in the trivial dissociation region. The RMSE for the combined probing setting (35, 5, 10) converges to the same limit of 20–50 μE_h, but the RMSE is relatively high before 150 samples. Note that 150 samples, with all their co-samples, actually require as many *ab initio* data as the grid itself, but the fact that local behaviors at the 150 samples can predict the global PES accurately shows that our nodal function is probably suitable for PES description. This sets the stage for investigations of higher dimensions.

Visualization of PES estimation. To visualize the quality of PES estimation and to understand the chemical distribution of the estimation error, we fix θ at 90° (*i.e.* fixing Ar on the xy-plane). We plot, in Fig. 3, the true *ab initio* energies and the estimation error against the position of Ar on the xy-plane. The PES was generated from 250 random samples. It can be seen that the error is high exclusively in the wall region, whereas the well and the dissociation region are well estimated. The errors at the wall are systematically positive, possibly due to extrapolation at configurations whose R are smaller than those of the samples.

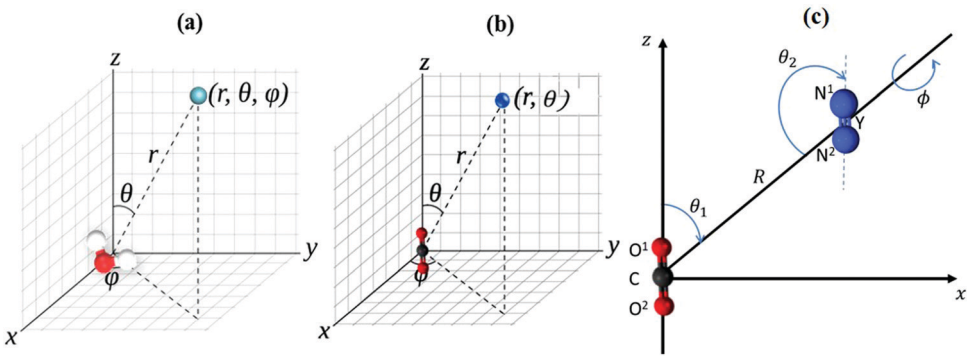

Fig. 1 The Jacobian coordinate systems of the *ab initio* data grid for each of the clusters studied in this work. (a) Ar–H_2O. (b) Ne–CO_2. (c) CO_2–N_2.

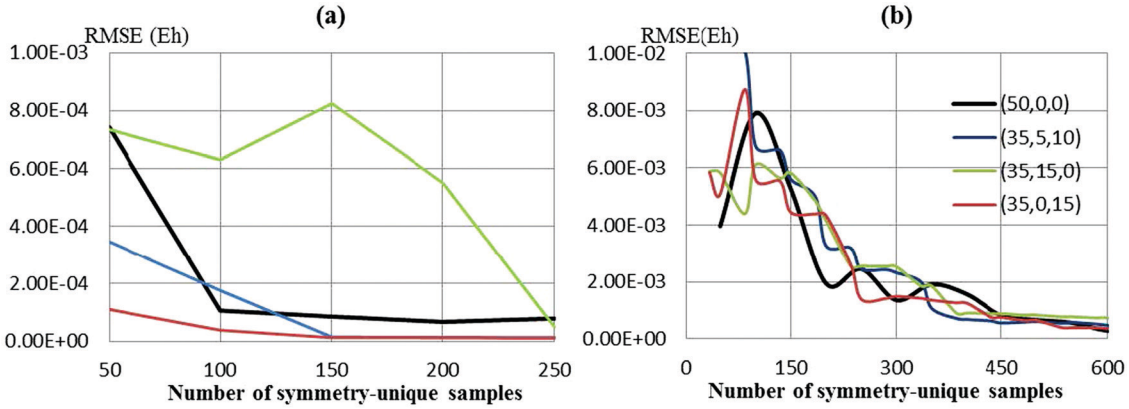

Fig. 2 Variation of RMSE during iterative sampling. The curves correspond to different iterative sampling schemes represented by (*S*, *D*, *E*) described in the text. (a) On a 100-point test set for Ar–H_2O. (b) On a 200-point test set for CO_2–N_2.

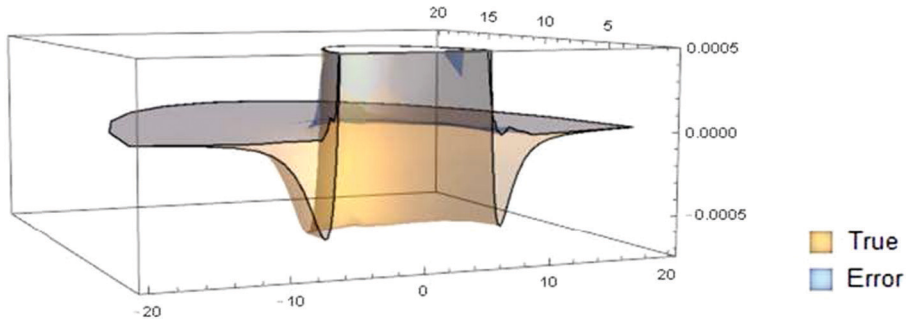

Fig. 3 *Ab initio* energy and estimation error for the "slice" $\theta = 90°$ of the PES of Ar–H_2O. PES generated by 250 samples obtained through the iterative sampling setting (*S*, *D*, *E*) = (35, 0, 15). Averaged over 3 runs. Energy unit is E_h and box dimensions are in a_0.

Ne-rigid CO_2

Fig. 1(b) shows the Jacobian coordinates (R, θ) defining the *ab initio* data grid, described in ref. 28. R is the distance between C and Ne. θ is the angle between the C→Ne vector and the CO_2 molecular axis. The geometry of rigid CO_2 is fixed at r(CO) = 2.196a_0. The *ab initio* data consists of 1200 symmetry-unique points, with ranging from 4.0 to 20.0a_0. The theory level is CCSD(T), and the basis set is AVTZ for C and O and atomic natural orbital (ANO) 6s5p3d2f for Ne. Standard counterpoise is used to correct for BSSE.

The internal coordinates (Ne–C, Ne–O^1, Ne–O^2) are used for our model. A random set of 40 symmetry-unique points from the grid with R between 4.5 and 15a_0 is chosen as the test set. Samples are chosen from the same grid and exclude the test points. Co-sample are randomly chosen around each sample, and their energies are obtained from the analytical form given in ref. 28 which achieves RMSEs less than 0.03 cm^{-1} (3.6×10^{-4} kJ mol^{-1}) compared to direct *ab initio* calculations.

The performance of the sampling schemes is similar to the case of Ar-rigid H_2O, but the RMSE drops much faster due to the simplicity of the actually two-dimensional PES. Even with random sampling, the RMSE can be reduced to 30–100 μE_h (0.08–0.26 kJ mol^{-1}) within 50 samples and converges to below 50 μE_h (0.13 kJ mol^{-1}) in 150 samples. Using the setting (*S*, *D*, *E*) = (35, 0, 15), the RMSE can converge to 20 μE_h (0.05 kJ mol^{-1}) in 150 samples. This RMSE value is comparable to that of the basic Gaussian Process interpolation described in ref. 40. Same as Ar–H_2O, the error increases with decreasing R on the wall region. Since there are three internuclear distances but only two actual dimensions, the results for Ne-rigid CO_2 show that our model works well with redundant internal coordinates.

Rigid CO_2-rigid N_2

Coordinate system and *ab initio* details. Fig. 1(c) shows the Jacobian coordinates (R, θ_1, θ_2, ϕ) defining the *ab initio* data grid, as described in ref. 11. R is the distance between C and the center of mass Y of N_2. $\theta_1 = \angle O^1CY$. $\theta_2 = \angle CYN^1$. ϕ is the dihedral angle O^1CYN^1. The configurations of CO_2 and N_2 are frozen at r(CO) = 2.21727a_0 and r(NN) = 2.10665a_0. The grid consists of 21 840 symmetry-unique points, with R ranging from 4.0 to 30.0a_0. However, we restrict our attention to the test configurations whose minimum intermolecular internuclear distances d_{min} lie between 4.5 and 15.0a_0. The sampling space consists of configurations with d_{min} between 4.45 and 15.1a_0. We want the sampling space to be slightly larger than the test space to reduce unfavorable extrapolation. The *ab initio* theory level and basis set are CCSD(T)-F12(a)/AVTZ, with standard counterpoise to correct for BSSE.

The internal coordinates (N^1–C, N^1–O^1, N^1–O^2, N^2–C, N^2–O^1, N^2–O^2) are used for our model. We use two kinds of test sets.

The first is a random set of 200 symmetry-unique points from the grid. The second is the entire "slice" on the PES with $\theta_2 = 0$, $\phi = 0$, used for visualization of PES. All samples are chosen from the grid, excluding the test points. Co-samples energies are obtained from the analytical form given in ref. 11, which achieves RMSEs less than 1.0 cm^{-1} (0.012 kJ mol^{-1}) compared to direct *ab initio* calculations.

Analysis of sampling. We tried the iterative sampling scheme discussed above in the section of Ar–H_2O. The performance of four settings within 600 symmetry-unique samples is given in Fig. 2(b). RMSE is calculated for a 200-point test set. Each curve is the average of 5 runs.

Unlike in Fig. 2(a), we see that here the sampling schemes do not show significant difference from each other in terms of RMSE variation. It can only be observed that, compared to the simple random sampling, the settings with error probing or distance probing are more reliable in reducing the error steadily during sampling. The random sampling curve $(S, D, E) = (50, 0, 0)$ has several significant rises, whereas the other curves almost monotonously decrease after 100 samples. However, as sampling goes on, the random scheme can also spot good samples and obtain an accuracy comparable with schemes using probing.

The RMSE converges at about 1500 samples. Even for the same test set, the final converged RMSE vary greatly, ranging from 70 to 300 μE_h (0.18–0.79 kJ mol^{-1}). This large range suggests that there may still be room for improvement of sampling.

"Divide and Conquer" for sampling. We found that the RMSE can be reliably limited to around 100 μE_h when we restricted the test set and the sample set to points whose d_{min} are greater than $5.0a_0$, but no significant improvement was observed if we only restricted the test set. We speculated that the configurations at the wall region, having very high energy, interfere with the sampling or the energy estimation at the well region. Conversely, the samples in the well region may influence the estimation of the wall configurations. Thus, the idea of a "Divide and Conquer" strategy seemed immediately promising: one could probably improve the full PES estimation by estimating disjoint regions first.

In light of this observation, we separated the PES into 3 pieces (units in a_0):

- "Wall" for test points with $4.5 < d_{min} < 5.5$, samples with $4.45 < d_{min} < 5.6$;
- "Well" for test points with $5.5 < d_{min} < 10.0$, samples with $5.4 < d_{min} < 10.1$;
- "Tail" for test points with $10.0 < d_{min} < 15.0$, samples with $9.9 < d_{min} < 15.1$.

The boundary d_{min} between Wall and Well ($5.5a_0$) was so chosen because it is approximately the sum of the van der Waals radii of N and C (or N and O), where we expect collisions to occur. The boundary d_{min} between Well and Tail ($10.0a_0$) corresponds to configurations with absolute values of energy below 100 μE_h. These choices make the unlikely assumption that whether a point belongs to Wall, Well or Tail depends only on d_{min}. However, as we shall see, it proves to be a simple and useful approximation.

We intend to sample separately for each piece. We therefore tested, for each range, how many samples are needed to bring down the error to a reasonably low value. The following is observed.

- Wall: the RMSE converges to 150–500 μE_h in 800–900 samples;
- Well: the RMSE converges to 20–80 μE_h in 500–600 samples;
- Tail: the RMSE converges to <15 μE_h in 600 samples, but just 200 samples are needed to bring it down to below 80 μE_h.

Therefore, we divide 1800 samples into 900 in Wall, 600 in Well and 300 in Tail. Sampling is done for each piece separately, with random sampling alternating with distance probing and error probing with $(S, D, E) = (35, 5, 10)$. Then we pool the 1800 samples together to form a final global PES according to (3). Note that (3) guarantees the final PES is smooth, *i.e.* we divided the configuration space only for better sampling, and did not generate separate PESs for different regions. However, there exist methods where PESs accurate for separate regions are obtained and combined, *e.g.* energy switching mentioned in Introduction.

To see how our piecewise algorithm performs for the four-dimensional PES of rigid CO_2-rigid N_2, again, 200-point random test sets are chosen. The estimation results on the test sets are shown in Fig. 4(a–c). The improvement of (c) over (b) is clear. The RMSE with piecewise sampling is 40–150 μE_h (0.10–0.39 kJ mol^{-1}), a significant improvement in stability from the 70–300 without the piecewise algorithm.

To visualize the estimated PES, we examine the "slice" with $\theta_2 = 0$, $\phi = 0$. Fig. 4(d and e) shows the performances of estimating this slice for sampling algorithms with and without the sampling algorithms. It is clear that the latter algorithm gives much better estimation at the wall and the well regions (better overlap of the yellow and blue surfaces), but the estimation is worse at the dissociation region because samples are relatively sparse there. Overall, the RMSE of the estimations of the slice is 50 *vs.* 80 μE_h with and without the piecewise algorithm, respectively. This "Divide and Conquer" strategy yields a peculiar three-piece RMSE decrease pattern as the sample size increases, shown in Fig. 4(f).

The piecewise algorithm may have reduced the error by the simple fact that it samples more intensely at the chemically important wall and well regions, instead of wasting too many samples at the dissociation region, or by the fact that it separately treats the wall and the well, preventing their cross-contamination by providing more focused spaces for random sampling and distance and error probing. Our conjecture is that the second factor at least plays more than a trivial role because, as we have observed, the well region PES can be much better estimated when the wall points are excluded from the sample set.

The potential shortcoming of the "Divide and Conquer" method is that for a larger system, there may be many more features that turn out to be relevant for the PES, and the boundary between Wall, Well and Tail may be much less well-defined by d_{min}. However, the generation of boundaries may be automated by considering the van der Waals radii of all atoms in a system.

Analysis of error. As in the section of Ar–H_2O, we want to know where large errors occur. Fig. 5(a and b) show the estimation error of 200 test points, plotted against their R and θ_1, (a) being the front view and (b) being the top view, where a darker area means a larger error. The underlying PES is generated with 1800 samples by the piecewise algorithm described above.

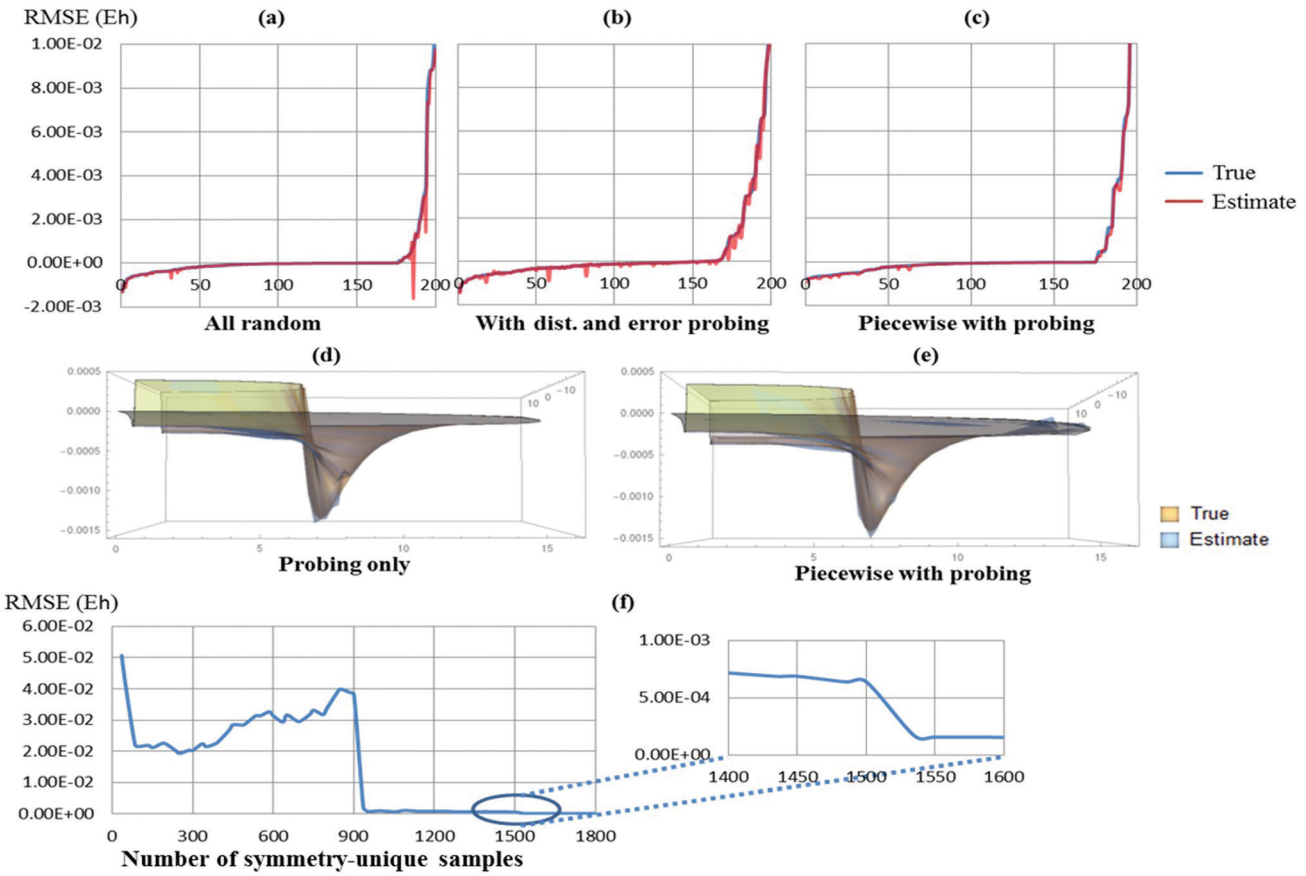

Fig. 4 The effect of the "Divide and Conquer" strategy in PES estimation of CO_2–N_2. For all runs, the sample size is 1800. (a–c) True and estimated energy at 200 test points with (a) random sampling, (b) iterative sampling with error probing and distance probing, and (c) piecewise sampling with iterative probing. Horizontal axes are the test point indices. All estimation curves (red) are results averaged over 3 runs. (d and e) True and estimated PES on the slice $\theta_2 = \phi = 0$, with and without the piecewise sampling algorithm. Averaged over 3 runs. Energy unit is E_h and box dimensions are in a_0. (f) Variation of RMSE during piecewise sampling. The horizontal axis is the number of samples. Two sharp drops corresponding to switches of sampling regions can be observed.

It is clear from Fig. 5(a and b) that the error does not simply increase with decreasing intermolecular distance. Some configurations with a smaller R are better estimated than configurations with a larger R. The large-error configurations appear as "spikes" in Fig. 5(a). According to Fig. 5(b), these spikes appear at intermolecular distances of 6–8a_0, as if forming a wall surrounding CO_2. We hypothesize that large errors are likely to occur at R values where some orientations are colliding but others with the same R are non-colliding. The "switch of nature" of the system puts it at the boundary of the wall and well regions, making it harder to make accurate energy predictions from samples.

To test this hypothesis, we pick the configurations with largest errors, fix their R, θ_1 and survey their *ab initio* energy as their θ_2 and ϕ vary. Fig. 5(c) shows the *ab initio* energy profile at $R = 7.25a_0$, $\theta_1 = 30^\circ$. (These values correspond to the highest spike in Fig. 5(a).) Apparently, the energy depends heavily on θ_2, dropping from more than 11 000 μE_h for $\theta_2 = 0^\circ$ to only about 200 μE_h for $\theta_2 = 90^\circ$. The dependence of energy on ϕ, on the other hand, is insignificant. We therefore fixed ϕ at 60° arbitrarily and visualized configurations with $R = 7.25a_0$, $\theta_1 = 30^\circ$, $\phi = 60^\circ$ at different values of θ_2 using space-filling models with van der Waals radii given by ref. 41. The models are shown in Fig. 5(d). As expected, a transition between colliding and non-colliding orientations is observed. Smaller values of θ_2 give colliding configurations, where the van der Waals surfaces of the molecules overlap, while $\theta_2 = 90^\circ$ is non-colliding. Analysis of other large-error configurations gives similar transitions between collision and non-collision. This discussion may have illustrated in a novel way the significance of the van der Waals radius as a reliable parameter for modeling intermolecular interactions.

Comparison with other interpolation methods

The number of necessary *ab initio* evaluations around each sample increases as the square of the number of dimensions for the Shepard interpolation, where Taylor expansion up to second order, and therefore the Hessian matrices, must be calculated. For our method, the number of necessary co-samples L increases linearly with the number of dimensions. Therefore, our method has great advantage in computational efficiency compared to the modified Shepard interpolation for higher-dimensional systems. We are developing flexible algorithms to apply the method on arbitrary dimers with up to 10 atoms.

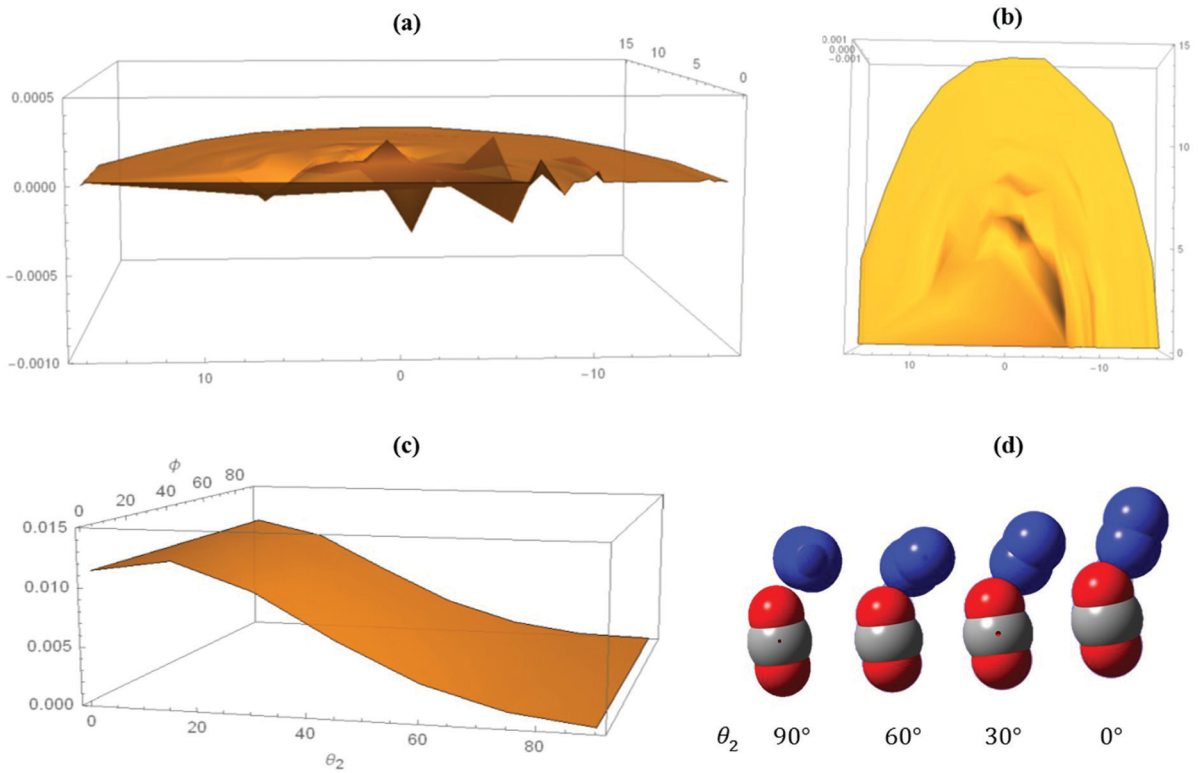

Fig. 5 Error of the estimated PES of CO_2–N_2. (a) Plot of error on a 200-point test set against R and θ_1 of the test points. Energy in E_h. Box dimensions in a_0. (b) The top view of the error plot; (c) *ab initio* energy of configurations with $R = 7:25a_0$, $\theta_1 = 30°$. The energy is sensitive to θ_2 much more than to ϕ. (d) Space-filling models of configurations with $R = 7:25a_0$, $\theta_1 = 30°$, $\phi = 60°$ and various θ_2 using van der Waals radii. Both colliding and non-colliding configurations are present.

To compare our Gaussian nodal functions with Taylor expansions up to the first order, we applied both methods on a simple hypothetical one-dimensional PES with the Morse potential (5).

$$V(\tilde{r}) = D_e(1 - e^{-a(\tilde{r}-r_e)})^2 - V(r_e) \quad (5)$$

where we set $r_e = 2a_0$, $D_e = V(r_e) = 500E_h$, $a = 0.5$. Four sample points are arbitrarily chosen at $r = 1.0, 2.0, 5.0, 10.0a_0$. On the one hand, first order Taylor expansions in $\tilde{r}^{-1} - r^{-1}$ are obtained. On the other hand, Gaussian nodal functions are generated with regression on 4 co-samples for each sample. For both types of nodal functions, inverse-distance weighing (3) and (4) are used with $p = 6$. Fig. 6 plots the hypothetical "true" PES as well as those estimated by both types of nodal functions. It is clear that Gaussian-form nodal functions are able to reduce "bumps" compared to first order Taylor expansions. This is because a Taylor expansion is, in a sense, strictly local, whereas a Gaussian-form nodal function, being fitted with a set of co-samples, promises to take care of a wider range of configurations around the sample. It is also interesting to observe that in Fig. 6, it seems that our method extrapolates beyond the leftmost sample at $1.0a_0$ better than first order Taylor expansions. The shortcoming of the Gaussian-form is that it does not promise to account for the first derivatives exactly, and, of course, it is more computationally expensive than first order Taylor expansions because of the required calculations on co-samples and the "brute force" nonlinear regression. More discussion will follow shortly.

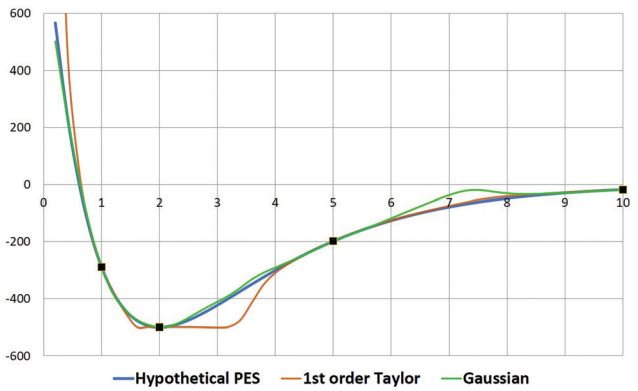

Fig. 6 Comparison of first order Taylor expansions and Gaussians as nodal functions in modelling a hypothetical one-dimensional PES. Horizontal axis in a_0.

Compared to the full Gaussian process, our method seems to be significantly better thanks to the asymmetric terms and the flexible width parameters in the nodal functions. In our preliminary studies on the flexible water dimer, the RMSE we achieved with simple GP was 650 μE_h (1.71 kJ mol^{-1}) with 1200 training data. Though the clusters in this study are all rigid, intramolecular vibrations are not expected to perturb the PES remarkably, since ground states' mean vibrational amplitudes should be on the order of 0.05 Å,[42] much less than the length

scale considered for the rigid-body PES. Therefore, we believe that our method presented here will give more accurate PES estimations than the full GP for flexible clusters. This conjecture, however, must be rigorously tested. In addition, we note that the computational effort of full GP increases as the cube of the number of training samples, because matrix inversions are required to compute the correlations between samples.[30,43] It is therefore very time consuming to use the GP once the training data size exceeds 1000. For our method, on the other hand, the computational cost scales linearly with respect to the size of *ab initio* data.

There are two major limitations to our approach. The first limitation is that the nonlinear regression does not search for the global minimum. To find a global minimum, ''multistart'', *i.e.* different sets of initial values of α_i and β_i, or general optimization techniques like simulated annealing[44] should be used. However, we have noted that the error of the final PES is not sensitive to β_i and not sensitive to α_i within an order of magnitude. With multistart experiments on our 1D hypothetical PES, we find that both multistart and one-off nonlinear regression can agree on the order of magnitude of α_i. The reason is partly that the linear regression step, with optimizations on A_i and B_i, usually does well to give a nodal function describing the local features around a sample. We assumed that this is true also for more than one dimensions, but this is admittedly a compromise because it is very computationally expensive to do multistart regression for all α_i and β_i.

The second major limitation is the large number of co-samples required, whose *ab initio* energies require huge computational costs if a PES is to be constructed *de novo*. This is also a limitation for Taylor expansions, as derivatives must be obtained numerically. Our method requires less co-sample data than second order Taylor expansions but more than first-order. We have the following suggestions regarding this limitation, but further testing must be done.

Firstly, if derivatives of the PES are also of interest, some data needed for computing derivatives can act as co-samples. Secondly, the co-samples can be shared between samples if the samples are close enough, *e.g.* the Cartesian distance between the two samples are less than $2.0a_0$. A sample can also act as a co-sample of another sample if they are close enough. This sharing can reduce the required number of *ab initio* calculations by at least half. Thirdly, the *ab initio* calculations of the co-samples do not need the same degree of accuracy as the samples, and may be accomplished by a faster method with a lower theory level or smaller basis set, probably with adjustments such that the energies of the samples match between theories.

Conclusions

We have presented a new interpolation method based on the modified Shepard interpolation, in which the nodal functions are composed of a symmetric Gaussian term and an asymmetric exponential term in each dimension. The parameters in the nodal functions are determined by regression on *ab initio* co-samples around each sample. An iterative sampling scheme has been developed to add samples where errors are expected to be large. We are able to achieve an RMSE below 0.13 kJ mol^{-1} in 150 samples for Ar-rigid H_2O and Ne-rigid CO_2, and below 0.39 kJ mol^{-1} in 1800 samples for rigid N_2-rigid CO_2. For the last system, we found that sampling separately in the wall, well and dissociation regions is very useful to reduce the error.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank M. Hochlaf and colleagues for the *ab initio* data and analytical code of the PES of CO_2–N_2.

References

1 N. Rekik, Toward accurate prediction of potential energy surfaces and the spectral density of hydrogen bonded systems, *Phys. B*, 2014, **436**, 164–176.
2 J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes, Theory of protein folding: the energy landscape perspective, *Annu. Rev. Phys. Chem.*, 1997, **48**, 545–600.
3 A. R. Dinner, A. Šali, L. J. Smith, C. M. Dobson and M. Karplus, Understanding protein folding *via* free-energy surfaces from theory and experiment, *Trends Biochem. Sci.*, 2000, **25**, 331–339.
4 K. Vengadesan and N. Gautham, Enhanced sampling of the molecular potential energy surface using mutually orthogonal Latin squares: application to peptide structures, *Biophys. J.*, 2003, **84**, 2897–2906.
5 J. M. Bowman, G. Czakó and B. Fu, High-dimensional *ab initio* potential energy surfaces for reaction dynamics calculations, *Phys. Chem. Chem. Phys.*, 2011, **13**, 8094.
6 G. C. Schatz, The analytical representation of electronic potential-energy surfaces, *Rev. Mod. Phys.*, 1989, **61**, 669–688.
7 M. A. Collins, Molecular potential-energy surfaces for chemical reaction dynamics, *Theor. Chem. Acc.*, 2002, **108**, 313–324.
8 K. C. Thompson, M. J. T. Jordan and M. A. Collins, Polyatomic molecular potential energy surfaces by interpolation in local internal coordinates, *J. Chem. Phys.*, 1998, **108**, 8302–8316.
9 A. Brown, B. J. Braams, K. Christoffel, Z. Jin and J. M. Bowman, Classical and quasiclassical spectral analysis of CH_5^+ using an *ab initio* potential energy surface, *J. Chem. Phys.*, 2003, **119**, 8790–8793.
10 X. Huang, B. J. Braams and J. M. Bowman, *Ab initio* potential energy and dipole moment surfaces for $H_5O_2^+$, *J. Chem. Phys.*, 2005, **122**, 044308.
11 S. Nasri, Y. Ajili, N.-E. Jaidane, Y. N. Kalugina, P. Halvick, T. Stoecklin and M. Hochlaf, Potential energy surface of the CO_2–N_2 van der Waals complex, *J. Chem. Phys.*, 2015, **142**, 174301.
12 J. M. Anglada, G. J. Hoffman, L. V. Slipchenko, M. Costa, M. F. Ruiz-López and J. S. Francisco, Atmospheric significance of water clusters and ozone-water complexes, *J. Phys. Chem. A*, 2013, **117**, 10381–10396.

13 V. Vaida, Perspective: water cluster mediated atmospheric chemistry, *J. Chem. Phys.*, 2011, **135**, 020901.
14 F. Negri, F. Ancilotto, G. Mistura and F. Toigo, *Ab initio* potential energy surfaces of He–CO_2 and Ne–CO_2 van der Waals complexes, *J. Chem. Phys.*, 1999, **111**, 6439–6445.
15 J. N. Murrell, S. Carter, S. C. Farantos, P. Huxley and A. J. C. Varandas, *Molecular Potential Energy Functions*, 1984.
16 A. J. Varandas and J. N. Murrell, A many-body expansion of polyatomic potential energy surfaces: application to H_n systems, *Faraday Discuss. Chem. Soc.*, 1977, **62**, 92–109.
17 A. Schmelzer and J. N. Murrell, The general analytic expression for S4-symmetry-invariant potential functions of tetra-atomic homonuclear molecules, *Int. J. Quantum Chem.*, 1985, **28**, 287–295.
18 B. J. Braams and J. M. Bowman, Permutationally invariant potential energy surfaces in high dimensionality, *Int. Rev. Phys. Chem.*, 2009, **28**, 577–606.
19 J. N. Murrell and S. Carter, Approximate single-valued representations of multivalued potential energy surfaces, *J. Phys. Chem.*, 1984, **88**, 4887–4891.
20 S. Y. Lin, P. Zhang and J. Z. Zhang, Hybrid many-body-expansion/Shepard-interpolation method for constructing ab initio potential energy surfaces for quantum dynamics calculations, *Chem. Phys. Lett.*, 2013, **556**, 393–397.
21 O. B. M. Teixeira, V. C. Mota, J. M. Garcia de La Vega and A. J. C. Varandas, Single-sheeted double many-body expansion potential energy surface for ground-state ClO_2, *J. Phys. Chem. A*, 2014, **118**, 4851–4862.
22 H. Liu, Y. Wang and J. M. Bowman, Quantum calculations of the IR spectrum of liquid water using *ab initio* and model potential and dipole moment surfaces and comparison with experiment, *J. Chem. Phys.*, 2015, **142**, 194502.
23 J. Ischtwan and M. A. Collins, Molecular potential energy surfaces by interpolation, *J. Chem. Phys.*, 1994, **100**, 8080–8088.
24 R. P. A. Bettens and M. A. Collins, Learning to interpolate molecular potential energy surfaces with confidence: a Bayesian approach, *J. Chem. Phys.*, 1999, **111**, 816–826.
25 K. Toyoura, D. Hirano, A. Seko, M. Shiga, A. Kuwabara, M. Karasuyama, K. Shitara and I. Takeuchi, Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: a case study on proton conduction in oxides, *Phys. Rev. B*, 2016, **93**, 54112.
26 Y. Guan, S. Yang and D. H. Zhang, Construction of reactive potential energy surfaces with Gaussian process regression: active data selection, *Mol. Phys.*, 2018, **116**, 823–834.
27 J. Cui and R. V. Krems, Efficient non-parametric fitting of potential energy surfaces for polyatomic molecules with Gaussian processes, *J. Phys. B: At., Mol. Opt. Phys.*, 2016, **49**, 224001.
28 R. Chen, E. Jiao, H. Zhu and D. Xie, A new *ab initio* potential energy surface and microwave and infrared spectra for the Ne–CO_2 complex, *J. Chem. Phys.*, 2010, **133**, 104302.
29 J. P. Alborzpour, D. P. Tew and S. Habershon, Efficient and accurate evaluation of potential energy matrix elements for quantum dynamics using Gaussian process regression, *J. Chem. Phys.*, 2016, **145**, 174112.
30 C. Qu, Q. Yu, B. L. Van Hoozen, J. M. Bowman and R. A. Vargas-Hernández, Assessing Gaussian process regression and permutationally invariant polynomial approaches to represent high-dimensional potential energy surfaces, *J. Chem. Theory Comput.*, 2018, **14**, 3381–3396.
31 T. Vanfleteren, T. Földes and M. Herman, Analysis of a perpendicular band in Ar-H_2O with origin close to the $\nu1+\nu3$, R(0) line in H2O, *Chem. Phys. Lett.*, 2015, **627**, 36–38.
32 A. van der Avoird and D. J. Nesbitt, Rovibrational states of the H_2O-H_2 complex: an *ab initio* calculation, *J. Chem. Phys.*, 2011, **134**, 044314.
33 A. J. C. Varandas, Energy switching approach to potential surfaces: an accurate single valued function for the water molecule, *J. Chem. Phys.*, 1996, **105**, 3524–3531.
34 A. J. C. Varandas, A. I. Voronin and P. J. S. B. Caridade, Energy switching approach to potential surfaces. III. Three-valued function for the water molecule, *J. Chem. Phys.*, 1998, **108**, 7623–7630.
35 A. J. C. Varandas, A realistic multi-sheeted potential energy surface for NO2(2A′) from the double many-body expansion method and a novel multiple energy-switching scheme, *J. Chem. Phys.*, 2003, **119**, 2596–2613.
36 T. Skopal, M. Krátký, J. Pokorný and V. Snášel, A new range query algorithm for Universal B-trees, *Information Systems*, 2006, **31**, 489–511.
37 E. Carlini, A. Lulli and L. Ricci, Dragon: multidimensional range queries on distributed aggregation trees, *Future Gener. Comput. Syst.*, 2016, **55**, 101–115.
38 Z. Homayoon, R. Conte, C. Qu and J. M. Bowman, Full-dimensional, high-level ab initio potential energy surfaces for H2(H2O) and H2(H2O)2 with application to hydrogen clathrate hydrates, *J. Chem. Phys.*, 2015, **143**, 084302.
39 A. J. C. Varandas, Accurate combined-hyperbolic-inverse-power-representation of *ab initio* potential energy surface for the hydroperoxyl radical and dynamics study of O + OH reaction, *J. Chem. Phys.*, 2013, **138**, 134117.
40 E. Uteva, R. S. Graham, R. D. Wilkinson and R. J. Wheatley, Interpolation of intermolecular potentials using Gaussian processes, *J. Chem. Phys.*, 2017, **147**, 161706.
41 A. Bondi, Van der Waals volumes and radii, *J. Phys. Chem.*, 1964, **68**, 441–451.
42 B. Cyvin, S. Cyvin and G. Hagen, Condensed values of mean and amplitudes of vibration, *Chem. Phys. Lett.*, 1967, **1**, 211–213.
43 J. Hensman, N. Fusi and N. D. Lawrence, Gaussian Processes for Big Data, UAI, 2013.
44 S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, Optimization by simulated annealing, *Science*, 1983, **220**, 671–680.

