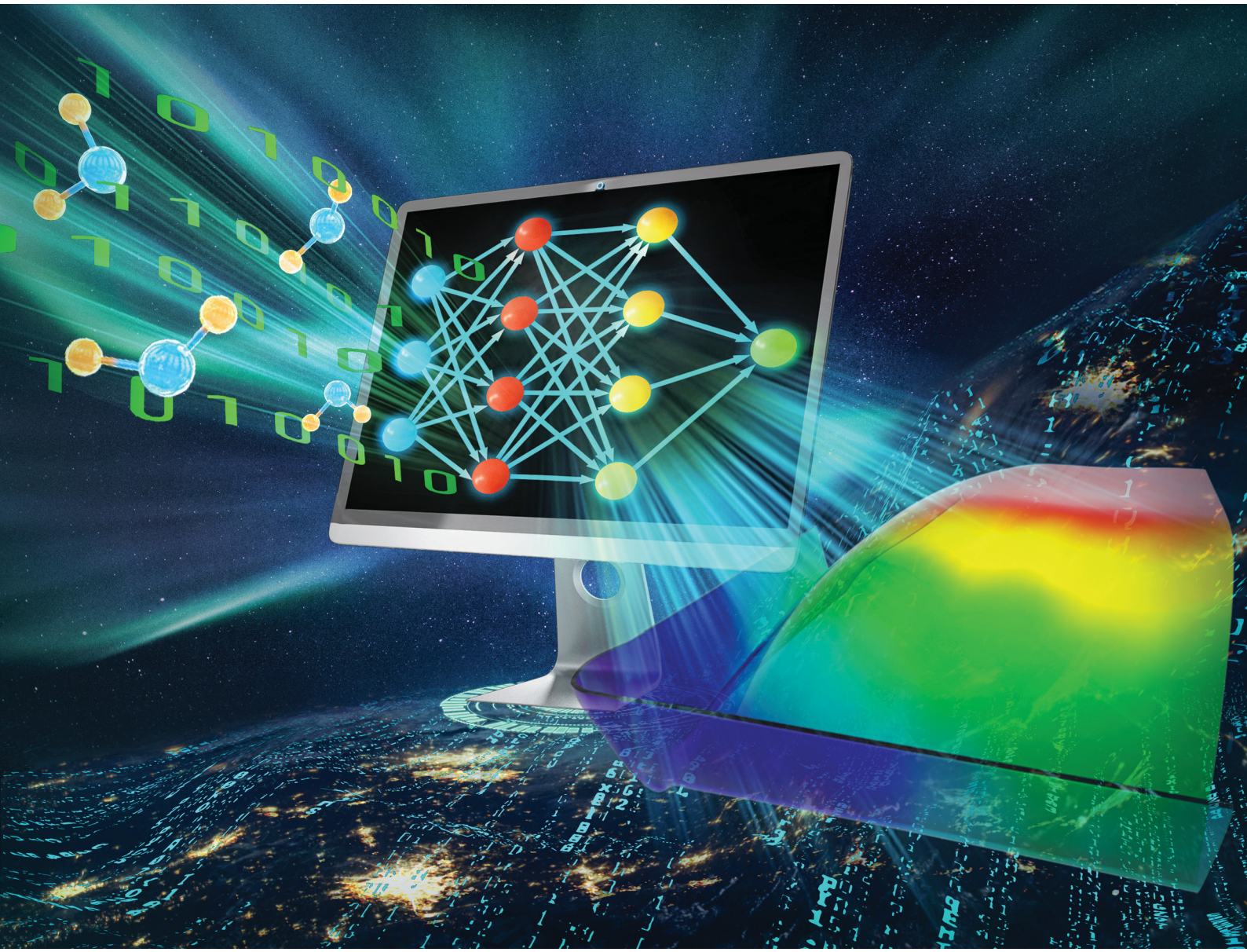
Volume 24
Number 21
7 June 2022
Pages 12661–13418

PCCP

Physical Chemistry Chemical Physics

rsc.li/pccp



ISSN 1463-9076

ROYAL SOCIETY OF CHEMISTRY

PAPER

Maodu Chen *et al.*

Representing globally accurate reactive potential energy surfaces with complex topography by combining Gaussian process regression and neural networks

PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2022, **24**, 12827

Representing globally accurate reactive potential energy surfaces with complex topography by combining Gaussian process regression and neural networks

Zijiang Yang, Hanghang Chen and Maodu Chen *

There has been increasing attention in using machine learning technologies, such as neural networks (NNs) and Gaussian process regression (GPR), to model multi-dimensional potential energy surfaces (PESs). A PES constructed using NNs features high accuracy and generalization capability, but a single NN cannot actively select training points as GPR does, resulting in expensive *ab initio* calculations as the molecular complexity increases. However, a PES constructed using GPR has a slow speed of evaluation and it is difficult to accurately describe a fast-changing potential. Herein, an efficient scheme for representing globally accurate reactive PESs with complex topography based on as few points as possible by incorporating active data selection of GPR into NN fitting is proposed. The validity of this strategy is tested using the BeH_2^+ system, and only 1270 points are automatically sampled. The generalization performance and speed of evaluation of the generated PES are much better than those of the GPR PES constructed using the same dataset. Moreover, an accurate NN PES is fitted by 12 122 points as a benchmark for comparison to further test the global accuracy of the PES obtained using this scheme, and the corresponding results present extremely consistent topography characteristics and calculated $Be^+(^2S) + H_2$ reaction probabilities.

Received 13th February 2022,
Accepted 11th April 2022

DOI: 10.1039/d2cp00719c

rsc.li/pccp

1. Introduction

Potential energy surface (PES), as a crucial concept in physical chemistry, was introduced by the separation of nuclear and electronic motions, which gives rise to molecular spectroscopies and chemical reactivity dynamics simulations. Since the quality of dynamical calculations is determined directly by the accuracy of potential energy, numerous numerical methods[1–9] for generating PESs have been developed over the past few decades to avoid excessive computational cost. Constructing an applicable PES of a molecular system generally requires two steps, namely, the calculations of energy points in the selected coordinate range and representing the mathematical relationship between the nuclear configurations and the corresponding potential energies from a mass of discrete data. The developments in the electronic structure theory and computing power have enabled high-precision *ab initio* energies for small systems. However, establishing a global PES of a system with multiple dimensions is still a challenge. This is particularly true for reactive PESs with complex topography in which all the reactant and product channels and the regions where reaction can access need to be included for accurately characterizing the classical trajectory or quantum scattering calculations. For example, it usually takes tens of thousands of points to build complex-forming reactive PESs dominated by a well even for the simplest triatomic systems.[10–12] Therefore, it is a key issue to efficiently sample data points in the nuclear configuration space to make the construction of polyatomic or complex structured PESs truly possible.

In recent years, there has been increasing attention in representing high-dimensional PESs using machine learning algorithms.[13–18] Among these methods, it is worth mentioning that the artificial neural network (NN)[19] is a powerful and robust tool for fitting high-quality PESs of reactive systems in the gas phase and the interaction of molecules with surfaces.[20–37] NNs can give an analytic form about the nuclear coordinates and energies by minimizing the cost function to obtain the optimal parameters of every neuron. However, the implementation of the NN approach is essentially based on the *ab initio* points selected in advance, and it becomes exponentially difficult to perform electronic structure calculations in enormous configurations as the complexity of the system increases. There have been widespread studies on how to saturate the energy points in a large configuration space based

Key Laboratory of Materials Modification by Laser, Electron, and Ion Beams (Ministry of Education), School of Physics, Dalian University of Technology, Dalian 116024, P. R. China. E-mail: mdchen@dlut.edu.cn

on the NN fitting. Raff *et al.* developed the use of trajectories and distance between *ab initio* points to choose new configurations.[38] Behler proposed the sampled scheme of the configuration space by the large discrepancy between two different NN fits.[39] Lin *et al.* put forward an uncertainty-driven strategy to automatically construct multidimensional NN PESs.[40,41] They used the weighted negative squared difference surface between two independent NN structures as the uncertainty metric to search and add new data at the less reliable region of PES. These ways in generating an optimal dataset automatically amount to active learning, a machine learning algorithm with an efficient decision on the selection of training data.

Another popular machine learning method in modeling PESs is the kernel-based Gaussian process regression (GPR).[42] Unlike NNs, GPR is a nonparametric model without a specific functional form, and it provides a statistical estimate of the energy value at an unknown configuration based on the pre-existing *ab initio* points. Benefiting from the Bayesian modeling, GPR can construct accurate PESs with fewer data points, and it has been successfully applied in multiple molecular systems.[43–56] An important advantage of the GPR model is that it can explicitly provide the predictive uncertainty at an unknown configuration *via* variance, providing a straightforward active learning scheme to sample data points. Guan *et al.* utilized this unique feature to reproduce the $H + H_2O \leftrightarrow H_2 + OH$ and $H + CH_4 \leftrightarrow H_2 + CH_3$ reactive PESs by continually adding new energy points with the maximum of variance,[57] and the two reliable PESs were obtained only by assembling 920 and 4000 points, respectively. Uteva *et al.* used three active learning strategies to determine the training sets of GPR in generating intermolecular PESs for CO_2–Ne, CO_2–H_2, and Ar_3 systems,[58] and their studies further demonstrate the high efficiency of GPR in sampling data. Therefore, using the GPR method to represent PESs can save a mass of *ab initio* calculations, and it is more convenient and efficient for data acquisition than the NN model.

On the other hand, the speed of evaluation is also a very important factor for the assessment of constructed PESs. Reactive PESs obtained by the GPR approach are much slower to evaluate than the NN fitting because the product of two vectors with the size of the number of training points n needs to be evaluated numerically, scaling as $O(n)$, which restrains the subsequent dynamics calculations to a great extent when the value of n is relatively large. In contrast, the numerical evaluation of NN PESs only depends on the number of layers and neurons of the network, thus it is very fast to access the energy values of the arbitrary configurations. Moreover, there exists the risk of ill-conditioned covariance matrices when the training set contains very close points or rapidly varying energy values, indicating that the GPR model is not suitable for constructing reactive PESs with obvious well or barrier structures owing to the requirement of dense points in these regions.

To sum up, a pure NN model can represent accurate PESs, but it cannot actively select training points as GPR does, thus the corresponding electronic structure calculations become much too expensive as the molecular complexity increases. PESs generated by the GPR method can significantly save the number of *ab initio* points compared with the NN fitting, but the speed of evaluation is slow and the regions with fast-changing energy cannot be accurately reproduced, thus the GPR model is usually used to construct PESs near the equilibrium geometry rather than global reactive PESs. Therefore, constructing globally accurate multi-dimensional PESs with a rapid speed of evaluation remains a challenge by using a single model. Combining the advantages of the above two machine learning algorithms could be a good idea to overcome this difficulty. In this work, we propose a new scheme for representing reactive PESs with complex topography by incorporating the active data selection of GPR into the NN fitting, which can generate accurate analytic PESs with a rapid speed of evaluation by sampling as few *ab initio* points as possible. First, a small initial dataset is selected to produce a rough GPR PES; and then the new *ab initio* data are added to the training set by searching the maximum in the variance space, and this procedure is automatically repeated until obtaining the convergent value of the highest predictive variance; finally, the NN method is used to construct the analytical PES based on the finally determined training points. We test the validity of this scheme in a simple triatomic $Be^+(^2S) + H_2$ reactive system. The main reason is that accurate *ab initio* data can be obtained due to the relatively small number of electrons of this system, which is critical for testing a new approach for constructing PESs. Moreover, compared to H_3, LiH_2, or other simpler systems, the ground state BeH_2^+ PES features more complex topography characteristics, including wells, barriers, cusps formed by the avoided crossing, and rapidly changing energies in some regions. Therefore, the $Be^+(^2S) + H_2$ system is a suitable candidate for examining the feasibility and universal applicability of this new strategy. The remainder of this paper is as follows: Section 2 gives the detailed methodology, including the *ab initio* calculations, searching points in GPR and NN fitting. In Section 3, the results and discussion are presented, which prove the effectiveness of this scheme in representing reactive PESs with few points. Section 4 concludes this article.

2. Methods

In this work, the *ab initio* calculations are performed using Molpro 2012 software.[59] The energy points of the ground state $(1^2A')$ BeH_2^+ are calculated at the internally contracted multi-reference configuration interaction[60,61] levels with the Davidson correction (MRCI + Q), and the augmented Dunning-type correlation consistent polarized quadruple basis set[62] is adopted for H and Be atoms. *Ab initio* energy calculations for this system are carried out in a wide configuration space, and the data points are sampled in the Jacobi coordinates (r, R, and θ). The reactive region of $Be^+(^2S) + H_2$ is defined as $0.4\ \text{Å} \leq r \leq 9.6\ \text{Å}$, $0.1\ \text{Å} \leq R \leq 13.2\ \text{Å}$, $0^\circ \leq \theta/\text{degree} \leq 90^\circ$, and the product region of $BeH^+ + H$ is defined as $1.0\ \text{Å} \leq r \leq 9.6\ \text{Å}$,

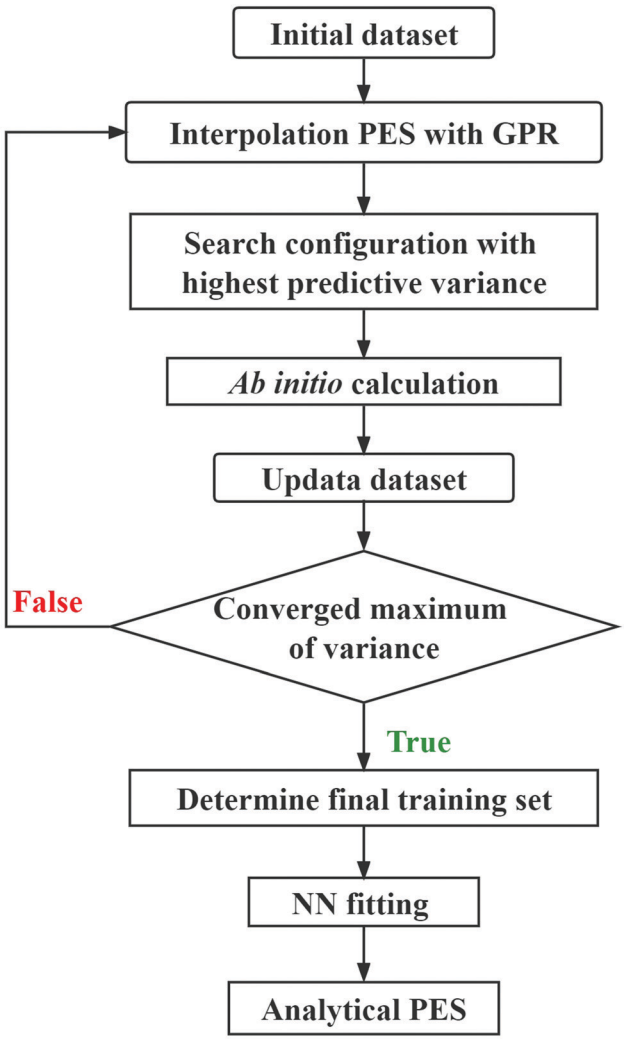


Fig. 1 The pragmatic procedure of the new scheme for representing multi-dimensional reactive PESs.

0.1 Å $\leq R \leq$ 13.2 Å, $0^\circ \leq \theta$/degree $\leq 180^\circ$. A total of 12 040 energy points covering the whole configuration region are selected as the reference set, which provides the candidate points added to the training database and serves as the test data to examine the generalization ability of the trained model.

This new scheme that combines the actively selecting points of GPR and the NN fitting is a universal approach for the construction of multidimensional reactive PESs, and the basic procedure of this strategy is illustrated in Fig. 1. For the triatomic system, the initial configurations can be sampled in the Jacobi coordinates by using the Latin hypercube sampling (LHS) approach,[63] which can avoid clustering and the samples cover the whole coordinate region in each degree of freedom. It is important to note that using the Jacobi coordinates to sample is not suitable for polyatomic reactive systems due to the significantly increasing complexity. In addition, a previous study[58] suggests that the LHS has limitations for large systems. As the system's dimensions increase, the initial configurations can be sampled along each reaction pathway with the other coordinates fixed at the corresponding transition states, which are determined by relatively low-level *ab initio* calculations, ensuring that the dynamically relevant configuration space is covered. A preliminary GPR PES is constructed based on the initial energy points. Next, the new *ab initio* data are added to the training set by searching the configuration with the highest predictive variance by this GPR PES in a large coordinate space, and then a new GPR PES is constructed using the updated training dataset. This process is repeated automatically until obtaining the convergent maximum of predictive variance, and the training set is eventually determined. In the final step, the analytical PES is represented by the NN fitting based on this training set. To ensure the important permutation symmetry of the generated PES, the permutation invariant polynomial (PIP)[9] is adopted as the input of GPR and NN models, namely, the PIP-GPR[53] and PIP-NN[21,24] methods are used in the process of constructing PES.

Next, the basic theory and important equations involved in this strategy are described. The Gaussian process is a kernel-based non-parametric supervised machine learning algorithm, which can be regarded as a limit of the Bayesian network with an infinite number of nodes. The detailed description of GPR can be found in the relevant literature,[42] and here we only give a brief introduction about its application in active data selection for representing PESs. Supporting the initial dataset containing n configurations $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, where the inputs of $\boldsymbol{x}_i$ are the PIPs constructed by the interatomic distance, and the outputs $\boldsymbol{y} = [y_1, \ldots, y_n]$ of the GPR model are the corresponding normalized potential energies. The joint multivariate Gaussian distribution can be expressed as follows:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})) \tag{1}$$

in which the mean function is set as zero for simplicity, and $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ is a $n \times n$ matrix with elements of kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, representing the covariance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Here, the type of anisotropic Matérn kernel with $\nu = 2.5$ is selected and written as follows:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = c\left(1 + \sqrt{5}\frac{\mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\boldsymbol{l}} + \frac{5}{3}\frac{\mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_j)^2}{\boldsymbol{l}^2}\right)\exp\left(-\sqrt{5}\frac{\mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\boldsymbol{l}}\right) + \delta_{ij}\sigma_n \tag{2}$$

where c is a constant to be optimized, and $\mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ represents the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$; here, $\boldsymbol{l} = [l_1, l_2, l_3]$ is the length-scale vector. $\delta_{ij}\sigma_n$ donates the noise term of following a Gaussian distribution that is added to the diagonal of $\boldsymbol{K}$, which can avoid the risk of the ill covariance matrix; c, $\boldsymbol{l}$ and σ_n form the set of hyperparameters, denoted as $\boldsymbol{\theta}$.

The goal of training a GPR is to obtain the parameters of kernel function by maximizing the following logarithm of the marginal likelihood,

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K}| - \frac{n}{2}\log(2\pi) \tag{3}$$

A new data point ($\boldsymbol{x}^*$, $\boldsymbol{y}^*$) to be predicted also follows the prior distribution of eqn (1),

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}^* \end{bmatrix} \sim \mathcal{N}\begin{bmatrix} \boldsymbol{0}, & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) & \boldsymbol{K}^{*\mathrm{T}}(\boldsymbol{x}^*,\boldsymbol{X}) \\ & \boldsymbol{K}^*(\boldsymbol{X},\boldsymbol{x}^*) & \boldsymbol{K}^{**}(\boldsymbol{x}^*,\boldsymbol{x}^*) \end{bmatrix} \quad (4)$$

where $\boldsymbol{K}^{**} = (\boldsymbol{x}^*, \boldsymbol{x}^*)$, and $\boldsymbol{K}^*$ is a vector that consists of the covariance between $\boldsymbol{x}^*$ and all of the training data. As a result, the predicted mean of $\boldsymbol{y}^*$ is given by:

$$\mu(\boldsymbol{x}^*) = [\boldsymbol{K}^*\boldsymbol{K} + \sigma_n{}^2\boldsymbol{I}]^{-1}\boldsymbol{y} \quad (5)$$

and its variance can be calculated as follows:

$$\sigma^2(\boldsymbol{x}^*) = \boldsymbol{K}^{**} - \boldsymbol{K}^{*\mathrm{T}}[\boldsymbol{K} + \sigma_n{}^2\boldsymbol{I}]^{-1}\boldsymbol{K}^* \quad (6)$$

The new *ab initio* data with the highest predictive variance are iteratively added to the GPR training dataset until the maximum of variance converges, and then the analysis PES is obtained by using the PIP-NN model based on the dataset determined by the GPR model.

NNs consist of layers of interconnected mathematical function simulating biological neurons, and it can present a flexible form with arbitrary precision. There has been increasing interest in generating multi-dimensional PESs with NNs, and for more details please refer to the relevant reviews.[14,15] For the studied system, the NN structure contains two hidden layers with 11 neurons. The analytical form in expressing the relationship between inputs and outputs is written as follows:

$$\boldsymbol{y} = \varphi^{(3)}\left(\sum_{i=1}^{11} w_i^{(3)}\varphi^{(2)}\left(\sum_{j=1}^{11} w_j^{(2)}\varphi^{(1)}\left(\sum_{k=1}^{3} w_k^{(1)}X_k + b_j^{(1)}\right) + b_i^{(2)}\right) + b^{(3)}\right) \quad (7)$$

where w and b represent the weights and bias, which are regulated by the Levenberg–Marquardt algorithm.[64] φ is the transfer function between the two adjacent layers, and the smooth hyperbolic tangent function and simple linear function are selected in the 1–2, 2–3 layers, and 3–4 layers, respectively. We perform the cross-validation approach to avoid overfitting, namely, 90% *ab initio* points are used to fit the NN PES, and the performance of the trained model is examined by the other data.

3. Results and discussion

To begin with, a total of 40 *ab initio* data are selected by the LHS strategy as the initial training set, and then the new data in the test dataset are added based on PIP-GPR interpolation. The highest predictive variance as a function of the number of training points is described in Fig. 2(a). As can be seen, the value of the highest variance decreases rapidly at first and then tends to be flat. The highest variance is convergent to 3.45 eV^2 when the number of training points reaches 1270, and these

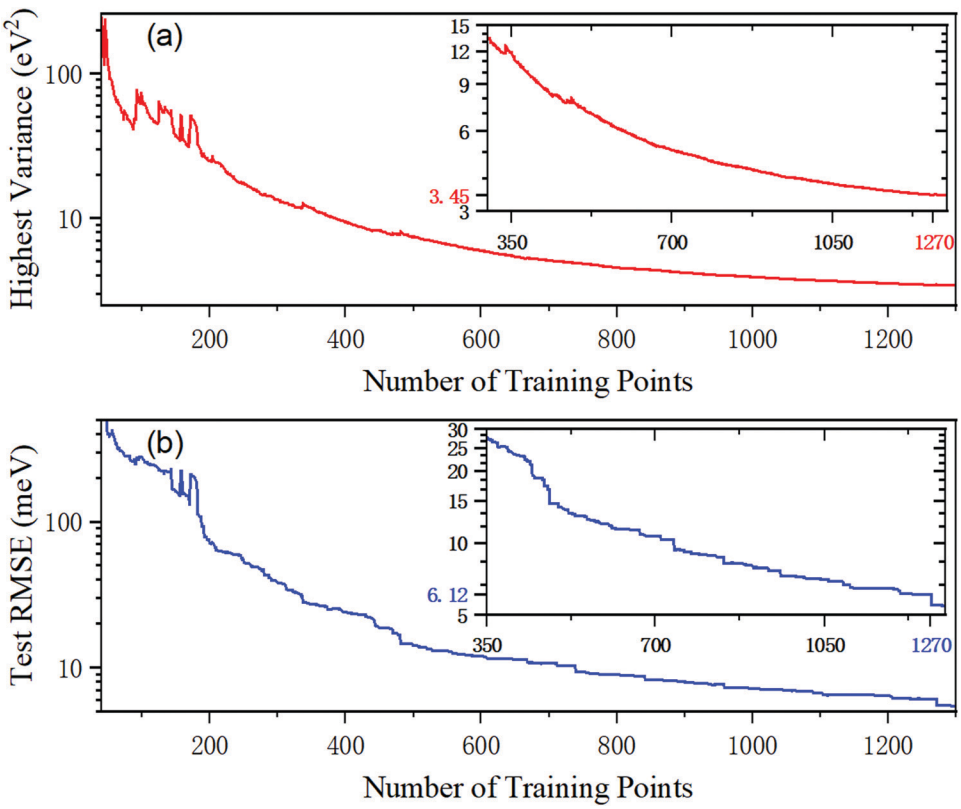


Fig. 2 The highest predictive variance and the RMSE of the test database as a function of the number of training points in the GPR training for the BeH_2^+ system.

actively sampling energy points are used in the next PIP-NN fitting. Fig. 2(b) gives the test error calculated on PIP-GPR PES as a function of the number of training data. The test error refers to the root mean square error (RMSE) of all the test data points. The RMSE value is very large when the model is trained with a small number of points, which is different from H_3[57] or Ar_3[58] PESs constructed using the GPR model. In these systems, the RMES can be decreased to the meV level merely by using dozens of samples. This is because the ground state BeH_2^+ PES contains abundant wells and barriers and the energy values change rapidly in some configurations; more points are required for accurately describing its structural characteristics even adopting the GPR method that greatly saves the electronic structure calculations. The test RMSE of the PIP-GPR PES trained by 1270 points is 6.12 meV.

Fig. 3 shows the difference between the predicted energies and the *ab initio* results in the test database, and the values of predictive energy are obtained on the PESs constructed using the PIP-GPR model and the proposed new approach with 1270 training data. The energy values of the abscissa axis are relative to the dissociation limit of triatomic Be^+–H–H. The test RMSEs of the PES represented by the pure PIP-GPR model and this new method are 6.12 meV and 1.80 meV, respectively, implying the accuracy of the PES is remarkably improved by using an additional NN fitting after actively selecting points with the GPR model. In the new method, the final analytic PES is fitted by the PIP-NN scheme, which can yield extremely accurate PESs, such as the recently reported ultracold reactive system of KRb + KRb → K_2 + Rb_2,[65] which presented an RMSE of only 1.86 cm^{-1}. The $Be^+(^2S)$ + H_2 reactive PES features multiple wells, barriers, and cusps formed by the avoided crossing effect of the first excited state,[66] which go against the fitting accuracy, and too small RMSE can also increase the risk of long-range potential for the tested system. Although the fitting RMSE does not reach the order of spectroscopic accuracy, the accuracy of the obtained PES is sufficient for dynamics studies on the endothermic reaction of $Be^+(^2S)$ + H_2. It can be seen from the distributions of the test errors that both the methods can give accurate predictions in the energy region below −4 eV. However, the PIP-GPR PES presents large predictive errors in the region of relatively high energy. On the contrary, the PES generated by the new strategy keeps a very small predictive error in the whole energy region. The absolute highest error values in the test dataset for the PIP-GPR method and the new model are 179.6 meV and 11.3 meV, respectively, suggesting that the generalization performance of the new approach is much better than that of the pure GPR model for the construction of PESs with complex topography.

The speed of evaluation is also a very important aspect to assess PESs, and it directly affects the efficiency of the subsequent quantum scattering calculations. The evaluation of NN PESs depends on the number of neurons, and any two layers are linked by a simple function; thus, the speed of evaluating a NN PES is very fast. For the GPR model, as shown in eqn (5), it requires to calculate the covariance matrixes between the predicted configurations and all of the training data and the

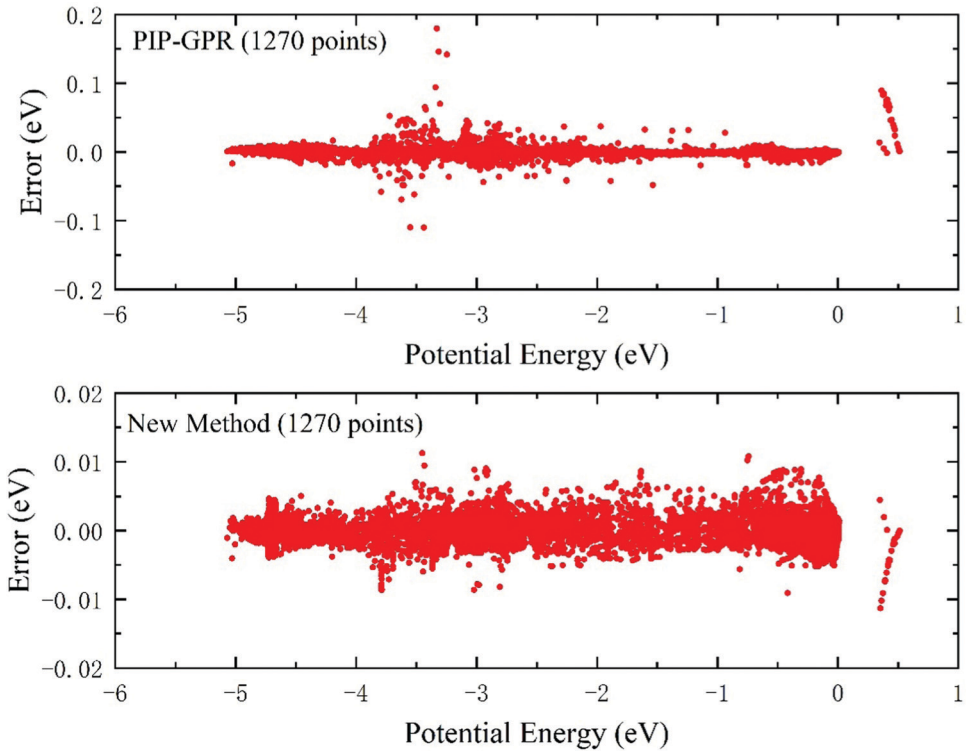


Fig. 3 Predictive error distributions in the test database of the PESs constructed by the PIP-GPR model (1270 points) and the new method (1270 points) for the BeH_2^+ system.

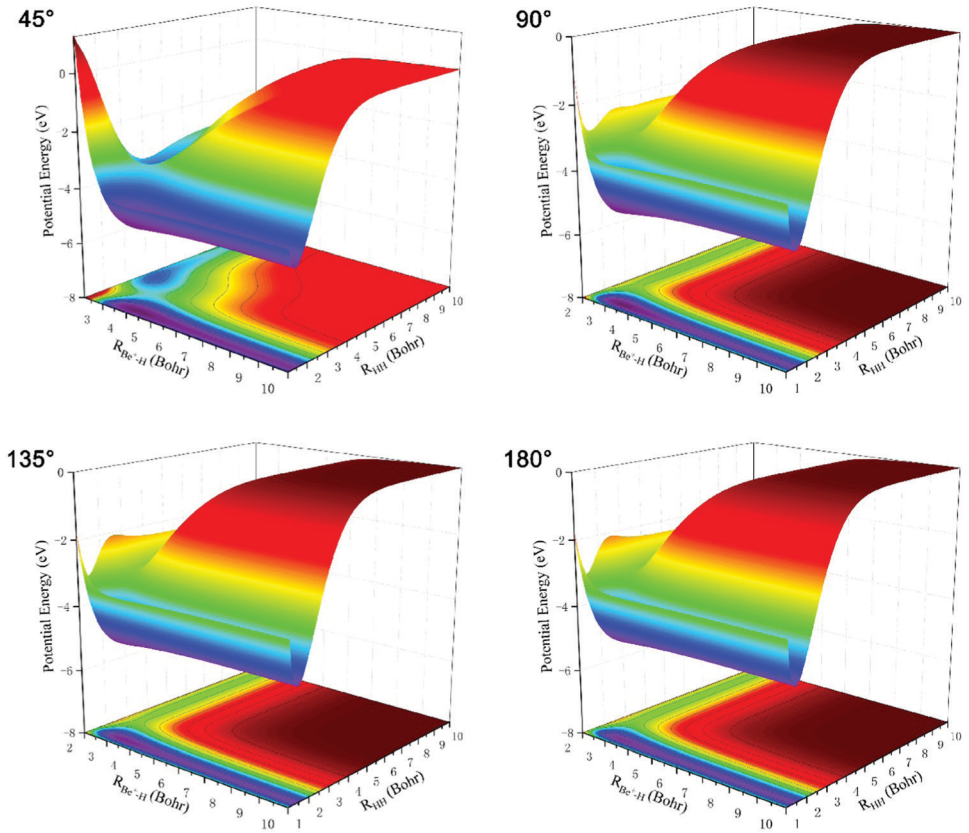


Fig. 4 Three-dimensional BeH_2^+ PESs constructed by the new method at four different Be^+–H–H angles (45°, 90°, 135°, and 180°).

product of vector–vector with n dimensions, so the efficiency of predicting a new data point is very low when the number of training data is relatively large. We use a single core on a single compute node (Intel(R) Core(TM) i5-1035G1 CPU@1.00 GHz) to evaluate 100 000 potential values on the PESs modeled using different schemes, and the evaluation times for the PIP-GPR PES and the PES constructed by the new method are 25.41 s and 7.37 s, respectively. The time cost is saved around three times than the pure GPR method, and this ratio will increase remarkably when the training set contains more data points.

Three-dimensional BeH_2^+ PESs constructed using the new scheme at four different Be^+–H–H angles (45°, 90°, 135°, and 180°) are plotted in Fig. 4. The PES is smooth in the whole coordination space and no non-physical structures are present for every angle, demonstrating that there is no overfitting during PIP-NN training. The right valley and left valley on the PES correspond to the $Be^+(^2S)$ + H_2 channel and BeH^+ + H channel, respectively. It can be seen that there exist a well and a barrier on the PES for each Be^+–H–H angle, and the two structures become less obvious with the increase of the Be^+–HH angle. The topography characteristics of BeH_2^+ PES imply the complicated changes between the nuclear configuration and energy.

The molecular constants of diatomic species $H_2(X^1\Sigma_g^+)$ and $BeH^+(X^1\Sigma^+)$ are listed in Table 1. The values of bond length R_e, dissociation energy D_e, vibrational frequencies ω_e, and anharmonicity constants $\omega_e x_e$ calculated on the PES generated by the new method coincide with the experimental values well,[67] indicating that the new PES can accurately describe the distributions of the rovibrational states of the reactant and product when the dynamics calculations of the $Be^+(^2S)$ + H_2 → BeH^+ + H reaction are implemented. Table 2 gives the geometries and energy values of stationary points for the ground state BeH_2^+, compared with the previous MRCI + Q results.[68] The energy values are relative to the Be^+ + H_2 dissociation limit. It can be seen that the equilibrium structure and saddle points obtained on the PES constructed using this new scheme are in good agreement with the high-level *ab initio* calculations.

Table 1 Molecular constants of $H_2(X^1\Sigma_g^+)$ and $BeH^+(X^1\Sigma^+)$

		This work	Experimental data[67]
$H_2(X^1\Sigma_g^+)$	R_e (bohr)	1.400	1.401
	D_e (eV)	4.743	4.747
	ω_e (cm^{-1})	4402.7	4401.2
	$\omega_e x_e$ (cm^{-1})	104.90	121.33
$BeH^+(X^1\Sigma^+)$	R_e (bohr)	2.489	2.480
	D_e (eV)	3.163	3.280
	ω_e (cm^{-1})	2220.5	2221.7
	$\omega_e x_e$ (cm^{-1})	40.74	39.79

To further verify the reliability of this scheme in representing global reactive PESs, we use all the test data and additional

Table 2 Stationary points for the ground state BeH_2^+

	r_{HH} (bohr)	R_{Be^+-HH} (bohr)	Energy (eV)
Minimum, $\theta = 90^\circ$			
This work	1.454	3.352	−0.380
MRCI + Q[68]	1.449	3.352	−0.374
Saddle point, $\theta = 0^\circ$			
This work	1.438	4.091	−0.129
MRCI + Q[68]	1.438	4.067	−0.133

82 points on the minimum energy path to fit an accurate PIP-NN PES as the benchmark for comparison, and the fitting RMSE is only 0.91 meV. Fig. 5(a and b) show the collinear and global minimum energy paths of the $Be^+(^2S) + H_2 \rightarrow BeH^+ + H$ reaction, respectively. For the collinear collision, there are a shallow well and a low barrier, and the reactive paths calculated on the PESs generated by the three approaches are indistinguishable. Compared to the collinear path, the global minimum energy path includes a more obvious well and barrier, and there exists a small hump behind the barrier. It can be seen that the path obtained on the PES generated by the new method

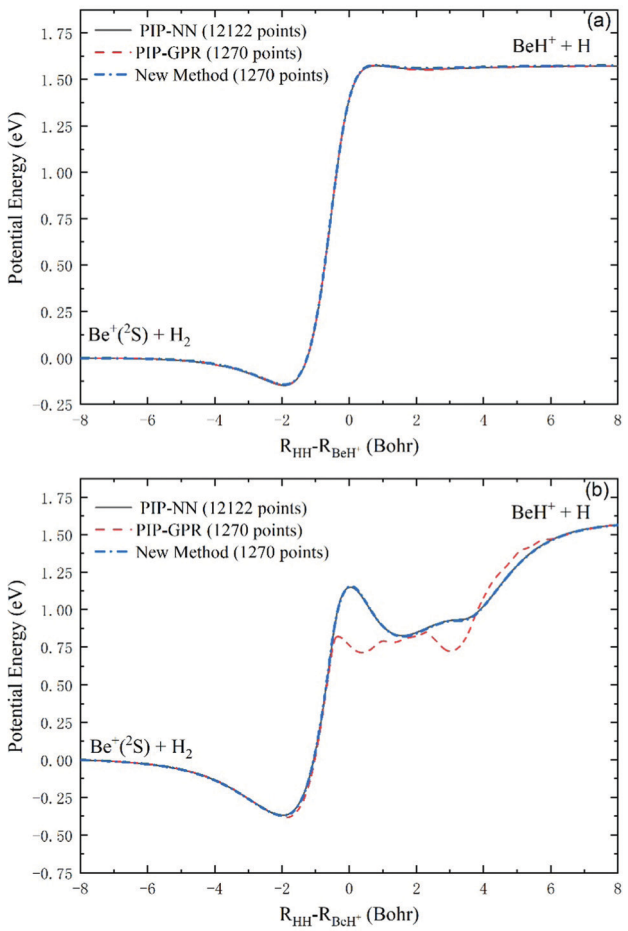


Fig. 5 Collinear (a) and global (b) minimum energy paths of the $Be^+(^2S) + H_2 \rightarrow BeH^+ + H$ reaction calculated on the PESs constructed by the PIP-NN model (12 122 points), the PIP-GPR model (1270 points), and the new method (1270 points).

is completely consistent with the PIP-NN PES constructed with 12 122 points, but the PIP-GPR PES produced by the same database cannot well reproduce the barrier and the complex shape of the product region. These results indicate that the GPR model only can produce sufficiently accurate results at the smooth parts of the PES, whereas the regions with fast-changing energy values cannot be described correctly for this system.

In addition to the assessments of the error, speed of evaluation, and topography characteristics, quantifying the quality of reactive PESs by quantum scattering is also very important. To prove that the new scheme can accurately character the dynamically relevant regions of PES, we perform the reactant coordinate based time-dependent wave packet[69,70] calculations on the $Be^+(^2S) + H_2 \rightarrow BeH^+ + H$ reaction based on the PESs produced by different strategies. The time evolution of the wave function is based on the second-order split operator method.[71] The main parameters used in the dynamical calculations are given in Table 3. In Fig. 6, the total reaction probabilities of the $Be^+(^2S) + H_2 \rightarrow BeH^+ + H$ reaction with the total angular momentum number $J = 0$ are displayed. The energy threshold value is 1.45 eV, which corresponds to the endothermicity of this reaction. The results obtained on PIP-NN PES with 12 122 points save as the benchmark for comparison. It is clear that the reaction probabilities calculated on PIP-GPR PES have substantial errors, and the key resonance characteristics formed by the intermediate complex on the wells are not presented. This suggests that BeH_2^+ PES modeled by the GPR method is extremely unreliable for dynamics studies even though the number of training points has reached thousands. The main reason for the almost negligible probabilities calculated on PIP-GPR PES is that the collision of $J = 0$ partial wave is dominated by the global minimum energy path, as shown in Fig. 5(b), and this PES does not reproduce the key activated barrier, which corresponds to the transition state of this reaction, thus the calculated reaction probabilities are significantly weakened. To prove that the strategy can quickly converge to an error range necessary to run dynamics, we represent a PES with only 600 points based on this new scheme, and the corresponding reaction probabilities are also presented in Fig. 6. For the new PES with 600 points, the reaction probabilities agree with the results calculated on PIP-NN PES generated with 12 122 points, indicating the high efficiency of the presented strategy in obtaining reliable PESs that can be used in dynamics studies. When the number of training data increases to 1270, the corresponding results are almost identical to those of the NN PES, suggesting that the PES obtained by this new scheme can accurately describe the dynamically relevant regions.

Our calculation results of the ground BeH_2^+ PES suggest that the standard GPR method could not correctly represent the reactive PESs with complex topography characteristics based on a small number of points. One important reason is that the GPR model inherently cannot give accurate prediction for the rapidly changing value, and the ill-conditioned covariance matrix may appear when the data distribution is relatively dense, resulting in numerical instability in those regions. But

Table 3 Main numerical parameters in the time-dependent wave packet calculations

$Be^+(^2S) + H_2 \rightarrow BeH^+ + H$	
Grid/basis range and size	R (bohr) $\in$ [0.1, 25.0], N_R^{tot} = 399, N_R^{int} = 269 r (bohr) $\in$ [0.01, 20.0], N_r^{tot} = 200, N_r^{int} = 99 N_j = 139
Initial wave packet $\exp\left(-\frac{(R-R_c)^2}{2\Delta_R^2}\right)\cos k_0 R$	R_c = 16.0 bohr, Δ_R = 0.20 bohr, $k_0 = (2E_0\mu_R)^{1/2}$ with E_0 = 4.0 eV
Total propagation time	30 000 a.u.

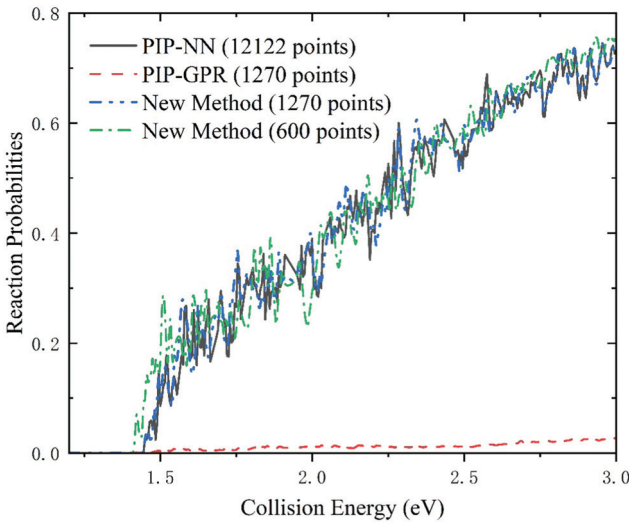


Fig. 6 The total reaction probabilities of the $Be^+(^2S) + H_2 \rightarrow BeH^+ + H$ reaction with J = 0 calculated on the PESs constructed by the PIP-NN model (12 122 points), the PIP-GPR model (1270 points), and the new method (600 and 1270 points).

the GPR model is a very convenient and reliable tool for sampling data by giving the uncertainties of the predicted points. NNs have excellent generalization ability, but the performance of NNs in constructing PESs depends strongly on the distribution of training data. Incorporating the actively selecting data of GPR into the NN fitting can produce globally accurate reactive PESs with a rapid speed of evaluation based on as few *ab initio* points as possible.

4. Conclusions and prospects

There has been fast developing interest in constructing multi-dimensional PESs using modern machine learning technologies, and most of them are implemented by the NN and GPR algorithms. NN is a powerful tool to generate accurate PESs, but the NN fitting is based on the points selected in advance, resulting in expensive electron structure calculations as the molecular complexity increases. Although GPR provides a direct approach in actively sampling data, the constructed PESs have a low speed of evaluation and the rapidly changing potential cannot be accurately characterized. Therefore, combining the advantages of the above two methods could be a good idea to obtain accurate PESs with highly efficient evaluation based on as few *ab initio* points as possible. In this work, an efficient scheme for representing globally accurate reactive PESs with complex topography by incorporating the actively selecting points of GPR into the NN fitting is proposed. This strategy is verified by the triatomic example of the BeH_2^+ system, and only 1270 points are assembled to construct the global PES. The accuracy, generalization performance, and speed of evaluation of the PES constructed by this new scheme are much better than those of the PES produced by the pure GPR with the same training data. To further test the global accuracy of the PES, an accurate PIP-NN PES is constructed with 12 122 points as the benchmark for comparison. The topography characteristics of the PES generated by this scheme and the calculated quantum reaction probabilities of the $Be^+(^2S) + H_2$ reaction are in good agreement with the corresponding results obtained on this PIP-NN PES.

The proposed scheme is an all-purpose method for representing global reactive PESs, characterized by high accuracy and rapid speed of evaluation, and its advantages are particularly prominent for systems with complex topography. For instance, the long-lived complex-forming reactive systems usually require a mass of *ab initio* points to model PESs and abundant energy grids to character the quantum dynamics. This method can greatly save the cost of electronic structure calculations and the generated PES can accelerate the quantum scattering calculations. The presented scheme seems to be very promising for constructing reactive PESs with more dimensionality. On the other hand, this scheme also has some limitations for particularly large systems. The inverse of the covariance matrix needs to be calculated in each GPR iteration, and the training complexity scales as $O(n^3)$, so the speed of actively selecting points dramatically decreases with the increase of dimension. This speed can be effectively improved by decreasing the number of training data at the cost of reducing the accuracy.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 11774043).

References

1 M. A. Collins, *Theor. Chem. Acc.*, 2002, **108**, 313–324.
2 T. Hollebeek, T. S. Ho and H. Rabitz, *Annu. Rev. Phys. Chem.*, 1999, **50**, 537–570.
3 C. Qu, Q. Yu and J. M. Bowman, *Annu. Rev. Phys. Chem.*, 2018, **69**, 151–175.
4 G. C. Schatz, *Rev. Mod. Phys.*, 1989, **61**, 669–688.
5 J. Ischtwan and M. A. Collins, *J. Chem. Phys.*, 1994, **100**, 8080–8088.
6 A. Aguado and M. Paniagua, *J. Chem. Phys.*, 1992, **96**, 1265–1275.
7 K. S. Sorbie and J. N. Murrell, *Mol. Phys.*, 1975, **29**, 1387–1407.
8 A. J. C. Varandas, *Adv. Chem. Phys.*, 1988, **74**, 255–338.
9 B. J. Braams and J. M. Bowman, *Int. Rev. Phys. Chem.*, 2009, **28**, 577–606.
10 W. T. Li, J. C. Yuan, M. L. Yuan, Y. Zhang, M. H. Yao and Z. G. Sun, *Phys. Chem. Chem. Phys.*, 2018, **20**, 1039–1050.
11 Z. J. Yang, S. F. Wang, J. C. Yuan and M. D. Chen, *Phys. Chem. Chem. Phys.*, 2019, **21**, 22203–22214.
12 J. C. Yuan, D. He and M. D. Chen, *Sci. Rep.*, 2015, **5**, 14594.
13 J. Behler, *J. Chem. Phys.*, 2016, **145**, 170901.
14 B. Jiang, J. Li and H. Guo, *J. Phys. Chem. Lett.*, 2020, **11**, 5120–5131.
15 S. Manzhos and T. Carrington, *Chem. Rev.*, 2021, **121**, 10187–10217.
16 G. Schmitz, I. H. Godtliebsen and O. Christiansen, *J. Chem. Phys.*, 2019, **150**, 244113.
17 T. T. Nguyen, E. Szekely, G. Imbalzano, J. Behler, G. Csanyi, M. Ceriotti, A. W. Gotz and F. Paesani, *J. Chem. Phys.*, 2018, **148**, 241725.
18 V. Vassilev-Galindo, G. Fonseca, I. Poltavsky and A. Tkatchenko, *J. Chem. Phys.*, 2021, **154**, 094119.
19 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533–536.
20 S. Manzhos, X. G. Wang, R. Dawes and T. Carrington, *J. Phys. Chem. A*, 2006, **110**, 5295–5304.
21 B. Jiang and H. Guo, *J. Chem. Phys.*, 2013, **139**, 054112.
22 C. J. Xie, X. L. Zhu, D. R. Yarkony and H. Guo, *J. Chem. Phys.*, 2018, **149**, 144107.
23 B. Jiang and H. Guo, *J. Chem. Phys.*, 2014, **141**, 034109.
24 B. Jiang, J. Li and H. Guo, *Int. Rev. Phys. Chem.*, 2016, **35**, 479–506.
25 J. C. Yuan, D. He and M. D. Chen, *Phys. Chem. Chem. Phys.*, 2015, **17**, 11732–11739.
26 K. J. Shao, J. Chen, Z. Q. Zhao and D. H. Zhang, *J. Chem. Phys.*, 2016, **145**, 071101.
27 J. Behler, *Int. J. Quantum Chem.*, 2015, **115**, 1032–1050.
28 B. Kolb, B. Zhao, J. Li, B. Jiang and H. Guo, *J. Chem. Phys.*, 2016, **144**, 224103.
29 T. G. Yang, A. Y. Li, G. K. Chen, C. J. Xie, A. G. Suits, W. C. Campbell, H. Guo and E. R. Hudson, *J. Phys. Chem. Lett.*, 2018, **9**, 3555–3560.
30 X. R. Zhang, L. L. Li, J. Chen, S. Liu and D. H. Zhang, *Nat. Commun.*, 2020, **11**, 223.
31 I. Miyazato, S. Nishimura, L. Takahashi, J. Ohyama and K. Takahashi, *J. Phys. Chem. Lett.*, 2020, **11**, 787–795.
32 Y. Y. Hong, Z. X. Yin, Y. F. Guan, Z. J. Zhang, B. N. Fu and D. H. Zhang, *J. Phys. Chem. Lett.*, 2020, **11**, 7552–7558.
33 X. X. Hu, Y. P. Zhou, B. Jiang, H. Guo and D. Q. Xie, *Phys. Chem. Chem. Phys.*, 2017, **19**, 12826–12837.
34 J. L. Chen, X. Y. Zhou, Y. L. Zhang and B. Jiang, *Nat. Commun.*, 2018, **9**, 4039.
35 J. Q. Han, L. F. Zhang, R. Car and E. Weinan, *Commun. Comput. Phys.*, 2018, **23**, 629–639.
36 K. T. Schutt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko and K. R. Muller, *J. Chem. Phys.*, 2018, **148**, 241722.
37 Z. X. Yin, B. J. Braams, Y. F. Guan, B. N. Fu and D. H. Zhang, *Phys. Chem. Chem. Phys.*, 2021, **23**, 1082–1091.
38 L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley and R. Komanduri, *J. Chem. Phys.*, 2005, **122**, 084104.
39 J. Behler, *Phys. Chem. Chem. Phys.*, 2011, **13**, 17930–17955.
40 Q. D. Lin, Y. L. Zhang, B. Zhao and B. Jiang, *J. Chem. Phys.*, 2020, **152**, 154104.
41 Q. D. Lin, L. Zhang, Y. L. Zhang and B. Jiang, *J. Chem. Theory Comput.*, 2021, **17**, 2691–2701.
42 C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, MIT Press, Cambridge, Mass., 2006, vol. 1.
43 A. Christianen, T. Karman, R. A. Vargas-Hernandez, G. C. Groenenboom and R. V. Krems, *J. Chem. Phys.*, 2019, **150**, 064106.
44 Q. Liu, L. Liu, F. An, J. Huang, Y. Z. Zhou and D. Q. Xie, *J. Chem. Phys.*, 2021, **155**, 084302.
45 B. Kolb, P. Marshall, B. Zhao, B. Jiang and H. Guo, *J. Phys. Chem. A*, 2017, **121**, 2552–2557.
46 J. Cui and R. V. Krems, *Phys. Rev. Lett.*, 2015, **115**, 073202.
47 A. Kamath, R. A. Vargas-Hernandez, R. V. Krems, T. Carrington and S. Manzhos, *J. Chem. Phys.*, 2018, **148**, 241702.
48 E. Uteva, R. S. Graham, R. D. Wilkinson and R. J. Wheatley, *J. Chem. Phys.*, 2017, **147**, 161706.
49 V. L. Deringer, A. P. Bartok, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csanyi, *Chem. Rev.*, 2021, **121**, 10073–10141.
50 J. Cui and R. V. Krems, *J. Phys. B: At., Mol. Opt.*, 2016, **49**, 224001.
51 R. V. Krems, *Phys. Chem. Chem. Phys.*, 2019, **21**, 13392–13410.
52 S. Venturi, R. L. Jaffe and M. Panesi, *J. Phys. Chem. A*, 2020, **124**, 5129–5146.
53 C. Qu, Q. Yu, B. L. Van Hoozen, J. M. Bowman and R. A. Vargas-Hernandez, *J. Chem. Theory Comput.*, 2018, **14**, 3381–3396.
54 J. Dai and R. V. Krems, *J. Chem. Theory Comput.*, 2020, **16**, 1386–1395.
55 H. Sugisawa, T. Ida and R. V. Krems, *J. Chem. Phys.*, 2020, **153**, 114101.
56 Q. F. Song, Q. Y. Zhang and Q. Y. Meng, *J. Chem. Phys.*, 2020, **152**, 134309.
57 Y. F. Guan, S. Yang and D. H. Zhang, *Mol. Phys.*, 2018, **116**, 823–834.

58 E. Uteva, R. S. Graham, R. D. Wilkinson and R. J. Wheatley, *J. Chem. Phys.*, 2018, **149**, 174114.
59 H. J. Werner, P. J. Knowles, G. Knizia, F. R. Manby and M. Schütz, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 242–253.
60 H. J. Werner and P. J. Knowles, *J. Chem. Phys.*, 1988, **89**, 5803–5814.
61 P. J. Knowles and H. J. Werner, *Chem. Phys. Lett.*, 1988, **145**, 514–522.
62 R. A. Kendall, T. H. Dunning and R. J. Harrison, *J. Chem. Phys.*, 1992, **96**, 6796–6806.
63 M. Stein, *Technometrics*, 1987, **29**, 143–151.
64 M. T. Hagan and M. B. Menhaj, *IEEE Trans. Neural Netw.*, 1994, **5**, 989–993.
65 D. Z. Yang, J. X. Zuo, J. Huang, X. X. Hu, R. Dawes, D. Q. Xie and H. Guo, *J. Phys. Chem. Lett.*, 2020, **11**, 2605–2610.
66 Z. J. Yang, J. C. Yuan, S. F. Wang and M. D. Chen, *RSC Adv.*, 2018, **8**, 22823–22834.
67 K. P. Huber and G. Herzberf, *Constants of Diatomic Molecules*, Springer, 1979.
68 A. J. Page, D. J. D. Wilson and E. I. von Nagy-Felsobuki, *Phys. Chem. Chem. Phys.*, 2010, **12**, 13788–13797.
69 S. Gómez-Carrasco and O. Roncero, *J. Chem. Phys.*, 2006, **125**, 054102.
70 Z. G. Sun, X. Lin, S. Y. Lee and D. H. Zhang, *J. Phys. Chem. A*, 2009, **113**, 4145–4154.
71 M. D. Feit, J. A. Fleck and A. Steiger, *J. Comput. Phys.*, 1982, **47**, 412–433.

