Molecular excited states through a machine learning lens

Pavlo O Dral, Mario Barbatti

This is a post-peer-review, pre-copyedit version of an article published in Nature Reviews Chemistry. The final authenticated version is available online at: http://dx.doi.org/10.1038/s41570-021-00278-1

Molecular excited states through a machine learning lens

Pavlo O. Dral[a†] and Mario Barbatti[b†]

[a]*State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Department of Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China*

[b]*Aix Marseille University, CNRS, ICR, Marseille, France*

E-Mail: dral@xmu.edu.cn

E-mail: mario.barbatti@univ-amu.fr

Abstract | Theoretical simulations of electronic excitations and associated processes in molecules are indispensable for fundamental research and technological innovations. However, such simulations are notoriously challenging to perform with quantum mechanical (QM) methods. Advances in machine learning (ML) open many new avenues for assisting molecular excited-state simulations. In this Review, we track such progress, assess the current state-of-the-art and highlight the critical issues to solve in the future. We overview a broad range of ML applications in excited-state research, which include the prediction of molecular properties, improvements of QM methods for the calculations of excited-state properties and the search of new materials. ML approaches can help us understand hidden factors that influence photo-processes, leading to a better control of such processes and new rules for the design of materials for optoelectronic applications.

The Schrödinger equation for a fixed nuclear configuration of a molecule predicts that the electrons can only occupy specific quantum configurations, named electronic states. The state corresponding to the lowest energy is called ground state and all the others are excited electronic states. In this Review, we will refer to the latter merely as excited states as we exclusively focus on this type of excitations.

Excited states are at the core of a myriad of phenomena.[1,2] They are crucial for Earth's biosphere, being responsible for the first step in harvesting the energy in sunlight through

photosynthesis.[3] They are equally vital for vision and bioluminescence.[4] They play a role in mutagenesis and carcinogenesis.[5] In technology, excited states are at the basis of optoelectronics, with applications in photovoltaics and light-emitting diodes (LEDs), for example.[6-8] In the lab, many advanced and routine analytic chemistry tools depend on them.[9,10]

Under low levels of radiation, molecules are usually in their ground state. However, excited states can be activated in diverse ways, through photo-absorption, particle collisions, or bond dissociation. Photo-excited states can be populated using a wide range of wavelengths. For example, in organometallic compounds, electrons may be excited by infrared (IR) and visible (vis) radiation,[11] whereas in organic materials, they are typically excited by radiation in the visible to ultraviolet (UV) region.[12] Inner-shell electrons are excited by X-rays.[13] Upon excitation, a molecule remains in the excited states' manifold until it either returns to the ground state or forms a new compound, through isomerization, dissociation, bond rearrangement, or reaction with other molecules. During this process, the photoenergy heats vibrational modes through either internal conversion or intersystem crossing and, occasionally, it may be reemitted through either fluorescence or phosphorescence. In many cases, the nuclear conformation can dramatically change, reaching regions where the Born–Oppenheimer approximation loses validity (conical intersections**[G]**).[14] Depending on the system, the excited-state relaxation may take from tens of femtoseconds (internal conversion) to a few seconds (phosphorescence).[15] Many computational methods, such as complete active space perturbation theory to second order (CASPT2) or time-dependent density-functional theory (TD-DFT), are specifically developed to predict molecular excited states (see Box 1 for a brief description of these and few other popular methods in excited-state research). When simulating excited states, we may be interested in time-independent features such as potential energy surfaces (PES), or in knowing how the molecule evolves with time.[16,17] We may also be interested in statistical treatments, for example, for the prediction of reaction rates.[18] Often, we need a combination of approaches, for instance, when we propagate dynamics using static properties.

The electronic density of excited states is exceedingly more complex than that of the ground state. It may involve unpaired electrons and multiple electron excitations. Moreover, while the ground state may be isolated from the other states by few eV, excited states tend to bundle in narrow spectral ranges, mixing their characters and exchanging their order in response to small nuclear displacements.[16] These features make predicting excited states strikingly more

challenging and costly than ground states (FIG. 1). Whenever the ground state shares such features, such as in radicals or metal complexes, its prediction becomes problematic, too.[19]

The accuracy of the QM calculations of molecular excited-state energies in routine calculations with commonly used methods is low, typically 0.2 eV.[20,21] Making things worse, the errors are not systematic and the accuracy may depend on the state type. For instance, while localized states may be predicted well, charge-transfer states are often misplaced by TD-DFT.[22] This type of imbalance may lead to wrong topographies of the excited-state PES. The accuracy of the simulations has to be further sacrificed when a great number of excited-state calculations need to be performed, such as in dynamics simulations or high-throughput virtual screening (HTVS) of new chromophores. In these cases, more accurate QM approaches are too computationally expensive, and more affordable but less accurate methods are required. Machine learning (ML) opens the perspective of drastically reducing these costs without compromising the accurate description of excited states in different systems and for different properties (FIG. 2). The field of ML is itself rapidly evolving, with the development of a variety of approaches for simulations, analysis and design of molecular systems (see Box 2).[23-25]

ML for excited states was used for the first time in 1990 to aid the identification of compounds structures from their UV absorption spectra.[26] At that time, neural networks (NN) were gaining attention in chemistry and related fields.[27] Around the same time, another study was reported in which fluorescence spectra were used for a similar task.[28] Unfortunately, this initial excitement about NNs was followed by a rather long hiatus of almost two decades (FIG. 3). The technology was apparently not ripe enough for widespread use, and the community was not convinced of the usefulness of ML in chemistry. Nevertheless, ML found its niche in the field of drug design, where it is currently used in the context of quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR).

The application of a special type of ML technique called supervised learning (Box 3) for drug design influenced later studies on excited states, in which ML was applied to find the relationship between molecular structure and excited-state properties ranging from atomic- (such as molecular excitation energy) to macroscopic-scale (such as power conversion efficiency of a solar cell) properties. Solving this task with ML dominates modern studies that appear at an ever-increasing pace since 2015, as evident from the literature surveyed in this Review.

Supervised learning enables extremely fast predictions of desired excited-state properties with little loss of accuracy.[29] For example, ML trained on excitation energies — the most fundamental excited-state property — as a function of structure can be used for subsequent spectrum calculations,[30-35] dynamics simulations,[23,29,36-46] or to search for new materials that emit at a desired wavelength.[47-49] The excited-state properties can be calculated with a QM method and then used to train a ML model, which then can be used as a surrogate for the QM method. ML can also be trained on experimentally measured data, making the ML approaches more accurate than affordable QM methods for the prediction of experimental properties.[50] Furthermore, experimental observables, such as the power conversion efficiency of a solar cell, which are extremely difficult (if not impossible) to simulate with QM methods, can be learned and predicted for new molecules with ML .[51-64] One particular ML application in quantum chemistry that stands out because it changes the very basis of excited-state simulations is the use of supervised learning for the improvement of QM methods themselves, even by learning the wavefunctions.[23,65] For example, speed and accuracy of configuration interaction (CI) approach can be drastically improved by using ML to select important configurations.[66] However, there are, not many studies yet reporting the use of such ML-improved QM methods for excited-state simulations. One notable example is a recent work in which kernel ridge regression (KRR, Box 3) was used to calculate atomic charge instead of self-consistent charges in TD-DFTB (see Box 1) that, in turn, was used to compute the excited-state potential energies.[67] Given the abundance of QM methods for excited states and various examples of ML-improved QM methods,[23,65] it is a matter of time until the latter become routine tools for computational chemists.

Excited-state research benefits not just from supervised learning but also from other types of ML, which also started gaining attention. Unsupervised learning (Box 2) is useful to gain new insights from existing data — experimental and theoretical — that otherwise would be impossible or cumbersome to obtain by manual inspection.[24,30,68-72] One of the first studies using unsupervised learning for excited states was reported in 2001,[68] in which it was applied to understand the influence of microenvironment factors such as solvent accessibility and packing density on the fluorescence of proteins. A special class of unsupervised learning techniques, generative models, as well as reinforcement learning can be used to generate new compounds with desired excited-state properties automatically.[73-78]

Semi-supervised learning[79] is another ML approach, which is currently underappreciated in excited-state research. It leverages the advantages of both supervised and unsupervised

learning and learns from the data set in which only part of the data is labeled.[79] The only example of its use in excited-state research is a recent model developed[80] for fast and accurate determination of the multireference character. Such a diagnostic is extremely useful for choosing an appropriate QM method.

In this Review, we discuss how ML techniques have been and can be used in excited-state simulations, particularly for predicting excited-state energies, simulating spectra, constructing surfaces and performing dynamics simulations, and designing optoelectronic materials. Our primary focus is on molecular systems, but we will adopt fuzzy boundaries, occasionally discussing supramolecular assemblies and solids.

[H1] Energies

Understanding how the energy of the electronic states depends on chemical composition and spatial arrangement of atoms is fundamental for a deep comprehension and control of the photo-processes. Thus, the development of ML models for learning excited-state energies as a function of the structure is getting attention. ML approaches not only can be incredibly faster than QM methods, but can also learn on experimental data and thus be more accurate than commonly used QM methods (assuming the experimental error is itself low enough).[49,81]

A universal ML model would instantaneously predict the energy of multiple electronic states for any molecule in any configuration with the same accuracy of a high-level QM method. Such a model is yet to be developed because there are obstacles on many fronts and only partial solutions have been suggested so far. This is in contrast with the much more substantial progress in the development of a universal ML model for ground-state energies (ground-state ML potentials), for which existing models approach coupled-cluster accuracy for typical organic molecules.[82]

In the next subsections, we discuss the main challenges and aspects impacting the development of ML models for excited-state energies.

[H2] Complexity of electronic density

One of the obstacles in learning excited-state energies is that excited-state electronic densities are much more complex and spatially spread than ground-state electronic density. Therefore, ML models developed for the prediction of ground-state energies will not necessarily work similarly well for the prediction of excited-state energies. Errors on the predicted excitation

energies can be 10 times larger than the errors on the ground-state energies when training various ML potentials with 3D structural descriptors (Box 2 and Box 3) on the same number of points, as it has been reported[83] for a large and diverse set of molecules (QM8 dataset)[84]. Thus, if attempting to use such ML potentials, one may need substantially larger training sets to attain excited-state energies with the same accuracy of ground-state energies.

The errors on the excitation energies can be further increased when employing ML potentials that use local descriptors and partition the energy into atomic contributions. In this case, the descriptor considers structural parameters within a cutoff distance. These models work rather well for ground-state energies and can be transferable, meaning that they can be trained on smaller molecules and make reasonable predictions for larger molecules.[23] In excited states, these models exhibit substantially worse accuracy than models using global molecular representations, as was shown on the same QM8 dataset in a couple of independent studies.[83,85]

[H2] Improving descriptors and models

The limitations of popular ground-state ML potentials for the prediction of excited-state energies can be overcome by developing alternative ML models and descriptors tailored to the treatment of excited states. One of the most important aspects is the selection of the appropriate descriptors that best represent molecules for learning excited-state energies. This is a topic of intensive ongoing research. A low computational cost solution is the use of cheminformatics descriptors typically employed in drug design. These descriptors can be quickly derived from molecular structures and, because they were developed for various complicated molecular properties, they also work remarkably well for excited-state energies of molecules of very different sizes and compositions.[49,81,86] Although these descriptors have not currently been used to learn different conformations of the same molecule, there is an indication that they can be combined with 3D information to obtain improved accuracy on the prediction of orbital energies.[87] It is very beneficial to include information about many-body structural parameters (such as angular information) in addition to two-body parameters (such as radial information) for training ML models to better capture subtle geometrical changes.[83] Descriptors inspired by drug-design studies can also include computationally affordable ground-state QM properties.[47,48]

To ensure greater transferability and accuracy, ML models can be combined with QM methods in different ways (Box 3). Any such combination using QM method would obviously result in increased computational cost compared to pure ML approaches, which may be an obstacle

when a large number of predictions are necessary, for example, in high-throughput screening. A powerful way to achieve such a combination is to use ML to improve the QM calculation implicitly, making faster and possibly more accurate predictions of excited-state energies.[67] Another powerful way is to use a Δ-learning scheme[88] to correct excited-state energies calculated with low-level and low-cost QM methods (Box 3). One advantage of the Δ-learning is that it explicitly corrects a QM method without the need to modify it. In this case, one is not limited to structural descriptors, but can also use readily available low-level QM descriptors such as orbital energies.[50,84,89-92] This was shown already in 2004 in the pioneering study by GuanHua Chen and co-workers.[89]

Alternatively, one can use transfer learning to train models on an abundant (easier available and therefore more plentiful) type of data and refine them with less abundant target type of data. An essential feature of transfer learning is that it can be applied to various combinations of abundant and targeted types of data, such as QM excitation energies for small (abundant) and large (target) molecules,[93] or QM (abundant) and experimental (target) data. The latter approach was only used to learn molecular orbital energies.[94] Co-kriging **[G]**, a related approach, also bears the potential for learning excited-state energies by exploiting data of different types and availability, although it has only been demonstrated for bandgaps in solids[95] rather than molecular excited-state energies.

[H2] Learning multiple states

A severe obstacle for developing ML models for excited states is that multiple states must be calculated with the same accuracy. Several approaches have emerged to tackle this problem.[45,46] One of them used a separate ML model for each state, as has been reported in a number of studies.[31,32,37-39,41,43,46,83,84,96,97] Although this is the simplest approach, it has two main drawbacks. First, it may predict the wrong position of states relative to each other. Second, it ignores any correlation between energies of different states. In principle, these problems can be solved by learning energy gaps between states, but then errors will accumulate when many states are considered. Alternatively, one may learn all energy levels simultaneously, which was shown to improve the accuracy of ML models, although this approach may require a longer training time.[35,43,44,46,98]

[H2] Reference data

The performance of ML algorithms hinges on the size and diversity of the training sets[25,99] and intensive investigations on how ML accuracy for excited-state properties depends on the

training set size are still required. Therefore, the limited availability of high-quality QM or experimental excited-state energies poses a big challenge for the performance of ML approaches and only a handful of studies reported learning curves for both ground- and excited-state properties.[43,83] These studies showed that both types of learning curves do behave similarly, but their exact shape strongly depends on the chosen model and data set and that, for some combinations of ML model and data set, learning ground-state properties is not easier than learning excited-state properties.[43] One of the few datasets of energies attained using QM methods (such as CC2 and TD-DFT; see Box 1) for several excited states of 22,000 molecules is the QM8 dataset[84], which is a subset of the popular QM9 benchmark set[100]. Recently, another dataset QM-symex with excited-state properties for ten excitations of 173,000 symmetric molecules was reported.[101]

As a solution to the limited data availability, ML models are often developed and trained on much more computationally affordable bandgaps or orbital energies from which bandgaps can be calculated.[85,95,98,102-115] Such studies are also facilitated by the availability of many big databases with these properties.[100,116-120] One should be aware that orbital energy gaps are a poor approximation for excitation energies. Kohn–Sham energy gaps, for instance, correspond to a zero-order expansion of TD-DFT results.[121]

[H2] Environmental effects

Another challenge for ML is the proper description of environmental effects (such as solvents) on excitation energies and two possible strategies have been suggested to address this problem.[34] One approach would be to completely ignore the environment in the ML model and only incorporate environmental effects implicitly through training on reference data that include such effects. This approach is equivalent to implicit solvent models in QM.[34] Alternatively, information about the environment can be included explicitly in the ML model, typically at the descriptor level.[34,122] In a broader context, it is known from the use of ML approaches for the study of ground-state properties that special treatment of long-range interactions, such as explicit inclusion of additional dispersion corrections or training separate ML models on charges for capturing electrostatic effects, is often necessary.[23,123] In any case, the generation of reference data that capture environmental effects requires additional computational cost.

[H2] Gauging errors and choosing models

We have discussed so far some of the general challenges related to the learning of excited-state energies. However, some of these challenges may not have a severe impact for the problem at hand. For example, ML models with local descriptors can still be successfully applied to predict excitation energies for large oligomers[93,124] or learn multiple energy levels of small molecules.[35,44] If the researcher is interested in a single molecule, then ML models that are not transferable to other molecules can be used.[31,32] Independent ML models can be trained for each excited state separately and still provide reliable results.[32,37,39,41] It is also often overlooked that for a well-trained ML model, the major source of error typically comes from the inaccuracies in the training data obtained with QM methods and therefore, the further reduction of the learning error will not significantly improve the accuracy of the final simulation result.[32]

In some cases it might be unclear which ML algorithm to adopt for a specific application, because various algorithms — either based on kernel methods or neural networks (Box 3) — have been used for similar or even the same tasks, delivering comparable results.[23,43,46] It should be good practice to benchmark several ML methods in the same way it is routinely done for QM methods before applying them to a particular research problem. This strategy brings us to another underappreciated approach: the use of an ensemble of ML models — weak learners — to obtain better overall prediction — strong learner (see Box 3).[92]

[H1] Spectroscopy

Spectroscopy to probe molecular excited-state properties is essential for elucidating the chemical structures and electronic dynamics. ML is emerging as a useful tool for spectroscopy and can assist in both predicting and analyzing the electronic spectrum for a given chemical structure and solving the inverse problem of determining a chemical structure for a given spectrum. A wide range of electronic spectroscopic methods exist, which measure various physicochemical phenomena,[9] and ML has been already applied to assist different spectroscopic techniques, including steady-state absorption (in both optical[26,32,34,35,125,126] and X-ray[127-129] domains), emission,[28] and multidimensional time-resolved spectroscopy.[34,126,130] Some of these studies[26-28] were the earliest works using ML for excited-state research, and we are currently experiencing a resurgence in interest in such methods.

[H2] Spectra from surrogate models

In most studies so far, ML has not directly been applied to predict the spectrum itself. Instead, ML is typically used as a surrogate model for a QM method, to predict transition energies and transition probabilities. Then, the spectrum is predicted with one of the techniques discussed in Box 4 using the ML predictions as if they were QM predictions. One particular problem is that learning quantities needed for transition probabilities, such as oscillator strengths or transition dipole moments, seems to be more challenging than learning energies.[84,124]

In the case of the single-point convolution approximation, which requires only information about the equilibrium geometry, ML can be used to learn oscillator strengths and excitation energies of electronic transitions of various molecules. This approach has been applied, with various degrees of success, to data sets containing only one or two lowest-lying excitations.[84,86,98]

In the nuclear ensemble approach (NEA, Box 4), energies and transition probabilities must be obtained for many geometries of the same molecule, which are generated either from molecular dynamics or some probability distribution, such as a Wigner distribution for the nuclear normal modes (FIG. 4a). ML has been used for creating surrogate models for excitation energies and, in most cases, for oscillator strengths or transition dipole moments, which were subsequently used for calculating spectra of various molecular systems (FIG. 4b).[30-33,35,124] These models were developed for different purposes and are based either on NN[31,35,124] or kernel methods.[32] Our ML-NEA implementation is based on the KREG model **[G]** and allows to significantly speed up the simulations of spectra for dozens of excitations of single small molecules or medium-sized molecules and only requires hundreds of QM training points (FIG. 4c and d).[32] It practically does not add any computational cost compared to the calculation of the spectra with the traditional NEA approach using the same number QM calculations because training on such a small number of points takes minutes. Developing ML models for the calculation of NEA spectra for many excitations (often few dozens are required for a spectrum simulation) and chemically diverse molecules remains an open challenge. One attempt to do so was limited to a few excited states in several small systems.[35] For larger molecules, even oligomers, in which excitations are well localized and could be learned with local descriptors and rather small cutoff distances, relatively large errors were observed for transition dipole moments, which could be adequately learned only with a very large cutoff approaching the length of the largest oligomer.[124] The large ensembles of nuclear geometries enabled by ML (it can be easily used to predict excited-state properties for tens of thousands or millions of ensemble points) have

the distinct advantage of producing high-precision, statistically converged results compared to the relatively small QM ensembles as has been verified in steady-state absorption spectra.[31,32,124]

ML starts to be used for high accuracy spectrum simulations (Box 4). Time-dependent approaches have been successfully applied to simulate linear[125,126,131] and multidimensional spectra.[34,126] These spectroscopic simulations should soon benefit from the development of ML methods for excited-state dynamics (discussed in section Dynamics), because such ML dynamics simulations allow to directly predict time-dependent signals. An application of ML that remains unexplored has been its use to predict Franck–Condon factors that would enable fast simulations of vibrationally resolved spectra within time-independent approaches.

[H2] Direct spectrum simulation

A fundamentally distinct approach for the simulation of spectra with ML is to directly predict them without using ML as a surrogate model for QM methods. Because one of the main strengths of ML is related to the creation of images, it can, in principle, be used directly to obtain spectral images for a given molecular structure. This research area awaits to be explored more in the coming years. It should be noted, however, that such a direct approach of simulating spectra with ML needs a training set of spectra in the first place. For example, this approach is not suited for accelerating calculations of a spectrum of a single molecule, for which non-direct methods, such as nuclear ensemble approach using ML as a surrogate model for QM methods, are more suitable.

One step towards the direct simulation of spectra with ML was made for the simulation of two-dimensional electronic spectra of a light-harvesting complex.[130] In that study, calculated pigment-site energies were used as descriptors, rather than the more conventional structural descriptors. An approach to directly simulate spectrum for a given structure was suggested recently for the prediction of K-edge X-ray absorption near-edge structure spectra of diverse molecules from the QM9 dataset[132] and solid materials.[133] This approach enables assignment of geometrical structures by comparing the experimental and simulated spectra, which brings us to the next important application of spectrum simulations: structure determination.

[H2] Structure determination

Spectrum simulations are an essential tool to answer which chemical structure corresponds to a given measured spectrum. As we have seen in this section, ML can alleviate the cost of spectrum simulations, accelerating structure determination. ML can help determine structures

in other ways too. One of them is to use ML to generate an extensive database of structures with associated spectra to which a given spectrum can be matched.[134] Another way is to use supervised and unsupervised learning to determine structural parameters from a given spectrum directly.[26-28,127-129] For example, unsupervised learning can be used to analyze how the spectrum depends on the structure and environment, enabling the formulation of rules for structure determination.[68,69,135,136] Such insight may also be useful for designing new materials, as we discuss in the section Design of optoelectronic materials.

[H1] Dynamics

Dynamics simulations provide a unique insight into how excited-state properties evolve with time, enabling the calculation of time-dependent properties from first principles. Excited-state dynamics can be used to simulate spectra, estimate reaction rates, determine time constants of radiationless processes and monitor exciton transfer times, among many applications. However, the computational cost of dynamics simulations is staggering as it requires hundreds of thousands or even millions of QM calculations (Box 5). Furthermore, the analysis of dynamics simulations may also require intense manual inspection.

ML is a very welcome addition to the computational toolbox for excited-state dynamics. As in the case of spectrum simulation, ML can assist in performing dynamics simulations by using ML as a surrogate for QM methods, or it can be used to calculate the ultimate property of interest directly, circumventing dynamics simulations altogether. Moreover, ML can help in the analysis of dynamics.

[H2] Nonadiabatic dynamics

ML can be used as a surrogate for QM in many conventional methods for dynamics simulations (we refer the reader to a recent Review[45] that provides an excellent description of many ML models and related technical issues). Propagating QM dynamics on high-accuracy excited-state PESs is only possible for a relatively small number of atoms treated quantum-mechanically, because it requires information about excited-state energies, energy gradients, and couplings that are computationally costly to calculate. This information is obtained either from precomputed PES[137] or by performing QM calculations on-the-fly during the time propagation (Box 5).[138] ML can bring these two conventional strategies together, because in this case the PES is always given as an interpolating function, rather than obtained from a direct solution of the Schrödinger equation.

Because of the development of fast ML approaches for the simulation of ground-state dynamics,[29] it may seem to be straightforward to extend these approaches to excited-state dynamics, particularly to one of the most popular types of dynamics: trajectory surface hopping (Box 5). However, besides the challenges related to learning excited-state energies, such as complex electronic density and multiple states (see the section Energies), one has to consider couplings between states and deal with the high density of excited states. A cost-efficient generation of the training set also becomes especially important in dynamics simulations, as one usually does not know a priori which regions of the configuration space are the most representative during the excited-state time evolution. This problem can be tackled using active learning, which is based on propagating many exploratory trajectories that iteratively identify low-confidence regions for ML.[41,43-45,97]

For dynamics propagated in the adiabatic representation (Box 5), the population transfer between electronic states is mediated by nonadiabatic coupling vectors. Learning these coupling vectors is particularly difficult as they are very narrow functions of the nuclear coordinates (FIG. 5) that become singularities at a conical intersection **[G]**. Moreover, the sign of coupling vectors (phase) is arbitrary, a feature inhered from the wavefunctions' arbitrary global phase. Despite these problems, even very narrow nonadiabatic coupling vectors can be learned. One has to ensure that the ML training set includes nuclear configurations featuring large coupling magnitudes,[29] which can be obtained through a careful selection of training points. These points can be handpicked,[29] or selected by using dense sampling of all points,[38] or sampled in specific regions,[139] or included through active learning.[41,43-45] The coupling vectors tend to infinity at zero energy gaps, therefore it is also helpful to train the model on these vectors multiplied by the energy gap.[41,43-45,140] The problem of the arbitrary phase of the coupling vectors can be addressed by either correcting the phase before the ML training[41,140] or using ML itself to select the phases that lead to best predictions during the training process.[44] For nonadiabatic dynamics involving intersystem crossing (transition between states of different spin multiplicities), one may also need to learn spin–orbit couplings.[44]

The difficulties involved with the learning of nonadiabatic couplings can be avoided altogether by adopting nonadiabatic dynamics methods that estimate the couplings from some features of the PES, such as energy gaps and energy gradients (see Box 5). Such approach has been used for the calculation of ML dynamics with the Zhu–Nakamura[37,39,141] and the Hessian[44] approximations. Alternatively, one can calculate nonadiabatic couplings in the vicinity of conical intersections by switching to conventional QM methods during ML dynamics. This

approach gives rise to mixed QM–ML dynamics, in which most of the dynamics is propagated on the ML PES, while in critical regions with small energy gap or regions of low-confidence in ML PES, propagation is switched to conventional QM dynamics.[38,39,96]

Both kernel methods and NNs have been used to simulate excited-state dynamics, in the same way these approaches can be used to learn excited-state energies (see section Energies). They are mostly based on the same ML models that are applied for learning ground-state energies. Some of them are used to learn each state separately,[37-39,41] while others are adapted to learn multiple states simultaneously,[43,44] with both types of approaches giving good agreement with the QM dynamics.[45] A comparison of NN and KRR methods for nonadiabatic excited-state dynamics did not highlight a significant difference in accuracy between the two classes of methods.[43] This confirms typical observation for ML in QM research that similar accuracy can often be achieved with different ML models, although each has its advantages and disadvantages. For example, kernel methods are more straightforward to train owing to their closed analytical solution of parameters, but a larger quantity of data can be treated with NNs (Box 3).[23]

Other types of excited-state dynamics can also benefit from ML. The emergence of full quantum wavepacket dynamics based on on-the-fly PES calculation[142] has found a powerful ally in ML.[125,143-151] In most of these approaches,[144-151] the global PES is built iteratively in an active learning manner from a collection of local ML PESs obtained during the time propagation.

Often, excited-state wavepacket dynamics is performed in the diabatic representation (Box 5), and ML algorithms have been intensely employed to aid the fitting and diabatization process.[152-168] Several recent works have extended the latter approaches to also use ML for learning transition dipole moments[164] and spin–orbit couplings.[169]

All ML works mentioned in this section were dealing with only a small number of excited states. However, dynamics often must be started with a large number of closely-lying excited states and it may be particularly challenging for ML to describe all of them adequately. A hybrid QM–ML approach can potentially solve this problem by using QM at the start of dynamics when it quickly transitions from high to low-lying states. This would make training ML models much easier because only low-lying states will need to be trained accurately.

Another challenge in the use of ML for excited-state dynamics is related to the description of rare events,[25] such as tunneling.[170] To capture these properly, the training set should include enough information about conditions for these rare events, such as representative geometries. If such information is known beforehand, it could be used to augment the training set. Otherwise, one must rely on active learning approaches to learn such conditions.[140]

Note that all studies surveyed above only reported test-systems results as part of proof-of-principle demonstrations of ML dynamics. The lack of specific applications signals the most critical challenge in this field: despite all the progress, using ML for dynamics is much more cumbersome than using QM because no automatic protocols exist to perform ML dynamics simulations from start to finish. The biggest problem is to obtain a cost-efficient training set. While active learning can help,[45,46] it is currently a rather slow approach itself, as it requires running multiple trajectories and re-training the ML model repeatedly. It is unclear what are the optimal settings for active learning (these include the initial training set, number of trajectories, how long to run trajectories and uncertainty quantification). We anticipate that much effort will be directed in the future to automate active learning approaches and the whole workflow of ML dynamics simulations.

[H2] Dynamics of large systems

There is a whole body of research devoted to the development of different ML models to tackle large molecular systems and assemblies beyond the capability of conventional QM methods. One of such ML approaches treats the relevant photoactive part of the system at the QM level, while ML is used for the surrounding environment.[40,171] Another approach uses ML to learn excitation energies for a photoactive region calculated within the conventional QM–molecular mechanics (MM) (see Box 1) approximation.[122] In a third approach, charge and exciton transfer are simulated based on MM trajectories using a ML model only to learn the parameters of the Frenkel Hamiltonian and fragment molecular orbital Hamiltonian.[172] Beyond explicit inclusion of environment atoms, dynamics can be propagated using model Hamiltonians for open quantum systems,[173] and ML has been applied to enable and accelerate the calculation of such dynamics.[126,174-176] In a final approach, nonadiabatic events in the excited state are evaluated with QM methods on top of a ML ground-state dynamics (this approximation can be justified for very large assemblies, in which excited-state geometry deformation can be neglected).[42] In this case, time-derivative nonadiabatic couplings are still calculated with a QM method. Learning these couplings remains an open challenge, although unsupervised learning has

shown that they may be learned just using static rather than time-dependent structural information.[72]

[H2] Direct simulations of dynamical properties

Ultimate properties related to excited-state dynamics can be directly predicted using ML approaches that bypass the calculation of actual dynamics, in a way similar to that discussed for spectrum simulations. For example, exciton-transfer times has been directly calculated using ML.[177] This kind of simulations can be instrumental in searching for new materials with the desired range of property values, as we discuss in the next section, Design of optoelectronic materials. Again, as in the case of spectrum simulation, the limitation of a direct approach is that it requires a training set for the properties under investigation. For example, for exciton-transfer times, many dynamics simulations had to be performed in the first place to feed enough data into the model. Thus, a direct approach cannot fully replace the use of ML as a surrogate for QM methods in dynamics simulations. Likely, in the future, surrogate and direct approaches will be used together so that a surrogate approach generates a cost-efficient training data set for the direct approach.

[H2] Dynamics analysis

The arduous task of analyzing dynamics results can be substantially eased by unsupervised learning. For example, unsupervised learning was used to identify and characterize meta-stable patterns in molecular conformations and transitions between them during excited-state dynamics.[30] Dominant active coordinates from excited-state dynamics simulations can be determined using unsupervised learning.[70] It can also assist the clustering of similar trajectories and structures.[71]

Unsupervised learning can lead to unexpected findings. For example, it has recently uncovered surprising relations between nonadiabatic couplings, and static and dynamic features in a perovskite material consisting of organic and inorganic parts.[72] This work showed that the static structure affects nonadiabatic time-derivative couplings more than atomic motions. In addition, charge recombination in the inorganic part was strongly influenced by motions of adjacent organic molecules that contribute to neither hole nor electron wavefunction. This knowledge may lead to practical design rules for improving the efficiency of perovskite solar cells, as it is apparent that this is influenced by geometrical changes that can be induced by chemical modification. Such design rules greatly assist in the design of optoelectronic materials — the topic our next section.

[H1] Design of optoelectronic materials

Training ML models for the prediction of excited-state properties can aid the search of optoelectronic materials.[178,179] For example, ML trained on excitation energies of fluorescent molecules can be used not only for emission spectrum simulation but also for the design of materials that emit at a required wavelength for application as LED or a fluorescent label in cell imaging.[47-49] More often than not, the compounds that are intended to be used in complex photodevices are selected based on macroscopic properties such as power conversion efficiency (PCE) or fluorescence quantum yield that ultimately depend on atomic-scale molecular excited-state properties. Calculating these properties with QM is often a formidable task and the beauty of ML is that it can be used for learning any of them.

ML has been intensively used for the discovery of materials to be employed in optoelectronics, which are mainly organic semiconductors,[74,180-184] components of bulk heterojunctions for organic photovoltaics (BHJ-OPV),[51-54,60-62,75,76] materials for dye-sensitized solar cells (DSSC)[55-59,63,185,186] and perovskites.[95,111,112,187] ML models have also been developed for the design of fluorescent organic molecules (proteins and others)[47-49,68,69,188,189] and photosynthetic complexes[177] that can be used in biological, biomedical and biotechnological research.

Direct applications of ML require that a specific ML model learns the ultimate target property of a material, such as the PCE of a solar cell built with it (FIG. 6). However, such learning typically requires a significant amount of experimental training data, the generation of which is very time and resource consuming, limiting the applicability of direct ML. An alternative approach is to learn an intermediate, but more accessible property related to the ultimate target property (FIG. 6). The intermediate properties are usually calculated with QM simulations and then used to train ML surrogate models for QM properties. Thus, supervised learning may be used to learn the energies of frontier orbitals and bandgaps known to be correlated to the PCE.[190] In the following two subsections, we see how various ultimate and intermediate properties can be learned. The most straightforward application of such ML models is for HTVS of potential candidates to identify the most promising materials (FIG. 6). ML can, however, also be used for an automated generation of structures of new compounds with desired properties (FIG. 6). We will discuss this strategy in the final subsection.

We should remark that whatever computational method is chosen for in silico materials design, the best performing molecules may be difficult to synthesize. Thus, an estimate of synthesizability alongside performance is highly desirable.[60,63,74] Furthermore, even if promising molecules can be prepared synthetically, their performance strongly depends on the photodevice setup they are used in. ML comes in handy also in this case, because it can be used to find the optimal device setup, process that usually requires lots of manual experimentation. Examples of such ML applications include the optimization of dye-sensitized solar cells with multiple linear regression and partial least squares regression.[191-194] and the construction of complex optoelectronic structures using a combination of NNs for forward modeling and inverse design.[195] ML can even be used to optimize device structure on a nanoscale, for example, by guiding experiments to create nanostructures of required shape[196] or to suggest experimental setups for controlling the growth of a 2D semiconducting material.[197]

[H2] Learning ultimate target properties

The ultimate target property of an optoelectronic material obviously depends on material's intended use. The property targeted by ML for applications in BHJ-OPV[51-54,61,62,64] and dye-sensitized solar cells[55-60,63] is PCE. In organic field-effect transistors, the target property is electron mobility,[184] whereas in fluorescent molecules, target properties are emission maximum,[47-49,188] colour[189] and photoluminescence quantum yield.[81] The data sets also greatly vary in size, composition and quality of reference values (owing to the various measurement conditions of the experimental values). Given the different nature of ultimate target properties and data sets, there is no single recipe for creating ML models. Thus, for each application, lots of human input is necessary for the selection of an appropriate combination of ML algorithms and descriptors. As for descriptors, one can categorize them into structural, physicochemical and QM descriptors. Often, the same descriptors that are used for drug design are included in ML models for optoelectronic materials design. QM descriptors are usually obtained from computationally affordable semiempirical methods. These descriptors can be electronic localization energy calculated using Hückel theory,[51] orbital energies,[59,60] or more computationally expensive descriptors derived from vibrational frequencies.[55]

ML algorithms reported in the literature include multiple linear regression, partial least squares regression, Lasso regression, NNs, kernel methods, random forest and gradient boosting regression (see Box 3). In the case of materials design, the ideal ML algorithm not only needs to deliver good prediction accuracy, but should also inform on which factors are the most significant to achieve the aimed goal, so that design rules for new materials can be derived.

There are several ways to get such interpretable models. One of them is to use supervised learning, which allows us to judge which descriptors are the most important, typically employing a random forest or Lasso regression and partial list squares. Such approaches, for example, gave evidence that important properties determining PCE or electron mobilities are orbital energies.[58-61,184] Another way is to use unsupervised learning. This approach can help to classify materials based on their properties , for example, by establishing a relationship between high PCE values and structural parameters.[198] Supervised and unsupervised learning techniques can also be combined, as was done to find a correlation between complex protein fluorescence spectra and protein structure.[68,69]

The algorithm accuracy also matters, of course. The big challenge is to create accurate and transferable models to be applied to HTVS of lead molecules from databases with millions of candidates, especially if only a small number of experimental reference values is available (for example, about 50 molecules).[51] Creating and curating such databases[199] is a crucial challenge on its own as it requires feedback from ML and human experts across disciplines, as was done in the Harvard Clean Energy Project (CEP).[73,178] In materials search, simply collecting training data for ML from existing literature has several drawbacks. The first of them is that ML predictions should be verified and refined, and this process requires new experiments. The second drawback is that materials performance in a device is influenced by many experimental conditions that vary from one study to another and can hardly be taken into account by ML. Even the current state-of-the-art ML models, which do not consider experimental factors, are challenging to use. As mentioned above, a plethora of ML models with very different algorithms and descriptors have been suggested in the literature. Such a lack of clear-cut guidelines on what method to use hampers new ML applications. Further development of integrated software platforms is required for at least an initial automatic assessment of a selection of typical ML models for learning ultimate target properties.

[H2] Learning intermediate properties

Ultimate target properties of optoelectronic materials are known to be correlated to many molecular features that can be used as intermediate properties. Therefore, ML has been used to learn various intermediate properties such as bandgaps,[95,111,112,200-202] band edges,[203] intramolecular reorganization energies,[180,182,183,204,205] delayed fluorescence rate constant,[181] decay rates of emitters,[206] exciton-transfer times and transfer efficiencies,[177] and electronic couplings.[172,207-214] All these properties were calculated with QM methods, and ML was later used as a surrogate model for QM to accelerate screening (FIG. 6). Moreover, ML can also

learn experimental intermediate properties, such as the absorption wavelength of photoswitch molecules.[215] QM methods could also be used to calculate this property (and ML would be then used as a surrogate model for QM), but these methods can be computationally demanding and can lead to less accurate data than those obtained with experimental measurments.[215]

Learning intermediate properties may also provide in-depth theoretical insight into the photophysical processes that takes place in the material and that can be experimentally measured. Therefore, ML can use experimental data as input and predict important intermediate properties, thus directly linking experiments to the theoretical description. A highly inventive example of such ML application is to learn the exciton wavefunctions using experimental near-field spectra as input.[163] Similarly, ML could take time-resolved photoluminescence data as input to predict decay rates of emitters, although this approach has only been applied on computer-generated data.[206]

The more common task of predicting intermediate properties from theoretical data faces similar challenges of the prediction of ultimate target properties previously discussed in relation to the selection of the appropriate ML algorithm and descriptors. Again, no single recipe exists and, for example, descriptors adopted for solving specific task can range from QM properties (such as those based on orbital energies[180] or even the whole spectrum simulated with TD-DFT[216]) to structural descriptors inspired by drug design, as well as combinations of them.[180-183,216]

However, because intermediate properties are mostly QM properties, it is natural to use common ML approaches that were specifically designed for learning other molecular QM properties. For example, ML models that use 3D structural descriptors for the calculation of ground-state energies often serve as the basis for models for the calculation of excited-state energies, electronic couplings[172,207,209-211,213] and absorption wavelengths (see the previous sections and Box 3).[215]

The main advantage of learning intermediate rather than ultimate properties is that it is usually easier to generate additional reference data for the former, opening much more efficient and economical ways for materials design. One of these ways is to use unsupervised learning to determine the most important structural motives, enabling the design of materials with desired intermediate properties.[204,205] Active learning used in the context of HTVS provides another approach to the design of new materials. Once the ML model is trained on an initial set of reference data, it can be used to predict intermediate properties for millions of candidates to identify the pool of potential leads. The QM methods can then recalculate these properties for

this pool of leads and refine the ML model with these additional reference data. This process is repeated until the final leads are obtained. This strategy resulted in the discovery of lead candidates that were synthesized and that exhibited high external quantum efficiencies.[181] An alternative way of using ML models for intermediate properties is to combine them with automated structure generation, which we discuss in the next section.

[H2] Automated structure generation

The approaches discussed so far in this section use supervised learning to predict ultimate or intermediate properties, useful in HTVS of an existing database of potential lead candidates. However, the pressing issue of the rational materials design is to build a database of promising compounds in an optimal way (FIG. 6). The dream scenario would be ML designing new materials automatically with as minimum human intervention as possible.[73] Several approaches have been suggested in this respect. One of them is based on the genetic algorithm, which selects molecules according to the evolutionary principle of the survival of the fittest. The fittest molecules can be defined as those that exhibit either the desired ultimate target or intermediate property. Thus, this approach can be combined with ML models created for the prediction of these properties, as has been done for PCE.[57,185,186]

Another approach that holds great promise is an effective unsupervised learning technique based on generative models such as variational autoencoder that can transform molecules into a continuous latent space, where the points with desired properties can be found and transformed back to molecules (Box 2). It was demonstrated that a variational autoencoder is superior to a genetic algorithm for such task.[74] In that study, SMILES **[G]** were used as input for an encoder and output of a decoder, while a NN-based predictor was used for the estimate of the HOMO and LUMO energies. Interestingly, the accuracy of the prediction of these electronic properties as tested on the popular QM9 benchmark set[100] was comparable with that achievable with supervised learning methods specifically designed for such tasks.[74] Analogous approaches were used to generate molecules with the desired range of LUMO and optical transition energies for the design of polymer solar cells[75], and HOMO, LUMO, and PCE of non-fullerene acceptors in solar cells.[76]

Properties such as HOMO and LUMO energies can be obtained if the wavefunction is known. With the emergence of machine-learned wavefunctions, it is possible to find structures with the desired range of these energies or HOMO–LUMO gaps, as was done in a proof-of-concept

study for a single molecule, in which the ground-state wavefunction was predicted on an atomic orbitals basis.[217]

Reinforcement learning[218] is another powerful approach for materials design[73] that also holds promise for searching novel optoelectronic materials. So far, this approach was used to find molecules that have desired absorption[77] and fluorescence[78] properties by recurrent neural networks **[G]** and Monte Carlo tree **[G]** search, where the search was guided by giving larger reward to molecules with TD-DFT excited-state properties closer to the target ones. In the future, reinforcement learning may be combined with supervised learning to speed up the calculation of properties defining the reward.

Finally, the ultimate open challenge is automating the entire workflow, from the definition of the optoelectronic materials with desired parameters to the production of working devices. This automatization can only be achieved by integrating ML with robotic laboratories, where it would be possible to synthesize the required compounds, build the devices, and measure their performance.[25] This futuristic goal is not that far from reality, as such robotic laboratories have been already built and used for the synthesis of new compounds[219] and discovery of new materials.[220]

[H1] Conclusions and outlook

ML entered the field of excited states relatively recently, but it is here to stay. ML is geared to become an integral part of excited-state simulations and the design of optoelectronic materials (FIG. 2). Supervised learning can either directly predict excited-state properties or improve the performances of QM methods, which can be subsequently used for excited-state simulations. Both supervised and unsupervised learning provide insights into structure–property relationships that lead to a better understanding of photophysical processes and rules for the design of new optoelectronic materials. Generative models and evolutionary algorithms can assist the automated rational materials design.

Despite the fast development of ML approaches for the study of excited states (FIG. 3), the number of reported studies using such ML models remains relatively small compared to other research areas. Most of the works covered here are proof-of-concept studies proposing new methods rather than applying ML for practical research and development (R&D).

One of the key aspects of the design of a ML model is the selection of an appropriate algorithm and descriptors. We have discussed how a ML model that gives good results for the prediction of one property (for example, ground-state atomization energies) is not necessarily the best model for the prediction of another property (for example, excitation energy, see section Energies).[83,108] An incredibly tricky problem to solve is the design of surrogate ML models for excited-state properties that are transferable to other systems, and it is currently unclear how transferable ML can be. One of the promising ways would be to use ML to improve QM methods, as it was done earlier when ML was used to accelerate semiempirical QM calculations, which was then applied to calculate excited-state PES.[67] We expect much of the research to be focused in this direction.[47]

Any ML model is only as good as the training data, which has several implications. One of them is that ML can often be used to increase the precision of simulations, for example, to obtain more precise absorption spectra,[32] but its accuracy will be limited by the accuracy of the reference QM method used to generate training data. In other words, the error of ML models with respect to the reference QM method is less than the error of the QM method with respect to the true value.[32,108] Another implication is that many research groups are developing special techniques to obtain proper training sets. For example, excited-state PES representation to be used in the calculation of molecular dynamics can be obtained by sampling from existing QM trajectories,[39] using farthest-point sampling **[G]**,[34,38] and by active learning,[45] but there is no single established technique for this.

The quality of training data has also an implication in materials design, which might require some change in the way results are reported in the literature. Usually, only successful device design attempts are reported, introducing a bias in the available training data. For ML materials design, bad results are as relevant as good ones because they allow ML to filter off materials with undesired performance easily.[59] Another problem is that experimentally measured device performance strongly depends on many factors, leading to very noisy training data.[59] Thus, measuring materials performance under the same conditions is highly desired. Especially in the field of materials design, interpretable ML is of high importance as it will allow researchers to understand which are the essential design rules for new materials.[59]

While continuing method development is a must, we are already at the point where ML can and should be used for practical R&D. This is exemplified by a couple of excited-state dynamics studies, in which ML uncovered the design rules of perovskite solar cells and

improved our understanding of plasmonic and catalytic processes in nanoclusters.[42,72] Other examples include the successful identification of new optoelectronic materials.[77,78,181,196] We anticipate the rapid rise of the number of studies using ML in excited-state research leading to practical implications. ML will be routinely used to calculate absorption and emission spectra, perform excited-state dynamics and design new materials. Optoelectronic materials design will greatly benefit from the continuing advancement of robotic laboratories. Unsupervised learning will become a standard tool for analyzing excited-state simulations, often carried out with supervised learning methods.

ML has a bright future in the field of excited-state research.

Acknowledgements

POD acknowledges funding by the National Natural Science Foundation of China (No. 22003051) and via the Lab project of the State Key Laboratory of Physical Chemistry of Solid Surfaces. MB acknowledges the support of the European Research Council (ERC) Advanced grant SubNano (Grant agreement 832237). The Review is dedicated to the 100th anniversary of Xiamen University.

Author contributions

All authors contributed equally to the preparation of this manuscript.

Competing interests

The authors declare no competing interests.

Glossary

Co-kriging

A kriging approach extended to treat heterogeneous data of various accuracy coming from different sources.

Conical intersections

Crossing between the adiabatic potential energy surfaces of two electronic states in a shape of a cone.

Farthest-point sampling

Sampling of points from the data set that maximizes the distance between them.

KREG model

Model based on kernel ridge regression that uses the Gaussian kernel function and takes as input the vector of inverted internuclear distances normalized with respect to distances in equilibrium geometry.

Monte Carlo tree search

Method for the selection of search directions in the decision tree using Monte Carlo.

Recurrent neural network (NN)

A NN that exploits the interdependences between elements of a sequence rather than treating them as independent inputs or outputs.

Simplified molecular-input line-entry system (SMILES)

Notation system designed to encode chemical structure in a sequence of text characters.

TOC Blurb

Machine learning is starting to reshape our approaches to excited-state simulations by accelerating and improving or even completely bypassing traditional theoretical methods. It holds big promises for taking the optoelectronic materials design to a new level.

References

1. Ponseca, C. S., Chábera, P., Uhlig, J., Persson, P. & Sundström, V. Ultrafast Electron Dynamics in Solar Energy Conversion. *Chem. Rev.* **117**, 10940–11024 (2017).
2. Brunk, E. & Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.* **115**, 6217–6263 (2015).
3. Zhang, B. & Sun, L. Artificial photosynthesis: opportunities and challenges of molecular catalysts. *Chem. Soc. Rev.* **48**, 2216–2264 (2019).
4. Gozem, S., Luk, H. L., Schapiro, I. & Olivucci, M. Theory and Simulation of the Ultrafast Double-Bond Isomerization of Biological Chromophores. *Chem. Rev.* **117**, 13502–13565 (2017).
5. Chakraborty, P., Karsili, T. N. V., Marchetti, B. & Matsika, S. Mechanistic insights into photoinduced damage of DNA and RNA nucleobases in the gas phase and in bulk solution. *Faraday Discuss.* **207**, 329–350 (2018).
6. Yang, Z. *et al.* Recent advances in organic thermally activated delayed fluorescence materials. *Chem. Soc. Rev.* **46**, 915–1016 (2017).
7. Kaloni, T. P., Giesbrecht, P. K., Schreckenbach, G. & Freund, M. S. Polythiophene: From Fundamental Perspectives to Applications. *Chemistry of Materials* **29**, 10248–10283 (2017).
8. Zhang, J. Z. & Reisner, E. Advancing photosystem II photoelectrochemistry for semi-artificial photosynthesis. *Nat. Rev. Chem.* **4**, 6–21 (2019).
9. Bennett, K., Kowalewski, M. & Mukamel, S. Probing Electronic and Vibrational Dynamics in Molecules by Time-Resolved Photoelectron, Auger-Electron, and X-Ray Photon Scattering Spectroscopy. *Faraday Discuss.* **177**, 405–428 (2015).
10. Gao, Y., Nie, W., Wang, X., Fan, F. & Li, C. Advanced space- and time-resolved techniques for photocatalyst studies. *Chem. Commun.* **56**, 1007–1021 (2020).
11. Mancuso, J. L., Mroz, A. M., Le, K. N. & Hendon, C. H. Electronic Structure Modeling of Metal–Organic Frameworks. *Chem. Rev.* **120**, 8641–8715 (2020).
12. Taniguchi, M., Du, H. & Lindsey, J. S. PhotochemCAD 3: Diverse Modules for Photophysical Calculations with Multiple Spectral Databases. *Photochem. Photobiol.* **94**, 277–289 (2018).
13. Norman, P. & Dreuw, A. Simulating X-ray Spectroscopies and Calculating Core-Excited States of Molecules. *Chem. Rev.* **118**, 7208–7248 (2018).
14. Yonehara, T., Hanasaki, K. & Takatsuka, K. Fundamental Approaches to Nonadiabaticity: Toward a Chemical Theory beyond the Born–Oppenheimer Paradigm. *Chem. Rev.* **112**, 499–542 (2011).
15. Baryshnikov, G., Minaev, B. & Ågren, H. Theory and Calculation of the Phosphorescence Phenomenon. *Chem. Rev.* **117**, 6500–6537 (2017).
16. Crespo-Otero, R. & Barbatti, M. Recent Advances and Perspectives on Nonadiabatic Mixed Quantum–Classical Dynamics. *Chem. Rev.* **118**, 7026–7068 (2018).
17. Curchod, B. F. E. & Martínez, T. J. Ab Initio Nonadiabatic Quantum Molecular Dynamics. *Chem. Rev.* **118**, 3305–3336 (2018).
18. Kumpulainen, T., Lang, B., Rosspeintner, A. & Vauthey, E. Ultrafast Elementary Photochemical Processes of Organic Molecules in Liquid Solution. *Chem. Rev.* **117**, 10826–10939 (2017).
19. Lischka, H. *et al.* Multireference Approaches for Excited States of Molecules. *Chem. Rev.* **118**, 7293–7361 (2018).
20. Kozma, B. *et al.* A New Benchmark Set for Excitation Energy of Charge Transfer States: Systematic Investigation of Coupled Cluster Type Methods. *J. Chem. Theory Comput.* **16**, 4213–4225 (2020).

21. Laurent, A. D. & Jacquemin, D. TD-DFT benchmarks: A review. *Int. J. Quant. Chem.* **113**, 2019–2039 (2013).
22. Peach, M. J., Benfield, P., Helgaker, T. & Tozer, D. J. Excitation energies in density functional theory: an evaluation and a diagnostic test. *J. Chem. Phys.* **128**, 044118 (2008).
23. Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **11**, 2336–2347 (2020).
24. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
25. von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
26. Otto, M. & Hörchner, U. in *Software Development in Chemistry 4* DOI: (ed J. Gasteiger) 377–384 (Springer, 1990).
27. Zupan, J. & Gasteiger, J. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **248**, 1–30 (1991).
28. Bos, M. & Weber, H. T. Comparison of the training of neural networks for quantitative x-ray fluorescence spectrometry by a genetic algorithm and backward error propagation. *Anal. Chim. Acta* **247**, 97–105 (1991).
29. Dral, P. O. in *Advances in Quantum Chemistry: Chemical Physics and Quantum Chemistry* Vol. 81 (eds Kenneth Ruud & Erkki J. Brändas) 291–324 (Academic Press, 2020).
30. Liu, F., Du, L., Zhang, D. & Gao, J. Direct Learning Hidden Excited State Interaction Patterns from ab initio Dynamics and Its Implication as Alternative Molecular Mechanism Models. *Sci. Rep.* **7**, 8737 (2017).
31. Ye, S. *et al.* A neural network protocol for electronic excitations of N-methylacetamide. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11612–11617 (2019).
32. Xue, B.-X., Barbatti, M. & Dral, P. O. Machine Learning for Absorption Cross Sections. *J. Phys. Chem. A* **124**, 7199–7210 (2020).
33. Zhang, Y. *et al.* Efficient and Accurate Simulations of Vibrational and Electronic Spectra with Symmetry-Preserving Neural Network Models for Tensorial Properties. *J. Phys. Chem. B* **124**, 7284–7290 (2020).
34. Chen, M. S., Zuehlsdorff, T. J., Morawietz, T., Isborn, C. M. & Markland, T. E. Exploiting Machine Learning to Efficiently Predict Multidimensional Optical Spectra in Complex Environments. *J. Phys. Chem. Lett.* **11**, 7559–7568 (2020).
35. Westermayr, J. & Marquetand, P. Deep learning for UV absorption spectra with SchNarc: First steps toward transferability in chemical compound space. *J. Chem. Phys.* **153**, 154112 (2020).
36. Carbogno, C., Behler, J., Reuter, K. & Gross, A. Signatures of nonadiabatic O_2 dissociation at Al(111): First-principles fewest-switches study. *Phys. Rev. B* **81**, 035410 (2010).
37. Chen, W.-K., Liu, X.-Y., Fang, W., Dral, P. O. & Cui, G. Deep Learning for Nonadiabatic Excited-State Dynamics. *J. Phys. Chem. Lett.* **9**, 6702–6708 (2018).
38. Dral, P. O., Barbatti, M. & Thiel, W. Nonadiabatic Excited-State Dynamics with Machine Learning. *J. Phys. Chem. Lett.* **9**, 5660–5663 (2018).
39. Hu, D., Xie, Y., Li, X., Li, L. & Lan, Z. Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *J. Phys. Chem. Lett.* **9**, 2725–2732 (2018).
40. Chen, W. K., Fang, W. H. & Cui, G. Integrating Machine Learning with the Multilayer Energy-Based Fragment Method for Excited States of Large Systems. *J. Phys. Chem. Lett.* **10**, 7836–7841 (2019).
41. Westermayr, J. *et al.* Machine learning enables long time scale molecular photodynamics simulations. *Chem. Sci.* **10**, 8100–8107 (2019).

42. Chu, W., Saidi, W. A. & Prezhdo, O. V. Long-Lived Hot Electron in a Metallic Particle for Plasmonics and Catalysis: Ab Initio Nonadiabatic Molecular Dynamics with Machine Learning. *ACS Nano* **14**, 10608–10615 (2020).
43. Westermayr, J., Faber, F. A., Christensen, A. S., von Lilienfeld, O. A. & Marquetand, P. Neural networks and kernel ridge regression for excited states dynamics of $CH_2NH_2^+$: From single-state to multi-state representations and multi-property machine learning models. *Mach. Learn.: Sci. Technol.* **1**, 025009 (2020).
44. Westermayr, J., Gastegger, M. & Marquetand, P. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett.* **11**, 3828–3834 (2020).
45. Westermayr, J. & Marquetand, P. Machine learning and excited-state molecular dynamics. *Mach. Learn.: Sci. Technol.* **1**, 043001 (2020).
46. Westermayr, J. & Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* DOI: 10.1021/acs.chemrev.0c00749 (2020).
47. Nantasenamat, C., Isarankura-Na-Ayudhya, C., Tansila, N., Naenna, T. & Prachayasittikul, V. Prediction of GFP spectral properties using artificial neural network. *J. Comput. Chem.* **28**, 1275–1289 (2007).
48. Nantasenamat, C. *et al.* Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine. *Chemometr. Intell. Lab. Syst.* **120**, 42–52 (2013).
49. Ye, Z.-R. *et al.* Predicting the emission wavelength of organic molecules using a combinatorial QSAR and machine learning approach. *RSC Adv.* **10**, 23834–23841 (2020).
50. Wang, J. N. *et al.* An accurate and efficient method to predict the electronic excitation energies of BODIPY fluorescent dyes. *J. Comput. Chem.* **34**, 566–575 (2013).
51. Olivares-Amaya, R. *et al.* Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **4**, 4849–4861 (2011).
52. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
53. Pyzer-Knapp, E. O., Simm, G. N. & Aspuru Guzik, A. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. *Mater. Horiz.* **3**, 226–233 (2016).
54. Lopez, S. A., Sanchez-Lengeling, B., de Goes Soares, J. & Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule* **1**, 857–870 (2017).
55. Venkatraman, V., Astrand, P. O. & Alsberg, B. K. Quantitative structure-property relationship modeling of Gratzel solar cell dyes. *J. Comput. Chem.* **35**, 214–226 (2014).
56. Venkatraman, V. & Alsberg, B. K. A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes. *Dyes Pigm.* **114**, 69–77 (2015).
57. Venkatraman, V., Foscato, M., Jensen, V. R. & Alsberg, B. K. Evolutionary de novo design of phenothiazine derivatives for dye-sensitized solar cells. *J. Mater. Chem. A* **3**, 9851–9860 (2015).
58. Li, H. *et al.* A cascaded QSAR model for efficient prediction of overall power conversion efficiency of all-organic dye-sensitized solar cells. *J. Comput. Chem.* **36**, 1036–1046 (2015).
59. Tortorella, S., Marotta, G., Cruciani, G. & De Angelis, F. Quantitative structure–property relationship modeling of ruthenium sensitizers for solar cells applications: novel tools for designing promising candidates. *RSC Adv.* **5**, 23865–23873 (2015).
60. Tortorella, S., De Angelis, F. & Cruciani, G. Quantitative structure-property relationship modeling of small organic molecules for solar cells applications. *J. Chemom.* **32**, e2957 (2018).

61. Sahu, H., Rao, W., Troisi, A. & Ma, H. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Adv. Energy Mater.* **8**, 1801032 (2018).
62. Padula, D., Simpson, J. D. & Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horiz.* **6**, 343–349 (2019).
63. Wen, Y., Fu, L., Li, G., Ma, J. & Ma, H. Accelerated Discovery of Potential Organic Dyes for Dye-Sensitized Solar Cells by Interpretable Machine Learning Models and Virtual Screening. *Sol. RRL* **31**, 2000110 (2020).
64. Lee, M.-H. Robust random forest based non-fullerene organic solar cells efficiency prediction. *Org. Electron.* **76**, 105465 (2020).
65. Manzhos, S. Machine learning for the solution of the Schrödinger equation. *Machine Learning: Science and Technology* **1** (2020).
66. Coe, J. P. Machine Learning Configuration Interaction. *J. Chem. Theory Comput.* **14**, 5739–5749 (2018).
67. Babaei, M., Azar, Y. T. & Sadeghi, A. Locality meets machine learning: Excited and ground-state energy surfaces of large systems at the cost of small ones. *Phys. Rev. B* **101** (2020).
68. Reshetnyak, Y. K., Koshevnik, Y. & Burstein, E. A. Decomposition of Protein Tryptophan Fluorescence Spectra into Log-Normal Components. III. Correlation between Fluorescence and Microenvironment Parameters of Individual Tryptophan Residues. *Biophys. J.* **81**, 1735–1758 (2001).
69. Hixon, J. & Reshetnyak, Y. Algorithm for the Analysis of Tryptophan Fluorescence Spectra and Their Correlation with Protein Structural Parameters. *Algorithms* **2**, 1155–1176 (2009).
70. Li, X., Xie, Y., Hu, D. & Lan, Z. Analysis of the Geometrical Evolution in On-the-Fly Surface-Hopping Nonadiabatic Dynamics with Machine Learning Dimensionality Reduction Approaches: Classical Multidimensional Scaling and Isometric Feature Mapping. *J. Chem. Theory Comput.* **13**, 4611–4623 (2017).
71. Li, X., Hu, D., Xie, Y. & Lan, Z. Analysis of trajectory similarity and configuration similarity in on-the-fly surface-hopping simulation on multi-channel nonadiabatic photoisomerization dynamics. *J. Chem. Phys.* **149**, 244104 (2018).
72. Zhou, G., Chu, W. & Prezhdo, O. V. Structural Deformation Controls Charge Losses in MAPbI3: Unsupervised Machine Learning of Nonadiabatic Molecular Dynamics. *ACS Energy Lett.* **5**, 1930–1938 (2020).
73. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).
74. Gomez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
75. Jørgensen, P. B. *et al.* Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **148**, 241735 (2018).
76. Peng, S. P. & Zhao, Y. Convolutional Neural Networks for the Design and Analysis of Non-Fullerene Acceptors. *J. Chem. Inf. Model.* **59**, 4993–5001 (2019).
77. Sumita, M., Yang, X., Ishihara, S., Tamura, R. & Tsuda, K. Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. *ACS Cent. Sci.* **4**, 1126–1133 (2018).
78. Sumita, M. *et al.* De Novo Creation of a Naked-Eye-Detectable Fluorescent Molecule Based on Quantum-Chemical Computation and Machine Learning. *ChemRxiv. Preprint* DOI: 10.26434/chemrxiv.14306522.v1 (2021).
79. Chapelle, O., Schölkopf, B. & Zien, A. *Semi-Supervised Learning*. DOI: (The MIT Press, 2006).
80. Duan, C., Liu, F., Nandy, A. & Kulik, H. J. Semi-Supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost. *J. Phys. Chem. Lett.* **11**, 6640–6648 (2020).

81. Ju, C.-W., Bai, H., Li, B. & Liu, R. Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields. *J. Chem. Inf. Model.* **61**, 1053–1065 (2021).
82. Smith, J. S. *et al.* Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 2903 (2019).
83. Pronobis, W., Schütt, K. T., Tkatchenko, A. & Müller, K.-R. Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning. *Eur. Phys. J. B* **91**, 178 (2018).
84. Ramakrishnan, R., Hartmann, M., Tapavicza, E. & von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **143**, 084111 (2015).
85. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
86. Kang, B., Seok, C. & Lee, J. Prediction of Molecular Electronic Transitions Using Random Forests. *J. Chem. Inf. Model.* **60**, 5984–5994 (2020).
87. Ma, J. *et al.* Transferable Multilevel Attention Neural Network for Accurate Prediction of Quantum Chemistry Properties via Multitask Learning. *J. Chem. Inf. Model.* **61**, 1066–1082 (2021).
88. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
89. Wang, X. *et al.* Improving the Accuracy of Density-Functional Theory Calculation: The Statistical Correction Approach. *J. Phys. Chem. A* **108**, 8514–8525 (2004).
90. Li, H. *et al.* Improving the accuracy of density-functional theory calculation: the genetic algorithm and neural network approach. *J. Chem. Phys.* **126**, 144101 (2007).
91. Gao, T. *et al.* An accurate density functional theory calculation for electronic excitation energies: the least-squares support vector machine. *J. Chem. Phys.* **130**, 184104 (2009).
92. Cui, J. *et al.* AdaBoost Ensemble Correction Models for TDDFT Calculated Absorption Energies. *IEEE Access* **7**, 38397–38406 (2019).
93. Lee, C.-K. *et al.* Transfer learning with graph neural networks for optoelectronic properties of conjugated oligomers. *J. Chem. Phys.* **154**, 024906 (2021).
94. Paul, A. *et al.* in *International Joint Conference on Neural Networks* DOI: (Budapest, Hungary, 2019).
95. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
96. Thompson, K. & Martıńez, T. J. Ab initio/interpolated quantum dynamics on coupled electronic states with full configuration interaction wave functions. *J. Chem. Phys.* **110**, 1376–1382 (1999).
97. Netzloff, H. M., Collins, M. A. & Gordon, M. S. Growing multiconfigurational potential energy surfaces with applications to $X+H_2$ (X=C,N,O) reactions. *J. Chem. Phys.* **124**, 154104 (2006).
98. Montavon, G. *et al.* Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **15**, 095003 (2013).
99. Stuke, A. *et al.* Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *J. Chem. Phys.* **150**, 204121 (2019).
100. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
101. Liang, J. *et al.* QM-symex, update of the QM-sym database with excited state information for 173 kilo molecules. *Sci. Data* **7**, 400 (2020).

102. Huang, B. & von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **145**, 161102 (2016).
103. Ramakrishnan, R. & von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *CHIMIA* **69**, 182–186 (2015).
104. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
105. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J Phys Chem Lett* **9**, 1668–1673 (2018).
106. Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
107. Pereira, F. *et al.* Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *J. Chem. Inf. Model.* **57**, 11–21 (2017).
108. Faber, F. A. *et al.* Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
109. Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148**, 241717 (2018).
110. Liu, F., Duan, C. & Kulik, H. J. Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening. *J. Phys. Chem. Lett.* **11**, 8067–8076 (2020).
111. Lu, S. *et al.* Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **9**, 3405 (2018).
112. Pilania, G. *et al.* Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
113. Ghosh, K. *et al.* Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Adv. Sci.* **6**, 1801367 (2019).
114. Pinheiro, G. A. *et al.* Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset. *J. Phys. Chem. A* **124**, 9854–9866 (2020).
115. Westermayr, J. & Maurer, R. J. Physically inspired deep learning of molecular excitations and photoemission spectra. *arXiv:2103.09948v1 [physics.chem-ph]* DOI: (2021).
116. Lopez, S. A. *et al.* The Harvard organic photovoltaic dataset. *Sci. Data.* **3**, 160086 (2016).
117. Hoja, J. *et al.* QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **8**, 43 (2021).
118. Stuke, A. *et al.* Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data.* **7**, 58 (2020).
119. Nakata, M., Shimazaki, T., Hashimoto, M. & Maeda, T. PubChemQC PM6: Data Sets of 221 Million Molecules with Optimized Molecular Geometries and Electronic Properties. *J. Chem. Inf. Model.* **60**, 5891–5899 (2020).
120. Snurr, R. Q. *et al.* Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery with a New Electronic Structure Database. *ChemRxiv. Preprint* DOI: 10.26434/chemrxiv.13147616.v1 (2020).
121. Gonze, X. & Scheffler, M. Exchange and Correlation Kernels at the Resonance Frequency: Implications for Excitation Energies in Density-Functional Theory. *Phys. Rev. Lett.* **82**, 4416–4419 (1999).
122. Häse, F., Valleau, S., Pyzer-Knapp, E. & Aspuru-Guzik, A. Machine learning exciton dynamics. *Chem. Sci.* **7**, 5139–5147 (2016).
123. Tong, Q. *et al.* Combining Machine Learning Potential and Structure Prediction for Accelerated Materials Design and Discovery. *J. Phys. Chem. Lett.* **11**, 8710–8720 (2020).

124. Lu, C. *et al.* Deep Learning for Optoelectronic Properties of Organic Semiconductors. *J. Phys. Chem. C* **124**, 7048–7060 (2020).
125. Richings, G. W. & Habershon, S. Direct Grid-Based Nonadiabatic Dynamics on Machine-Learned Potential Energy Surfaces: Application to Spin-Forbidden Processes. *J. Phys. Chem. A* **124**, 9299–9313 (2020).
126. Ueno, S. & Tanimura, Y. Modeling and simulating excited-state dynamics of a system in condensed phases: Machine Learning approach. *arXiv:2102.02427v1 [physics.chem-ph]* DOI: (2021).
127. Zheng, C. *et al.* Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Computational Materials* **4**, 12 (2018).
128. Torrisi, S. B. *et al.* Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Comput. Mater.* **6**, 109 (2020).
129. Timoshenko, J. & Frenkel, A. I. "Inverting" X-ray Absorption Spectra of Catalysts by Machine Learning in Search for Activity Descriptors. *ACS Catal.* **9**, 10192–10211 (2019).
130. Rodríguez, M. & Kramer, T. Machine learning of two-dimensional spectroscopic data. *Chem. Phys.* **520**, 52–60 (2019).
131. Xie, C., Zhu, X., Yarkony, D. R. & Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. IV. Coupled diabatic potential energy matrices. *J. Chem. Phys.* **149**, 144107 (2018).
132. Carbone, M. R., Topsakal, M., Lu, D. & Yoo, S. Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy. *Phys. Rev. Lett.* **124**, 156401 (2020).
133. Rankine, C. D., Madkhali, M. M. M. & Penfold, T. J. A Deep Neural Network for the Rapid Prediction of X-ray Absorption Spectra. *J. Phys. Chem. A* **124**, 4263–4270 (2020).
134. Aarva, A., Deringer, V. L., Sainio, S., Laurila, T. & Caro, M. A. Understanding X-ray Spectroscopy of Carbonaceous Materials by Combining Experiments, Density Functional Theory, and Machine Learning. Part I: Fingerprint Spectra. *Chemistry of Materials* **31**, 9243–9255 (2019).
135. Aarva, A., Deringer, V. L., Sainio, S., Laurila, T. & Caro, M. A. Understanding X-ray Spectroscopy of Carbonaceous Materials by Combining Experiments, Density Functional Theory, and Machine Learning. Part II: Quantitative Fitting of Spectra. *Chemistry of Materials* **31**, 9256–9267 (2019).
136. Kolesnichenko, P. V., Zhang, Q., Zheng, C., Fuhrer, M. S. & Davis, J. A. Multidimensional analysis of excitonic spectra of monolayers of tungsten disulphide: toward computer-aided identification of structural and environmental perturbations of 2D materials. *Mach. Learn.: Sci. Technol.* **2**, 025021 (2021).
137. Mukherjee, B. *et al.* Beyond Born–Oppenheimer theory for spectroscopic and scattering processes. *Int. Rev. Phys. Chem.* **38**, 287–341 (2019).
138. Nelson, T., Naumov, A., Fernandez-Alberti, S. & Tretiak, S. Nonadiabatic excited-state molecular dynamics: On-the-fly limiting of essential excited states. *Chem. Phys.* **481**, 84–90 (2016).
139. Ardiansyah, M. & Brorsen, K. R. Mixed Quantum-Classical Dynamics with Machine Learning-Based Potentials via Wigner Sampling. *J. Phys. Chem. A* **124**, 9326–9331 (2020).
140. Li, J. *et al.* Automatic discovery of photoisomerization mechanisms with nanosecond machine learning photodynamics simulations. *Chem. Sci.* DOI: 10.1039/d0sc05610c (2021).
141. Posenitskiy, E. *Non-adiabatic molecular dynamics of PAH-related complexes* PhD thesis, Université Paul Sabatier, (2020).
142. Richings, G. W. *et al.* Quantum dynamics simulations using Gaussian wavepackets: the vMCG method. *Int. Rev. Phys. Chem.* **34**, 269–308 (2015).
143. Koch, W. & Zhang, D. H. Communication: separable potential energy surfaces from multiplicative artificial neural networks. *J. Chem. Phys.* **141**, 021101 (2014).

144. Frankcombe, T. J., Collins, M. A. & Worth, G. A. Converged quantum dynamics with modified Shepard interpolation and Gaussian wave packets. *Chem. Phys. Lett.* **489**, 242–247 (2010).
145. Alborzpour, J. P., Tew, D. P. & Habershon, S. Efficient and accurate evaluation of potential energy matrix elements for quantum dynamics using Gaussian process regression. *J. Chem. Phys.* **145**, 174112 (2016).
146. Richings, G. W. & Habershon, S. Direct grid-based quantum dynamics on propagated diabatic potential energy surfaces. *Chem. Phys. Lett.* **683**, 228–233 (2017).
147. Richings, G. W. & Habershon, S. Direct Quantum Dynamics Using Grid-Based Wave Function Propagation and Machine-Learned Potential Energy Surfaces. *J. Chem. Theory Comput.* **13**, 4012–4024 (2017).
148. Richings, G. W. & Habershon, S. MCTDH on-the-fly: Efficient grid-based quantum dynamics without pre-computed potential energy surfaces. *J. Chem. Phys.* **148**, 134116 (2018).
149. Polyak, I., Richings, G. W., Habershon, S. & Knowles, P. J. Direct quantum dynamics using variational Gaussian wavepackets and Gaussian process regression. *J. Chem. Phys.* **150**, 041101 (2019).
150. Zhang, D. H., Collins, M. A. & Lee, S. Y. First-principles theory for the H + H_2O, D_2O reactions. *Science* **290**, 961–963 (2000).
151. Crespos, C., Collins, M. A., Pijper, E. & Kroes, G. J. Application of the modified Shepard interpolation method to the determination of the potential energy surface for a molecule-surface reaction: H_2 + Pt(111). *J. Chem. Phys.* **120**, 2392–2404 (2004).
152. Evenhuis, C. R. & Collins, M. A. Interpolation of diabatic potential energy surfaces. *J. Chem. Phys.* **121**, 2515–2527 (2004).
153. Evenhuis, C. R., Lin, X., Zhang, D. H., Yarkony, D. & Collins, M. A. Interpolation of diabatic potential-energy surfaces: quantum dynamics on ab initio surfaces. *J. Chem. Phys.* **123**, 134110 (2005).
154. Godsi, O., Evenhuis, C. R. & Collins, M. A. Interpolation of multidimensional diabatic potential energy matrices. *J. Chem. Phys.* **125**, 104105 (2006).
155. Evenhuis, C. & Martinez, T. J. A scheme to interpolate potential energy surfaces and derivative coupling vectors without performing a global diabatization. *J. Chem. Phys.* **135**, 224110 (2011).
156. Lenzen, T. & Manthe, U. Neural network based coupled diabatic potential energy surfaces for reactive scattering. *J. Chem. Phys.* **147**, 084105 (2017).
157. Guan, Y., Fu, B. & Zhang, D. H. Construction of diabatic energy surfaces for LiFH with artificial neural networks. *J. Chem. Phys.* **147**, 224307 (2017).
158. Yuan, J., He, D., Wang, S., Chen, M. & Han, K. Diabatic potential energy surfaces of MgH_2^+ and dynamic studies for the $Mg^+(3p)$ + H_2 → MgH^+ + H reaction. *Phys. Chem. Chem. Phys.* **20**, 6638–6647 (2018).
159. Williams, D. M. G. & Eisfeld, W. Neural network diabatization: A new ansatz for accurate high-dimensional coupled potential energy surfaces. *J. Chem. Phys.* **149**, 204106 (2018).
160. Guan, Y., Zhang, D. H., Guo, H. & Yarkony, D. R. Representation of coupled adiabatic potential energy surfaces using neural network based quasi-diabatic Hamiltonians: 1,2 $^2A'$ states of LiFH. *Phys. Chem. Chem. Phys.* **21**, 14205–14213 (2019).
161. Guan, Y., Guo, H. & Yarkony, D. R. Neural network based quasi-diabatic Hamiltonians with symmetry adaptation and a correct description of conical intersections. *J. Chem. Phys.* **150**, 214101 (2019).
162. Yin, Z., Guan, Y., Fu, B. & Zhang, D. H. Two-state diabatic potential energy surfaces of ClH_2 based on nonadiabatic couplings with neural networks. *Phys. Chem. Chem. Phys.* **21**, 20372–20383 (2019).
163. Zheng, F., Gao, X. & Eisfeld, A. Excitonic Wave Function Reconstruction from Near-Field Spectra Using Machine Learning Techniques. *Phys. Rev. Lett.* **123**, 163202 (2019).

164. Guan, Y., Guo, H. & Yarkony, D. R. 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.* **16**, 302–313 (2020).
165. Shen, Y. & Yarkony, D. R. 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 Full 39 Degrees of Freedom. *J. Phys. Chem. A* **124**, 4539–4548 (2020).
166. Shu, Y. & Truhlar, D. G. Diabatization by Machine Intelligence. *J. Chem. Theory Comput.* **16**, 6456–6464 (2020).
167. Shu, Y., Varga, Z., Sampaio de Oliveira-Filho, A. G. & Truhlar, D. G. Permutationally Restrained Diabatization by Machine Intelligence. *J. Chem. Theory Comput.* **ASAP** (2021).
168. Ha, J. K., Kim, K. & Min, S. K. Machine Learning-Assisted Excited State Molecular Dynamics with the State-Interaction State-Averaged Spin-Restricted Ensemble-Referenced Kohn-Sham Approach. *J. Chem. Theory Comput.* **17**, 694–702 (2021).
169. Guan, Y. & Yarkony, D. R. Accurate Neural Network Representation of the Ab Initio Determined Spin-Orbit Interaction in the Diabatic Representation Including the Effects of Conical Intersections. *J. Phys. Chem. Lett.* **11**, 1848–1858 (2020).
170. Zheng, J., Xu, X., Meana-Pañeda, R. & Truhlar, D. G. Army ants tunneling for classical simulations. *Chem. Sci.* **5**, 2091–2099 (2014).
171. Chen, W. K., Zhang, Y., Jiang, B., Fang, W. H. & Cui, G. Efficient Construction of Excited-State Hessian Matrices with Machine Learning Accelerated Multilayer Energy-Based Fragment Method. *J. Phys. Chem. A* **124**, 5684–5695 (2020).
172. Krämer, M. *et al.* Charge and Exciton Transfer Simulations Using Machine-Learned Hamiltonians. *J. Chem. Theory Comput.* **16**, 4061–4070 (2020).
173. Moiseyev, N. *Non-Hermitian Quantum Mechanics*. DOI: (Cambridge University Press, 2011).
174. Yang, B., He, B., Wan, J., Kubal, S. & Zhao, Y. Applications of neural networks to dynamics simulation of Landau-Zener transitions. *Chem. Phys.* **528**, 110509 (2020).
175. Herrera Rodriguez, L. E. & Kananenka, A. A. Convolutional neural networks for long-time dissipative quantum dynamics. *J. Phys. Chem. Lett.* **12**, 2476–2483 (2021).
176. Ueno, S. & Tanimura, Y. Modeling Intermolecular and Intramolecular Modes of Liquid Water Using Multiple Heat Baths: Machine Learning Approach. *J. Chem. Theory Comput.* **16**, 2099–2108 (2020).
177. Häse, F., Kreisbeck, C. & Aspuru-Guzik, A. Machine learning for quantum dynamics: deep learning of excitation energy transfer properties. *Chem. Sci.* **8**, 8419–8426 (2017).
178. Hachmann, J. *et al.* The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2**, 2241–2251 (2011).
179. Zhuo, Y. & Brgoch, J. Opportunities for Next-Generation Luminescent Materials through Artificial Intelligence. *J. Phys. Chem. Lett.* **12**, 764–772 (2021).
180. Misra, M., Andrienko, D., Baumeier, B., Faulon, J. L. & von Lilienfeld, O. A. Toward Quantitative Structure-Property Relationships for Charge Transfer Rates of Polycyclic Aromatic Hydrocarbons. *J. Chem. Theory Comput.* **7**, 2549–2555 (2011).
181. Gomez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
182. Atahan-Evrenk, S. A quantitative structure–property study of reorganization energy for known p-type organic semiconductors. *RSC Adv.* **8**, 40330–40337 (2018).

183. Atahan-Evrenk, S. & Atalay, F. B. Prediction of Intramolecular Reorganization Energy Using Machine Learning. *J. Phys. Chem. A* **123**, 7855–7863 (2019).
184. Lee, M. H. Machine Learning for Understanding the Relationship between the Charge Transport Mobility and Electronic Energy Levels for n-Type Organic Field-Effect Transistors. *Adv. Electron. Mater.* **5**, 1900573 (2019).
185. Kar, S., Roy, J., Leszczynska, D. & Leszczynski, J. Power Conversion Efficiency of Arylamine Organic Dyes for Dye-Sensitized Solar Cells (DSSCs) Explicit to Cobalt Electrolyte: Understanding the Structural Attributes Using a Direct QSPR Approach. *Computation* **5**, 2 (2017).
186. Kar, S., Roy, J. K. & Leszczynski, J. In silico designing of power conversion efficient organic lead dyes for solar cells using todays innovative approaches to assure renewable energy for future. *npj Comput. Mater.* **3**, 22 (2017).
187. Li, Z. *et al.* Robot-Accelerated Perovskite Investigation and Discovery. *Chem. Mater.* **32**, 5650–5663 (2020).
188. Chudakov, D. M., Matz, M. V., Lukyanov, S. & Lukyanov, K. A. Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol. Rev.* **90**, 1103–1163 (2010).
189. da Silva, R. S., Marins, L. F., Almeida, D. V., Dos Santos Machado, K. & Werhli, A. V. A comparison of classifiers for predicting the class color of fluorescent proteins. *Comput. Biol. Chem.* **83**, 107089 (2019).
190. Scharber, M. C. *et al.* Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10 % Energy-Conversion Efficiency. *Adv. Mater.* **18**, 789–794 (2006).
191. Bella, F., Nair, J. R. & Gerbaldi, C. Towards green, efficient and durable quasi-solid dye-sensitized solar cells integrated with a cellulose-based gel-polymer electrolyte optimized by a chemometric DoE approach. *RSC Adv.* **3**, 15993–16001 (2013).
192. Pugliese, D. *et al.* A chemometric approach for the sensitization procedure of ZnO flowerlike microstructures for dye-sensitized solar cells. *ACS Appl. Mater. Interfaces* **5**, 11288–11295 (2013).
193. Bella, F., Sacco, A., Pugliese, D., Laurenti, M. & Bianco, S. Additives and salts for dye-sensitized solar cells electrolytes: what is the best choice? *J. Power Sources* **264**, 333–343 (2014).
194. Bella, F., Mobarak, N. N., Jumaah, F. N. & Ahmad, A. From seaweeds to biopolymeric electrolytes for third generation solar cells: An intriguing approach. *Electrochim. Acta* **151**, 306–311 (2015).
195. Liu, D., Tan, Y., Khoram, E. & Yu, Z. Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures. *ACS Photonics* **5**, 1365–1369 (2018).
196. Liu, R. *et al.* Causal Inference Machine Learning Leads Original Experimental Discovery in CdSe/CdS Core/Shell Nanoparticles. *J. Phys. Chem. Lett.* **11**, 7232–7238 (2020).
197. Hong, S. *et al.* Defect Healing in Layered Materials: A Machine Learning-Assisted Characterization of MoS_2 Crystal Phases. *J. Phys. Chem. Lett.* **10**, 2739–2744 (2019).
198. Huang, Y. *et al.* Structure–Property Correlation Study for Organic Photovoltaic Polymer Materials Using Data Science Approach. *J. Phys. Chem. C* **124**, 12871–12882 (2020).
199. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Adv. Sci.* **6**, 1900808 (2019).
200. Huwig, K., Fan, C. & Springborg, M. From properties to materials: An efficient and simple approach. *J. Chem. Phys.* **147**, 234105 (2017).
201. Kanal, I. Y., Owens, S. G., Bechtel, J. S. & Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **4**, 1613–1623 (2013).
202. Rajan, A. C. *et al.* Machine-Learning-Assisted Accurate Band Gap Predictions of Functionalized MXene. *Chem. Mater.* **30**, 4031–4038 (2018).

203. Mishra, A. *et al.* Accelerated Data-Driven Accurate Positioning of the Band Edges of MXenes. *J. Phys. Chem. Lett.* **10**, 780–785 (2019).
204. Kunkel, C., Schober, C., Oberhofer, H. & Reuter, K. Knowledge discovery through chemical space networks: the case of organic electronics. *J. Mol. Model.* **25**, 87 (2019).
205. Kunkel, C., Schober, C., Margraf, J. T., Reuter, K. & Oberhofer, H. Finding the Right Bricks for Molecular Legos: A Data Mining Approach to Organic Semiconductor Design. *Chemistry of Materials* **31**, 969–978 (2019).
206. Đorđević, N. *et al.* Machine Learning for Analysis of Time-Resolved Luminescence Data. *ACS Photonics* **5**, 4888–4895 (2018).
207. Musil, F. *et al.* Machine learning for the structure-energy-property landscapes of molecular crystals. *Chem. Sci.* **9**, 1289–1300 (2018).
208. Lederer, J., Kaiser, W., Mattoni, A. & Gagliardi, A. Machine Learning-Based Charge Transport Computation for Pentacene. *Adv. Theory Simul.* **2**, 1800136 (2018).
209. Çaylak, O., Yaman, A. & Baumeier, B. Evolutionary Approach to Constructing a Deep Feedforward Neural Network for Prediction of Electronic Coupling Elements in Molecular Materials. *J. Chem. Theory Comput.* **15**, 1777–1784 (2019).
210. Wang, C. I., Braza, M. K. E., Claudio, G. C., Nellas, R. B. & Hsu, C. P. Machine Learning for Predicting Electron Transfer Coupling. *J. Phys. Chem. A* **123**, 7792–7802 (2019).
211. Wang, C. I., Joanito, I., Lan, C. F. & Hsu, C. P. Artificial neural networks for predicting charge transfer coupling. *J. Chem. Phys.* **153**, 214113 (2020).
212. Bag, S., Aggarwal, A. & Maiti, P. K. Machine Learning Prediction of Electronic Coupling between the Guanine Bases of DNA. *J. Phys. Chem. A* **124**, 7658–7664 (2020).
213. Rinderle, M., Kaiser, W., Mattoni, A. & Gagliardi, A. Machine-Learned Charge Transfer Integrals for Multiscale Simulations in Organic Thin Films. *J. Phys. Chem. C* **124**, 17733–17743 (2020).
214. Miller, E. D., Jones, M. L., Henry, M. M., Stanfill, B. & Jankowski, E. Machine learning predictions of electronic couplings for charge transport calculations of P3HT. *AIChE Journal* **65**, e16760 (2019).
215. Thawani, A. R. *et al.* The Photoswitch Dataset: A Molecular Machine Learning Benchmark for the Advancement of Synthetic Chemistry. *ChemRxiv. Preprint* DOI: 10.26434/chemrxiv.12609899.v1 (2020).
216. Roch, L. M. *et al.* From Absorption Spectra to Charge Transfer in Nanoaggregates of Oligomers with Machine Learning. *ACS Nano* **14**, 6589–6598 (2020).
217. Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
218. Kaelbling, L. P., Littman, M. L. & Moore, A. W. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* **4**, 237–285 (1996).
219. Granda, J. M., Donina, L., Dragone, V., Long, D. L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
220. MacLeod, B. P. *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
221. Bolovinos, A., Philis, J., Pantos, E., Tsekeris, P. & Andritsopoulos, G. The methylbenzenes vis-à-vis benzene. *J. Mol. Spectrosc.* **94**, 55–68 (1982).
222. Casida, M. E. & Huix-Rotllant, M. Progress in Time-Dependent Density-Functional Theory. *Annu. Rev. Phys. Chem.* **63**, 287–323 (2012).
223. Pulay, P. A perspective on the CASPT2 method. *Int. J. Quant. Chem.* **111**, 3273–3279 (2011).

224. Hättig, C. in *Computational Nanoscience: Do It Yourself!* DOI: *NIC Series* (eds J. Grotendort, S. Blügel, & D. Marx) 1–34 (Forschungszentrum Jülich, 2006).
225. Dreuw, A. & Wormit, M. The Algebraic Diagrammatic Construction Scheme for the Polarization Propagator for the Calculation of Excited States. *WIREs: Comput. Mol. Sci.* **5**, 82–95 (2015).
226. Golze, D., Dvorak, M. & Rinke, P. The GW Compendium: A Practical Guide to Theoretical Photoemission Spectroscopy. *Front. Chem.* **7**, 377 (2019).
227. Kranz, J. J. *et al.* Time-Dependent Extension of the Long-Range Corrected Density Functional Based Tight-Binding Method. *J. Chem. Theory Comput.* **13**, 1737–1747 (2017).
228. Thiel, W. Semiempirical quantum-chemical methods. *WIREs: Comput. Mol. Sci.* **4**, 145–157 (2014).
229. Weingart, O. Combined Quantum and Molecular Mechanics (QM/MM) Approaches to Simulate Ultrafast Photodynamics in Biological Systems. *Curr. Org. Chem.* **21**, 586–601 (2017).
230. Plasser, F., Gómez, S., Menger, M. F. S. J., Mai, S. & González, L. Highly efficient surface hopping dynamics using a linear vibronic coupling model. *Phys. Chem. Chem. Phys.* **21**, 57–69 (2019).
231. Niu, Y., Peng, Q., Deng, C., Gao, X. & Shuai, Z. Theory of Excited State Decays and Optical Spectra: Application to Polyatomic Molecules. *J. Phys. Chem. A* **114**, 7817–7831 (2010).
232. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn, DOI: 763 (Springer-Verlag, 2009).
233. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv. Preprint.* DOI, arXiv:1312.6114 [stat.ML] (2013).
234. Settles, B. Active Learning Literature Survey. Computer Sciences Technical Report 1648. (2009).
235. Dral, P. O., Owens, A., Yurchenko, S. N. & Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **146**, 244108 (2017).
236. Schütt, K. T. *et al.* SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural. Inf. Process. Syst.* **30**, 992–1002 (2017).
237. Zhang, L., Han, J., Wang, H., Car, R. & E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
238. Zhang, L. F. *et al.* End-To-End Symmetry Preserving Inter-Atomic Potential Energy Model for Finite and Extended Systems. *Adv. Neural. Inf. Process. Syst.* **31**, 4436–4446 (2018).
239. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
240. Hansen, K. *et al.* Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
241. Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
242. Dral, P. O. *MLatom*: A Program Package for Quantum Chemical Research Assisted by Machine Learning. *J. Comput. Chem.* **40**, 2339–2347 (2019).
243. Schütt, K. T. *et al.* SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **15**, 448–455 (2019).
244. Wang, H., Zhang, L., Han, J. & E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun*, **228**, 178–184 (2018).
245. Himanen, L. *et al.* DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
246. Heller, E. J. *The Semiclassical Way to Dynamics and Spectroscopy*. DOI: (Princeton University Press, 2018).

247. Mukamel, S. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. *Annu. Rev. Phys. Chem.* **51**, 691–729 (2000).
248. Bai, S., Mansour, R., Stojanovic, L., Toldo, J. M. & Barbatti, M. On the origin of the shift between vertical excitation and band maximum in molecular photoabsorption. *J. Mol. Model.* **26**, 107 (2020).
249. Crespo-Otero, R. & Barbatti, M. Spectrum simulation and decomposition with nuclear ensemble: formal derivation and application to benzene, furan and 2-phenylfuran. *Theor. Chem. Acc.* **131**, 1237 (2012).
250. Segarra-Marti, J. *et al.* Modeling multidimensional spectral lineshapes from first principles: application to water-solvated adenine. *Faraday Discuss.* **221**, 219–244 (2019).
251. Biczysko, M., Bloino, J., Santoro, F. & Barone, V. in *Computational Strategies for Spectroscopy: From Small Molecules to Nano Systems* DOI: 10.1002/9781118008720.ch8 (ed Vincenzo Barone) (John Wiley & Sons, Inc., 2011).
252. Petrenko, T. & Neese, F. Analysis and prediction of absorption band shapes, fluorescence band shapes, resonance Raman intensities, and excitation profiles using the time-dependent theory of electronic spectroscopy. *J. Chem. Phys.* **127**, 164319 (2007).
253. Tanimura, Y. & Kubo, R. Time Evolution of a Quantum System in Contact with a Nearly Gaussian-Markoffian Noise Bath. *J. Phys. Soc. Jpn.* **58**, 101–114 (1989).
254. Worth, G. A. & Cederbaum, L. S. Beyond Born-Oppenheimer: molecular dynamics through a conical intersection. *Annu. Rev. Phys. Chem.* **55**, 127–158 (2004).
255. Worth, G. A., Meyer, H. D., Köppel, H., Cederbaum, L. S. & Burghardt, I. Using the MCTDH Wavepacket Propagation Method to Describe Multimode Non-Adiabatic Dynamics. *Int. Rev. Phys. Chem.* **27**, 569–606 (2008).
256. Barbatti, M. & Crespo-Otero, R. in *Density-Functional Methods for Excited States* DOI: 10.1007/128_2014_605 (eds Nicolas Ferré, Michael Filatov, & Miquel Huix-Rotllant) 415–444 (Springer International Publishing, 2016).
257. Nelson, T. R. *et al.* Non-adiabatic Excited-State Molecular Dynamics: Theory and Applications for Modeling Photophysics in Extended Molecular Materials. *Chem. Rev.* **120**, 2215–2287 (2020).
258. Yue, L. *et al.* Benchmark Performance of Global Switching versus Local Switching for Trajectory Surface Hopping Molecular Dynamics Simulation: Cis↔Trans Azobenzene Photoisomerization. *ChemPhysChem* **18**, 1274–1287 (2017).
259. Suchan, J., Janos, J. & Slavicek, P. Pragmatic Approach to Photodynamics: Mixed Landau-Zener Surface Hopping with Intersystem Crossing. *J. Chem. Theory Comput.* **16**, 5809–5820 (2020).
260. Köppel, H. in *Conical Intersections* DOI: 10.1142/9789812565464_0004 (eds Wolfgang Domcke, David R. Yarkony, & Horst Köppel) 175–204 (World Scientific, 2004).
261. Li, S. L., Truhlar, D. G., Schmidt, M. W. & Gordon, M. S. Model space diabatization for quantum photochemistry. *J. Chem. Phys.* **142**, 064106 (2015).
262. Zhu, X. & Yarkony, D. R. 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.* **132**, 104101 (2010).
263. Wittenbrink, N., Venghaus, F., Williams, D. & Eisfeld, W. A new approach for the development of diabatic potential energy surfaces: Hybrid block-diagonalization and diabatization by ansatz. *J. Chem. Phys.* **145**, 184108 (2016).

Display Items

Figures

FIG. 1. Potential energy surfaces (PES) of molecules. Quantum mechanical (QM) modeling of excited-state PES faces many challenges owing to the description of the diverse electronic-structure characters in different regions of the surface (represented by different colors). The figure schematically illustrates a one-dimensional PES, but for a molecule with N atoms, the full PES dimensionality is $3N–6$. This example shows three types of electronic-structure characters, nπ* (indicated in orange), ππ* (blue), and closed shell (cs, dark green). The main electronic configuration of each character is given in the lower panel.

FIG. 2. Machine learning (ML) for the study of excited states. ML can be used to predict properties associated to the excited states of a molecule fast after learning on quantum mechanical (QM) or experimental data. ML models can also be used to improve QM methods, analyze data and discover new compounds. The final deliverables of ML are then spectra, electronic property surfaces (such as potential energy, transition dipole moments, nonadiabatic, and spin–orbit couplings surfaces), dynamics simulations or new materials.

FIG. 3. Timeline of pioneering developments in the field of machine learning for excited-state research.

FIG. 4. Machine learning (ML) spectra calculated with the nuclear ensemble approach (NEA). a | NEA absorption spectrum requires quantum mechanical (QM) calculations of transition energy $\Delta E(\mathbf{R})$ and oscillator strength $f(\mathbf{R})$ for many (several hundreds to few thousands) nuclear geometries $\mathbf{R}$. The spectrum $\sigma(\Delta E)$ is given as a distribution of $\Delta E(\mathbf{R})$ weighted by $f(\mathbf{R})$. **b** | Absorption spectrum can be obtained without much loss of accuracy by using ML trained on a relatively small number of QM data, to predict $\Delta E(\mathbf{R})$ and $f(\mathbf{R})$. **c** | Example of an absorption spectrum of benzene calculated with ML trained on 250 points (ML–NEA, red), which is in very good agreement with the spectrum calculated from 50 thousand time-dependent density functional theory (TD-DFT) calculations (TDDFT–NEA, blue). ML–NEA can also accurately reproduce low-energy absorption bands present in the experimental spectrum (from Ref. [221]) that single-point convolution (TDDFT–SPC, Gaussian broadening of line spectrum, green) fails to describe.[32] **d** | Example of ML–NEA emission spectrum (red) compared to reference TD-DFT emission spectrum (blue) of sinapic acid in a proof-of-concept study.[30]

FIG. 5. Machine learning for surface-hopping excited-state dynamics. a | The calculation of excited-state dynamics with ML requires models for energies and gradients. Couplings between states are also needed. These are narrow functions, which are difficult but possible to learn.[29,45] Thus, dynamics is sometimes run with QM couplings calculated at small energy gaps between states[38] or using models without explicit couplings (like the Zhu–Nakamura approach).[37,39] **b** | A recipe for learning narrow couplings is the inclusion of points with large couplings into the training set. When the maximum point (red circle) is not among the training points (magenta circles) fed to ML, it erroneously predicts almost zero coupling (blue curve). After the maximum is included in the training set, ML faithfully reproduces the narrow coupling (red curve).[29]

FIG. 6. Strategies for optoelectronic materials design with machine learning (ML). a | ML can be used to learn ultimate target properties, such as power conversion efficiency (PCE) of a solar cell (top). Alternatively, it can be used to learn intermediate properties such

as HOMO–LUMO gap correlated with PCE (bottom). Intermediate properties are usually easier to obtain. Topologies of the materials space for both ultimate and intermediate properties are similar, but not identical. Thus, the search for optimal materials based on intermediate properties may lead to selection of a different compound, but with similar performance to the compound with the optimal ultimate property. **b** | ML methods, such as kernel ridge regression (KRR), neural networks (NN), can make fast predictions that facilitate the high-throughput virtual screening (HTVS) of materials (top). ML can also assist automated structure generation to design materials with desired excited-state properties (bottom). In case of HTVS, extensive screening of large number of materials in the existing database is required. In case of automated structure generation, ML may directly design materials with ever increasing performance.

Boxes

Box 1. Quantum chemical methods for excited states.

Task zero in excited-state simulations is the electronic-structure calculation. That is, the computation of the potential energy and other electronic properties (such as transition dipole moment) of the excited states for a fixed nuclear geometry. The quantum-chemical method to execute such task can either work under adiabatic approximation (single reference) or be tailored to assess nonadiabatic regimes (multireference).[19]

Two of the most common methods for electronic-structure calculations are the linear-response time-dependent density functional theory (TD-DFT)[222] and the complete active space perturbation theory to the second-order (CASPT2).[223] Other popular single-reference methods for electronic-structure calculations are couple cluster to approximated second-order (CC2)[224] and algebraic diagrammatic construction to second-order (ADC(2)).[225] For chemically complex systems (such as radicals), the expensive but reliable multireference configuration interaction (MRCI)[19] may be the best choice. The workhorse for the study of excitations in materials is the GW approximation.[226] Many of these methods have parametrized versions that replace some of their most expensive routines with empirical or precomputed values. Such is the case of time-dependent density functional tight binding (TD-DFTB)[227] and MRCI based on the orthogonalization method 2 (OM2/MRCI).[228] Realistic modeling of excited-state phenomena may require considering the full molecular environment (solvents, surfaces, proteins). Such a task is usually only affordable through hybrid methods that treat the core molecule quantum mechanically and the remaining system with molecular mechanics (QM/MM).[229]

For many applications, we must go beyond the electronic structure for a single geometry. We may need to know how the electronic properties change along the reaction pathway. Such investigations may require building the potential energy surface (PES) for each state along selected nuclear coordinates or using reasonable functional guesses for the potential energy dependencies.[230] After the relevant sections of the PES are known, they can be used for estimating the probability for radiative and radiationless transitions between states, which enables the simulation of non-adiabatic dynamics evolution,[16] reaction

rates[231] and diverse types of spectra (Box 4).[9] Many of these applications have been adapted to work with on-the-fly electronic structure calculations, with the advantage of dismissing PES pre-calculation (Box 5).

Box 2. Machine learning methods for excited states.

Depending on the task at hand, the appropriate type of ML technique can be chosen. In excited-state simulations, supervised and unsupervised learnings are currently the most used approaches.

Supervised learning uses a statistical approach to find (that is, 'to learn') the relationship between features (**x**, often called descriptors in chemistry) and labels (**y**), given a labeled data set (training set).[232] The goal is then to make predictions for new features as reliable as possible. The accuracy of supervised learning strongly depends on many factors, such as the learning algorithm, choice of features and the training set's composition.[29] Many of these problems in excited-state research were surveyed in a recent Review.[46] General aspects of supervised learning applied to the study of excited states is provided in Box 3.

While supervised learning can also be used to uncover dependencies in data sets, it requires labeled data. Often, it is desirable to analyze unlabeled data sets and, for this purpose, unsupervised learning is perfectly suited.[232] For example, in excited-state simulations, unsupervised learning is used to classify similar geometries and understand key geometrical changes in photophysical processes. Unsupervised learning includes methods like the principal component analysis and k-mean clustering. Another particular class of this ML type is generative models, such as variational autoencoder, in which an encoder converts molecular representation into latent representation and a decoder converts the latent variables back to the molecular representation.[233] When generative models are coupled with a predictor (supervised learning model), they can also estimate properties of interest. It is a powerful technique for materials design.[74]

[PE: Please insert here figure box 2]

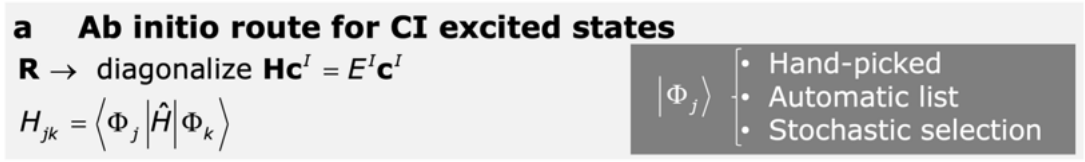
Box 3. Supervised learning for excited states.

In excited-state research, supervised learning algorithms range from relatively simple multiple linear regression, Lasso regression, and partial least squares regression with few parameters, to many variants of more complex neural networks (NNs) and kernel methods with many parameters. Gaussian process regression (also called kriging), kernel ridge regression (KRR) and support vector regression are examples of so-called kernel methods.[232] Most of the parameters of kernel methods can be calculated analytically, whereas NN parameters are subject to elaborate fitting procedures. Several ML models

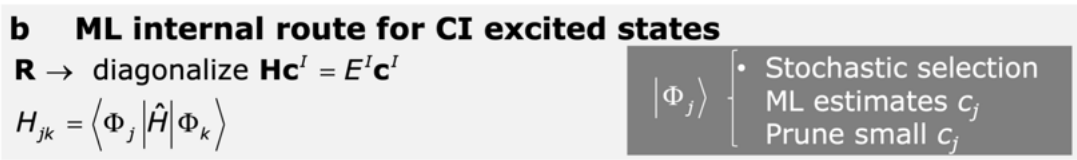
(weak learners) can be combined in an ensemble to get better accuracy (strong learner), as for example in AdaBoost, gradient boosting regression and random forests.[232] Because labeling data for supervised learning is resource-intensive in excited-state research, active learning[234] is often applied to choose the points to label with the aim of minimizing their number.

Specialized supervised learning methods were developed for creating surrogate models for the calculation of QM properties such as potential energies. Examples of such methods are the KREG **[G]**,[32,235] SchNet,[236] deep potential molecular dynamics (DPMD),[237] and deep potential–smooth edition (DeepPOT-SE)[238] models. Also, special 3D structural descriptors such as the Coulomb matrix,[239] bag-of-bonds[240] and smooth overlap of atomic positions (SOAP),[241] which can be used as input to various ML algorithms,[25] are popular in this research field. Specialized software, such as MLatom,[242] SchNetPack,[243] DeePMD-kit,[244] and DScribe[245] implementing such ML models and descriptors are also available.

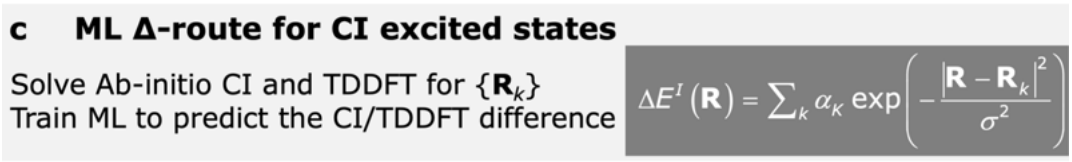
The advantages of supervised learning compared to quantum mechanical (QM) methods are that ML can learn practically any property without explicitly implementing the physical model and make incredibly fast predictions.[23,25] The disadvantages are that it is very difficult to create a ML model as transferable as QM methods.[23] These disadvantages may be overcome in hybrid QM–ML methods exploiting the transferability of lower-level QM methods and accuracy and speed of ML methods. Here we compare different approaches to calculating QM properties on an example of predicting excited-state energies at the level of configuration interaction (CI).



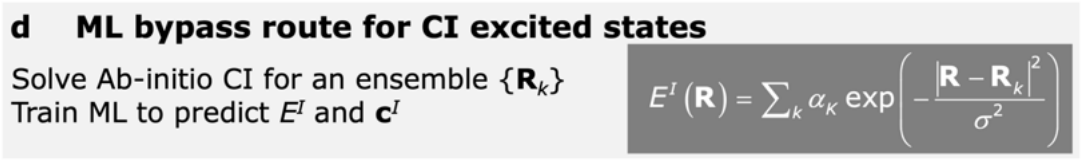
The pure QM approach to find the energies for each state I of a molecule with geometry **R** is based on describing the wavefunctions as a linear combination of configurations Φ_J (each one representing a possible electronic distribution) with coefficients **c** calculated through the diagonalization of the CI Hamiltonian **H**. This approach is very computationally costly, and its cost and accuracy are reduced by making approximations such as selecting only a subset of all possible configurations.



An implicit QM–ML approach improves the approximated CI algorithm's accuracy by optimizing the configuration selection.



Whereas, an explicit QM–ML approach (Δ-learning)[88] improves the accuracy of a lower-level approximated quantum mechanical approach, for example, time-dependent density functional theory (TDDFT), to the level of CI accuracy by learning the difference in energies ΔE between them. In case of a typical KRR-based ML model, the regression coefficients α_k are needed to be found by training on a set with nuclear geometries $\mathbf{R}_k$ (σ is a model hyperparameter). Δ-learning is very robust and generic and does not require modifications to the QM method in contrast to the implicit approach.



A pure ML approach completely bypasses the CI calculations, directly predicting properties at CI level after being trained on a precomputed ensemble of CI results. It usually requires much more training points than Δ-learning approach.[88]

Box 4. Electronic spectroscopy.

There are many different types of electronic spectroscopy, but their simulation always requires two essential elements: excited-state energies and transition probabilities. The transition probabilities require the calculation of quantities expressing state couplings and radiation–matter couplings, such as Franck–Condon factors, correlation functions, or transition dipole moments.[246,247]

Spectra can be simulated at different levels. The lowest, most straightforward approach is based on a single point convolution. Within this approximation, a steady-state spectrum is approximated as the sum over Gaussian bands centered at each excitation energy with the areas proportional to the respective oscillator strengths, all computed at the ground-state minimum geometry.[248] Although crude, this approximation delivers valuable information for the experimental assignment.

Nuclear ensemble approach (NEA; FIG. 4),[249] enables the prediction of information about the band envelopes using an ensemble of different nuclear geometries in the source state. In the NEA, excitation energies and transition probabilities are calculated for each point in the ensemble (FIG. 4a). An incoherent sum over these points convoluted by a line-shape function yields the spectral profile (FIG. 4b). NEA accounts for post-Condon effects, but completely neglects vibronic effects. It can be adapted to predict linear and multidimensional steady-state and time-resolved spectra.[250]

High-accuracy spectrum simulations, including detailed vibronic information, can be simulated either by using time-independent[251] or time-dependent[247] approaches. Time-independent approaches, useful for the calculation of steady-state spectra, depend on the extensive calculation of Franck–Condon factors. Time-dependent approaches involve the Fourier transform of time-domain signals and can be used for both linear and multidimensional steady-state and time-resolved spectra. The time-dependent signals can be explicitly computed through dynamics simulations or evaluated for model systems.[252,253] For steady-state absorption, the time-dependent signal is the overlap between the excited wavepacket and the ground state wavefunction.[246]

Box 5. Excited-state dynamics.

When a molecule is promoted to an excited state, its wavefunction is, in general, not an eigenvector of the Hamiltonian. For this reason, the wavefunction evolves with time according to the time-dependent Schrödinger equation.[254] This time evolution, which encompasses the transfer of the molecule between different states (nonadiabatic process), continues until the energy excess is dissipated either by heating vibrational modes, re-emitting a photon, or rearranging the bonds: the description of these processes is the goal of excited-state dynamics simulations. There are many different methods for this, which we can be split into two big classes: methods that use pre-computed potential energy surfaces (PESs) (either in the form of a grid or a functional model)[255] and methods that compute these surfaces on-the-fly during the dynamics (also known as direct dynamics).[17,256,257]

Dynamics simulations of excited states usually require electronic structure calculations for many nuclear geometries, often hundreds of thousands of them. Take, for instance, trajectory surface hopping, one of the most popular methods for excited-state dynamics with on-the-fly PES calculation.[16] In this method, we run an ensemble of classical trajectories on a PES of an excited state and use a stochastic algorithm to allow each trajectory to change to another surface during the propagation. For propagating each trajectory, we compute the energy and force for each electronic state at the current geometry to predict the next one. We also calculate couplings between states to estimate the hopping probability to other surfaces. Alternatively, these probabilities can be estimated only from the PES by using the Zhu–Nakamura[258] and Belyaev–Lebedev[259] approximations. Either way, a modest dynamics simulation may require 100 trajectories of 1 ps each, with 0.5 fs time steps, which implies that we must perform 200k electronic structure calculations.

Dynamics based on precomputed PES often work on diabatic representation.[254] Because quantum mechanical (QM) methods predict adiabatic quantities, one additional step is required: the diabatization. In such transformation for two states, two adiabatic PESs, for instance, S_1 and S_2, give rise to two diabatic PES, for instance, $n\pi^*$ and $\pi\pi^*$ (plus the diabatic coupling surface between them). By definition, the electronic character ($n\pi^*$, for instance) remains constant over the entire diabatic PES. For this reason, the diabatic representation eliminates discontinuities in property surfaces (transition

dipole moment surfaces, for example) and singularities in the PES owing to conical intersections.[260] Despite the many available methods for diabatization,[260-263] this is still a challenging and costly task in computational chemistry.