PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2024, **26**, 4306

Fast and accurate excited states predictions: machine learning and diabatization†

Štěpán Sršeň, *[ab] O. Anatole von Lilienfeld [cde] and Petr Slavíček *[a]

Received 22nd November 2023,
Accepted 2nd January 2024

DOI: 10.1039/d3cp05685f

rsc.li/pccp

The efficiency of machine learning algorithms for electronically excited states is far behind ground-state applications. One of the underlying problems is the insufficient smoothness of the fitted potential energy surfaces and other properties in the vicinity of state crossings and conical intersections, which is a prerequisite for an efficient regression. Smooth surfaces can be obtained by switching to the diabatic basis. However, diabatization itself is still an outstanding problem. We overcome these limitations by solving both problems at once. We use a machine learning approach combining clustering and regression techniques to correct for the deficiencies of property-based diabatization which, in return, provides us with smooth surfaces that can be easily fitted. Our approach extends the applicability of property-based diabatization to multidimensional systems. We utilize the proposed diabatization scheme to achieve higher prediction accuracy for adiabatic states and we show its performance by reconstructing global potential energy surfaces of excited states of nitrosyl fluoride and formaldehyde. While the proposed methodology is independent of the specific property-based diabatization and regression algorithm, we show its performance for kernel ridge regression and a very simple diabatization based on transition multipoles. Compared to most other algorithms based on machine learning, our approach needs only a small amount of training data.

1 Introduction

Machine learning (ML) has been recently experiencing tremendous expansion in various fields of science and computational chemistry is not an exception.[1] The motivation for using ML approaches is the high computational cost of quantum chemical calculations. We usually know how to obtain accurate results; however, such calculations are often computationally intractable and we have to settle with less accurate methods. ML can help us to shift the balance in favour of accuracy. Unfortunately, the applications of ML methods to electronically excited states have not yet reached the level of accuracy as the more common problem of dealing with ground-state properties.[2] The fact that excited states are still an outstanding problem for ML is due to the high complexity of reference quantum calculations, high densities of states, and the fact that the predicted properties are not smooth in the vicinity of state crossings and conical intersections.[3] We tackle here the problem of the low smoothness of excited-state properties.

Eigenfunctions and eigenvalues of the electronic Hamiltonian, which we usually get from electronic structure calculations, correspond to the so-called adiabatic representation. The states are ordered by their electronic energy for each nuclear configuration, resulting in non-crossing potential energy surfaces (PESs). While adiabatic states might become degenerate, they never truly cross if they have the same multiplicity. Electronic energies and other properties are then highly curved and non-differentiable. Low smoothness of the adiabatic basis represents a major problem for ML regression. Using a smooth diabatic basis, which allows for state crossings, seems like a natural solution how to improve ML efficiency. The two representations are connected through a geometry-dependent unitary transformation. Unfortunately, finding the diabatic basis is an outstanding problem itself. While the adiabatic basis can be obtained from a diabatic basis simply by diagonalization, the inverse procedure is highly complex as the diabatic basis is not uniquely defined. Even state-of-the-art methods such as fitting-while-diabatizing[4–6] procedure usually require expert knowledge about the system and lots of manual work and expensive calculations. Dozens of various diabatization schemes based on

[a] *Department of Physical Chemistry, University of Chemistry and Technology, Technická 5, 162 28 Prague, Czech Republic. E-mail: stepan.srsen@vscht.cz, petr.slavicek@vscht.cz*
[b] *Institute of Theoretical Chemistry, Faculty of Chemistry, University of Vienna, Währinger Str. 17, 1090 Wien, Austria*
[c] *Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada*
[d] *Departments of Chemistry, Materials Science and Engineering, and Physics, University of Toronto, St. George Campus, Toronto, ON, Canada*
[e] *Machine Learning Group, Technische Universität Berlin and Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3cp05685f

nonadiabatic couplings (NACs) elimination, wavefunction smoothness, or properties smoothness have been proposed.[7–17] To date, diabatization has been mostly limited to low-dimensional systems or specific wavefunction-based methods. However, several works attempting to solve the problem of automatic data-driven determination of the diabatic basis emerged during the last few years, predominantly based on neural networks.[18–22] Very recently, an approach for the fitting of adiabatic energies of coupled surfaces avoiding diabatization by fitting coordinate-dependent coefficients of the characteristic polynomial of a potential matrix decomposition was suggested.[23]

While classical diabatization schemes are usually system-specific and very laborious, current ML-based approaches usually require lots of expensive training data and often a manual selection of reference geometries where adiabatic and diabatic bases coincide. We aim to combine the best of both worlds: we augment simple property-based diabatization schemes with an ML algorithm that corrects their deficiencies. As a result, we can obtain a smooth diabatic representation already with dozens or hundreds of samples. Note that our goal is not to compete with complex state-of-the-art diabatization schemes trained on huge samples in the accuracy of diabatic states. Instead, we utilize the proposed diabatization to improve the prediction accuracy in the adiabatic basis while using small training datasets. Therefore, smoothness is more important than for example exact locations of conical intersections, *etc.* Property-based diabatization is arguably the simplest category of diabatization techniques.[7] It uses pairwise properties of adiabatic states such as transition dipole moments to obtain diabatic states whose characters change as slowly as possible. Unfortunately, there are some problems connected with this category of diabatization methods, which prevent their widespread application to larger molecules with multiple electronic states involved. First, we need to select such properties that allow the discrimination of all the involved electronic states. Second, ordering/labels of the states are not consistent throughout the configuration space as each nuclear geometry is diabatized separately: we get a set of diabatic energies and couplings (off-diagonal elements) for a given geometry and we have to assign them to the global diabatic states. The third issue arises from random signs of the electronic wavefunctions, which lead to random signs of the pairwise properties and further to random signs of the diabatic couplings. While the latter two issues can be easily resolved manually by inspection in one or two dimensions, it is impossible for a general multi-dimensional system.

The so-called cluster-growing algorithm has been previously proposed to correct the signs of diabatic couplings obtained with a different diabatization method.[24,25] It uses a greedy ML-based approach and it gradually corrects the signs of neighbouring geometries, starting from a manually corrected initial cluster. While it proved to be useful for sign correction, it has not been used for the simultaneous correction of signs and state ordering, which is a significantly more complex problem. We identify two main problems connected with such applications: first, the manual correction of the initial cluster becomes cumbersome when dealing also with state permutations, especially for high-dimensional systems. Second, poor sign or state assignments can lead to a cascade of more wrong assignments as it is a greedy algorithm that makes only locally optimal choices at each stage. As we aim at smaller training samples, we can afford to overcome these limitations by employing a stochastic iterative procedure to reach convergence in our approach.

Within the proposed framework, we diabatize each geometry separately using property-based diabatization, and correct for inconsistent signs and ordering of diabatic states with the ML approach. The general idea of our approach is simple: properties in the diabatic basis should be smooth and smooth properties are easy to fit so we change the ordering and signs so that the properties are well-fitted with our ML model based on a combination of kernel ridge regression (KRR) and clustering. As a result, our methodology can extend the applicability of the whole category of property-based diabatization schemes to multidimensional systems with multiple states with as little as dozens of training samples. At the same time, we get an efficient way how to predict adiabatic energies, which can be obtained from the fitted diabatic states and couplings simply by diagonalization, and therefore save time on expensive *ab initio* calculations. While our ML algorithm can be in principle applied to any property-based diabatization, we propose here a series of simple diabatization methods based on transition multipole moments from the ground state as a byproduct. We also test the direct application of our ML algorithm without prior property-based diabatization, that is, testing whether ML prediction capabilities can be improved by simple reordering of adiabatic states. For example, recent research showed on the prediction of the energy gap between the highest occupied and the lowest unoccupied molecular orbital that prior classification can improve the smoothness of the fitted property and therefore ML performance.[26]

We focus here on the prediction of PESs but other properties can be predicted as well: atomic forces and approximate NACs can be directly obtained from the diabatic representation and other properties such as dipole moments can be fitted separately in the diabatic basis.[24,25] We show the performance of the proposed methodology by reconstructing global PESs of excited states of nitrosyl fluoride and formaldehyde in thermally reachable regions at 300 K as we aim mainly at the application in modeling electronic spectroscopies. Using these small molecules for testing purposes allows us to use overlaps between all the states of all the sampled geometries for analysis, visualization, and benchmarking.

2 Computational methods

2.1 Property-based diabatization

We coupled our ML algorithm with property-based diabatization as it in principle the most straightforward approach to diabatization. Moreover, we propose here a series of very simple property-based diabatization methods which are easy to

implement. Within the Born–Oppenheimer approximation, the eigenvectors of the electronic Hamiltonian are called electronically adiabatic states and the eigenvalues are called adiabatic PESs. The seam space of a conical intersection between two interacting states, formed by geometries with degenerate adiabatic energies, has $N_{\mathrm{int}} - 2$ dimensions with N_{int} being the number of internal coordinates.[27] It might seem that such a small subspace cannot play a significant role but the Born–Oppenheimer approximation breaks already for geometries in the vicinity of conical intersections and this is where radiationless transitions take place. When we aim to describe processes involving excited states, we usually have to go beyond the Born–Oppenheimer approximation. To do so, we have to calculate the probabilities of radiationless transitions usually expressed *via* NACs for states of the same spin multiplicity. However, these couplings are expensive to compute, often difficult to converge and exhibit singularities at conical intersection seams.

To get rid of the cuspidal ridges in PESs and other properties and singularities in NACs near conical intersections, we can switch to a different representation by applying a geometry-dependent unitary transformation matrix $\mathbf{T}(\mathbf{R})$:[28]

$$\Psi_i^{\mathrm{d}}(\mathbf{r};\mathbf{R}) = \sum_j T_{ji}(\mathbf{R})\Psi_j^{\mathrm{ad}}(\mathbf{r};\mathbf{R}) \quad (1)$$

$$\mathbf{U}(\mathbf{R}) = \mathbf{T}(\mathbf{R})^{\mathrm{T}}\mathbf{V}(\mathbf{R})\mathbf{T}(\mathbf{R}) \quad (2)$$

where $\Psi_j^{\mathrm{ad}}(\mathbf{r};\mathbf{R})$ are the original adiabatic wavefunctions, $\Psi_i^{\mathrm{d}}(\mathbf{r};\mathbf{R})$ are the transformed diabatic wavefunctions, $\mathbf{V}(\mathbf{R})$ is the diagonal matrix of adiabatic PESs and $\mathbf{U}(\mathbf{R})$ is the transformed potential energy matrix (PEM) which is not diagonal anymore. The so-called strict diabatic basis would be obtained by such a transformation which would completely remove NACs. However, Mead and Truhlar[29] showed in 1982 that the strictly diabatic electronic basis does not in general exist. Therefore, we have to settle with a basis that provides smooth elements of PEM and removes singularities in NACs. We call such a basis diabatic even though, strictly speaking, we should use the term pseudo-diabatic basis.

The diabatic basis is very convenient for ML applications as the diabatic PEM and also other properties are supposed to evolve smoothly with geometrical coordinates. At the same time, we can switch back to the adiabatic basis at any time simply by diagonalization. However, the non-existence of the strictly diabatic basis also means that the diabatic basis is not uniquely defined. Property-based diabatization schemes based on property unblending are the simplest and cheapest to apply. As diabatic wavefunctions are supposed to be smooth functions of geometry, we expect their properties to change smoothly as well. While enforcing global smoothness is a difficult problem, we can redefine the problem locally. Two crossing states become blended in the vicinity of a conical intersection and so do their properties. Property-unblending diabatization methods use this observation and make properties of the transformed diabatic states as different as possible which corresponds to the maximization of the following objective function:[7]

$$f = \sum_{i,j} |\langle\Psi_i^{\mathrm{d}}(\mathbf{r};\mathbf{R})|\hat{P}|\Psi_i^{\mathrm{d}}(\mathbf{r};\mathbf{R})\rangle - \langle\Psi_j^{\mathrm{d}}(\mathbf{r};\mathbf{R})|\hat{P}|\Psi_j^{\mathrm{d}}(\mathbf{r};\mathbf{R})\rangle|^2 \quad (3)$$

where $\hat{P}$ is the property operator. It has been shown that this maximization is equivalent to the maximization of the following objective function in the adiabatic basis:[30]

$$f(\mathbf{T}) = \sum_i \left| \sum_{j,k} T_{ji}(\mathbf{R})T_{ki}(\mathbf{R})\langle\Psi_j^{\mathrm{ad}}(\mathbf{r};\mathbf{R})|\hat{P}|\Psi_k^{\mathrm{ad}}(\mathbf{r};\mathbf{R})\rangle \right|^2 \quad (4)$$

Many different property-unblending methods have been proposed using different properties to differentiate the states.[11,13,30–33] It is important to note that the separation of matrix eigenvalues can be achieved by diagonalization.[7] Therefore, the matrix formed by the eigenvectors of the property matrix corresponding to the $\hat{P}$ operator in the adiabatic basis can be used for diabatization. Unfortunately, this procedure leads to the above-mentioned problems with inconsistent ordering of the diabatic states and random signs of diabatic couplings.

The methodology proposed in this paper can be in principle connected with an arbitrary property-based diabatization method to extend its applicability to multidimensional problems. Nevertheless, we also propose here a series of simple and pragmatic property-based diabatization methods. The reasoning behind our methods is similar to the dipole-quadrupole[30] (DQ) diabatization: we want to distinguish the electronic states based on their transition multipole moments. However, the DQ and similar methods require transition multipole moments between all pairs of states, which are not always easily available from electronic-structure calculations.[11,13,30–32] For example, the popular TDDFT method based on the linear-response theory does not usually even yield the full matrix of (transition) dipole moments. One has to usually perform a separate calculation for each electronic state, which is both computationally demanding and laborious. It is even more problematic for higher multipole moments. We instead propose to form the property matrix based on inner products between transition multipoles from the ground electronic state, which are usually readily available.

This way, we form a series of methods, which we call transition dipole (tD), transition dipole and quadrupole (tDQ), and transition dipole, quadrupole and octupole (tDQO) diabatization depending on the highest multipole included. The property matrix $\mathbf{P}$ is then formed according to the following formulas, respectively:

$$P_{ab}^{\mathrm{tD}} = \boldsymbol{\mu}_{0a}\cdot\boldsymbol{\mu}_{0b} \quad (5)$$

$$P_{ab}^{\mathrm{tDQ}} = \boldsymbol{\mu}_{0a}\cdot\boldsymbol{\mu}_{0b} + \omega_{\mathrm{Q}}\langle\mathbf{Q}_{0a},\mathbf{Q}_{0b}\rangle_{\mathrm{F}} \quad (6)$$

$$P_{ab}^{\mathrm{tDQO}} = \boldsymbol{\mu}_{0a}\cdot\boldsymbol{\mu}_{0b} + \omega_{\mathrm{Q}}\langle\mathbf{Q}_{0a},\mathbf{Q}_{0b}\rangle_{\mathrm{F}} + \omega_{\mathrm{O}}\langle\mathbf{O}_{0a},\mathbf{O}_{0b}\rangle_{\mathrm{F}} \quad (7)$$

where $\boldsymbol{\mu}_{0a}$, $\mathbf{Q}_{0a}$ and $\mathbf{O}_{0a}$ are the transition dipole, transition quadrupole and transition octupole moments, respectively, as implemented in the PySCF[34,35] code, version 2.0.1. $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ is the Frobenius inner product, that is, the sum over the element-wise product. The weights ω_{Q} and ω_{O} can be set by hand or

optimized within cross-validation or a similar procedure. However, for simplicity, we do not use here this flexibility and set all weights to 1. Also, we did not observe a significant improvement when tweaking the coefficients for our test molecules.

We do not claim these methods to be universal but they are pragmatic as they can be employed and tested very quickly. We can simply form the property matrix **P**, calculate the matrix of eigenvectors, and use it as the transformation matrix in eqn (2). The employment of these methods is reasonable as long as the ground is sufficiently separated from the other electronic states within the sampled configuration space.

2.2 ML-based reordering

Eventually, we want to correct the deficiencies of property-based diabatization but we start with a simpler problem: can we reorder the adiabatic energies for each geometry so that they form smoother surfaces than the original adiabatic PESs? We simply want to reorder the adiabatic electronic energies of each nuclear configuration so that they form new PESs that can cross where it is advantageous for learning. If the answer were positive, then we would be able to get better ML predictions of adiabatic energies without any underlying property-based diabatization. Also, such an algorithm can directly diabatize states of different symmetry since they cross without mixing, that is with zero NACs. Yet another motivation is the benchmark of our optimization procedure because we devised an alternative approach to solving this problem based on wavefunctions overlaps as described below, to which we can compare the results.

Direct optimization of the state ordering by the minimization of the prediction error is problematic as the variable state order introduces too much variability to the model, resulting in difficult optimization and overfitting problems. Overfitting might be reduced by introducing a regularization term penalizing the higher roughness/curvature of the predicted PESs. Nevertheless, we propose here a simpler clustering approach based on the expectation-maximization (EM) algorithm on which many common clustering algorithms such as *k*-means are based as well. By clustering, we refer here to the assignment of adiabatic energies of individual geometries to global states and their PESs. The main difference between our clustering and *k*-means is that we cluster the data by minimizing the prediction errors for each geometry instead of the distance to the centroid. Also, we impose the restriction that each adiabatic energy of a single geometry is assigned to a different global PES.

A simplified flowchart of the optimization procedure is depicted in Fig. 1. We start the optimization from an initial ordering/clusters corresponding to some PESs, that is, either original energy-ordered adiabatic states or randomly shuffled states. The order of states for individual geometries can be seen as model parameters and we can use the EM algorithm to optimize them. We fix the state clusters and set KRR model hyperparameters in the expectation step and we use these fixed state clusters to estimate new ordering for each geometry separately in the maximization step. The excited states of each geometry are iteratively reassigned to the clusters in the maximization step by training a KRR model for each cluster corresponding to a single PES with the fixed ordering and hyperparameters but without the geometry which is currently being assigned. A distance matrix for the left-out geometry is then formed by calculating the prediction errors for its states using all the cluster KRR models. So we have a distance matrix between the energies of a single molecule and the state clusters and we want to find the best assignment so that the total prediction error is minimized. This is a common linear sum assignment problem, also known as the minimum (here maximum) weight matching. We solved this matching problem by the modified Jonker–Volgenant algorithm[36] minimizing the mean squared error as implemented in the SciPy[37] python package. We repeat this process of fixing the clusters, setting hyperparameters and estimating new state orders geometry by geometry until the clusters do not change anymore. Note that we observed better convergence by updating the clusters after the assignment of each geometry, a modification also applicable to *k*-means.[38]

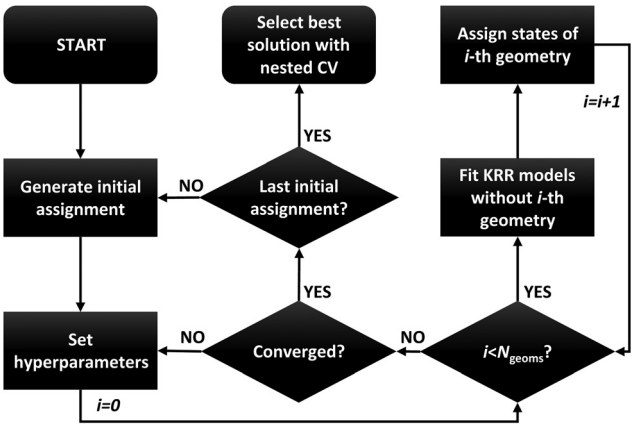

Fig. 1 Flowchart of the proposed ML-based reordering algorithm.

Since the proposed clustering algorithm is stochastic and does not guarantee the global minimum, we start the optimization procedure many times from the original and also different randomly generated initial orderings. The number of initial conditions for the ML reordering optimization procedure was selected to obtain reasonably converged results and also to approximately match the results of the wavefunction-based reordering described below, that is, 1000 optimization runs. Since the results for different initial conditions are independent, the whole procedure can be efficiently parallelized. The obtained solutions are then compared by using cross-validation prediction errors and the best one is selected. However, the performance evaluated simply by the cross-validation prediction errors from KRR hyperparameters tuning (described below) is optimistically biased. The problem is when the same data are used to both select the model and tune the hyperparameters. We overcome this limitation by using nested (double) cross-validation, that is, the hyperparameters are optimized for each ordering in inner nested cross-validation. This way, we avoid the leakage of information from the training set to the test set.

Note that the proposed clustering algorithm is just one of the possibilities for how to perform the optimization. Alternatively, it is possible to optimize the ordering for instance by some metaheuristics such as simulated annealing or genetic algorithms. The advantage of the proposed ordering is its simplicity.

2.3 ML for property-based diabatization

The ML framework for state assignment outlined above is directly applicable to crossings between states of different symmetry which do not form conical intersections. Such states do not mix and the couplings are zero by definition. The seam has then the dimensionality of $N_{\text{int}} - 1$ with N_{int} being the number of internal coordinates and simple reordering of states is the optimal solution. The proposed algorithm, as defined in the previous section, can even improve the learning of states forming conical intersections with $N_{\text{int}} - 2$ dimensional seam as the algorithm can find a route through the conical intersections which provides smoother surfaces with more slowly changing characters of the involved states. In one dimension, for example, when following a trajectory or a scan, it simply decides whether it is advantageous for the learning to switch adiabatic states in the vicinity of the conical intersection depending on the number of training nuclear geometries. Nevertheless, the most efficient learning for conical intersections can be achieved in a diabatic basis.

We first apply a property-based diabatization yielding adiabatic PEMs with inconsistent state ordering and couplings' signs. We now want to modify the assignment step of the ML-based algorithm described above to obtain consistent order of states and signs based not only on diabatic PESs (diagonal elements) but also on diabatic couplings (off-diagonal elements). Mathematically speaking, for each iteration and nuclear geometry, we want to find such an assignment of its PEM **B** represented by a signed permutation matrix **S**, which minimizes the Frobenius norm to the predicted PEM **A** from ML models trained without that particular geometry:

$$\min_{\mathbf{S}\epsilon\mathcal{S}} \left\|\mathbf{A} - \mathbf{S}\mathbf{B}\mathbf{S}^{\mathrm{T}}\right\|_{\mathrm{F}} = \max_{\mathbf{S}\epsilon\mathcal{S}} \mathrm{Tr}(\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{B}\mathbf{S}^{\mathrm{T}}) \tag{8}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and $\mathcal{S}$ is the set of all signed permutation matrices. Unfortunately, this is not a linear sum assignment problem anymore because of the off-diagonal elements which couple the rows and columns together. This problem corresponds to the quadratic assignment problem (except the permutation matrices are signed) which is an NP-complete problem so there is no known algorithm for solving it in polynomial time. In fact, there are $2^{n-1}n!$ signed permutational matrices for n states.

We can get an approximate solution by neglecting the arguably small diabatic couplings and using only the diagonal PESs; the problem then reduces to the linear assignment problem described in the previous section. However, we still need to correct the signs of diabatic couplings. The simplest approach is to compare all 2^{n-1} possible sign combinations for n states of each geometry and select the combination with the minimum error, an approach similar to phase-free learning of spin–orbit and nonadiabatic couplings by Westermayr *et al.*[39] The assumption that the diabatic couplings are completely negligible compared to the diagonal terms is unnecessarily strict. We can use the result from such simplified optimization as a starting point for further optimization taking into account even the diabatic couplings. We use here an exhaustive search: we iteratively test all permutations of states and signs for every single nuclear configuration and choose the best-performing permutation with the smallest loss function. Note, that the search is exhaustive only in terms of states but it is iterative in terms of nuclear configurations. Also, the exhaustive search can be replaced by a 2-opt optimization if too many states were included.

Note again that different optimization procedures can be used. However, the main advantage of the iterative assignment on the leave-one-out basis is its simplicity and its reasonable resistance to overfitting.

2.4 Wavefunction-based reordering

To benchmark the ML algorithm and analyze the test cases, we propose yet another reordering algorithm based on wavefunctions, yet it is applicable only to direct reordering of adiabatic states and it cannot be used for the diabatic basis. The proposed wavefunction-based ordering is based on the assumption that the states preserve, at least to some extent, their character through the state crossings and conical intersections. As a result, wavefunction descriptors can be used to reorder the excited states of the sampled nuclear geometries in order to obtain states most preserving their characters. The most natural criterion for the similarity of electronic states is their overlap. Using wavefunction overlaps, we can define distances between all the electronic states of all the nuclear configurations representing the nuclear density.

As we have distances not only between nuclear configurations but also between the excited states, we can directly cluster the states. We propose here a clustering procedure based on the direct maximization of the silhouette coefficient. However, note that other clustering techniques can be applied as well; one has to only incorporate the condition that each state of a single nuclear configuration is assigned to a different cluster. The silhouette coefficient measures how similar are data points to other points within their own cluster compared to data points in other clusters. The silhouette coefficient can be calculated with any distance metric. In contrast, the most popular k-means algorithm cannot be used to cluster states based on overlaps as it requires the calculation of cluster centres.

We first define the distance of point i to its own cluster C_I and the closest different cluster, respectively:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I} d(i,j) \tag{9}$$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j) \tag{10}$$

where $d(i,j)$ is the distance between points i and j and $|C_I|$ is the size of the cluster C_I. The silhouette for the given point is then given by these two quantities:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{11}$$

The mean silhouette over all states of all the sampled nuclear configurations represents our objective function to be maximized. Since the wavefunction overlap is a similarity metric, we define the distance by its complement to one:

$$d_{\mathrm{ol}}(i,j) = 1 - \left|\langle \boldsymbol{\Psi}_i | \boldsymbol{\Psi}'_j \rangle\right| \tag{12}$$

where $\boldsymbol{\Psi}_i$ and $\boldsymbol{\Psi}'_j$ are the wavefunctions of the two electronic states of two possibly different nuclear configurations. As wavefunctions can have arbitrary signs, we use the absolute value of the overlap. Alternatively, it is possible to use squared values or apply a phase correction.

We work here with CI-type wavefunctions which can be expressed as an expansion into Slater determinants:

$$\boldsymbol{\Psi}_i = \sum_k c_{ik} \boldsymbol{\Phi}_k \tag{13}$$

where c_{ik} are the CI expansion coefficients into Slater determinants $\boldsymbol{\Phi}_k$. Note that this group of methods includes also popular time-dependent density functional theory (TDDFT), which can be written in the form of CI singles (CIS) expansion. The overlap is then given by the overlaps between the two sets of Slater determinants:

$$\langle \boldsymbol{\Psi}_i | \boldsymbol{\Psi}'_j \rangle = \sum_k \sum_l c_{ik} c'_{jl} \langle \boldsymbol{\Phi}_k | \boldsymbol{\Phi}'_l \rangle \tag{14}$$

The overlap between two Slater determinants can be in turn expressed as a determinant containing overlaps between the constituting molecular orbitals (MOs):[40,41]

$$\langle \boldsymbol{\Phi}_k | \boldsymbol{\Phi}'_l \rangle = \begin{vmatrix} \langle \phi_{k1} | \phi'_{l1} \rangle & \cdots & \langle \phi_{k1} | \phi'_{ln} \rangle \\ \vdots & \ddots & \vdots \\ \langle \phi_{kn} | \phi'_{l1} \rangle & \cdots & \langle \phi_{kn} | \phi'_{ln} \rangle \end{vmatrix} \tag{15}$$

The calculation of wavefunction overlaps can be quite laborious and we need overlaps between all the states of all the geometries but this procedure serves here only to provide insight and validate the ML algorithm. Also, the geometries have to be aligned first in order to obtain meaningful values.

We start the optimization from the initial ordering/clusters, that is, the energy-ordered adiabatic states. Analogically to the ML reordering, we iteratively calculate the silhouette coefficient for each possible cluster assignment of each state separately for the selected geometry given the fixed clusters from the previous iteration. This way, we obtain a square matrix of silhouette coefficients between the states of the given geometry and the clusters and we select the best assignment again by solving the linear sum assignment problem. We iteratively repeat this procedure geometry by geometry until the clusters do not change anymore.

2.5 Regression model

Our ML model serves two purposes: we want to reconstruct diabatic PEMs and we want to predict adiabatic energies to reduce the number of expensive *ab initio* calculations. Many different regression models have been developed and their applications to excited-state simulations have been discussed.[2] We make our models reasonably simple mainly for two different reasons: we want to keep our methodology clear and reproducible, and we need to perform the training many times during the correction procedure of the property-based diabatization so it has to be cheap. Therefore, we train a separate ML model for each adiabatic PES or each element of the diabatic PEM using the KRR method. KRR is a simple kernel method frequently used in quantum chemistry.[42,43] Kernel methods use the so-called kernel trick that allows using linear regression algorithms to model nonlinear problems through an implicit transformation of the input data into a higher-dimensional space.[44] In our case, the KRR method is the favourable choice because of its simplicity and efficiency for small training samples.

In KRR, the quantity of interest is predicted for feature vector $\mathbf{x}$ (molecular representation) using training samples $\mathbf{x}_i$ in the following way:[45]

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \tag{16}$$

where $k(\mathbf{x}_i,\mathbf{x})$ is a kernel function providing a similarity measure between the two vectors and α_i are the regression coefficients. We use here the Gaussian kernel which is especially popular in chemistry:[42]

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2\right) \tag{17}$$

where $\|\cdot\|_2$ is the Euclidean norm and σ is a model hyperparameter. The regression coefficients are obtained from the training data by the following minimization:[45]

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda \sum_{i=1,j=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j) \alpha_j \right) \tag{18}$$

The first term is a common residual sum of squares. The second term including another hyperparameter λ is responsible for the regularization which should prevent overfitting of the training data. This minimization has a closed-form solution:

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y} \tag{19}$$

where $\mathbf{y}$ is the vector of known solutions for the training data and $\mathbf{K}$ is a kernel matrix with elements $\mathbf{K}_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$. This equation is in practice solved by the Cholesky decomposition. Within our approach, we need to solve this equation a lot of times, very often for the same or slightly modified kernel but with different $\mathbf{y}$. This can be done efficiently by caching and/or updating the intermediate results of the Cholesky decomposition, effectively reducing the formal $O(n^3)$ scaling with the number of samples up to quadratic dependence. The hyperparameters σ and λ are selected on a grid using 10-fold cross-validation.

A crucial ingredient for the prediction of molecular properties is a molecular representation or molecular descriptors, that is, the feature vector $\mathbf{x}$ encoding the system, usually *via* the molecular structure.[43,46,47] It should fulfil some basic requirements for ML to be efficient: it should usually possess translational, rotational, and permutational invariance.[48,49] By working with nuclear configurations of a single molecular entity, some of the desired properties are automatically fulfilled. Namely, the number of atoms is constant, resulting in a constant-size molecular representation. However, as opposed to adiabatic properties, the diabatic PEM is not in general invariant with respect to permutations and inversion. Instead, it follows the symmetry of the corresponding complete nuclear permutation inversion (CNPI) group.[10,50,51] While permutations do not play any role in the FNO molecule, formaldehyde belongs to the C_{2v}(M) CNPI group, which is isomorphic to the C_{2v} point group.[52] As a result, its diabatic states carry irreducible representations of the C_{2v}(M) CNPI group. The symmetries of the involved system-specific irreducible representations can be directly incorporated into the fitted model.[5,52–54] However, this can be a rather difficult task. We opted for a simpler option as we are interested mainly in more accurate predictions of adiabatic properties rather than accurate diabatic PEMs. By using a representation invariant to inversion and permutation of equivalent atoms, we effectively limit our model to a subspace of the whole configuration space. However, the rest of the configuration space is still given by corresponding CNPI symmetries. Moreover, the fitted subspace is sufficient if we are interested only in efficient prediction for adiabatic states as these are invariant with respect to both inversion and permutations.

We used a simple vector of normalized inverted internuclear distances as the molecular representation:[55,56]

$$\mathbf{x} = \left(\frac{r_{i,j}^{\mathrm{ref}}}{r_{i,j}} \quad \text{for } 1 < i \leq N \quad \text{for } j < i \right) \tag{20}$$

where $r_{i,j}$ is the Euclidean distance between atoms i and j and $r_{i,j}^{\mathrm{ref}}$ is the reference value. The reference values are usually taken from the minimal geometry but we used here average values sampled in the nuclear ensemble. This representation is simple yet efficient for our small molecules and it possesses both translational and rotational invariance. While it is also invariant with respect to inversion, it is not permutationally invariant. Same as Guan *et al.*,[57] we enforce the permutational invariance for the formaldehyde molecule by permuting the hydrogen atoms so that the bond distances follow $r_{\mathrm{CH}_1} < r_{\mathrm{CH}_2}$. Alternatively, we could use for example the permutationally invariant kernel.[58]

2.6 Computational details

The molecules were optimized at the B3LYP/6-31g* level with subsequent vibrational analysis on the same level using Gaussian G09,[59] revision D.01. 1000 nuclear configurations for each molecule were subsequently sampled using the harmonic approximation and the temperature-dependent Wigner quasiprobability distribution:[60,61]

$$P_{\mathrm{W}}(\mathbf{q},\mathbf{p},T) = \prod_i \frac{1}{\pi\hbar}\tanh\left(\frac{\hbar\omega_i}{2k_{\mathrm{B}}T}\right) \times \exp\left(\tanh\left(\frac{\hbar\omega_i}{2k_{\mathrm{B}}T}\right)\left(-\frac{p_i^2}{\mu_i\hbar\omega_i} - \frac{\mu_i\omega_i q_i^2}{\hbar}\right)\right) \tag{21}$$

where q_i is the deviation along the i-th normal mode and p_i, ω_i and μ_i are the corresponding momentum, angular frequency, and reduced mass, respectively. T is the temperature set to 300 K and k_{B} is the Boltzmann constant.

As described above, the hydrogen atoms were permuted for the formaldehyde molecule so that $r_{\mathrm{CH}_1} < r_{\mathrm{CH}_2}$ to ensure permutational invariance. Moreover, we inverted the geometries so that the oxygen atom was always located on the same side of the CH_1H_2 plane for the calculation of overlaps to ensure invariance with respect to inversion. All the nuclear configurations for each molecule were geometrically aligned to one reference minimizing the mean square error between atomic centres *via* translation and rotation in order to obtain reasonable wavefunction overlaps needed for the analysis. Subsequently, the excited-state calculations for the sampled geometries were performed again at the B3LYP/6-31g* level of theory within the Tamm–Dancoff approximation in the PySCF[34,35] code, version 2.0.1. Note that this level of theory does not provide quantitative results and the present calculations serve only to show the performance of the proposed algorithms. However, this level of theory combined with small test molecules allows us to calculate overlaps between all pairs of states of all sampled geometries, which is vital for the analysis and tuning of the optimization procedure.

3 Results and discussion

We chose nitrosyl fluoride (FNO) as the first example to show how the proposed methodology works. The first reason is that it is small so it can be easily analyzed but it is already a 3D problem that cannot be simply corrected by hand. The second reason is its C_s point group resulting in two sets of electronic states with either A' or A'' symmetry. We can therefore examine the behaviour of the algorithm both when two states of different symmetry cross without mixing and when states of the same symmetry form conical intersections. The second test case is the formaldehyde molecule which represents already a 6D problem but it is still possible to calculate pairwise wavefunction overlaps for analytical and benchmark purposes. Also, both molecules contain a set of singlet states that do not interact with other higher or lower-lying states at the employed level of theory, which is a prerequisite for efficient diabatization. States entering and leaving the predefined manifold represent a general problem for diabatization methods. Sampled geometries for both molecules, training indices, and calculated excitation energies and transition moments are included in ESI.†

3.1 Nitrosyl fluoride: 1D scan

Let us first look at the 1D scan of the FNO molecule along the NO bond to demonstrate how the proposed methodology works. The first three excited singlet states are all energetically well separated and do not mix or cross. We, therefore, focus on the next three states S_4–S_6 which cross and mix within the sampled configuration space. Note that these three states actually include the brightest states of the FNO molecule. We can see that while two states of the same A'' symmetry form an avoided crossing, the third state has a different A' symmetry a crosses them without any interaction (see Fig. 2a). We can directly apply the ML reordering algorithm without prior diabatization (see Fig. 2b). Such treatment correctly reconstructs the non-mixing diabatic state of different symmetry as the off-diagonal elements are zero and reordering actually represents the exact diabatization. The two states of the same symmetry switch their order in the centre of the avoided crossing resulting in two almost linear curves only with a small disruption located at the avoided crossing. While these states are not properly diabatic, they are much easier to fit than the original ones. Such a result looks encouraging; however, note that the 1D picture might be a bit misleading. The avoided crossing is caused by a conical intersection which cannot be displayed in one dimension. The reordering based on wavefunction overlaps is not plotted separately as it provides here the same result as the ML-based reordering but their agreement shows that the clustering works properly.

As a next step, we apply a simple tD diabatization scheme as outlined in Section 2.1. In this case, we need to distinguish only two states of the same symmetry along one coordinate so the property-unblending diabatization using just the transition dipole moments from the ground state is sufficient. Fig. 2c displays the diagonal elements of the diabatic PEM while Fig. 2d displays the off-diagonal elements, that is, the diabatic couplings. We can directly see the two problems of property-based diabatization: the ordering of the diabatic states is not consistent along the coordinate and the diabatic couplings have random signs. By the subsequent application of our algorithm, we get both smooth diabatic PESs and couplings

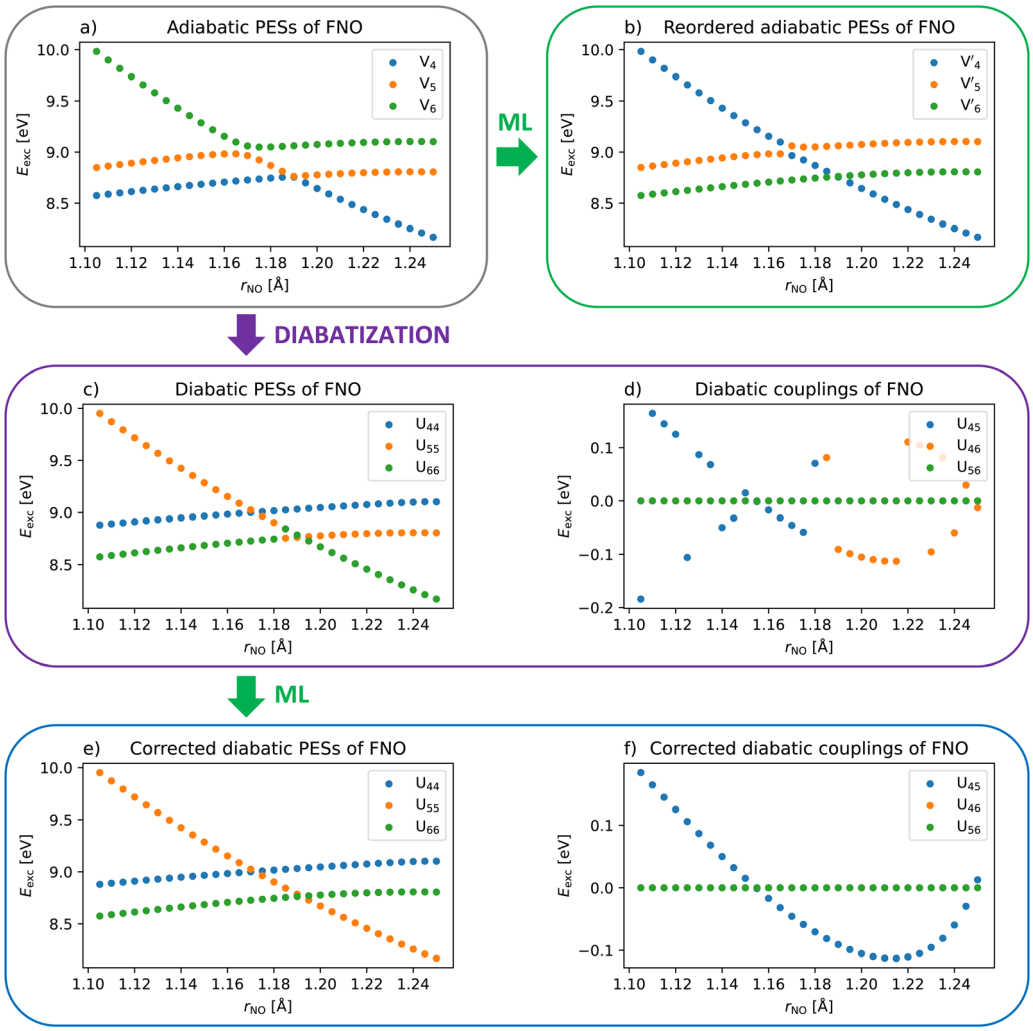

Fig. 2 Excited states of the FNO molecule along the NO bond in the (a) adiabatic basis, (b) reordered adiabatic basis, (c) and (d) diabatic basis, and (e) and (f) reordered and sign-corrected diabatic basis.

(see Fig. 2e and f). One might point out that the correct ordering and signs are obvious. This is true in one dimension but the ordering and signs cannot be easily corrected by hand in a multidimensional space. Our algorithm allows applying property-based diabatization to multidimensional problems as shown below.

3.2 Nitrosyl fluoride: 3D case

Let us now move to the full 3D space of the FNO molecule. In the full space, we have to include another two higher-lying states which interact with the three already included states. There are now two states of A' symmetry and three states of A'' symmetry. Fig. 3a presents a 2D multidimensional scaling projection of the five excited states for 100 nuclear configurations. Multidimensional scaling forms a low-dimensional representation of the data, in which the distances respect the distances in the original high-dimensional space as well as possible.[62] We defined the distances the same way as in eqn (12) so the overlaps are also reasonably preserved given the limitations of a 2D plot. The excited states form five clusters corresponding to five diabatic states and none of them coincides with a single adiabatic state plotted with different colours. It can be clearly seen that the three states of A'' symmetry mix together as there are samples connecting these clusters. On the contrary, the two A' states do not mix suggesting that they are well separated within the sampled space.

To provide insight, let us first look at wavefunction-based clustering which serves here for visualization and benchmark purposes. Fig. 3b shows the same projection after we applied the wavefunction-based clustering described in Section 2.4. The adiabatic states of each geometry are now assigned to the clusters as well as possible. We can now create an ML model for each of these clusters instead of the original adiabatic states. The geometrical topologies of conical intersections are of course still present but we might hope that the new clusters present a better way through them. Nevertheless, these models serve mainly as a benchmark to test our ML reordering on an adiabatic basis before switching to a diabatic basis. Similarly, we reordered the adiabatic states using our ML approach to see whether such treatment is sufficient.

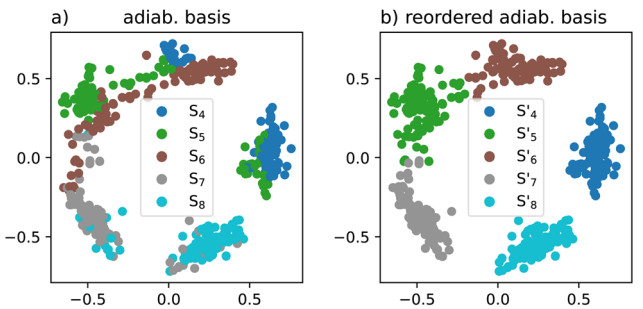

Fig. 3 Multidimensional scaling projection of excited state clusters (a) before and (b) after reordering based on wavefunction overlaps for 100 nuclear configurations and 5 excited states. The projection corresponds to a 2D space in which the wavefunction overlaps are preserved as well as possible.

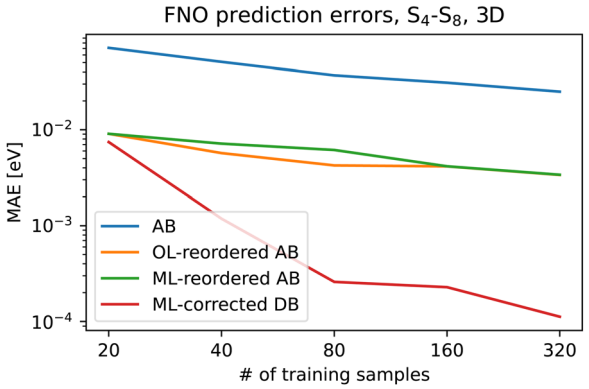

Fig. 4 The mean absolute error of the kernel ridge regression for the FNO molecule as a function of training set size for adiabatic basis (AB), adiabatic basis reordered using wavefunction overlaps (OL), ML-reordered adiabatic basis, and values obtained from the diagonalization of ML-corrected diabatic basis (DB).

Finally, we applied the property-based diabatization and corrected the signs and ordering with our ML approach. The tD diabatization is not sufficient anymore as we need to differentiate 5 states. Therefore, we use here the tDQ diabatization. Let us now compare the accuracy of the ML prediction before and after applying all these methods, that is, original adiabatic states, adiabatic states reordered using wavefunction overlaps, ML-reordered adiabatic states, and ML-corrected diabatic states. The results are plotted for different training set sizes in Fig. 4. We always selected a training set of a given size, reordered/corrected it with the proposed algorithms, and used it to train a separate KRR model for each PES, and also each diabatic coupling in the case of the diabatic basis. We subsequently used these models to predict PESs for the rest of the 1000 geometries, which were not selected for the training set and evaluated the prediction error by means of the mean absolute error (MAE). In the case of the diabatic basis, the predicted PEMs are diagonalized and the resulting adiabatic energies are compared to the other models. Note, that the results are plotted on the log–log scale.

We can see that the improvement in accuracy is enormous for all the proposed approaches. Both adiabatic reordering approaches improve learning consistently almost by one order of magnitude. Also, both reordering approaches provide comparable results which suggest that our ML reordering procedure is sufficient. By switching to the diabatic basis and correcting the signs and ordering, we get another significant increase in accuracy. Not only that the absolute errors are much smaller but also the slope is better. The MAE is smaller by two orders of magnitude already with 80 samples.

To inspect how the diabatic states look like, we plot their PESs in Fig. 5 for a fixed bonding angle using ML models trained on 320 geometries. While it is difficult to plot five surfaces at once in a clear way, the PESs are clearly smooth and cross each other without forming conical intersections.

3.3 Formaldehyde: 6D case

We repeated the whole procedure for the formaldehyde molecule where we selected the tDQO property-based diabatization

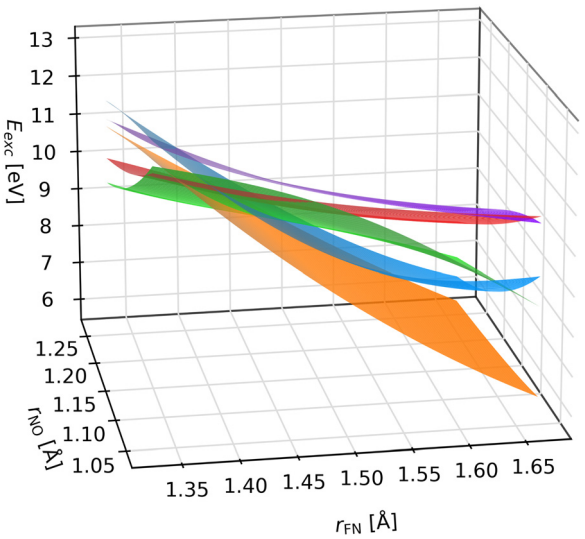

Fig. 5 Potential energy surfaces obtained by the proposed ML approach based on the tDQ diabatization of the S_4–S_8 adiabatic states of the FNO molecule. The bond angle is fixed to 110°. ML models were trained on 320 geometries from the Wigner distribution.

as the tDQ diabatization did not improve the learning. We could have in principle diabatized states of different symmetries separately for the FNO molecule but this is not the case for the formaldehyde molecule; while formaldehyde belongs to the C_{2v} point group in the minimal geometry, the symmetry is broken virtually for all the geometries. We included the S_2–S_5 states as those mix in the sampled region and are energetically well separated from both the S_1 state and the higher-lying states at the employed level of theory. The MAEs for both adiabatic reordering approaches and the diabatic ML approach are presented in Fig. 6. We observe again a major improvement in prediction accuracy by up to one order of magnitude with 320 training geometries. While the improvement is not as remarkable as for the FNO molecule, one order of magnitude is still a huge improvement. It is important to realize that the final diabatic ML models are always limited by the underlying property-based diabatization. Both adiabatic reordering approaches decrease the prediction errors by up to half an order of magnitude and provide again very similar results.

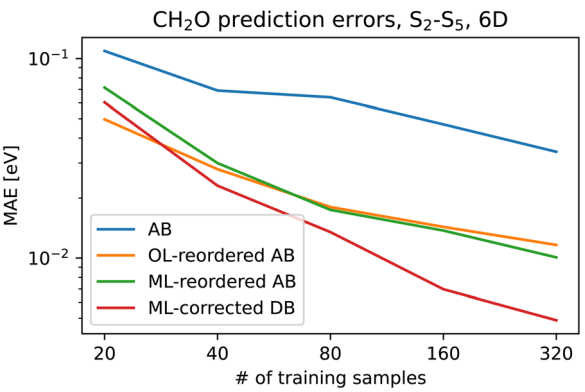

Fig. 6 The mean absolute error of the kernel ridge regression for the formaldehyde molecule as a function of training set size for adiabatic basis (AB), adiabatic basis reordered using wavefunction overlaps (OL), ML-reordered adiabatic basis, and values obtained from the diagonalization of ML-corrected diabatic basis (DB).

4 Conclusions

We tackled two different problems at once: efficient machine learning for excited-state properties and diabatization. We proposed and tested methodology for correcting deficiencies of property-based diabatization techniques including random signs of the diabatic couplings and inconsistent ordering of the diabatic states throughout the configuration space, which prohibited the wider deployment of these methods to multi-dimensional systems. To this end, we developed a stochastic ML optimization procedure based on the combination of KRR and clustering. The optimization provided us with smooth diabatic states which are also easy to fit and predict. The set of adiabatic energies can be then easily obtained by diagonalization of the predicted diabatic PEMs. This way, we were able to improve the prediction accuracy by about 2 orders of magnitude in terms of MAE for the adiabatic energies of the FNO molecule and almost 1 order of magnitude for the formaldehyde molecule. We managed to efficiently utilize unprecedentedly small training sets including from dozens up to hundreds of nuclear geometries. However, it is important to note that the quality and performance of the final ML models are heavily dependent on the underlying property-based diabatization. Our ML approach corrects inconsistent state ordering and sings but it cannot correct for improperly chosen diabatization properties or state manifolds.

Our ML approach is applicable to any property-based diabatization. However, we also proposed a series of simple property-based diabatization schemes that are easily applicable even to single-reference methods such as TDDFT. These schemes are based only on transition multipoles from the ground state which makes them pragmatic and easily applicable but also not universal. The algorithm can be in principle applied to conical intersections of three or more adiabatic states occupying even lower-dimensional space whenever the underlying property-based diabatization is able to distinguish them. The direct application of our reordering algorithms without prior diabatization also improved the learning significantly: up to one order of magnitude for the FNO molecule and up to half an order of magnitude for the formaldehyde molecule. However, such behaviour cannot be probably expected for much more complex PESs of large systems.

Overall, we developed a methodology making diabatization more accessible for quantum-chemistry practitioners as it is based on the simplest category of diabatization methods, that is, property-based diabatization. The ML-corrected diabatic basis can save us many computationally expensive *ab initio* calculations as we can use much smaller training samples to achieve the same prediction accuracy. We also kept our optimization procedure as simple as possible for the sake of better

transferability and reproducibility. Nevertheless, more efficient optimization procedures can be used for example by merging our algorithm with some metaheuristics; also, the cluster-growing[24] algorithm can be used as the initial solution for the proposed optimization if we find a way how to form the initial cluster automatically. The methodology can be in principle used with different ML models instead of KRR. However, the ML model has to be reasonably efficient as it gets retrained many times during the optimization procedure.

This work opens the way to various applications. While we used the presented ML-corrected diabatization to fit PESs of two simple molecules, an analogous approach can be used to efficiently model electronic spectra using the nuclear ensemble method or any other property reflecting the ground-state geometry distribution.[25,63–65] The proposed ML algorithm can be also directly used as an alternative to the cluster-growing algorithm to correct signs within other categories of diabatization methods as this particular problem is not specific only to property-based diabatization. Moreover, wrong state ordering was identified as a possible problem when learning differences between two electronic structure methods within Δ-ML.[3] The basic reordering algorithm could resolve the issue caused by inconsistent ordering of adiabatic states at the two employed levels of theory. The present approach might be extended in the future to tackle also the problem with states entering and leaving the predefined excited-state manifold for diabatization by fitting a larger number of diabatic states (or predicting a smaller number of adiabatic states) than the number of input adiabatic states. Implicitly fitting a larger number of diabatic states within neural network architecture has been already shown to improve prediction accuracy.[22] Eventually, the proposed diabatization might be in principle also used for efficient nonadiabatic dynamics simulations but one would have to take care that the configuration space is properly sampled and it might be advantageous to include gradients and NACs to the loss function if they are available for training. However, this application is yet to be explored. Also, such an application is in general more prone to the problem of states entering and leaving the predefined manifold.

Data availability

Sampled geometries for both molecules, training indices, and calculated excitation energies and transition moments have been uploaded as part of the ESI.†

Author contributions

Š. S.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft, writing – review and editing. O. A. V. L.: conceptualization, methodology, supervision, validation, writing – original draft, writing – review and editing. P. S.: conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing – original draft, writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Š. S. and P. S. gratefully acknowledge the support of the Czech Science Foundation, project no. 20-15825S. Š. S. also acknowledges the support of the Czech Science Foundation, project no. 22-13489O, after project no. 20-15825S finished. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID: 90140). O. A. v. L. has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement no. 772834). O. A. v. L. has received support as the Ed Clark Chair of Advanced Materials and as a Canada CIFAR AI Chair.

References

1 P. O. Dral, Quantum Chemistry in the Age of Machine Learning, *J. Phys. Chem. Lett.*, 2020, **11**, 2336–2347.
2 J. Westermayr and P. Marquetand, Machine Learning for Electronically Excited States of Molecules, *Chem. Rev.*, 2021, **121**, 9873–9926.
3 R. Ramakrishnan, M. Hartmann, E. Tapavicza and O. A. Von Lilienfeld, Electronic spectra from TDDFT and machine learning in chemical space, *J. Chem. Phys.*, 2015, **143**, 084111.
4 X. Zhu and D. R. Yarkony, Fitting coupled potential energy surfaces for large systems: Method and construction of a 3-state representation for phenol photodissociation in the full 33 internal degrees of freedom using multireference configuration interaction determined data, *J. Chem. Phys.*, 2014, **140**, 024112.
5 C. L. Malbon, B. Zhao, H. Guo and D. R. Yarkony, On the nonadiabatic collisional quenching of OH(A) by H_2: a four coupled quasi-diabatic state description, *Phys. Chem. Chem. Phys.*, 2020, **22**, 13516–13527.
6 Y. Shen and D. R. Yarkony, Construction of Quasi-diabatic Hamiltonians That Accurately Represent ab Initio Determined Adiabatic Electronic States Coupled by Conical Intersections for Systems on the Order of 15 Atoms. Application to Cyclopentoxide Photoelectron Detachment in the Ful, *J. Phys. Chem. A*, 2020, **124**, 4539–4548.
7 Y. Shu, Z. Varga, S. Kanchanakungwankul, L. Zhang and D. G. Truhlar, Diabatic States of Molecules, *J. Phys. Chem. A*, 2022, **126**, 992–1018.
8 R. Abrol and A. Kuppermann, An optimal adiabatic-to-diabatic transformation of the $1^2A'$ and $2^2A'$ states of H_3, *J. Chem. Phys.*, 2002, **116**, 1035–1062.

9 H. Köppel, Regularized diabatic states and quantum dynamics on intersecting potential energy surfaces, *Faraday Discuss.*, 2004, **127**, 35–47.
10 X. Zhu and D. R. Yarkony, Toward eliminating the electronic structure bottleneck in nonadiabatic dynamics on the fly: An algorithm to fit nonlocal, quasidiabatic, coupled electronic state Hamiltonians based on ab initio electronic structure data, *J. Chem. Phys.*, 2010, **132**, 104101.
11 C. E. Hoyer, K. Parker, L. Gagliardi and D. G. Truhlar, The DQ and DQΦ electronic structure diabatization methods: Validation for general applications, *J. Chem. Phys.*, 2016, **144**, 194101.
12 Z. Varga, K. A. Parker and D. G. Truhlar, Direct diabatization based on nonadiabatic couplings: the N/D method, *Phys. Chem. Chem. Phys.*, 2018, **20**, 26643–26659.
13 J. E. Subotnik, S. Yeganeh, R. J. Cave and M. A. Ratner, Constructing diabatic states from adiabatic states: Extending generalized Mulliken-Hush to multiple charge centers with Boys localization, *J. Chem. Phys.*, 2008, **129**, 244101.
14 T. Pacher, H. Köppel and L. S. Cederbaum, Quasidiabatic states from ab initio calculations by block diagonalization of the electronic Hamiltonian: Use of frozen orbitals, *J. Chem. Phys.*, 1991, **95**, 6668–6680.
15 N. Wittenbrink, F. Venghaus, D. Williams and W. Eisfeld, A new approach for the development of diabatic potential energy surfaces: Hybrid block-diagonalization and diabatization by ansatz, *J. Chem. Phys.*, 2016, **145**, 184108.
16 G. J. Atchity and K. Ruedenberg, Determination of diabatic states through enforcement of configurational uniformity, *Theor. Chem. Acc.*, 1997, **97**, 47–58.
17 K. R. Yang, X. Xu and D. G. Truhlar, Direct diabatization of electronic states by the fourfold-way: Including dynamical correlation by multi-configuration quasidegenerate perturbation theory with complete active space self-consistent-field diabatic molecular orbitals, *Chem. Phys. Lett.*, 2013, **573**, 84–89.
18 Y. Shu and D. G. Truhlar, Diabatization by Machine Intelligence, *J. Chem. Theory Comput.*, 2020, **16**, 6456–6464.
19 Y. Shu, Z. Varga, A. G. Sampaio De Oliveira-Filho and D. G. Truhlar, Permutationally Restrained Diabatization by Machine Intelligence, *J. Chem. Theory Comput.*, 2021, **17**, 1106–1116.
20 Y. Guan, D. H. Zhang, H. Guo and D. R. Yarkony, Representation of coupled adiabatic potential energy surfaces using neural network based quasi-diabatic Hamiltonians: 1,2 ^{2}A' states of LiFH, *Phys. Chem. Chem. Phys.*, 2019, **21**, 14205–14213.
21 D. M. Williams and W. Eisfeld, Neural network diabatization: A new ansatz for accurate high-dimensional coupled potential energy surfaces, *J. Chem. Phys.*, 2018, **149**, 204106.
22 S. Axelrod, E. Shakhnovich and R. Gómez-Bombarelli, Excited state non-adiabatic dynamics of large photoswitchable molecules using a chemically transferable machine learning potential, *Nat. Commun.*, 2022, **13**, 3440.
23 T. Y. Wang, S. P. Neville and M. S. Schuurman, Machine Learning Seams of Conical Intersection: A Characteristic Polynomial Approach, *J. Phys. Chem. Lett.*, 2023, **14**, 7780–7786.
24 Y. Shu, J. Kryven, A. G. Sampaio de Oliveira-Filho, L. Zhang, G.-L. Song, S. L. Li, R. Meana-Pañeda, B. Fu, J. M. Bowman and D. G. Truhlar, Direct diabatization and analytic representation of coupled potential energy surfaces and couplings for the reactive quenching of the excited $^2\Sigma^+$ state of OH by molecular hydrogen, *J. Chem. Phys.*, 2019, **151**, 104311.
25 Y. Guan, H. Guo and D. R. Yarkony, Extending the Representation of Multistate Coupled Potential Energy Surfaces to Include Properties Operators Using Neural Networks: Application to the 1,2^{1}A States of Ammonia, *J. Chem. Theory Comput.*, 2020, **16**, 302–313.
26 B. Mazouin, A. A. Schöpfer and O. A. von Lilienfeld, Selected machine learning of HOMO-LUMO gaps with improved data-efficiency, *Mater. Adv.*, 2022, **3**, 8306–8316.
27 D. R. Yarkony, Conical Intersections: The New Conventional Wisdom, *J. Phys. Chem. A*, 2001, **105**, 6277–6293.
28 F. T. Smith, Diabatic and Adiabatic Representations for Atomic Collision Problems, *Phys. Rev.*, 1969, **179**, 111–123.
29 C. A. Mead and D. G. Truhlar, Conditions for the definition of a strictly diabatic electronic basis for molecular systems, *J. Chem. Phys.*, 1982, **77**, 6090–6098.
30 C. E. Hoyer, X. Xu, D. Ma, L. Gagliardi and D. G. Truhlar, Diabatization based on the dipole and quadrupole: The DQ method, *J. Chem. Phys.*, 2014, **141**, 114104.
31 D. A. Kleier, T. A. Halgren, J. H. Hall and W. N. Lipscomb, Localized molecular orbitals for polyatomic molecules. I. a comparison of the Edmiston-Ruedenberg and Boys localization methods, *J. Chem. Phys.*, 1974, **61**, 3905–3919.
32 H. J. Werner and W. Meyer, MCSCF study of the avoided curve crossing of the two lowest $^1\Sigma^+$ states of LiF, *J. Chem. Phys.*, 1981, **74**, 5802–5807.
33 T. Karman, A. Van Der Avoird and G. C. Groenenboom, Communication: Multiple-property-based diabatization for open-shell van der Waals molecules, *J. Chem. Phys.*, 2016, **144**, 121101.
34 Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters and G. K. Chan, PySCF: the Python-based simulations of chemistry framework, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2018, **8**, e1340.
35 Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z. H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu, N. Mardirossian, J. D. McClain, M. Motta, B. Mussard, H. Q. Pham, A. Pulkin, W. Purwanto, P. J. Robinson, E. Ronca, E. R. Sayfutyarova, M. Scheurer, H. F. Schurkus, J. E. Smith, C. Sun, S. N. Sun, S. Upadhyay, L. K. Wagner, X. Wang, A. White, J. D. Whitfield, M. J. Williamson, S. Wouters, J. Yang, J. M. Yu, T. Zhu, T. C. Berkelbach, S. Sharma, A. Y. Sokolov and G. K. L. Chan, Recent developments in the PySCF program package, *J. Chem. Phys.*, 2020, **153**, 024109.
36 D. F. Crouse, On implementing 2D rectangular assignment algorithms, *IEEE Trans. Aerosp. Electron. Syst.*, 2016, **52**, 1679–1696.

37 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa and P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat. Methods*, 2020, **17**, 261–272.

38 P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Pearson Education Limited, Harlow, 2014.

39 J. Westermayr, M. Gastegger and P. Marquetand, Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics, *J. Phys. Chem. Lett.*, 2020, **11**, 3828–3834.

40 P. O. Löwdin, Quantum Theory of Many-Particle Systems. I. Physical Interpretations by Means of Density Matrices, Natural Spin-Orbitals, and Convergence Problems in the Method of Configurational Interaction, *Phys. Rev.*, 1955, **97**, 1474.

41 F. Plasser, M. Ruckenbauer, S. Mai, M. Oppel, P. Marquetand and L. González, Efficient and Flexible Computation of Many-Electron Wave Function Overlaps, *J. Chem. Theory Comput.*, 2016, **12**, 1207–1219.

42 M. Rupp, Machine learning for quantum mechanics in a nutshell, *Int. J. Quantum Chem.*, 2015, **115**, 1058–1073.

43 M. Rupp, O. A. von Lilienfeld and K. Burke, Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry, *J. Chem. Phys.*, 2018, **148**, 241401.

44 B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, Cambridge, 2002.

45 K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko and K.-R. Müller, Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies, *J. Chem. Theory Comput.*, 2013, **9**, 3404–3419.

46 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error, *J. Chem. Theory Comput.*, 2017, **13**, 5255–5264.

47 F. A. Faber, A. S. Christensen, B. Huang and O. A. von Lilienfeld, Alchemical and structural distribution based representation for universal quantum machine learning, *J. Chem. Phys.*, 2018, **148**, 241717.

48 O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp and A. Knoll, Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties, *Int. J. Quantum Chem.*, 2015, **115**, 1084–1093.

49 B. Huang and O. A. von Lilienfeld, Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity, *J. Chem. Phys.*, 2016, **145**, 161102.

50 H. C. Longuet-Higgins, The symmetry groups of non-rigid molecules, *Mol. Phys.*, 1963, **6**, 445–460.

51 S. P. Keating and C. A. Mead, Conical intersections in a system of four identical nuclei, *J. Chem. Phys.*, 1985, **82**, 5102–5117.

52 Y. Guan, C. Xie, H. Guo and D. R. Yarkony, Neural Network Based Quasi-diabatic Representation for S_0 and S_1 States of Formaldehyde, *J. Phys. Chem. A*, 2020, **124**, 10132–10142.

53 D. M. Williams and W. Eisfeld, Complete Nuclear Permutation Inversion Invariant Artificial Neural Network (CNPI-ANN) Diabatization for the Accurate Treatment of Vibronic Coupling Problems, *J. Phys. Chem. A*, 2020, **124**, 7608–7621.

54 Z. Yin, B. J. Braams, Y. Guan, B. Fu and D. H. Zhang, A fundamental invariant-neural network representation of quasi-diabatic Hamiltonians for the two lowest states of H_3, *Phys. Chem. Chem. Phys.*, 2021, **23**, 1082–1091.

55 P. O. Dral, A. Owens, S. N. Yurchenko and W. Thiel, Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels, *J. Chem. Phys.*, 2017, **146**, 244108.

56 P. O. Dral, MLatom: A program package for quantum chemical research assisted by machine learning, *J. Comput. Chem.*, 2019, **40**, 2339–2347.

57 Y. Guan, C. Xie, H. Guo and D. R. Yarkony, Enabling a unified description of both internal conversion and intersystem crossing in formaldehyde: A global coupled Quasi-Diabatic hamiltonian for its S_0, S_1, and T_1 states, *J. Chem. Theory Comput.*, 2021, **17**, 4157–4168.

58 A. P. Bartók and G. Csányi, Gaussian approximation potentials: A brief tutorial introduction, *Int. J. Quantum Chem.*, 2015, **115**, 1051–1057.

59 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09, Revision D.01*, 2013.

60 M. Hillery, R. O'Connell, M. Scully and E. Wigner, Distribution functions in physics: Fundamentals, *Phys. Rep.*, 1984, **106**, 121–167.

61 J. P. Zobel, J. J. Nogueira and L. González, Finite-temperature Wigner phase-space sampling and temperature effects on the excited-state dynamics of 2-nitronaphthalene, *Phys. Chem. Chem. Phys.*, 2019, **21**, 13906–13915.

62 A. Mead, Review of the Development of Multidimensional Scaling Methods, *J. R. Stat. Soc. Ser. D*, 1992, **41**, 27–39.
63 Š. Sršeň, J. Sita, P. Slavíček, V. Ladányi and D. Heger, Limits of the Nuclear Ensemble Method for Electronic Spectra Simulations: Temperature Dependence of the (E)-Azobenzene Spectrum, *J. Chem. Theory Comput.*, 2020, **16**, 6428–6438.
64 R. Crespo-Otero and M. Barbatti, Spectrum simulation and decomposition with nuclear ensemble: formal derivation and application to benzene, furan and 2-phenylfuran, *Theor. Chem. Acc.*, 2012, **131**, 1237.
65 B. X. Xue, M. Barbatti and P. O. Dral, Machine Learning for Absorption Cross Sections, *J. Phys. Chem. A*, 2020, **124**, 7199–7210.