

Fast and accurate excited states predictions: Machine learning and diabaticization

Štěpán Sršeň,^{*,†,‡} O. Anatole von Lilienfeld,^{¶,§,||} and Petr Slavíček^{*,†}

[†]*Department of Physical Chemistry, University of Chemistry and Technology, Prague
16628, Czech Republic*

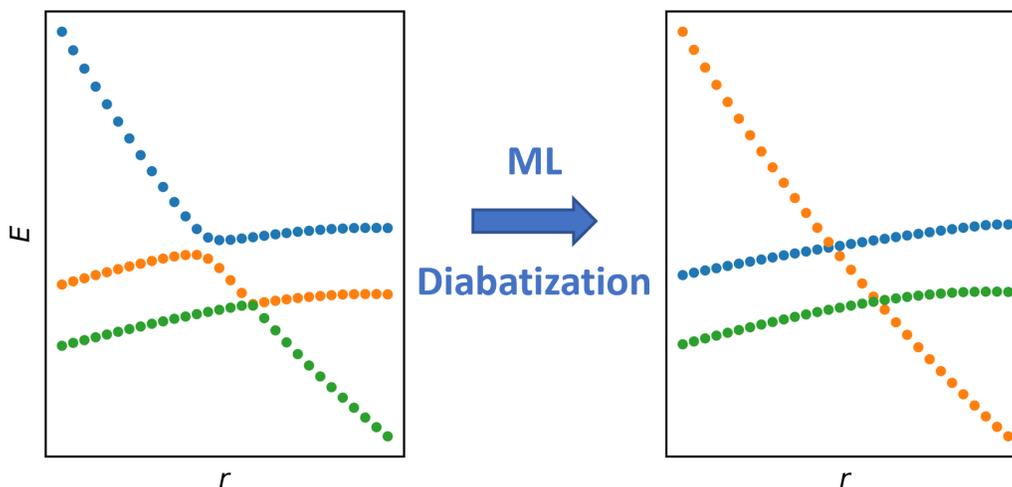
[‡]*Institute of Theoretical Chemistry, Faculty of Chemistry, University of Vienna,
Währinger Str. 17, 1090 Wien, Austria*

[¶]*Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada*

[§]*Departments of Chemistry, Materials Science and Engineering, and Physics, University of
Toronto, St. George Campus, Toronto, ON, Canada*

^{||}*Machine Learning Group, Technische Universität Berlin and Institute for the Foundations
of Learning and Data, 10587 Berlin, Germany*

E-mail: stepan.srsen@vscht.cz; petr.slavicek@vscht.cz



Abstract

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, diabaticization 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 diabaticization which, in return, provides us with smooth surfaces that can be easily fitted. Our approach extends the applicability of property-based diabaticization to multidimensional systems. We show the performance of the proposed methodology 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 diabaticization and regression algorithm, we show its performance for kernel ridge regression and a very simple diabaticization based on transition multipoles. Compared to most other algorithms based on machine learning, our approach needs only a small amount of training data.

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.

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 At the same time, excited states are part of many fundamental processes such as photosynthesis³ light-induced reactions⁴ or photovoltaics.⁵ 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.⁶ However, some research in this area has been performed just over the last few years and has brought indications that the problems can be tackled.² The majority of these studies deal with photodynamical simulations including predictions of electronic energies, atomic forces, and nonadiabatic or spin-orbit couplings.⁷⁻⁹ Regarding excited states, ML has been also applied to multireference character estimation,¹⁰ active space selection,¹¹ quantum yields prediction,¹² and also to diabaticization.¹³

We tackle here the problem of the low smoothness of excited-state properties in the vicinity of state crossings and conical intersections. Adiabatic representation is what we usually get from electronic structure calculations, that is, eigenfunctions and eigenvalues of the electronic Hamiltonian. The states are then ordered by their electronic energy for each nuclear configuration, resulting in non-crossing potential energy surfaces (PESs). While the 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 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: the diabatic basis is not unique, it usually needs to be determined from the adiabatic basis and it most often requires expert knowledge about the system.¹⁴ To date, only small systems have been accurately represented in a diabatic basis. However, several works emerged during the last few years which try to solve the

problem of automatic data-driven determination of the diabatic basis.^{13,15–19} While classical diabaticization schemes usually need lots of expert knowledge, manual work, and expensive calculations, ML-based approaches usually require lots of 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 diabaticization schemes with an ML algorithm that corrects their deficiencies.

Property-based diabaticization is arguably the simplest category of diabaticization techniques. It is based on the observation that the adiabatic states are nonsmooth as a consequence of the rapid change of the wavefunction character near conical intersections where the states mix. If the transformed diabatic wavefunctions change slowly as functions of geometry then their properties are expected to change slowly as well and vice versa. This can be achieved by maximizing the differences between properties of the transformed diabatic states and, therefore, suppressing the mixing of the states. In practice, a matrix of pairwise properties, such as for example (transition) dipole moments, is formed for the adiabatic states and the transformation matrix is obtained by diagonalization of the property matrix or a similar procedure, which ensures the separation of the properties in the diabatic basis.¹⁴ Unfortunately, there are some problems connected with this category of diabaticization methods, which prevent their widespread application to larger molecules with multiple electronic states involved. First, we need to select the property matrix in a way that allows discrimination of different electronic states. If we manage to do so, there are still two technical problems specific to property-based diabaticization. First, the order of the diabatic states might not be consistent throughout the whole configuration space. In other words, we get a set of diabatic energies and couplings (off-diagonal elements) in the diabatic basis for a given nuclear geometry, and we have to decide to which global diabatic states they belong (see figure 2c). The second issue arises from random signs of wavefunctions which lead to random signs of the off-diagonal elements of the property matrix and further to random signs of the diabatic couplings (see figure 2d). While both these issues can be easily resolved manually by

inspection in one or two dimensions, it is impossible for a general multidimensional system.

Within our framework, we diabaticize each geometry separately using property-based diabaticization, 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 diabaticization 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 diabaticization, we propose here a series of simple diabaticization 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 diabaticization, 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.²⁰

We focus here on the prediction of PESs but other properties can be treated in the same fashion. 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. While this is just a simple example, 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.²¹⁻²³ 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.

Computational methods

Property-based diabaticization

Within the Born-Oppenheimer approximation, the total Hamiltonian is separated into the nuclear kinetic energy and the rest called the electronic Hamiltonian. The eigenvectors of the electronic Hamiltonian for fixed positions of nuclei are called electronically adiabatic states and the eigenvalues are called adiabatic PESs. Nuclear geometries with degenerate adiabatic energies of two interacting states form a seam, a conical intersection. For two states of the same spin multiplicity, the dimensionality of the seam space is smaller by two than the dimensionality of the system represented in internal coordinates, that is, $N - 8$ for a nonlinear molecule with N atoms.²⁴ 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 between electronic states take place. What we usually call an avoided crossing when following a trajectory, is nothing but a shoulder of a conical intersection.²⁵

When we aim to describe photochemical processes or processes involving excited states in general, we usually have to go beyond the Born-Oppenheimer approximation. To do so, we need to calculate the probabilities of radiationless transitions which are usually expressed via nonadiabatic couplings (NACs) for states of the same spin multiplicity and via spin-orbit couplings (SOCs) for states of different multiplicity. However, these couplings are often expensive to compute and difficult to converge. Another problem is that NACs exhibit singularities at conical intersection seams.²⁶

Since the adiabatic representation shows cuspidal ridges in PESs and singularities in NACs in the vicinity of conical intersections, it might be advantageous to switch to a different representation by applying a geometry-dependent unitary transformation matrix $\mathbf{T}(\mathbf{R})$.²⁷

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

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

where $\Psi_j^{\text{ad}}(\mathbf{r}; \mathbf{R})$ are the original adiabatic wavefunctions, $\Psi_i^{\text{d}}(\mathbf{r}; \mathbf{R})$ 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 Truhlar²⁸ showed in 1982 that strictly diabatic electronic basis cannot, in general, exist for polyatomic molecules with three or more atoms. Therefore, we have to settle with a basis that provides smooth elements of PEM and sufficiently small NACs. We call this 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. There are dozens of various diabatization schemes based on NACs elimination, wavefunction smoothness, or properties smoothness.^{14,29-39} Unfortunately, the current schemes are usually limited to low-dimensional systems or a specific wavefunction-based method, require expert knowledge about the system and lots of manual work or computational resources. Lately, another category of ML-based diabatization schemes emerged.^{13,15-19} These approaches are usually based on neural networks with implicit diabatization layer and they learn the diabatization globally from a large training dataset. Also, a manual selection of reference geometries is often necessary.

Property-based diabatization schemes based on property unblending are the simplest and cheapest to apply. However, they have several problems which prevent their wider use. First of all, it is necessary to select such properties that can distinguish between all the included states. One can assume that a sufficiently large set of properties should

be able to identify the states and a large variety of different methods have been actually proposed.^{33,35,40–43} Another problem is connected with random signs of properties arising from two different states as mentioned earlier. This is a general problem present also for other diabaticization schemes. Here, it results in random sign changes of the obtained diabatic couplings, that is, the off-diagonal elements of diabatic PEM. There is also a third problem specific to property-unblending diabaticization: since each nuclear geometry is diabaticized separately via diagonalization or a similar procedure, the order of diabatic states is not consistent throughout the configuration space. The latter two problems can be easily solved manually by inspection in one dimension. However, they present a sore issue for a general multidimensional system. Here, we combine property-based diabaticization with ML-based clustering to overcome these limitations.

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 diabaticization 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:¹⁴

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

where \hat{P} is the property operator and the objective function includes squared differences between its expectation values.

It has been shown that the maximization in the eq. 3 is equivalent to the maximization of the sum of expectation values of the diabatic states:⁴⁰

$$f = \sum_i \left| \langle \Psi_i^d(\mathbf{r}; \mathbf{R}) | \hat{P} | \Psi_i^d(\mathbf{r}; \mathbf{R}) \rangle \right|^2 \quad (4)$$

which can be in turn expressed in the adiabatic basis using the expansion from eq. 1:

$$f(\mathbf{T}) = \sum_i \left| \sum_{j,k} T_{ji}(\mathbf{R}) T_{ki}(\mathbf{R}) \langle \Psi_j^{\text{ad}}(\mathbf{r}; \mathbf{R}) | \hat{P} | \Psi_k^{\text{ad}}(\mathbf{r}; \mathbf{R}) \rangle \right|^2 \quad (5)$$

It is important to note that the separation of matrix eigenvalues can be achieved by diagonalization.¹⁴ 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 the diabaticization.

The methodology proposed in this paper can be in principle connected with an arbitrary property-based diabaticization method to extend its applicability to multidimensional problems. Nevertheless, we also propose here a series of simple and pragmatic property-based diabaticization methods. The reasoning behind our methods is similar to the dipole-quadrupole⁴⁰ (DQ) diabaticization: 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.^{33,35,40-42} 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 is 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) diabaticization depending on the highest multipole included. The property matrix \mathbf{P} is then formed according to the following formulas, respectively:

$$P_{ab}^{\text{tD}} = \boldsymbol{\mu}_{0a} \cdot \boldsymbol{\mu}_{0b} \quad (6)$$

$$P_{ab}^{\text{tDQ}} = \boldsymbol{\mu}_{0a} \cdot \boldsymbol{\mu}_{0b} + w_{\text{Q}} \langle \mathbf{Q}_{0a}, \mathbf{Q}_{0b} \rangle_{\text{F}} \quad (7)$$

$$P_{ab}^{\text{tDQO}} = \boldsymbol{\mu}_{0a} \cdot \boldsymbol{\mu}_{0b} + w_{\text{Q}} \langle \mathbf{Q}_{0a}, \mathbf{Q}_{0b} \rangle_{\text{F}} + w_{\text{O}} \langle \mathbf{O}_{0a}, \mathbf{O}_{0b} \rangle_{\text{F}} \quad (8)$$

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^{44,45} code, version 2.0.1. $\langle \cdot, \cdot \rangle_{\text{F}}$ is the Frobenius inner product, that is, the sum over the element-wise product. The weights w_{Q} and w_{O} can be set by hand or optimized within cross-validation or a similar procedure. However, we do not use here this flexibility and set all weights to 1.

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 \mathbf{P} , calculate the matrix of eigenvectors, and use it as the transformation matrix in eq. 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.

A separate problem for diabaticization in general is the selection of the excited-state manifold which does not interact with other lower or higher-lying states. While this issue is not sufficiently discussed in the community, only small archetypal systems are often used to test diabaticization schemes. However, we now know that conical intersections are not rare at all, on the contrary.²⁴ It is, therefore, very often difficult or even impossible to select a reasonable manifold of states for the given configuration space of interest. While the number of interacting states might be decreased by the reduction of the active space for multireference methods, it doesn't really solve the problem. The selections of the excited-state manifold and the active space are indeed analogical and interconnected. While we do not solve here this separate problem, our ML algorithm might be a first step toward this goal as we might fit more diabatic states than the number of input adiabatic states in the future. Implicitly fitting a larger number of diabatic states within neural network architecture has been already shown to improve prediction accuracy.¹⁹

ML-based reordering

Eventually, we want to correct the deficiencies of property-based diabaticization 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 (see figure 2a and 2b for a real 1D example)? 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 was positive, then we would be able to get better ML predictions without any underlying property-based diabaticization. Also, such an algorithm can directly diabaticize 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 figure 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

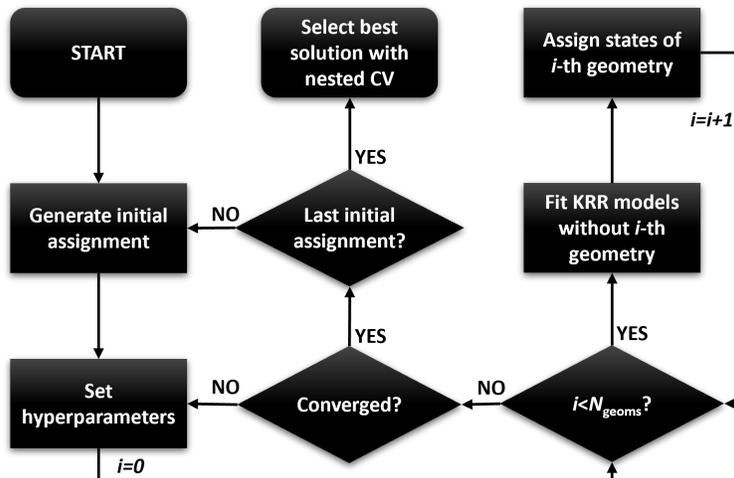


Figure 1: Flowchart of the proposed ML-based reordering algorithm.

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 algorithm⁴⁶ minimizing the mean squared error as implemented in the SciPy⁴⁷ 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.⁴⁸

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 dif-

ferent 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.

ML for property-based diabaticization

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 diabaticization 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 \mathbf{B} represented by a signed permutation matrix \mathbf{S} , which minimizes the Frobenius norm to the predicted PEM \mathbf{A} from ML models trained without that particular geometry:

$$\min_{\mathbf{S} \in \mathcal{S}} \|\mathbf{A} - \mathbf{SBS}^T\|_F = \max_{\mathbf{S} \in \mathcal{S}} \text{Tr}(\mathbf{A}^T \mathbf{SBS}^T) \quad (9)$$

where $\|\cdot\|_F$ 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.⁸ 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.

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:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I} d(i, j) \quad (10)$$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \quad (11)$$

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:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (12)$$

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:

$$d_{\text{ol}}(i, j) = 1 - |\langle \Psi_i | \Psi_j' \rangle| \quad (13)$$

where Ψ_i and Ψ_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:

$$\Psi_i = \sum_k c_{ik} \Phi_k \quad (14)$$

where c_{ik} are the CI expansion coefficients into Slater determinants Φ_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:

$$\langle \Psi_i | \Psi'_j \rangle = \sum_k \sum_l c_{ik} c'_{jl} \langle \Phi_k | \Phi'_l \rangle \quad (15)$$

The overlap between two Slater determinants can be in turn expressed as a determinant containing overlaps between the constituting molecular orbitals (MOs):^{49,50}

$$\langle \Phi_k | \Phi'_l \rangle = \begin{vmatrix} \langle \phi_{k1} | \phi'_{l1} \rangle & \dots & \langle \phi_{k1} | \phi'_{ln} \rangle \\ \vdots & \ddots & \vdots \\ \langle \phi_{kn} | \phi'_{l1} \rangle & \dots & \langle \phi_{kn} | \phi'_{ln} \rangle \end{vmatrix} \quad (16)$$

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.

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. 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.

We can identify two major categories of ML algorithms for nonlinear regression in the field.^{2,51} One category consists of kernel methods.⁵² Kernel methods use the so-called kernel trick, which allows using linear regression algorithms to model nonlinear problems through an implicit transformation of the input data into a higher-dimensional space.⁵³ Among these methods, the KRR method is quite simple and frequently used in quantum chemistry.^{53,54} Neural networks, popularized due to their versatility and amenability for GPU hardware, represent another valid category.² However, the design of their architecture as well as the identification of proper training protocols is more complicated and requires substantial experience and time investment.⁵⁵ Also, kernel methods are often more suitable for small datasets and we want to work with as few as hundreds of samples.

A crucial ingredient for the prediction of molecular properties is a molecular representation or molecular descriptors, that is, a feature vector encoding the system, usually via the molecular structure.^{55,56} It should fulfil some basic requirements for ML to be efficient: it should be efficient to calculate (i.e. compact) and devoid of redundant information: it should possess translational, rotational, and permutational invariance.^{57,58} We usually want it to be also differentiable, unique, and universal. When possible, constant size descriptors are desirable as they offer improved scalability throughout the chemical space.⁵⁹

Molecular representations typically encode either the entire system or each atom in its chemical environment.⁵⁷ One of the simplest representations from the first category is the Coulomb matrix⁶⁰ and the bag of bonds.⁶¹ Popular representations from the second cat-

egory are, for example, the smooth overlap of atomic positions,⁶² atom-centre symmetry functions,⁶³ or the Faber–Christensen–Huang–Lilienfeld⁵⁶ representation. Note that these are just examples as many alternatives have been presented.^{58,64,65} One can even generate functional descriptors automatically via deep learning during the training process⁶⁶ or separately by using a special type of neural network called autoencoder.⁶⁷

In our case, the KRR method is the favourable choice because of its simplicity and efficiency for small training samples. We create a separate KRR model for each PES or PEM element separately. 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:⁶⁸

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}) \quad (17)$$

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. The Gaussian and Laplace kernels are especially popular in chemistry.⁵³ We use here the Gaussian kernel which has the following form:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right) \quad (18)$$

where $\|\mathbf{x}\|_2$ is the Euclidean norm and σ is a model parameter.

The regression coefficients can be obtained from the training data by the following minimization:⁶⁸

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 + \lambda \sum_{i=1, j=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}_j) \alpha_j \right) \quad (19)$$

The first term is a common residual sum of squares. The second term including another model parameter λ is responsible for the regularization: it is supposed to prevent overfitting of the training data. This minimization has a closed-form solution:

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y} \quad (20)$$

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 parameters σ and λ are called hyperparameters and they are usually tuned on a grid using either the cross-validation or a separate validation set.⁵⁴ If we have enough data, we can use a separate validation set (different from the training set) on which we evaluate KRR models trained with different combinations of hyperparameters. The combination which provides the smallest error is then selected. Alternatively, it is possible to divide the training set into k subsets, use $(k - 1)$ subsets for the training, and evaluate the performance for the last one.⁶⁸ By repeating this process k times, each time leaving out a different subset, we get the k -fold cross-validation. We select the combination of parameters that provides the smallest average error over the left-out subsets. The 10-fold cross-validation is used here.

The crucial part of chemical applications of ML algorithms is the selection of molecular representation.⁵⁴ By working with nuclear configurations of a single molecular entity, some of its desired properties are automatically fulfilled. Namely, the number of atoms is constant, resulting in a constant-size molecular representation. Also, we can simply choose a fixed order of atoms when sampling the configuration space, ensuring permutational invariance with respect to different elements. With fixed atomic order, permutations do not play any role in the FNO molecule. The results for the formaldehyde molecule could be slightly better when taking into account two possible permutations of the hydrogen atoms using for example the permutationally invariant kernel⁶⁹ but we opted here for simplicity. We used a simple vector of normalized inverted internuclear distances:^{70,71}

$$\mathbf{x} = \left(\begin{array}{l} r_{i,j}^{\text{ref}} \\ r_{i,j} \end{array} \text{ for } 1 < i \leq N \text{ for } j < i \right) \quad (21)$$

where $r_{i,j}$ is the Euclidean distance between atoms i and j and $r_{i,j}^{\text{ref}}$ 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.

Computational details

The molecules were optimized at the B3LYP/6-31g* level with subsequent vibrational analysis on the same level using Gaussian G09,⁷² revision D.01. 1000 nuclear configurations for each molecule were subsequently sampled using the harmonic approximation and the temperature-dependent Wigner quasiprobability distribution:^{73,74}

$$P_{\text{W}}(\mathbf{q}, \mathbf{p}, T) = \prod_i \frac{1}{\pi \hbar} \tanh\left(\frac{\hbar \omega_i}{2k_{\text{B}}T}\right) \exp\left(\tanh\left(\frac{\hbar \omega_i}{2k_{\text{B}}T}\right) \left(-\frac{p_i^2}{\mu_i \hbar \omega_i} - \frac{\mu_i \omega_i q_i^2}{\hbar}\right)\right), \quad (22)$$

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.

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 which are 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^{44,45} 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.

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 which do not interact with other higher or lower-lying states at the employed level of theory.

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 S₄-S₆ 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 and crosses them without any interaction (see figure 2a). We can directly apply the ML reordering algorithm without prior diabaticization (see figure 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 diabaticization. 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

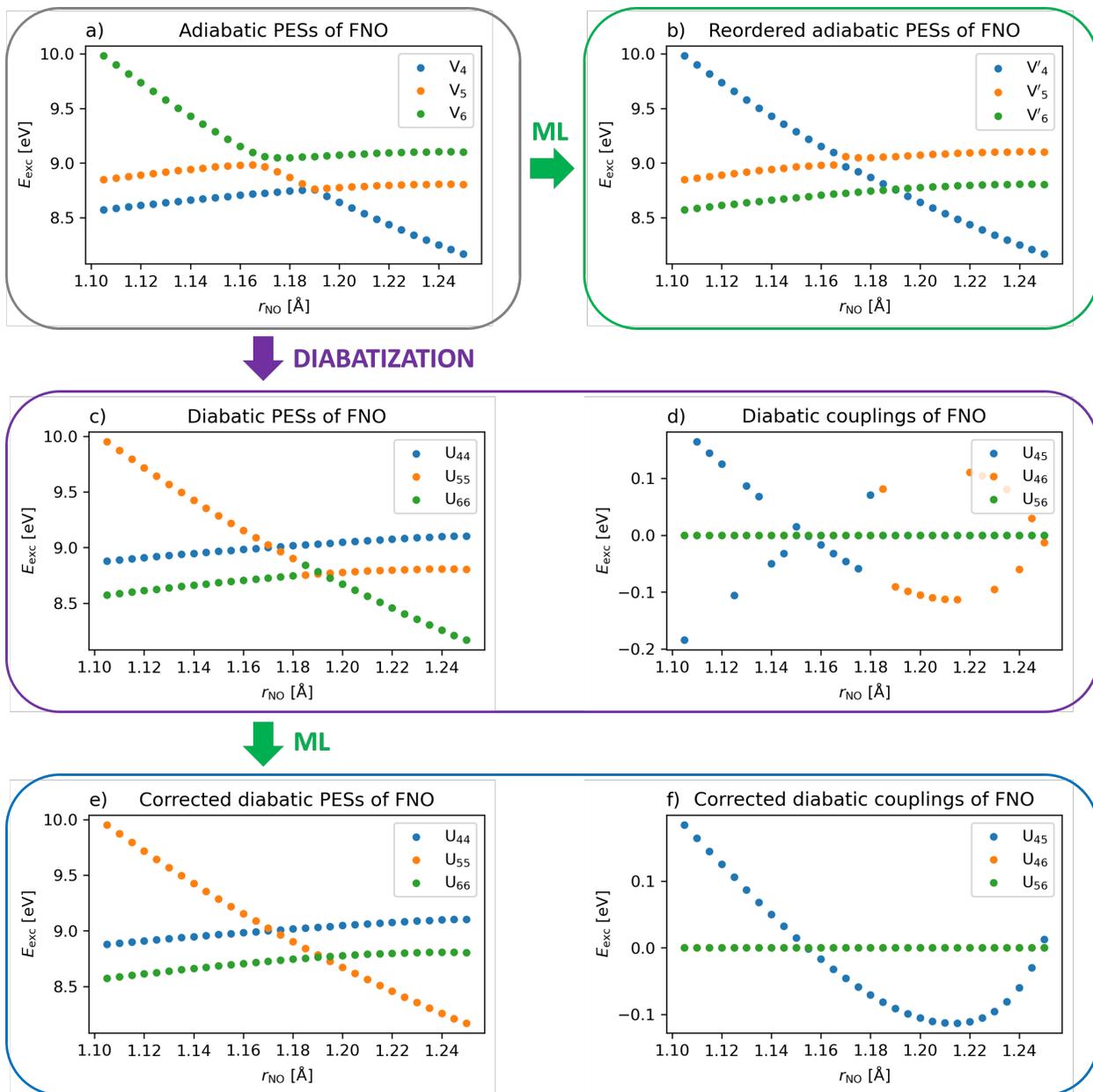


Figure 2: Excited states of the FNO molecule along the NO bond in the a) adiabatic basis, b) reordered adiabatic basis, c-d) diatomic basis, and e-f) reordered and sign-corrected diatomic basis.

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 diabaticization scheme as outlined in section . In this case, we need to distinguish only two states of the same symmetry along one coordinate so the property-unblending diabaticization using just the transition dipole moments from the ground state is sufficient. Figure 2c displays the diagonal elements of the diabatic PEM while figure 2d displays the off-diagonal elements, that is, the diabatic couplings. We can directly see the two problems of property-based diabaticization: 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 (see figures 2e and 2f). 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 diabaticization to multidimensional problems as shown below.

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. Figure 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.⁷⁵ We defined the distances the same way as in eq. 13 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.

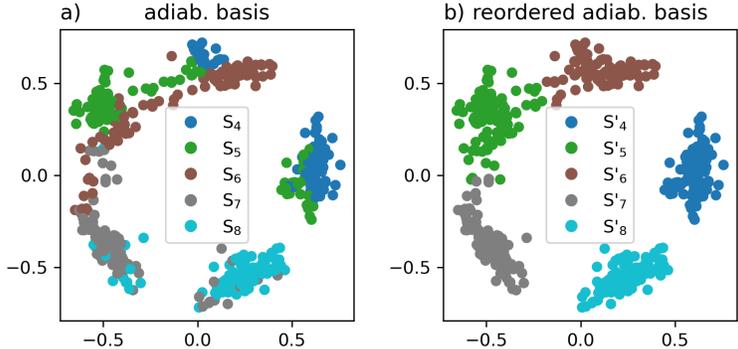


Figure 3: Multidimensional scaling projection of excited state clusters before and after re-ordering based on wavefunction overlaps for 100 nuclear configurations and 5 excited states. The projection corresponds to a 2D space in which the wavefunction overlaps are preserved as well as possible.

To provide insight, let us first look at wavefunction-based clustering which serves here for visualization and benchmark purposes. Figure 3b shows the same projection after we applied the wavefunction-based clustering described in section . 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.

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

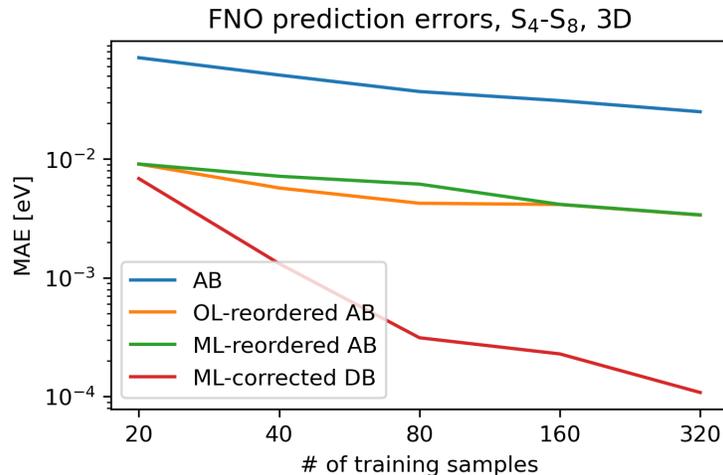


Figure 4: The mean absolute error of the kernel ridge regression for the FNO molecule as a function of training set size for adiabatic basis (AB), adiabatic basis reordered using wavefunction overlaps (OL), ML-reordered adiabatic basis, and values obtained from the diagonalization of ML-corrected diabatic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

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 figure 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.

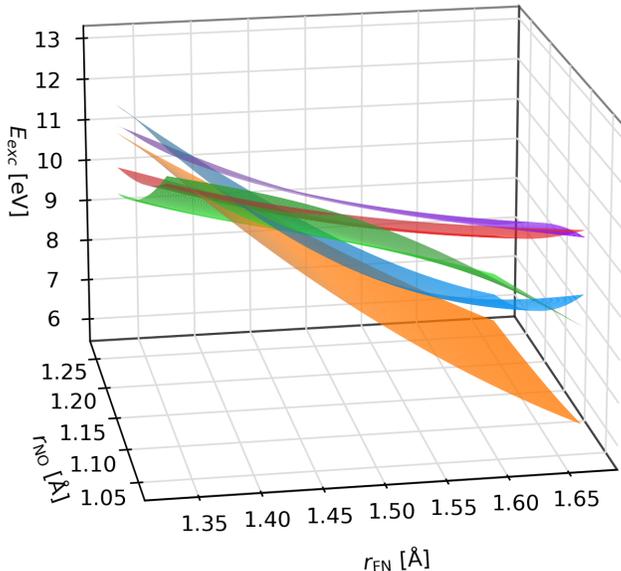


Figure 5: Potential energy surfaces obtained by the proposed ML approach based on the tDQ diabaticization of the S4-S8 adiabatic states of the FNO molecule. The bond angle is fixed to 110° . ML models were trained on 320 geometries from the Wigner distribution.

To inspect how the diabatic states look like, we plot their PESs in figure 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.

Formaldehyde: 6D case

We repeated the whole procedure for the formaldehyde molecule where we selected the tDQO property-based diabaticization as the tDQ diabaticization did not improve the learning. We could have in principle diabaticized 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. The MAEs for both adiabatic reordering approaches and the diabatic ML approach are presented in figure 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 diabaticization. Both adiabatic reordering approaches decrease the prediction errors by up to half an order of magnitude and provide again very similar results.

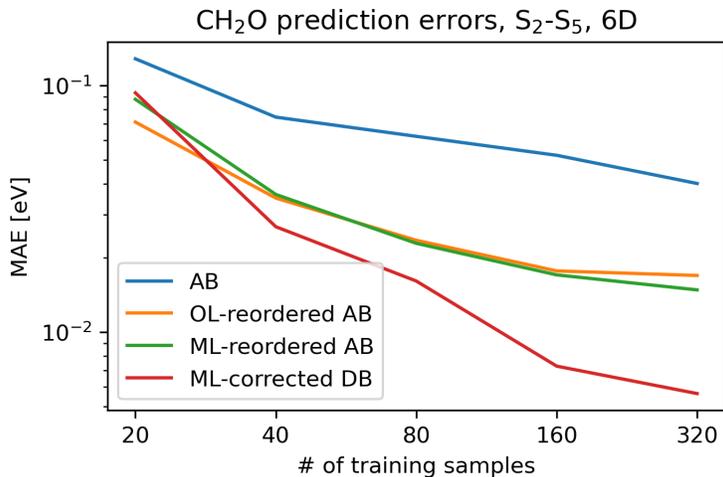


Figure 6: The mean absolute error of the kernel ridge regression for the formaldehyde molecule as a function of training set size for adiabatic basis (AB), adiabatic basis reordered using wavefunction overlaps (OL), ML-reordered adiabatic basis, and values obtained from the diagonalization of ML-corrected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

Conclusions

We tackled two different problems at once: efficient machine learning for excited-state properties and diabaticization. We proposed and tested methodology for correcting deficiencies of property-based diabaticization 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 multidimensional 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 PESs and couplings. 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. 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. We focused on small training sets including from dozens up to hundreds of nuclear geometries.

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 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 also 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 based on metaheuristics could be used. The methodology

can be also 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 also opens the way to other possible applications. Wrong state ordering was identified as a possible problem when learning differences between two electronic structure methods within Δ -ML.⁶ 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 the selection of an independent excited-state manifold for diabatization by fitting a larger number of diabatic states than the number of input adiabatic states.

Acknowledgement

The support of Czech Science Foundation project No. 20-15825S is gratefully acknowledged. 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 Union’s 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.

Supporting Information Available

Sampled geometries for both molecules, training indices, and calculated excitation energies and transition moments.

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) G. Cerullo, D. Polli, G. Lanzani, S. De Silvestri, H. Hashimoto and R. J. Cogdell, Photosynthetic light harvesting by carotenoids: Detection of an intermediate excited state, *Science*, 2002, **298**, 2395–2398.
- (4) N. M. O’Boyle, C. M. Campbell and G. R. Hutchison, Computational Design and Selection of Optimal Organic Photovoltaic Materials, *J. Phys. Chem. C*, 2011, **115**, 16200–16210.
- (5) C. Rauer, J. J. Nogueira, P. Marquetand and L. González, Cyclobutane Thymine Photodimerization Mechanism Revealed by Nonadiabatic Molecular Dynamics, *J. Am. Chem. Soc.*, 2016, **138**, 15911–15916.
- (6) 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.
- (7) J. Westermayr, M. Gastegger, M. F. Menger, S. Mai, L. González and P. Marquetand, Machine learning enables long time scale molecular photodynamics simulations, *Chem. Sci.*, 2019, **10**, 8100–8107.
- (8) 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.

- (9) D. Hu, Y. Xie, X. Li, L. Li and Z. Lan, Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation, *J. Phys. Chem. Lett.*, 2018, **9**, 2725–2732.
- (10) C. Duan, F. Liu, A. Nandy and H. J. Kulik, Semi-supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost, *J. Phys. Chem. Lett.*, 2020, **11**, 6640–6648.
- (11) W. S. Jeong, S. J. Stoneburner, D. King, R. Li, A. Walker, R. Lindh and L. Gagliardi, Automation of Active Space Selection for Multireference Methods via Machine Learning on Chemical Bond Dissociation, *J. Chem. Theory Comput.*, 2020, **16**, 2389–2399.
- (12) A. W. Kohn, Z. Lin and T. Van Voorhis, Toward Prediction of Nonradiative Decay Pathways in Organic Compounds I: The Case of Naphthalene Quantum Yields, *J. Phys. Chem. C*, 2019, **123**, 15394–15402.
- (13) D. M. Williams and W. Einfeld, Neural network diabaticization: A new ansatz for accurate high-dimensional coupled potential energy surfaces, *J. Chem. Phys.*, 2018, **149**, 204106.
- (14) Y. Shu, Z. Varga, S. Kanchanakungwankul, L. Zhang and D. G. Truhlar, Diabatic States of Molecules, *J. Phys. Chem. A*, 2022, **126**, 992–1018.
- (15) Y. Shu and D. G. Truhlar, Diabatization by Machine Intelligence, *J. Chem. Theory Comput.*, 2020, **16**, 6456–6464.
- (16) 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 $1,2^1A$ States of Ammonia, *J. Chem. Theory Comput.*, 2020, **16**, 302–313.

- (17) 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.
- (18) 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 $^2A'$ states of LiFH, *Phys. Chem. Chem. Phys.*, 2019, **21**, 14205–14213.
- (19) 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.
- (20) 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.
- (21) 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.
- (22) 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.
- (23) B. X. Xue, M. Barbatti and P. O. Dral, Machine Learning for Absorption Cross Sections, *J. Phys. Chem. A*, 2020, **124**, 7199–7210.
- (24) D. R. Yarkony, Conical Intersections: The New Conventional Wisdom, *J. Phys. Chem. A*, 2001, **105**, 6277–6293.
- (25) D. G. Truhlar and C. A. Mead, Relative likelihood of encountering conical intersections and avoided intersections on the potential energy surfaces of polyatomic molecules, *Phys. Rev. A*, 2003, **68**, 032501.

- (26) B. K. Kendrick, C. Alden Mead and D. G. Truhlar, Properties of nonadiabatic couplings and the generalized Born-Oppenheimer approximation, *Chem. Phys.*, 2002, **277**, 31–41.
- (27) F. T. Smith, Diabatic and Adiabatic Representations for Atomic Collision Problems, *Phys. Rev.*, 1969, **179**, 111–123.
- (28) 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.
- (29) 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.
- (30) H. Köppel, Regularized diabatic states and quantum dynamics on intersecting potential energy surfaces, *Faraday Discuss.*, 2004, **127**, 35–47.
- (31) 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.
- (32) 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.
- (33) C. E. Hoyer, K. Parker, L. Gagliardi and D. G. Truhlar, The DQ and DQ Φ electronic structure diabaticization methods: Validation for general applications, *J. Chem. Phys.*, 2016, **144**, 194101.
- (34) Z. Varga, K. A. Parker and D. G. Truhlar, Direct diabaticization based on nonadiabatic couplings: the N/D method, *Phys. Chem. Chem. Phys.*, 2018, **20**, 26643–26659.

- (35) 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.
- (36) 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.
- (37) N. Wittenbrink, F. Venghaus, D. Williams and W. Eisfeld, A new approach for the development of diabatic potential energy surfaces: Hybrid block-diagonalization and diabaticization by ansatz, *J. Chem. Phys.*, 2016, **145**, 184108.
- (38) G. J. Atchity and K. Ruedenberg, Determination of diabatic states through enforcement of configurational uniformity, *Theor. Chem. Acc.*, 1997, **97**, 47–58.
- (39) K. R. Yang, X. Xu and D. G. Truhlar, Direct diabaticization 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.
- (40) 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.
- (41) 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.
- (42) 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.
- (43) T. Karman, A. Van Der Avoird and G. C. Groenenboom, Communication: Multiple-

- property-based diabaticization for open-shell van der Waals molecules, *J. Chem. Phys.*, 2016, **144**, 121101.
- (44) 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, *WIREs Comput. Mol. Sci.*, 2018, **8**, e1340.
- (45) 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.
- (46) D. F. Crouse, On implementing 2D rectangular assignment algorithms, *IEEE Trans. Aerosp. Electron. Syst.*, 2016, **52**, 1679–1696.
- (47) 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, Í. 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.

- (48) P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Pearson Education Limited, Harlow, 2014.
- (49) 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.
- (50) F. Plasser, M. Ruckebauer, 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.
- (51) J. Westermayr, F. A. Faber, A. S. Christensen, O. A. von Lilienfeld and P. Marquetand, Neural networks and kernel ridge regression for excited states dynamics of CH₂NH₂⁺: From single-state to multi-state representations and multi-property machine learning models, *Mach. Learn. Sci. Technol.*, 2020, **1**, 025009.
- (52) B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, Cambridge, 2002.
- (53) M. Rupp, Machine learning for quantum mechanics in a nutshell, *Int. J. Quantum Chem.*, 2015, **115**, 1058–1073.
- (54) M. Rupp, O. A. von Lilienfeld and K. Burke, Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry, *J. Chem. Phys.*, 2018, **148**, 241401.
- (55) 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.
- (56) 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.
- (57) 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.
- (58) 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.
- (59) C. R. Collins, G. J. Gordon, O. A. von Lilienfeld and D. J. Yaron, Constant size descriptors for accurate machine learning models of molecular properties, *J. Chem. Phys.*, 2018, **148**, 241718.
- (60) M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, *Phys. Rev. Lett.*, 2012, **108**, 058301.
- (61) K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.
- (62) A. P. Bartók, R. Kondor and G. Csányi, On representing chemical environments, *Phys. Rev. B*, 2013, **87**, 184115.
- (63) J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, *J. Chem. Phys.*, 2011, **134**, 074106.
- (64) C. Liang, G. Tocci, D. M. Wilkins, A. Grisafi, S. Roke and M. Ceriotti, Solvent fluctua-

- tions and nuclear quantum effects modulate the molecular hyperpolarizability of water, *Phys. Rev. B*, 2017, **96**, 041407.
- (65) W. Pronobis, A. Tkatchenko and K.-R. Müller, Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules, *J. Chem. Theory Comput.*, 2018, **14**, 2991–3003.
- (66) 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.*, 2018, **148**, 241722.
- (67) R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *ACS Cent. Sci.*, 2018, **4**, 268–276.
- (68) 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.
- (69) A. P. Bartók and G. Csányi, Gaussian approximation potentials: A brief tutorial introduction, *Int. J. Quantum Chem.*, 2015, **115**, 1051–1057.
- (70) 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.
- (71) P. O. Dral, MLatom: A program package for quantum chemical research assisted by machine learning, *J. Comput. Chem.*, 2019, **40**, 2339–2347.

- (72) 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.
- (73) M. Hillery, R. O’Connell, M. Scully and E. Wigner, Distribution functions in physics: Fundamentals, *Phys. Rep.*, 1984, **106**, 121–167.
- (74) 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.
- (75) A. Mead, Review of the Development of Multidimensional Scaling Methods, *J. R. Stat. Soc. Ser. D*, 1992, **41**, 27–39.