Using principal component analysis for neural network high-dimensional potential energy surface

Bastien Casier, Stéphane Carniato, Tsveta Miteva, Nathalie Capron, Nicolas Sisourat

Using Principal Component Analysis for Neural Network High-Dimensional Potential Energy Surface.

Bastien Casier,[a] Stéphane Carniato,[a] Tsveta Miteva,[a] Nathalie Capron,[a] and Nicolas Sisourat[b]

(Dated: 22 September 2020)

Potential energy surfaces (PESs) play a central role in our understanding of chemical reactions. Despite the impressive development of efficient electronic structure methods and codes, such computations still remain a difficult task for the majority of relevant systems. In this context, artificial neural networks (NNs) are promising candidates to construct the PES of a wide range of systems. However, the choice of suitable molecular descriptors remains a bottleneck for these algorithms. In this work, we show that a principal components analysis (PCA) is a powerful tool to prepare an optimal set of descriptors and to build an efficient NN: this protocol leads to a substantial improvement of the NNs in learning and predicting a PES. Furthermore, the PCA provides a means to reduce the size of the input space (i.e. number of descriptors) without losing accuracy. As an example, we applied this novel approach to the computation of the high-dimensional PES describing the keto-enol tautomerism reaction occurring in the acetone molecule.

[a]Sorbonne Université, CNRS, Laboratoire de Chimie Physique Matière et Rayonnement, UMR 7614, F-75005 Paris, France

[b]Electronic mail: nicolas.sisourat@sorbonne-universite.fr; Sorbonne Université, CNRS, Laboratoire de Chimie Physique Matière et Rayonnement, UMR 7614, F-75005 Paris, France

I. INTRODUCTION

Potential energy surfaces (PESs) play a central role in our understanding of chemical reactions. For example, molecular dynamics (MD) simulations are currently the most employed methods to investigate the dynamical behaviour of atomic and molecular complex systems. However, the results of these simulations strongly depend on the quality of the PESs on which the propagation is performed[1]. The computations of accurate PESs represent a difficult task for the majority of relevant systems. Furthermore, MD simulations require single point energy calculations for a very large number of molecular geometries. Therefore, *ab initio* MD[2] are generally limited to systems with a small number of atoms using density functional theory[3,4] (DFT). A good alternative can be found in the accurate interpolation of the PESs. However, PESs are complex hypersurfaces for which it may be difficult to find a physics-based analytical description.

An elegant and promising approach to fit a PES is provided by artificial neural networks[5] (NNs) which are popular Machine Learning (ML) algorithms. Ideally, the resulting PES should be accurate, rapid to evaluate, analytically differentiable, scalable, and applicable to bond-breaking/bond-formation problems[6,7]. Moreover, it should be transferable between different systems and configurations. Unfortunately, a ML algorithm that would fulfill all these requirements does not exist yet[7,8]. Several NN architectures have been developed, such as the High-Dimensional NN (HDNN) introduced by Behler and Parinello in 2007[9], the Deep Tensor NN[10] (DTNN) and the SchNet architecture[11] both developed by Schütt *et al.*. While the HDNNs are perfectly optimized and efficient to describe the configuration space of a given molecular system, they are not transferable across the chemical space[7]. Inversely, the DTNN or the SchNet are well defined to describe the chemical space but usually these methods do not reach high accuracy in the configurations space[8] – *i.e.* within the chemical accuracy (< 1 kcal/mol or 0.05 eV). Obtaining a high accuracy in both spaces simultaneously remains an active research topic[8].

Another difficulty in applying NNs to the computations of PES is the choice of suitable molecular descriptors. In our study, we have developed an original protocol to prepare an optimal set of descriptors and to build an efficient NN. Using a principal component analysis[12] (PCA) a single feedforward neural network (FNN) architecture[5,6,12,13] can be employed to compute a reactive high-dimensional PES below the chemical accuracy. PCA has been used in the prediction of physical properties, like exciton transfer times[14] for example. In this context, PCA improves significantly the prediction accuracy[14,15]. Preprocessing NNs using PCA was investigated in the context

of diffraction tomography[16]. We mention also that PCA have been employed to select the most relevant degrees of freedom along a reaction path[17,18]. However, PCA has never been exploited in the fitting of high-dimensional PES. We show that this method i) improves the accuracy of the NN to learn and predict a PES and ii) may be used to reduce the size of the input space. This dimensionality reduction is essential to overcome the "curse of dimensionality"[19]. In fact, a large input space induces a large volume where the sampling density may be small. In other words, the data are described in an empty space and are strongly scattered. This can lead to a poor fitting of the PES if the training set is not large enough. Because computing points of the PESs for the training set is a demanding task, it is therefore essential to optimize the dimension of the input space. PCA allows us to select the most relevant principal components in order to maximize the sampling density[12].

We demonstrate the efficiency of this novel approach with the development of a FNN for fitting a high-dimensional PES which describes the tautomerism reaction occurring in the acetone molecule. When the PCA conditioning is used, a fitted PES below the chemical accuracy is obtained. Furthermore, our results suggest that a significant improvement of the NN thanks to the PCA protocol is generally achieved. PCA could therefore be employed to improve other NN architectures such as the HDNN, DTNN and SchNet ones.

The NN architecture and the learning process that we have implemented are given in sections II A and II B. In section II C, we present in details the PCA algorithm. Finally, in section III we discuss our sampling of the PES for the tautomerism reaction in the acetone molecule and we report the results of the PCA conditioning for two types of molecular descriptors defined by the Coulomb matrix: its eigenvalues and its off-diagonal elements.

II. METHODS

A. Feedforward Neural Network Architecture

In this work, we use a feedforward neural network[5,6,12,13] composed of three layers (see Figure 1). The first one is formed by a set of descriptors x_i, the second one is a hidden layer composed by N neurons and the last one is an output layer consisting of a single value energy. The energy computed by the FNN is given (in matrix form) by:

$$E = \mathbf{w_2} \cdot s\left(\mathbf{W_1} \cdot \mathbf{x} + \mathbf{b_1}\right) + b_2 \,, \tag{1}$$

where $\mathbf{x}$ represents the input vector, $\mathbf{W_1}$ the matrix of the weights between the input layer and the hidden layer, and $\mathbf{w_2}$ the vector of the weights between the unique neuron of the output layer and the neurons of the hidden layer. Finally, $\mathbf{b_1}$ and b_2 are, respectively, the bias vector of the hidden layer and the bias of the output layer. The activation function $s(x)$ is chosen to be a sigmoid function:

$$s(x) = \frac{1}{1+e^{-x}} \,. \tag{2}$$

In the following, a neural network is denoted as X-Ns-E, where X is the number of descriptors, N the number of neurons in the hidden layer and E is the energy given in the output layer.

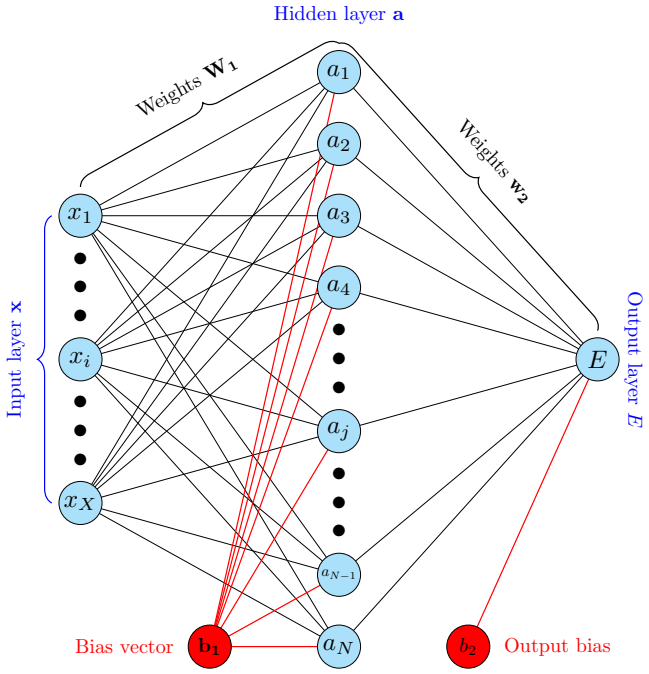


FIG. 1. Architecture of the feedforward neural network (FNN) used in this work. The input layer is composed of X descriptors represented by a vector $\mathbf{x}$. The hidden layer is described by a vector $\mathbf{a}$ where the N components are neurons activated by a sigmoid function $s(x)$. The output layer is a unique neuron E, which yields the energy of a particular geometry on the PES.

B. Feedforward Neural Network Training

The FNNs are trained to reproduce a set of geometries and energies, denoted as :

$$\mathscr{E} = \left\{ \bar{\mathbf{R}}(n), E_{\text{ref}}(n) \right\}_{n=1}^{N_{\text{ref}}} \tag{3}$$

where n is an element of the training examples. Each element n is characterized by a molecular geometry $\bar{\mathbf{R}}(n)$, which is associated with an energy $E_{\text{ref}}(n)$ (the computations of $E_{\text{ref}}(n)$ are described in section III A). In the training process, we have scaled these energies such that they are restricted to the interval $[0;1]$:

$$\tilde{E}_{\text{ref}}(n) = \frac{E_{\text{ref}}(n) - E_{\text{ref}}^{\min}}{E_{\text{ref}}^{\max} - E_{\text{ref}}^{\min}} \tag{4}$$

The training consists in the minimization of a cost function. We have chosen a batch learning model[12] in which the entire training set is fitted simultaneously through the quadratic error between the predicted energies $E(n)$ and the reference energies $\tilde{E}_{\text{ref}}(n)$ averaged on the training set:

$$\Gamma = \frac{1}{2N_{\text{ref}}} \sum_{n=1}^{N_{\text{ref}}} \left(E(n) - \tilde{E}_{\text{ref}}(n) \right)^2 . \tag{5}$$

In the above equation, Γ is the cost function, which depends on the weights $\mathbf{W_1}$ and $\mathbf{w_2}$, as well as on the biases $\mathbf{b_1}$ and b_2,

$$\Gamma\left(\mathbf{W_1}, \mathbf{w_2}, \mathbf{b_1}, b_2\right) = \frac{1}{2N_{\text{ref}}} \sum_{n=1}^{N_{\text{ref}}} \left(\left(\mathbf{w_2} \cdot s\left(\mathbf{W_1} \cdot \mathbf{x}(n) + \mathbf{b_1} \right) + b_2 \right) - \tilde{E}_{\text{ref}}(n) \right)^2 . \tag{6}$$

The cost function was minimized with respect to these parameters using the L-BFGS[20] optimizer selected in the library `NLOpt`[21] patched on `Python`. To initialize the learning process, we have started with a randomly determined set of initial parameters $\{\mathbf{W_1}, \mathbf{w_2}, \mathbf{b_1}, b_2\}$. They were selected through a central normal distribution function strongly squeezed around the average (standard deviation $\sigma \sim 0.01$).

In order to avoid overfitting, the parameters are checked using a test set whose size is half the one of the training set (the procedure to build the two sets is described in section III A). Furthermore, the quality evaluation of the learning and of the accuracy of the FNNs have been done using the root mean square error (RMSE) of the energies on the training and test sets, respectively:

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{ref}}} \sum_{n=1}^{N_{\text{ref}}} \left(E(n) - \tilde{E}_{\text{ref}}(n) \right)^2} . \tag{7}$$

C. Principal Component Analysis

One of the bottlenecks of machine learning algorithms lies in the representation of the input data $\mathbf{x}(n)$[22]. We propose here to use a principal component analysis[12] (PCA) to prepare the input vectors – *i.e.* the descriptors of molecular structures. While PCA is largely employed in image

processing and other domains[23,24], to our knowledge it has never been used in the computation of PESs. As shown below, this conditioning improves substantially the learning process.

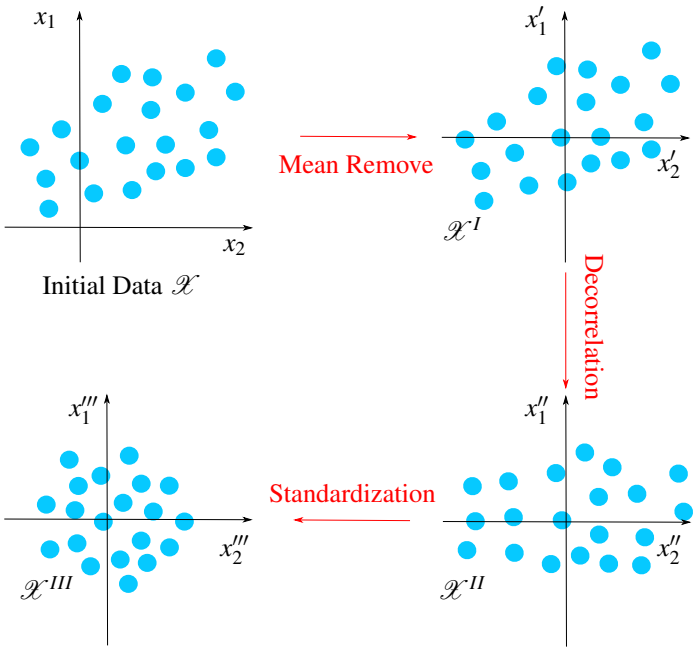


FIG. 2. Illustration of the three steps of the PCA method[12].

The PCA method is based on a statistical treatment of the data and particularly on the use of the covariance matrix of the input vectors $\mathbf{x}(n)$ defined from the training set $\mathscr{E}$.

We recall that $\mathbf{x}(n)$ is a vector of the input space describing a set of descriptors $x_{i,n}$ for a molecular structure $\bar{\mathbf{R}}(n)$ of the training set:

$$\mathbf{x}(n) = (x_{1,n}, x_{2,n}, x_{3,n}, \ldots, x_{i,n}, \ldots, x_{X,n})^T \ , \tag{8}$$

where X is the number of descriptors in the input layer (Figure 1).

1. The first step of the PCA consists in setting the mean value of each descriptor to zero. We therefore define a new set of descriptors as:

$$\mathbf{x}'(n) = \mathbf{x}(n) - \tilde{\mathbf{x}} \ , \tag{9}$$

where $\tilde{\mathbf{x}}$ represents the mean vector of the descriptors. This leads to a new set of vectors denoted as $\mathscr{X}^I = \{\mathbf{x}'(n)\}_{n=1}^{N_{\text{ref}}}$ (see Figure 2).

2. The second step is the calculation of the covariance matrix **D**. This matrix is built as the average, over the entire set $\mathscr{X}^I$, of the outer products of vectors $\mathbf{x}'(n)$ with themselves:

$$\mathbf{D} = \mathbb{E}\left[\mathbf{x}'\mathbf{x}'^T\right] . \tag{10}$$

The off-diagonal elements D_{ij} are the covariances of the descriptors $x'_{i,n}$ and $x'_{j,n}$. The off-diagonal elements of **D** are non zero due to linear correlations between the descriptors. The aim of the PCA is to remove these correlations. This is achieved by diagonalizing the matrix **D**:

$$\mathbf{D}\mathbf{q}_i = \lambda_i \mathbf{q}_i . \tag{11}$$

The eigenvectors matrix **Q**:

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \ldots, \mathbf{q}_i, \ldots, \mathbf{q}_X] , \tag{12}$$

is then used to define a new space, called feature space. The vectors $\mathbf{x}'(n)$ of the input space can then be represented without correlation (see Figure 2):

$$\mathbf{x}''(n) = \mathbf{Q}^{-1}\mathbf{x}'(n) . \tag{13}$$

If some eigenvalues λ_i are close to zero or small compared to the other ones, the corresponding descriptors do not contain useful information. They can be excluded from the feature space, thus reducing its size. The eigenvectors matrix **Q** can be rewritten as:

$$\tilde{\mathbf{Q}} = [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \ldots, \mathbf{q}_i, \ldots, \mathbf{q}_{X'}] , \tag{14}$$

where X' is the number of relevant eigenvalues (usually X' is smaller than X).

3. The final step, called standardization (see Figure 2), ensures that all descriptors have equal importance in the learning process. In other words, no descriptor is more important than another. It consists in fixing the variance of each descriptor to unity:

$$x'''_{i,n} = \frac{x''_{i,n}}{\sqrt{\lambda_i}} . \tag{15}$$

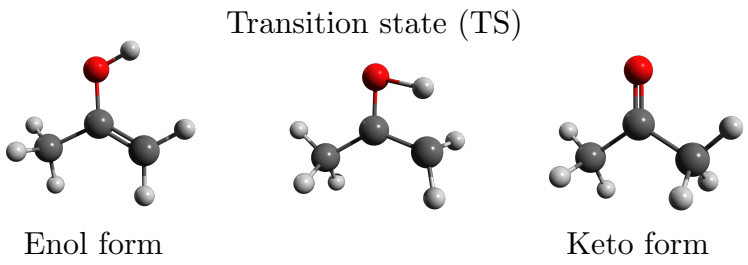
As a result, the weights learn at the same rate[25], which makes the training of the NNs faster.

III. RESULTS AND DISCUSSION

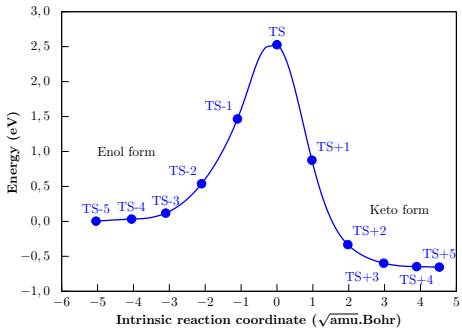
A. Training And Test Sets

The geometries and energies used in the training and test sets were obtained as follows: the transition state (TS) of the reaction was determined with the Climbing-Image Nudged Elastic Band (CI-NEB) method introduced by Henkelman *et al.*[26,27] as implemented in VASP 4.6[28,29]. We used DFT with the projector augmented-wave[30] (PAW) method and generalized gradient approximation (GGA) in Perdew-Wang[31] (PW91) parametrization. It should be mentioned that the VASP is a periodic code based on the use of plane waves. We have therefore selected a large cubic unit cell (lattice parameter $a = 10$ Å) with the unique Γ-point to consider an isolated molecule. Eight images in the elastic band were used to reach a good convergence of the forces. Note that the Quick-Min algorithm was applied to perform the CI-NEB relaxation of the band. The intrinsic reaction coordinate (IRC) is computed from this TS at the DFT/B3LYP/6-31G level (see below). To ensure, the TS found in the CI-NEB calculations are also that at the DFT/B3LYP/6-31G level, we have performed a frequency analysis from the transition state geometry at the PAW/PW91/PW and at the DFT/B3LYP/6-31G levels. An unique imaginary frequency was obtained, this ensure that the latter is caracteristic of a saddle point in both DFT levels. Furthermore, as shown in Table I, our results agree well with previous works[32,33]. We employed a similar approach to investigate the tautomerism reaction in acetylacetone (see Ref.[34] and Ref.[35] for further details).

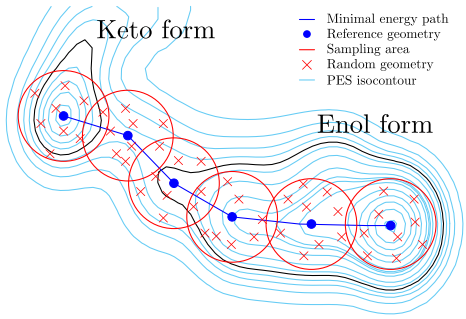
To construct the training and test sets, we performed a steepest descent relaxation along both IRC directions from the transition state geometry to recover the full minimum energy path (MEP) at the DFT/B3LYP/6-31G level of theory. After this, we selected 11 geometries along the obtained MEP and a normal modes analysis was done for each of them. The PES is then sampled by computing the geometries – as linear combinations of normal modes – and energies at randomly chosen points around these reference geometries (see Figure 3). From this set of structures, we selected a subset of geometries with energies less than or equal to 5.0 eV relative to the most stable structure – *i.e.* the keto form. These energies are above the transition state located at 3.18 eV above the keto form and thus describe properly the PES of the reaction. All the energies were computed at the DFT/B3LYP/6-31G level of theory using GAMESS-US[36]. Finally, 1500 and 850 points approximately were taken around each of the 11 reference points along the MEP to build the training and test sets, respectively.



(a)



(b)



(c)

FIG. 3. (a) Geometries of the enol and keto forms and of the transition state. (b) Minimum energy path (MEP) along the intrinsic reaction coordinate (IRC) obtained at the DFT/B3LYP/6-31G level – The IRC is given with mass-weighted coordinates in atomic unit. (c) Schematic illustration of our sampling along the MEP to build the training and test sets.

	Imaginary frequency (cm^{-1})	**Energy $\Delta E^{\ddagger}$ (eV)**
PAW/PW91/PW	*i* 1958	2.57 (2.47[33])
DFT/B3LYP/6-31G	*i* 2188	3.18 (2.82[32])

TABLE I. Imaginary frequencies and activation energies $\Delta E^{\ddagger}$ obtained by CI-NEB (PAW/PW91/PW) and IRC path (DFT/B3LYP/6-31G). The values in brackets come from the literature: Cucinotta *et al.*[33] (UPP/BLYP/PW) and Kaweetirawatt *et al.*[32] (DFT/B3LYP/6-31G*).

B. Choice of the descriptors

The choice of suitable descriptors is one of the bottlenecks of any machine learning algorithm development. As noted in a recent review[22], the search of descriptors has fed the literature for the

last twenty years. Nevertheless, a consensus is reached that the descriptors must respect two principal features. First, an unambiguous correspondence must exist between the descriptors and the molecular structures. In other words, each unique structure must have a unique set of descriptors. Second, the descriptors must be invariant to translation and rotation of the molecular structures and to permutation of any atoms[8,22].

A molecular system is uniquely defined by its nuclear charges Z_I and atomic positions $\mathbf{R}_I$[37,38]. The information contained in $\{Z_I, \mathbf{R}_I\}$ is sufficient to write down the non-relativistic Hamiltonian which corresponds to the energy of the system for a given geometry. The nuclear charges and the atomic positions can therefore be employed to uniquely map a given geometry onto the corresponding energy as: $f : \{Z_I, \mathbf{R}_I\} \mapsto E$. Based on this relationship, we used the Coulomb matrix to define molecular descriptors[37,38]. The Coulomb matrix is given by:

$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} , & \text{for} \quad I = J \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} , & \text{for} \quad I \neq J \end{cases} \tag{16}$$

The dimension of this matrix is equal to $N_{\text{at}} \times N_{\text{at}}$ for a system composed of N_{at} atoms. The diagonal elements represent a polynomial fit of the atomic energies to the nuclear charge Z_I and the off-diagonal elements take the form of the Coulomb repulsive potential between the nuclei[38,39]. Through these considerations the Coulomb matrix provides a global descriptor for the molecular structure. However, as reported in Ref.[22], the Coulomb matrix has not been used so far in the construction of PESs. We show below that this matrix combined with PCA can be efficiently employed to build an accurate FNN-PES.

1. Eigenvalues of the Coulomb matrix as descriptors

The eigenvalues of the Coulomb matrix are invariant with respect to rotation, translation and permutation. They were successfully used to predict molecular properties through the chemical space[37,38,40]. We first employed them as descriptors to illustrate the importance of using the PCA conditioning. However, as shown below, these descriptors alone do not lead to a sufficiently accurate PES to be used in practice. The main issue is that for a system with N_{at} atoms, the N_{at} eigenvalues represent only a subset of the information in the $3 \cdot N_{\text{at}} - 6$ dimensional conformation space[22] and this loss of information does not allow to correctly distinguish between different geometries.

FNNs	RMSE training set (eV)	RMSE test set (eV)
Without PCA		
10-44*s*-E	0.400	0.370
With PCA		
9-42*s*-E	0.320	0.320

TABLE II. Lowest RMSEs of the training and test sets without and with PCA using the eigenvalues of the Coulomb matrix as descriptors. (Training set: 17021 points ; Test set: 9257 points)

The blue line in Figure 4 represents the RMSE of the training set at the end of the learning process relative to the number of neurons N in the hidden layer using as descriptors the 10 eigenvalues of the Coulomb matrix (for the acetone C_3H_6O, $N_{at} = 10$). We can observe that the value of the RMSE strongly oscillates with respect to the number of neurons in the hidden layer. The lowest RMSE value was obtained for the 10-44*s*-E architecture (44 hidden neurons). The other architectures lead to RMSE values larger than 0.40 eV (see Table II and Figure 4). The learning process is therefore poorly achieved.

We then applied the PCA conditioning on the training set. One eigenvalue of the **D** matrix is always null, indicating that one of the N_{at} descriptors can be removed. This observation is a consequence of the trace invariance of the Coulomb matrix associated to molecular structure of the same composition. In fact, the sum of the eigenvalues is equal to the trace of the matrix, thus there are only $N_{at} - 1$ independent variables. Hence, the pertinent information can be just included in $N_{at} - 1$ new descriptors leading to a reduction of the input space.

The RMSEs obtained after applying the PCA on the training and test sets are shown in Figure 4. The RMSEs decrease smoothly with increasing number of neurons. Furthermore, the PCA conditioning leads to a much smaller RMSE for a given number of neurons. The PCA therefore improves significantly the learning process. It is worth noting that the RMSE associated with the test set is constant above 42 neurons and that the difference between the RMSE of the training and the test sets increases. This is characteristic of overfitting – *i.e.* the machine begins to be over parametrized. Hence, the best architecture is obtained for 42 neurons in the hidden layer.

Figure 4b represents the predicted energies according to the reference energies of the test set. Ideally, the blue points should be aligned along the red linear curve. Unfortunately, the points are strongly dispersed along this line revealing a large set of molecular structures with wrong energies,

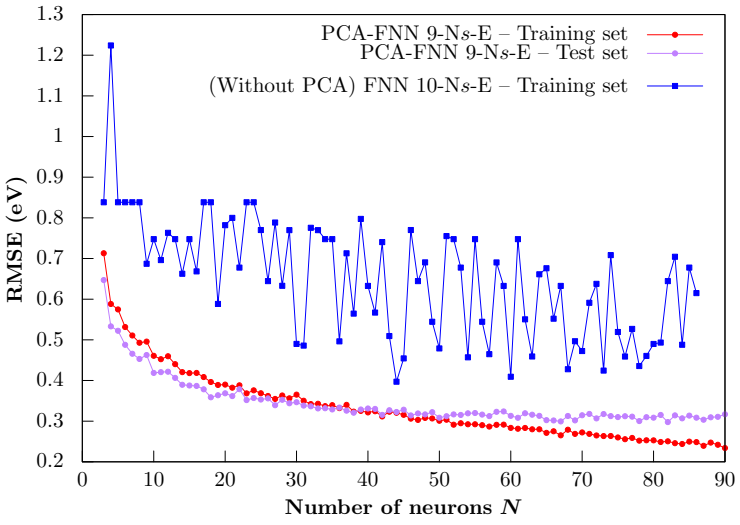
particularly for the energies higher than 3.0 eV – *i.e.* the activation barrier of the tautomerism process. Hence, despite a substantial improvement of the learning process through the PCA method, the PES is still not accurate enough. These results confirm that the eigenvalues of the Coulomb matrix cannot be used to build a reactive PES, even if PCA is applied.

2. Off-diagonal elements of the Coulomb matrix as descriptors

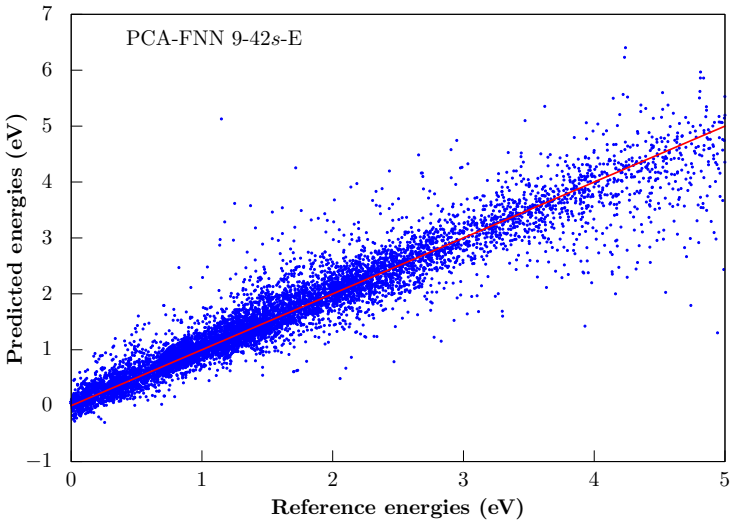
In what follows, we show that the Coulomb matrix can be exploited to build a PES if one considers the upper off-diagonal elements (M_{IJ} with $I > J$) as descriptors. In fact, these elements represent the full connected molecular graph through the inverse distances between atoms and include the nature of the nuclei in interaction. Hence, these descriptors can be used to describe any structures on the PES. Moreover, the number of these descriptors is larger or equal to the degrees of freedom of the PES. In the case of the acetone molecule there are 45 elements, thus each geometry can be distinguished[6]. To our knowledge, this is the first time the off-diagonal elements of the Coulomb matrix have been used as descriptors to fit a PES.

Figure 5a represents the RMSEs of the training and test sets at the end of the learning process relative to the number of neurons N in the hidden layer. First of all, we can observe that the RMSEs are in general smaller when the off-diagonal elements are used as descriptors compared to the eigenvalues of the Coulomb matrix. Nevertheless, when the training data are not conditioned by the PCA method, the RMSE reported as a function of the number of neurons presents some oscillations. However, their amplitudes are less important compared with the use of the eigenvalues. As in the latter case, we observe that the behavior of the RMSEs is smoother and that their values are in general smaller when PCA is applied compared to no PCA conditioning. For example, the quality of the PES is improved by a factor of ten with the optimal number of neurons for each machine using 45 descriptors (see Table III). In fact, the best trained FNNs provides a RMSE of the test set below the spectroscopic accuracy (~ 0.01 eV). Figure 5b illustrates the accuracy of our FNN obtained through the off-diagonal elements of the Coulomb matrix and the PCA protocol: the predicted energies of the test set are correctly aligned along the red linear curve, over the whole energy range. We note that beyond 38 neurons overfitting is observed: the difference between the RMSEs of the training and the test sets increases (Figure 5a). 38 neurons in the hidden layer is therefore an optimal choice.

We now show that PCA allows us to reduce the size of the input space by removing the de-



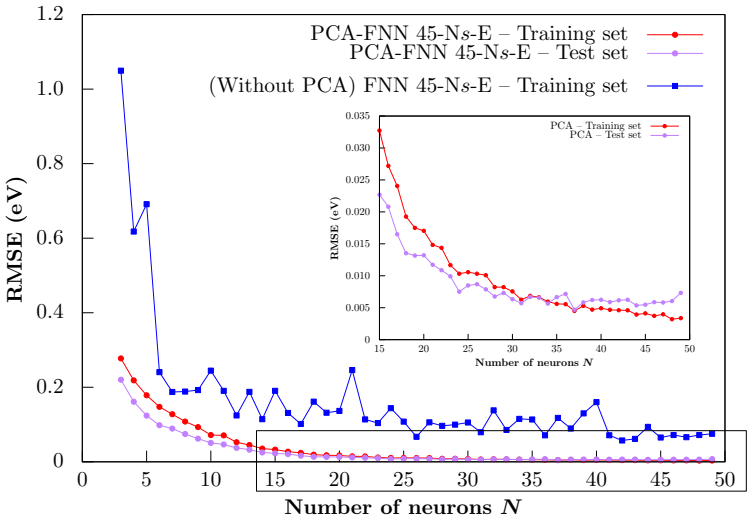
(a)



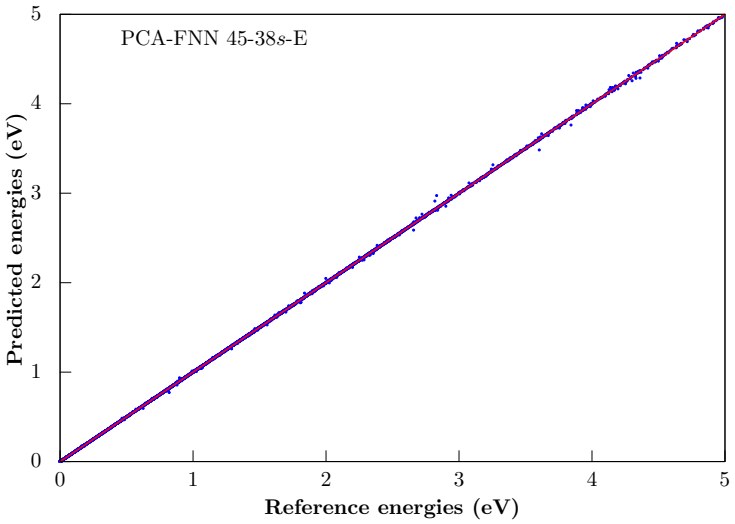
(b)

FIG. 4. (a) RMSEs of the training and test sets relative to the number of neurons N. The blue line shows the RMSEs obtained for the 10 eigenvalues of the Coulomb matrix without PCA. The red and purple lines are the RMSEs of training and test sets obtained with the PCA method. (b) Representation of the predicted energies obtained with our PCA-FNN 9-42s-E as functions of the reference energies computed at DFT/B3LYP/6-31G level for the test set. (Training set: 17021 points ; Test set: 9257 points)

scriptors for which the eigenvalues of the **D** matrix are comparatively small to the other ones. We demonstrate as well that this reduction does not lead to a loss of accuracy of the PES. Figure 6a



(a)



(b)

FIG. 5. (a) RMSEs of the training and test sets relative to the number of neurons N. The blue curve shows the RMSEs obtained for the 45 off-diagonal elements of the Coulomb matrix without PCA. The red and purple lines are the RMSEs of training and test sets obtained with the PCA method. (b) Representation of the predicted energies obtained with our PCA-FNN 45-38s-E as functions of the reference energies computed at DFT/B3LYP/6-31G level. (Training set: 17021 points ; Test set: 9257 points)

represents the RMSE of the test set (9257 points) relative to the number of neurons N in the hidden layer for different number of descriptors. We first discuss the case of 9 descriptors and compare

with the results obtained with the eigenvalues of the Coulomb matrix: the RMSE of the test set is 0.211 eV for the optimal neural network 9-48s-E (Table III). The RMSE has therefore the same magnitude than in the case with 9 eigenvalues. Hence, this supports our conclusions that more descriptors are needed to distinguish between the different geometries. We then investigate the case of 24 descriptors. The current feature space is therefore described by the same dimension as the physical PES ($3 \cdot N_{\text{at}} - 6$). This was done by removing the independent components i for which the eigenvalues λ_i are less than 10^{-4}. In this case, we see that the RMSE decreases according to the number of neurons without oscillation and reaches the value of 0.030 eV beyond 40 neurons (Table III). This RMSE is smaller than that obtained in the case without PCA. However, the RMSE is higher than the chemical accuracy. Therefore, to recover the same accuracy as with 45 descriptors, one needs to increase the number of principal components.

We have selected the 26 highest principal components. The green line in Figure 6a shows that in this case the lowest RMSE is found to be 0.010 eV (which equals the chemical accuracy). The quality of the fit is thus greatly improved. Now, if we select the 30 highest eigenvalues λ_i, we observe that the behavior of the RMSE relative to the number of neurons recovers the behavior obtained for 45 descriptors with PCA. In this case, a similar accuracy is reached for the same number of neurons into the hidden layer, but with less descriptors (Table III).

The reduction of the number of descriptors can be quantified by the amount of variance accounted for, or in other words, by taking the percentage of remaining eigenvalues with respect to the sum of all eigenvalues. In all cases considered above, the percentage of variance accounted for is more than 99%.

Moreover, note that the reduction of the number of descriptors allows us to use less free parameters in the FNN – their number is around O(XN), where X is the number of descriptors and N the number of neurons. For example, for 45 descriptors and 38 hidden neurons we have 1710 parameters, while with 30 descriptors we have 1140. This is particularly relevant since it was shown that a smaller number of descriptors leads to a better generalisation of the fit[12]: the error estimate on the test set is bounded by the ratio of the total number of free parameters to the number of training data points[12]. To illustrate the impact of the size of the feature space on the "curse of dimensionality" we show in Figure 6b the RMSE of the test set (9 257 points) obtained for different sizes of training set (20%, 60% and 100% – where 100% = 17 021 points). As shown before, when 24 descriptors are used for the feature space the RMSE is rather large. This is true for any size of the training set. When 45 descriptors are employed, the chemical accuracy is only reached for a large

FNNs	RMSE training set (eV)	RMSE test set (eV)
Without PCA		
45-45*s*-E	0.070	0.050
With PCA		
45-38*s*-E	0.005	0.007
9-48*s*-E	0.256	0.211
24-46*s*-E	0.034	0.027
26-45*s*-E	0.012	0.010
30-38*s*-E	0.007	0.007

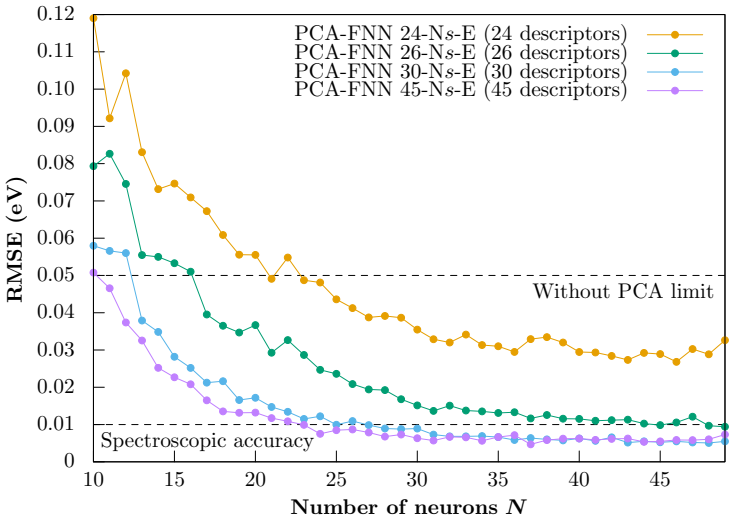
TABLE III. Lowest RMSEs of the training and test sets without and with PCA for the 45 off-diagonal elements of the Coulomb matrices. The three last rows provide the lowest RMSEs for different feature spaces obtained with PCA and 24, 26 and 30 descriptors, respectively. (Training set: 17021 points ; Test set: 9257 points)

training set (i.e. 60% case and above) owing to the small density of points in this large feature space. When the latter is reduced to 30 descriptors, the RMSE is the lowest for any size of the training set. Comparing the results with 45 and 30 descriptors, a smaller training set in the latter case can therefore be used while providing a similar accuracy. This is a relevant result because the computation of the training data points represents the most costly part of the development of PESs. We mention here that the cost of the PCA step is negligible compared with that of the NN training. As an example, for a 30-38s-E NN with 17021 training points, the training part takes about 1h while the PCA step takes 12sec (i.e. less than 0.4%). Moveover, our calculations suggest that the PCA cost scales approximately linearly with the size of the training set.

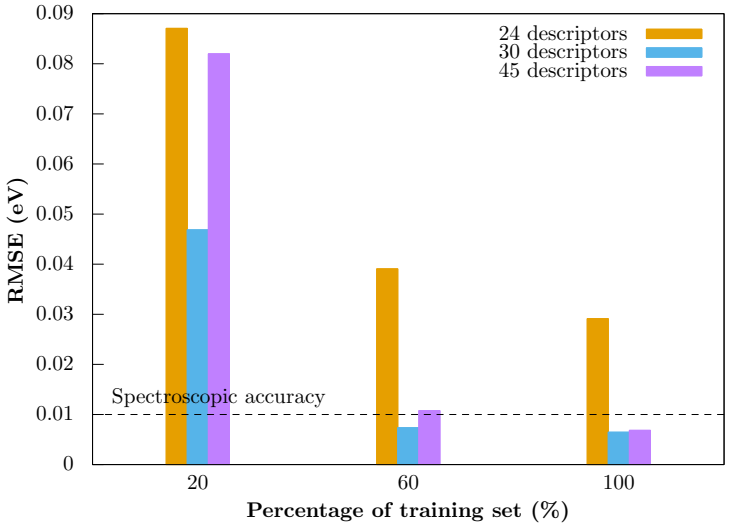
IV. CONCLUSION

In this work we have demonstrated the efficiency of a principal component analysis protocol to fit a high-dimensional PES. Our results demonstrate that this approach is a powerful tool to improve the learning and predicting of the NN and to reduce the size of the input space through a new feature space.

Both advantages were illustrated with two kinds of molecular descriptors: the eigenvalues of the



(a)



(b)

FIG. 6. (a) RMSEs of the test set according to the number of neurons N for different feature spaces obtained by PCA (Table III). (Training set: 17021 points ; Test set: 9257 points) (b) RMSEs of the test set according to the size of the training set for different sizes of feature space (Test set: 9257 points).

Coulomb matrix and its off-diagonal elements. We showed that in both cases the PCA improves the representation of the molecular structures encoded and hence the quality of the fit of the PES. Furthermore, our results indicate that the off-diagonal elements of the Coulomb matrix provide suitable descriptors to obtain fits of high-dimensional PESs below the chemical accuracy.

The efficiency of our approach was illustrated on the example of a feedforward neural network with a single hidden layer for fitting a high-dimensional PES describing the tautomerism reaction of the acetone molecule. However, we think that this novel method can be used for different systems and other neural network architectures. For example, a neural network using two hidden square unit augmented layers and 25 000 training geometries was recently employed to describe the hydrogen transfer between two oxygen atoms in the malonaldehyde[7]. A RMSE of 0.021 eV was thus obtained. In our work, we have achieved a lower RMSE with a simpler neural network architecture and fewer training points. Furthermore, HDNN is an elegant and efficient method to fit a PES but is currently limited to systems containing about three to four chemical elements, owing to the use of too many so-called symmetry functions. Using our approach, it should be possible to reduce the number of symmetry functions while maintaining the performance of the algorithm. Such achievement would expand the current HDNNs to other chemical environment.

V. ACKNOWLEDGEMENT

N.S. acknowledges the financial support from the LabEx MiChem part of French state funds managed by the ANR within the 'Investissements d'Avenir' program under reference ANR-11-IDEX-0004-02.

VI. DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]L. Raff, M. Malshe, M. Hagan, D. Doughan, M. Rockley, and R. Komanduri, "*Ab Initio* Potential Energy Surfaces for Complex, Multichannel Systems Using Modified Novelty Sampling and Feedforward Neural Networks," J. Chem. Phys. **122**, 84104 (**2005**).

[2]D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods* (Cambridge University Press, **2009**).

[3]P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas," Phys. Rev. **136**, 864 (**1964**).

[4]R. Car and M. Parrinello, "Unified Approach for Molecular Dynamics and Density Functional Theory," Phys. Rev. Lett. **55**, 2471 (**1985**).

[5]W. McCulloch and W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," Bull. Math. Biol. **5**, 115 (**1943**).

[6]J. Behler, "Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations," Phys. Chem. Chem. Phys. **13**, 17930–17955 (**2011**).

[7]O. T. Unke and M. Meuwly, "A Reactive, Scalable, and Transferable Model for Molecular Energies From a Neural Network Approach Based on Local Information," J. Chem. Phys. **148**, 241708 (**2018**).

[8]F. Noé, A. Tkatchenko, K. Müller, and C. Clementi, "Machine Learning for Molecular Simulation," Ann. Rev. Phys. Chem. **71**, (in press) (**2020**).

[9]J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential Energy Surfaces," Phys. Rev. Lett. **98**, 146401 (**2007**).

[10]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. Müller, and A. Tkatchenko, "Quantum-Chemical Insights From Deep Tensor Neural Networks," Nat. Commun. **8**, 13890 (**2017**).

[11]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet – A Deep Learning Architecture for Molecules and Materials," J. Chem. Phys. **148**, 241722 (**2018**).

[12]S. Haykin, *Neural Networks and Learning Machines, Third Edition* (Pearson Education Inc, **2009**).

[13]C. Handley and P. Popelier, "Potential Energy Surfaces Fitted by Artificial Neural Networks," J. Phys. Chem. A **114**, 3371 (**2010**).

[14]F. Häse, S. Valleau, E. Pyzer-Knapp, and A. Aspuru-Guzik, "Machine Learning Exciton Dynamics," Chem. Sci. **7**, 5139 (**2016**).

[15]F. Häse, I. Fdez. Galván, A. Aspuru-Guzik, R. Lindh, and M. Vacher, "How Machine Learning Can Assist the Interpretation of Ab Initio Molecular Dynamics Simulations and Conceptual Understanding of Chemistry," Chem. Sci. **10**, 2298–2307 (**2019**).

[16]E. M. Azoff, "Neural Network Principal Components Preprocessing and Diffraction Tomography," Neural Comput Appl. **1**, 107–114 (**1993**).

[17]S. R. Hare, L. A. Bratholm, D. R. Glowacki, and B. K. Carpenter, "Low Dimensional Representations Along Intrinsic Reaction Coordinates and Molecular Dynamics Trajectories Using Interatomic Distance Matrices," Chem. Sci. **10**, 9954–9968 (**2019**).

[18]A. B. Birkholz and H. B. Schlegel, "Coordinate Reduction for Exploring Chemical Reaction Paths," Theor Chem Acc **131** (**2012**).

[19]R. Bellman, *Adaptive Control Processes: A Guided Tour* (Princeton University Press, **1961**).

[20]J. Nocedal, "Updating Quasi-Newton Matrices with Limited Storage," Math. Comput. **35**, 773 (**1980**).

[21]S. Johnson, "The NLOpt NonLinear-Optimization Package," .

[22]J. Behler, "Perspective: Machine Learning Potentials for Atomistic Simulations," J. Chem. Phys. **145**, 170901 (**2016**).

[23]M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, "Principal Component Analysis and Long Time Protein Dynamics," J. Phys. Chem. **100**, 2567–2572 (**1996**).

[24]I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," Philos. Trans. R. Soc. A-Math. Phys. Eng. Sci. **374**, 20150202 (**2016**).

[25]S. Haykin, Chap. Multilayer Perceptrons.

[26]G. Henkelman, B. Uberuaga, and H. Jònsson, "A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths," J. Chem. Phys. **113**, 9901 (**2000**).

[27]G. Henkelman and H. Jònsson, "Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points," J. Chem. Phys. **113**, 9978 (**2000**).

[28]G. Kresse and J. Hafner, "*Ab Initio* Molecular Dynamics for Liquid Metals," Phys. Rev. B: Condens. Matter Mater. Phys. **47**, 558 (**1993**).

[29]G. Kresse and J. Furthmüller, "Efficient Iterative Schemes for *Ab Initio* Total-Energy Calculations Using a Plane Wave Basis Set," Phys. Rev. B: Condens. Matter Mater. Phys. **54**, 11169 (**1996**).

[30]P. Blöchl, "Projector Augmented Wave Method," Phys. Rev. B: Condens. Matter Mater. Phys. **50**, 17953 (**1994**).

[31]J. Perdew, J. Chevary, S. Vosko, K. Jackson, M. Pederson, D. Singh, and C. Fiolhais, "Atoms, Molecules, Solids, and Surfaces : Applications of the Generalized Gradient Approximation for Exchange and Correlation," Phys. Rev. B: Condens. Matter Mater. Phys. **46**, 6671 (**1992**).

[32]T. Kaweetirawatt, T. Yamaguchi, T. Higashiyama, M. Sumimoto, and K. Hori, "Theoretical Study of Keto-Enol Tautomerism by Quantum Mechanical Calculations," J. Phys. Org. Chem. **25**, 1097 (**2012**).

[33]C. Cucinotta, A. Ruini, A. Catellani, and A. Stirling, "*Ab Initio* Molecular Dynamics Study of the Keto-Enol Tautomerism of Acetone in Solution," Chem. Phys. Chem. **7**, 1229 (**2006**).

[34]N. Capron, B. Casier, N. Sisourat, M. Piancastelli, M. Simon, and S. Carniato, "Probing Keto-Enol Tautomerism Using Photoelectron Spectroscopy," Phys. Chem. Chem. Phys. **17**, 19991 (**2015**).

[35]B. Casier, N. Sisourat, S. Carniato, and N. Capron, "Keto-Enol Tautomerism in Micro-Hydrated Acetylacetone: An Atoms in Molecules Study," Theor. Chem. Acc. **137**, 1 (**2018**).

[36]M. Schmidt, K. Baldridge, J. Boatz, S. Elbert, M. Gordon, J. Jensen, S. Koseki, N. Matsunaga, K. Nguyen, S. Su, T. Windus, M. Dupuis, and J. J. Montgomery, "General Atomic and Molecular Electronic Structure System," J. Comput. Chem. **14**, 1347 (**1993**).

[37]M. Rupp, A. Tkatchenko, K. Müller, and O. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning," Phys. Rev. Lett. **108**, 58301 (**2012**).

[38]G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. von Lilienfeld, and K. Müller, "Learning Invariant Representations of Molecules for Atomization Energy Prediction," in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Curran Associates, Inc., **2012**) p. 440.

[39]M. Rupp, "Machine Learning for Quantum Mechanics in a Nutshell," Int. J. Quantum Chem. **115**, 1058 (**2015**).

[40]G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K. Müller, and O. von Lilienfeld, "Machine Learning of Molecular Electronic Properties in Chemical Compound Space," New J. Phys. **15**, 95003 (**2013**).