Perspective on integrating machine learning into computational chemistry and materials science

Julia Westermayr,[1] Michael Gastegger,[2] Kristof T. Schütt,[2,3] and Reinhard J. Maurer[1,a)]
[1)] *Department of Chemistry, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom*
[2)] *Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany*
[3)] *Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*

(Dated: 24 May 2021)

Machine learning (ML) methods are being used in almost every conceivable area of electronic structure theory and molecular simulation. In particular, ML has become firmly established in the construction of high-dimensional interatomic potentials. Not a day goes by without another proof of principle being published on how ML methods can represent and predict quantum mechanical properties – be they observable, such as molecular polarizabilities, or not, such as atomic charges. As ML is becoming pervasive in electronic structure theory and molecular simulation, we provide an overview of how atomistic computational modeling is being transformed by the incorporation of ML approaches. From the perspective of the practitioner in the field, we assess how common workflows to predict structure, dynamics, and spectroscopy are affected by ML. Finally, we discuss how a tighter and lasting integration of ML methods with computational chemistry and materials science can be achieved and what it will mean for research practice, software development, and postgraduate training.

Keywords: electronic structure theory, quantum chemistry, artificial intelligence, molecular dynamics simulation, materials discovery

I. INTRODUCTION

Atomistic and electronic structure simulations based on quantum theoretical calculations form a central aspect of modern chemistry and materials research. They enable the prediction of molecular and materials properties from first-principles as well as the simulation of atomic-scale dynamics. On this basis, computational chemists and physicists in academia and industry contribute to fundamental mechanistic understanding of chemical processes, to the identification of novel materials, and the optimization of existing ones. Over the last few decades, computational molecular simulation has been firmly established in the chemical sciences as an important part of the method portfolio. This was accompanied by a move to streamline and optimize common workflows for model building and simulation (see Figure 1). Algorithms for molecular geometry optimization, efficient molecular dynamics simulations, and electronic structure calculations perform highly specialized tasks while being massively scalable and parallelized across a diverse range of hardware architectures.[1,2] Simultaneously, PhD graduates in the field have been trained to be expert users of existing and developers of new simulation workflows. This is the *status quo* at the time when machine learning (ML) methods enter the stage.

The application of ML to atomistic simulation and electronic structure theory has been developing rapidly since its earliest works in a modern context.[3–11] A number of excellent reviews have recently been written to highlight progress in various contexts including the role of ML in catalyst design,[12,13] in the development of force-fields and interatomic potentials for ground state properties[14–19] and excited states, [20–22] in quantum chemistry,[23,24] in finding solutions to the Schrödinger equation,[25] and the role of unsupervised learning in atomistic simulation[26] (see Table I for a non-exhaustive list).

An excellent retrospective of the last decade of ML in the context of chemical discovery has recently been published by von Lilienfeld and Burke,[27] predicting a bright future in the context of ML for quantum chemistry that lies ahead. Indeed, not a day goes by without another novel ML approach being published, which promises to predict atomic and electronic properties of molecules and materials at ever greater accuracy and efficiency. A main goal of many ML models is the parametrization of analytical models to represent electronic structure. These ML models can then be evaluated extremely fast. Thus ML models can speed up simulations to achieve longer time and length scales. Their efficiency depends strongly on the design of descriptors or neural network architectures that optimally chart the vast space of chemical compounds and materials.[28,29] These approaches have the potential to fundamentally change day-to-day practices, workflows and paradigms in atomistic and quantum simulation as they become more tightly integrated with existing tools. **But how exactly will ML affect the method portfolio of future computational scientists working in electronic structure theory and molecular simulation?** How will this affect a practitioner who wants to determine the equilibrium structure and ground-state energy of a molecular system using electronic structure theory? How will it change the required expertise and demands on PhD graduates?

[a)] Electronic mail: r.maurer@warwick.ac.uk

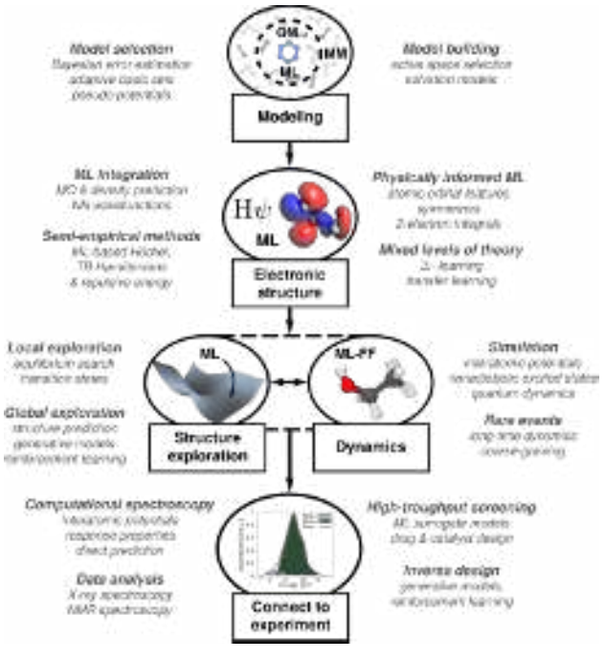


FIG. 1. Schematic depiction of the key workflow steps in computational molecular and materials modeling: Model building and method choice, electronic structure calculations, structure exploration and dynamics, and connection to experiment. All of these steps can benefit from ML models. In many cases ML methods do not just enhance existing approaches, but also open avenues toward new workflows.

For the uninitiated, it is easy to get lost in the vast array of ML models, which might soon be comparable to the zoo of exchange-correlation functionals available in density functional theory (DFT).[30] What will become the ML equivalent of go-to DFT functionals for practitioners? At the moment, there are relatively few examples where ML models have become generally applicable to researchers outside the immediate circle of developers. In this perspective, we are discussing recent advances through the lens of their potential benefit to a wide community of computational molecular scientists who are not ML experts. Our goal is to identify future possibilities of permanent integration of ML-based approaches into workflows and electronic structure and simulation software packages. This can for example involve a common code base and data structure for ML and simulation algorithms or bidirectional data exchange between workflows based on ML or physical simulation. Central to this perspective is the question how ML can effectively address the computational bottlenecks and capability gaps in electronic structure calculations and molecular simulations and what are the steps needed to make ML an integral part of the method portfolio of this field.

Our goal is to make this account as accessible as possible and to highlight applications and approaches that the community might want to keep track of in the future. We stress that our aim is not to provide a comprehensive review of existing ML descriptors, representations, and approaches, which is beyond the scope of this perspective and well covered by further reading material in Table I. Following the key steps of molecular modeling shown in Figure 1, each section focuses on how ML methods can benefit a central workflow or aspect of computational molecular and material science (*cf.* highlighted sentences in each paragraph). We place a particular focus on approaches that have the potential to augment existing or introduce new prevalent approaches.

Year	References	Topic of ML Review
2017	Behler [14]	Interatomic Potentials
2018	Goldsmith *et al.* [31]	ML in Catalysis
2019	Carleo *et al.* [32]	ML in Physical Sciences
2019	Yang *et al.* [33]	Drug Discovery
2019	Elton *et al.* [13]	Molecular Design
2019	Schleder *et al.* [34]	ML in Materials Science
2019	Ceriotti [26]	Unsupervised Learning
2020	Dral [23]	ML in Quantum Chemistry
2020	Noé *et al.* [35]	Molecular Simulation
2020	von Lilienfeld *et al.* [24]	Chemical Space
2020	Mueller *et al.* [15]	Interatomic Potentials
2020	Manzhos *et al.* [16]	Small Molecules and Reactions
2020	P. Gkeka et al. [17]	Force Fields & Coarse Graining
2020	Unke *et al.* [18]	Force Fields
2020	Toyao *et al.* [36]	Catalysis Informatics
2020	Manzhos [25]	ML in Electronic Structure
2020	Westermayr *et al.* [20]	ML for Excited States
2021	Behler [37]	Neural Network Potentials

TABLE I. Overview of recent reviews of machine learning methods in electronic structure theory and atomistic simulation. This is not intended to be a complete list of all reviews on the subject, but a selection of suggested further reading.

II. MACHINE LEARNING PRIMER

We start by introducing basic terminology and concepts of ML that will be used in the remaining sections of the perspective. ML is concerned with algorithms that improve with increasing amount of available data under some performance measure. Statistical learning theory offers a general framework to find predictive functions $f : \mathcal{X} \to \mathcal{Y}$ mapping an input space $\mathcal{X}$ to a target space $\mathcal{Y}$.[38] In contrast to conventional physical models, where one often starts with clear assumptions about the system to be modelled, ML focuses on *universal approximators*. These are able to represent any function with arbitrary accuracy, when given enough training data and parameters. Examples for this class of models are Gaussian processes (GPs) or neural networks (NNs).[39] GPs are defined by linear combinations of the covariances between data points. These are given by a suitable (nonlinear) kernel function. NNs consist of a sequence of multiple linear transforms, alternated with nonlinear *activation functions*. This is also referred to as *deep learning*, where

each set of transform and nonlinearity is called a *layer*.

The functional relationship to be found is specified by choosing a suitable *loss function*. If the loss $\ell(f(x), y)$ requires knowledge of the targets $y \in \mathcal{Y}$, this is called **supervised learning**. This includes classification and regression for categorical and continuous target spaces $\mathcal{Y}$, respectively (see also Fig. 2). ML force fields are examples of regression tasks (see section IV),[40] where often the squared error is used as loss function. For instance, classifiers can be used to automatically select appropriate quantum chemistry methods for a given system (see section III). In contrast, **unsupervised ML** aims to find patterns in the data that are specified by a loss function without having access to the ground truth targets y. Tasks falling under this category include clustering, dimensionality reduction, or density estimation of the data distribution. In the context of computational chemistry, unsupervised ML finds application in post-processing and analysis of molecular simulation data, *e.g.*, in identifying collective variables (CVs)and reaction pathways that will be discussed in section VI (see also Fig. 2).

The optimal predictive function minimizes the *expected risk*, *i.e.*, the expectation of the loss function weighted by the probability distribution over the data.[41] However, the data distribution is usually unknown and, in supervised learning, the loss requires access to the targets. Thus, one instead optimizes the *empirical risk*, *i.e.*, the expectation over a training set sampled from the data distribution. This could for example consist of electronic structure calculations of systems $x \in \mathcal{X}$ with properties $y \in \mathcal{Y}$. Since there typically exist many possible approximates that fit a finite training set, one introduces regularizer terms to the optimization problem, which punish complex solutions. This avoids *overfitting*, *i.e.*, an increased error on unseen data due to approximating a simple functional relationship with an overly complex function on the training set.

Another important aspect to consider is the selection of training examples, which should be representative of the distribution encountered when applying the ML model. This requires not only a sufficient number of training examples, but also sufficient coverage of the input space. If an ML model is applied outside of its training domain, *i.e.*, if it is used for extrapolation, its predictions quickly become unreliable. *Active learning* aims to detect this and acquire additional training data in the corresponding regions. Similarly, *Bayesian optimization* is an approach for global search that obtains additional examples where there is a high probability to optimize a given criterion based on the current model and its uncertainty. ML models are typically evaluated on a separate test set that is not used during the training process, i.e., also not for controlling overfitting. To get a better measure of the reliability of ML models in different regions and to detect holes, additional sampling of data can be carried out with *e.g.* enhanced sampling techniques.[42,43] Alternatively, when using two NNs, minima of their negative squared difference surface can be used to detect sparse

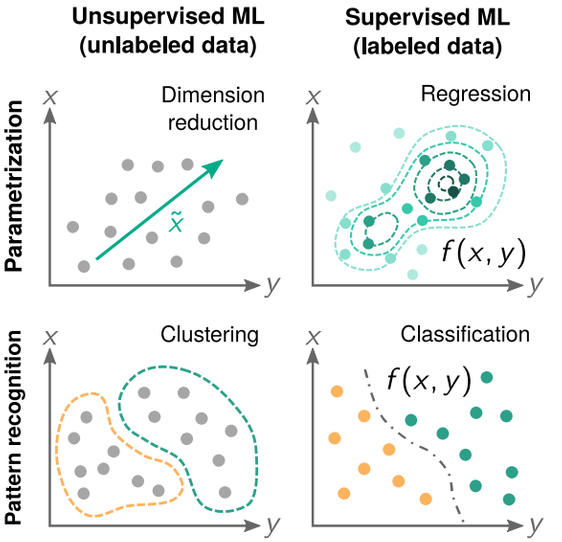


FIG. 2. Schematic depiction of different ML model categories. Unsupervised learning techniques use unlabeled data and are often used for dimensionality reduction or clustering, whereas supervised ML models perform regression or classification tasks on labelled data.

conformational regions.[44]

To design accurate and data-efficient ML models, it is important to be aware of the structure of the input space and how it is represented. Encoding prior knowledge in the model reduces the effective space to cover and, thus, the required amount of training data. Examples include the use of convolutions to encode roto-translational invariances[45] or delta learning, where only the difference to a baseline is learned.[46]. Beyond that, *transfer learning* studies how knowledge contained in models trained on one task can be reused for related tasks. This also means that, the question of whether a prediction is an extrapolation depends not only on the given training data, but also on the prior knowledge built into the ML model.

By employing a probabilistic input space and a structured target space, one obtains a model that can, *e.g.*, be used to generate novel molecular structures. The probability distribution over molecular space can be modeled explicitly, for example using variational autoencoders,[47] or implicitly, *e.g.*, by generative adversarial networks[48] that provide access to the distribution only through sampling. In a supervised setting, generative models can facilitate inverse design by learning a probability distribution of chemical structures conditioned on a desired target range of one or multiple properties.

Finally, **reinforcement learning** is concerned with learning the optimal action in a given state to maximize a specified, future reward. An example for this is an unfolded protein (state), where one applies changes to the geometry (action) in order to come closer to the folded structure with minimum energy (future reward).[49] Reinforcement learning includes an exploration strategy such

that more data is collected during the training process. Therefore, it can, for example, be used for molecular design without requiring a representative set of reference structures before training.

III. ML IMPROVES MODEL BUILDING, METHOD CHOICE, AND OPENS NEW MULTI-SCALE APPROACHES

The first task one faces when investigating a chemical problem *in silico* is to determine a suitable computational model. The modeling process involves the design of the atomistic structural model and the choice of computational method for calculating the properties of interest. Both choices traditionally are based on achieving a balance between a sufficiently accurate description of the chemical phenomena to be studied and limited computational effort that renders the calculations feasible.

Computational methods can range from electronic structure theory methods (*e.g.*, correlated wavefunction or density functional approaches) to more approximate empirical force fields. Depending on the level of approximation, a method can be appropriate for modeling certain phenomena, while being less reliable for others. One example are classical empirical force fields, which sacrifice the ability to model chemical bond breaking in favor of computational speed, but yield excellent predictions for ensemble averages of macromolecular systems. Different applications also place different accuracy requirements on the reference method. A concept often mentioned in the context of ML in chemistry is chemical accuracy, which originally specified that the energy error of a computational method deviates at most 1 kcal/mol from experiment. This accuracy requirement was coined by Pople in his Noble lecture[50] for thermodynamic properties, where it allows reliable comparison with experiment.[51] However, other applications may necessitate significantly more rigorous error limits. In the field of high-resolution vibrational spectroscopy for example, reliable predictions require so-called spectroscopic accuracy, which corresponds to an energy error smaller than 1 cm^{-1} or 0.003 kcal/mol.[52] The model building stage furthermore involves a range of decisions on how to represent the system, for example, how to treat environments such as solvents, what size the simulation cell should have, or which atoms to model explicitly. All these decisions can influence the quality of results at a fundamental level and hence need to be considered carefully.

Unfortunately, choices are often ambiguous and different strategies can still yield similar results or may only work in certain combinations. The associated design choices typically require a mix of expertise and chemical intuition of experienced practitioners. This makes it hard to see how ML could help to automate this process. Nevertheless, ML models can, *e.g.*, learn to infer decision rules or categorize complex patterns in a purely data driven fashion. This makes them a promising tool to provide support during the model building stage, making balanced model building choices more widely available and potentially achieving fully automated decision making in the future.

Transparent method selection protocols can be based on **uncertainty quantification**.[53,54] Currently, theoretical predictions tend to be reduced to a single number, without considering the spread due to, *e.g.*, method-specific modeling errors. Access to confidence intervals can provide several key advantages beyond determining how well a particular method is suited for a task. Trends in method predictions can be analyzed in a more general manner, going beyond the snapshots provided by traditional benchmark studies. When combined with experiment, uncertainties assigned to theoretical predictions allow for a better separation of error sources and interpretation of results. Recently, some progress has been made in tackling this problem with ML algorithms and Bayesian approaches in particular. Bayesian error estimation has been successfully used to construct multiple density functionals. Wellendorff *et al.*[55] reported a Bayesian functional with a non-local van der Waals (vdW) correlation term. This so-called BEEF-vdW functional provides predictions as well as computational error estimates. They demonstrated the utility of BEEF-vdW based on two surface science problems, modeling graphene adsorption on a Ni(111) surface and the binding of CO to Pt(111) and Rh(111) substrates. Bayesian frameworks for density functionals were also developed by Aldegunde, Kermode, and Zabaras[56] and Simm and Reiher[57]. All these approaches allow for the construction of specialized density functionals which yield confidence intervals for computed energies. This makes it possible to automatically probe the reliability of the method for different compounds and structures and identify problematic situations. Simm and Reiher[57] used their approach to estimate the errors associated with different reaction barriers along the catalytic cycle of Yandulov–Schrock catalyst, where they demonstrated that even similar reaction steps can exhibit very different confidence levels due to shortcomings of the computational method. By applying this approach to chemical reaction networks, Proppe *et al.*[58] demonstrated how this method can further be used to provide uncertainty estimates for chemical reaction rates.

Beyond error estimates, ML has been employed to automatically construct basis sets for electronic structure methods.[59] Usually, pre-defined basis sets are used for electronic structure computations, which aim to provide reasonable accuracy over a wide range of compounds. As such they use higher radial and angular resolution than might be necessary for certain molecules. Schütt and VandeVondele[59] have shown how ML can be used to generate an adaptive basis set tailored to a specific system based only on local structural information. Using liquid water as example, their adaptive basis set was able to reduce computational cost by up to a factor of 200. Similarly, local pseudopotentials have been constructed based on kernel ridge regression.[60] Another important

decision in method selection is whether the problem of interest exhibits strong electron correlation (also referred to as multi-reference character or static correlation). In this case, a single antisymmetric product wave function is no longer sufficient to describe the electronic system and single-reference methods (*e.g.*, semi-local approximations to DFT, single-reference coupled-cluster (CC) theory) yield inconsistent performance across configurational space and fail to describe bond breaking. Duan *et al.*[61] have proposed a semi-supervised ML approach to automatically classify chemical systems according to their multi-reference character in an efficient manner. This makes it possible to identify problematic systems without the need to carry out expensive high-level calculations and thus aid in the method selection process. In some situations, it can be advantageous to rely not on a single method, but instead employ a combination of electronic structure theories and basis set levels. Such composite methods have a long history in computational chemistry, with the Gaussian methods for thermochemistry (G2-G4)[62–64] being some of the most prominent examples. All composite methods have in common, that they profit from the cancellation of errors at different levels of theory and can offer improved accuracy at lower computational cost. Zaspel *et al.*[65] have leveraged ML and combination techniques to derive a composite method in a data driven fashion. They could demonstrate that their method achieved CC accuracy using only lower levels of theory.

The **model building** process encompasses many other aspects apart from method selection. This includes decisions on which structural aspects of the system need to be considered explicitly or only accounted for in their implicit effect on the system (*e.g.*, implicit versus explicit solvation models), whether periodic boundary conditions are required or which boundary box shapes and sizes are appropriate. Other aspects concern the electronic structure, especially in the context of multi-reference methods. Most of these approaches require decisions on which particular electronic reference configurations, often referred to as active space, to include in the description of a system. This problem is highly nontrivial, as it not only depends on the intrinsic electronic structure of a system but also on the chemical reaction to be studied. As a consequence, these methods (*e.g.*, Complete Active Space Self Consistent Field (CASSCF)) have been hard to use by non-expert users in a black box manner in the past. Jeong *et al.*[66] recently introduced a ML protocol based on decision trees for active space selection in bond dissociation studies. Their approach is able to predict active spaces able to reproduce the dissociation curves of diatomic molecules with a success rate of approximately 80 percent precision compared to random selection. This constitutes an important step toward black box applications of multi-reference methods.

ML approaches further show great potential in the context of **multi-scale modeling**. Multi-scale approaches combine information from different levels of theory to bridge different physical scales. Examples include hybrid quantum mechanics/molecular mechanics (QM/MM) simulations[67]. For example, Zhang, Shen, and Yang[68] have shown how a simple Δ-learning based model can improve the accuracy of solvent free energy calculations, where they could reach hybrid DFT accuracy using a semi-empirical DFTB baseline. A similar scheme has been employed by Böselt, Thürlemann, and Riniker[69] to simulate the interactions of organic compounds in water. Gastegger, Schütt, and Müller[70] used a ML/MM approach where a ML model completely replaced the QM region to model solvent effects on molecular spectra and reactions. This made it possible to achieve an acceleration of up to four orders of magnitude, while still retaining the accuracy of the hybrid functional reference method. Combining fragment methods with ML techniques, Chen, Fang, and Cui[71] were able to investigate excited states in extended systems in an efficient manner by only treating the photochemically active region with a multi-reference method while the environment is modeled with ML. Finally, Caccin *et al.*[72] have introduced a general framework for leveraging multi-scale models using ML to simulate crack propagation through materials, thus enabling simulations which would otherwise be impossible using either classical force-fields or electronic structure methods alone.

Future directions: While a complete automation of the model building stage has not yet been achieved, ML based algorithms have nevertheless led to significant progress toward this endeavor. Due to the complexity of the model building process, there still is a large number of untouched subjects which may serve as fruitful substrate for future ML research. Potential avenues include the automated selection of suitable levels of correlation methods for specific problems and using ML to automatically generate partitions in multi-scale approaches.

IV. ML IN ELECTRONIC STRUCTURE THEORY

The solution to the electronic Schrödinger equation can be approximated in various ways, where a tug-of-war between accuracy and computational efficiency is crucial to any choice of method. The bottlenecks that need to be addressed to achieve more efficient electronic structure calculations are mainly:

(1) the evaluation of multi-centre and multi-electron interaction integrals, which requires optimally-tuned basis representations to construct Hamiltonians and sets of secular equations and

(2) the (iterative) solution of coupled sets of equations to predict total energies, wave functions, electron densities, and other properties derived thereof.

To overcome these bottlenecks, developments of correlated wave-function-based methods, exchange-correlation

functionals within DFT, and methods based on many-body perturbation theory must go hand in hand with algorithmic advances. Progress on challenge (2) has been propelled by algorithmic ingenuity and a collective community effort to develop massively scalable linear algebra algorithms to be collected in central libraries such as the Electronic Structure Library (ESL[1]) and the Electronic Structure Interface (ELSI[2]). It is challenge (1), where ML methods can potentially have the biggest impact in eliminating computational bottlenecks while maintaining high predictive power.

Currently, the most pervasive application of ML is **to replace *ab-initio* electronic structure calculations with *ab-initio*-quality interatomic potentials**. In doing so, ML methods also significantly improve the predictive capabilities of molecular dynamics (MD) simulations by enabling *ab-initio*-accuracy at computational costs comparable to classical force fields (*cf.* section VI). In principle, ML models can parametrize any smooth function, such as the ground-state total energy, the forces, and other derived properties obtained from a first-principles calculation. Related ML models for interatomic potentials have already been reviewed extensively (see Table I for example). We therefore focus on ML representations of electronic structure quantities beyond ground-state energies and forces in the following.

Many ML representations of excited state properties, such as HOMO-LUMO gaps,[73–75] excited-state energies,[21,76–78] or band gaps[79–82] have been proposed and were mainly based on NNs or kernel methods. Recently, ML models have also been applied to derive excited-state or response properties explicitly by learning the density of states[83] or orbital energies,[74,78] respectively. These models have further been applied to obtain excitation spectra. However, a main challenge that is frequently encountered when fitting many energy levels is the non-smoothness of the target functions, which is true for orbital energies as well as adiabatic potential energy surfaces (PESs).[61,84] Avoided crossings at conical intersections in adiabatic potential energy landscapes represent a good example for this behaviour: When two potential energies become degenerate and form a cusp, the respective coupling values become singular at this point in the conformational space. Consequently, a direct learning of such properties is prohibited in many cases, making a smoothing of the target property or novel fitting approaches preferable. Approaches to achieve better learning behaviour strongly depend on the purpose of the ML model. For instance, in case of spectroscopic predictions it is sufficient to learn the spectral shape directly instead of the energy levels. This has been done with Gaussian Approximation Potentials for the density-of-states[83] and with NNs for X-ray spectroscopy[85,86] or for excitation spectra.[74] In the latter case, NNs could describe spectral intensities with deviations of 0.03 arb.u.. The same authors also fitted orbital energies of the QM9 data set comprising 134k organic molecules with a mean average error of 0.186 eV.[74] Alternatively, a diabatic[87] or latent Hamiltonian matrix[78] can be learned and used to obtain orbital energies or adiabatic energies as eigenvalues of the matrix, respectively. The latter approach was shown to improve the accuracy of orbital energy predictions by a factor of 2 compared to direct learning.[78]

ML parametrization of excited states is especially challenging when multi-reference methods are required, because states can switch their character along certain reaction paths, which leads to jumps in the PESs. While this can also be the case for ground-state PESs, this problem is more pronounced for higher-lying excited states in regions where the density of states is high, leading to significant higher noise in excited-state PESs and consequently, more difficult learning.[84]

While ML parametrization of electronic structure data is well established, it is intrinsically limited in its application range by the unfavorable scaling associated with bottleneck (1), *i.e.*, many highly accurate electronic structure methods are too computationally costly to generate sufficiently large training datasets that enable reliable parametrization. Sometimes, **better accuracy can be achieved with Δ-ML approaches**. This approach is based on the assumption that the difference in energy between two electronic structure methods - a low-level one and a high-level one - is easier to represent than either one of the two methods.[46] An alternative to the Δ-learning approach is **transfer learning**,[88] where a model is trained on data from a low level of theory and retrained with less data points of a more accurate method. A rule for determination of the number of data points needed in consecutive Δ-learning approaches that takes computational cost and prediction accuracy into account is proposed by Dral *et al.*[89]. Many studies use about 10% of the original training data for Δ-learning[76,78,90,91] and transfer learning.[92–96] In both cases, the ML model ideally yields an accuracy that is comparable to the higher-level theory. The prediction of energies with CC accuracy for the QM data sets was shown by Smith, Isayev, and Roitberg[97] using transfer learning and mostly range-separated semi-local DFT data (5 million DFT data points compared to 500,000 CC data points). Very recently, Bogojeski *et al.*[98] have demonstrated that with Δ-ML a model with CC accuracy was generated by using mostly semi-local DFT reference data and only a few data points calculated with CC theory. For instance, MD of resorcinol ($C_6H_4(OH)_2$) could be achieved with 1004 data points at DFT and CC accuracy. While the DFT ML model had mean absolute errors of 2-3 kcal/mol compared to CC, the Δ-ML model could achieve already 1 kcal/mol accuracy with respect to CC with as few as 25 data points.[98]

Data efficiency can also be improved by designing NN architectures that implicitly satisfy symmetry constraints (*i.e.*, rotational equivariance and permutational invariance) and, as a consequence, require much fewer data points to achieve a given accuracy.[99,100] This is only one of many possible strategies to **include more physical information into ML model architectures**.

Including the mathematical structures and the physical boundary conditions relevant to electronic structure methods into deep learning models leads to a further boost of data efficiency and model transferability. This has recently been shown with reproducing kernels optimized for long-range intermolecular forces[101] and with an ML-based parametrization of Density Functional Tight-binding (DFTB). The latter model provided error reductions of up to 67% for test molecules containing 8 heavy elements compared to existing DFTB parametrizations.[102] Similarly, the MOB-ML approach uses localized 2-electron interaction integrals from Hartree-Fock calculations as input to construct a highly accurate and transferable GPR model. This is applied to the prediction of CCSD correlation energies for a diverse range of molecular systems.[103–105] The MOB-ML approach for instance reaches chemical accuracy by using three times fewer training data points for organic molecules with up to 7 heavy atoms compared to Δ-ML approaches. Transferability was tested with molecules with up to 13 heavy atoms and MOB-ML could achieve chemical accuracy with 36 times fewer data points compared with Δ-ML.[104]

Alternatively, rather than circumventing the solution of iterative equations of correlated wavefunction methods, ML models may also be used to facilitate faster convergence. On average about 40% reduction of the number of iterations for different basis sets could be achieved by Townsend and Vogiatzis[106]. They trained an ML model to facilitate the convergence of CC methods based on lower-level theory electronic structure data. Besides ML models being powerful to accelerate the computation of target properties, they can also be used to predict correlated total energies of molecules based on Hartree-Fock or DFT results. Examples are NeuralXC,[107] DeepHC,[108] and OrbNet[109] which provide NN representations based on atomic orbital features.

ML becomes increasingly important as an integrated element of solving quantum many-body problems. First attempts to solve non-homogeneous ordinary and partial differential equations using ML algorithms[6,110–112] already date back to more than 20 years ago for model systems and have recently been applied to solve the quantum many-body problem for small organic molecular systems.[113–120] These efforts have recently been summarized in a comprehensive review[16] and perspective.[25] While they are conceptually exciting and potentially transformative in solving the many body problem, their integration into existing, widely accessible electronic structure software may not be fully practicable yet as existing models are limited to small system sizes and not yet transferable.

Rather than using ML methods to learn a representation of quantum states, they can also be used to parametrize electronic structure in an already known representation that is compatible with well-established electronic structure packages. Such **ML models are on their way to becoming an integrated element of electronic structure codes**. The resulting surrogate models, thereby, provide not only predictions of total energies and their derivatives, but further enable the derivation of many additional properties. One such example is the SchNOrb model (SchNet for Orbitals),[75] which is based on the deep tensor NN SchNet.[121,122] SchNOrb predicts Hamiltonians and overlap matrices in local atomic orbital representation compatible with most quantum chemistry software packages. Thus, it can be trained with data from quantum chemistry codes and its prediction can directly enter further quantum chemical calculations, *e.g.*, as an initial guess of the wave functions in self-consistent field calculations or to perform perturbation theory calculations of correlation energies. Self-consistent field iterations could be reduced by an average of 77% when using the SchNOrb wave function as an initial guess. Beyond that, it has been shown that the model can represent interaction integrals in localized effective minimal basis representations, which benefits the prediction accuracy for larger systems.[123]

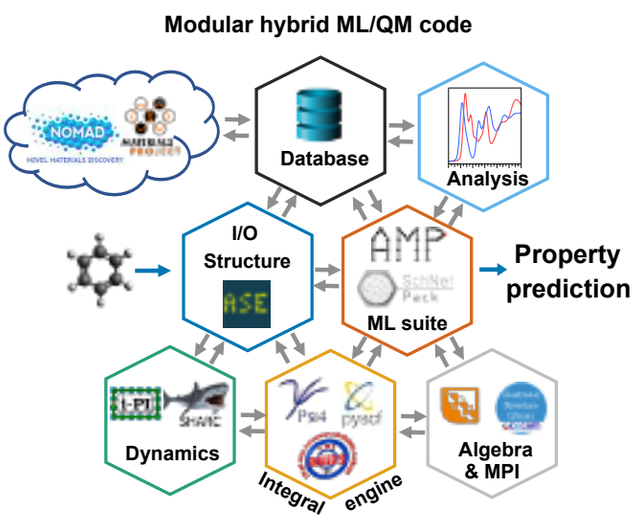


FIG. 3. Electronic structure software is increasingly becoming more modular. By moving away from monolithic (all-in-one) code models to a modular design, atomistic ML toolkits and data repositories, together with other standardized libraries, can be more tightly integrated into electronic structure workflows.

Alternatively, an ML model may predict **the electron density or a density functional**.[16,83,107,124] A recent example of a deep learning framework to predict the electronic density or properties related to the density of a reference DFT method is DeepDFT.[125] A symmetry-adapted method that considers geometrical covariance was proposed by Fabrizio *et al.*[126] and Grisafi *et al.*[127] to learn the charge density of different organic molecules via Gaussian Process Regression (GPR) models.[126,127] This model is physically inspired and learns the charge density via a sum of atom-centered basis functions with the coefficients of these functions being predicted by the ML model. The authors achieve linear scaling with respect to the number of atoms and allow for size-extensive

transferability. The latter was showcased by training the density of butadiene and butane and predicting the density of octa-tetraene and octane.[127] Fabrizio *et al.*[128] have further shown on the example of organic molecules that ML can be used to predict the on-top pair density in combination with a newly developed basis set. The on-top pair density can be used to assess electron correlation effects of a target compound, which most often cannot be described accurately using DFT. However, its evaluation requires post-HF or multi-reference calculations, which could be avoided due to the use of ML.

A **universal density functional provided by an ML model** could potentially eliminate the need for exhaustively comparing different types of functionals for a given chemical problem. So far, ML has been used to generate new DFT functionals or to adjust the energy functional, bypassing the need to solve the iterative Kohn-Sham equations and accelerating simulations for the ground state[104,107,129–134] and excited states[135] significantly. These models further promise better transferability for different types of molecular systems. Orbital-free DFT is another effort that allows for more reliable DFT calculations, but it requires the kinetic energy density functional.[136] However, various approaches have been put forward to parametrize the kinetic energy density functional with different kernel-based and deep learning methods.[137–140] Li *et al.*[124] recently presented an approach that integrates the iterative self-consistent field algorithm into an ML model to construct a learned representation of the exchange-correlation potential for 1D model systems of H_2 and H_4.

The concept of ML-based Hamiltonian and density-functional surrogate models directly leads to the construction of **approximate electronic structure models based on ML**. Recently reported approaches include an ML-based Hückel model,[141] parametrized Frenkel[102,142–145] and Tight-binding (TB) Hamiltonians[146] as well as semi-empirical methods with ML-tuned parameters.[147,148]. Beyond that, several groups have proposed to combine established DFTB Slater-Koster parametrizations with kernel ridge regression or NN representations of the repulsive energy contributions to improve the accuracy and transferability of DFTB.[149,150] On the example of the QM7-X data set[151], a mean absolute error of 0.5 kcal/mol could be achieved on the atomization energies of the DFTB-ML model compared to hybrid DFT reference values.[149]

Future directions: We expect a vivid development regarding the tight integration of ML within electronic structure software - an approach that some package developers already pursue (*e.g.*, in the case of *entos*[152] and DFTB+[153]). Already in recent years, electronic structure software has started to move away from monolithic (all-in-one) software to more modular designs with interfaces to general-purpose standalone libraries[154] (see Fig. 3). These developments will be helpful in the future to achieve integrated ML/QM solutions in computational workflows. As can be seen in Fig. 3, existing atomistic ML packages such as AMP,[155] sGDML[156] or SchNetPack[45,121] could be interfaced with electronic structure packages that heavily expose internal routines (*e.g.*, FHI-aims,[157] PSI4,[158] or PySCF[159]) and be used alongside dynamics packages such as i-Pi[160] and SHARC,[161,162] as well as algebra and electronic structure libraries such as ELSI[2] and ESL.[1] The structure generation, workflow and parser tool Atomic Simulation Environment (ASE)[163] is for example already interfaced with the above examples of AMP and SchNetPack. This could also involve a closer integration with existing data repositories such as NOMAD,[164,165] the Materials Project,[164,165] the MolSSI QC Archive[166] or the Quantum Machine repository.[167] Universal data communication standards between quantum chemistry and ML will play an important role in the future. Efficient and scalable multi-language interoperability would further be needed to pursue the goal of tight integration of ML in electronic structure theory. In the future, we believe that ML will be part of many electronic structure codes to enable highly accurate electronic structure predictions at generally low computational costs. In this regard, data-efficient ML models are highly beneficial. Many recent works have shown that incorporation of symmetries and physical information into ML representations improves data efficiency, *e.g.*, via the use of features derived from efficient low-level methods such as Hartree-Fock or MP2 theory to predict observables at high level of theories.[103] Existing electronic structure software may further benefit from latent ML representations to mitigate existing bottlenecks in integral evaluations or to efficiently represent scalar and vector field quantities.

V. ML WILL IMPROVE OUR ABILITY TO EXPLORE MOLECULAR STRUCTURE AND MATERIALS COMPOSITION

A key objective of computational chemistry and materials science is the prediction of new stable structures and viable reaction pathways to synthesize them. Beyond the significance to the discovery of new drugs and materials, finding stable equilibrium geometries and accessible transition states is a crucial element of computational molecular and materials discovery that typically involves tailored workflows.[168] As shown in Fig. 4, optimization problems in atomistic simulation span different scales from searching stable molecules across chemical space to charting the global energy landscape spanned by the chemical coordinates of a given molecule down to local structure relaxation and transition state search. Even without considering the computational cost of electronic structure calculations, high-dimensional structure search is uniquely challenging and can be greatly facilitated by ML methods.

Efficient chemical exploration methods need to be able to identify CVs in high-dimensional spaces that are associated with relevant reaction events that occur at vastly different time scales ranging from the femtosec-

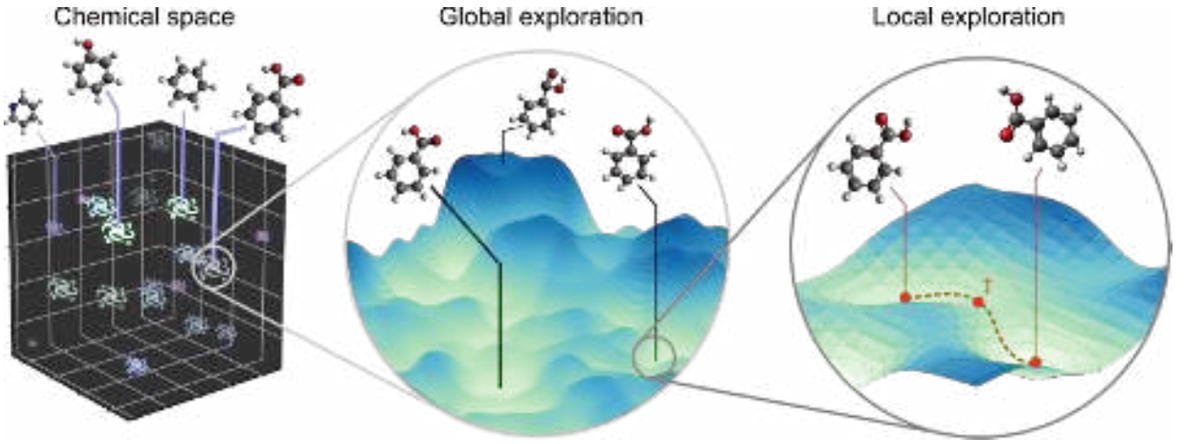


FIG. 4. Exploration methods can target different scales of molecular and material space. At the highest level, chemical space, both chemical composition and structure are varied. Global exploration targets a single PES with constant chemical composition and explores different structural conformations and their relative stability. At the lowest level, local details of the PES such as reaction pathways and transition states are investigated.

ond regime (electron transfer and vibrational motion) to multiple nanoseconds (configurational dynamics of biomolecules)[169]. It is therefore not surprising that the use of a variety of methods that fall under the umbrella of ML, has led to a significant boost in the capability to explore chemical structure space.

Even a task that is nominally as simple as **finding the nearest equilibrium structure**, *i.e.*, the local minimum of the potential energy landscape, can benefit from ML approaches. The most common geometry optimization algorithms are based on quasi-Newton methods that determine trial steps based on an approximate Hessian. Finding optimal initial guesses and preconditioners for the Hessian is key to minimizing the number of geometry optimizations that are required. Recently, several more sophisticated preconditioning schemes have been proposed based on GPR that, compared to established quasi-Newton algorithms, significantly reduce the required number of geometry optimization steps for molecules and transition metal complexes[170–172], for correlated quantum chemistry methods that require numerical differentiation[173], and for bulk materials and molecules adsorbed at surfaces.[174,175] Furthermore, unsupervised ML can be used to automatically identify if geometry optimization has failed or led to an irrelevant outcome as recently shown for transition metal complexes.[61]

ML methods have also recently been used to accelerate the search of first-order saddle points or transition states. Denzel and Kästner have used GPR to speed-up gradient-based transition state search starting from an equilibrium structure (one-ended search) by a factor of 2 compared to conventional methods.[170,176] Simultaneously, several approaches have been proposed to incorporate aspects of ML into double-ended transition state search based on the Nudged Elastic Band (NEB) method.[177–179] Garrido Torres *et al.*[179] have proposed a surrogate GPR model to accelerate a NEB method, leading to a factor of 5 to 25 fewer energy and force evaluations when compared to the conventional NEB method.

One of the most challenging tasks, namely **identifying the global minimum of a potential energy landscape associated with the most stable structure**, can be significantly facilitated by the use of ML. Established methods to perform global optimization are often evolutionary algorithms or stochastic methods. Examples for the former are genetic algorithms[180] and for the latter random structure search[181] or basin hopping.[182,183] A prominent example of a global optimisation problem on a complex high-dimensional energy landscape is protein folding. Here, the alphaFold[184] and alphaFold2[185] deep NN models were recently able to show what can be achieved when ML and structure optimisation methods are combined. In alphaFold, the ML model predicts residue distances and torsional angle distributions. On the basis of this, a coarse-grained potential is constructed to perform a sequence of random structure search and optimization cycles. Hammer and coworkers have proposed a global structure prediction algorithm, called ASLA, based on image recognition and reinforcement learning.[186,187] The use of image recognition to identify structural characteristics removes the need for encoding strings such as SMILES or descriptors of the atomic environment. The approach is applicable to molecules as well as materials and has been showcased on graphene formation, and oxide surface reconstructions.[188] In the case of graphene, the method is able to generate graphene as the most stable two-dimensional phase starting from initially random atom placement. Bayesian optimisation has become a common tool to achieve efficient structure prediction for crystals,[189,190] surface reconstructions,[191] and hybrid organic/inorganic interfaces to name just a few examples.[192,193] They often outperform evolutionary algorithms in terms of efficiency.

As shown in Figure 4, one level above the search for stable structures in energy landscapes lies the search for possible stable molecular compositions in chemical space. Generative ML models have recently shown great utility to predict molecules with tailored properties[194,195], for example using SMILES representation[196] or molecular graphs[197]. While these are supervised approaches that require reference data for training, several related approaches have been proposed that use reinforcement learning.[198,199] These models can further be constrained to only predict SMILES strings that are chemically valid.[200,201] Well beyond providing stability ranking, this approach can be used to generate molecules with arbitrary target properties to be used in drug and materials discovery. Unfortunately, molecular graph-based generative models are limited in their applicability, since they can not distinguish between different conformations that lead to the same graph. However, for applications such as protein folding, optimizing reaction environments or finding reaction paths, it is paramount to have full access to conformation space. Mansimov *et al.*[202] proposed a generative model to sample 3d conformations from SMILES. This approach suffers from the same limitations as the graph representation it is built upon when properties are directly related to the 3d structure. There have been several recent efforts to directly generate 3d molecular structures: Köhler, Klein, and Noé[203] proposed equivariant normalizing flows, which are able to estimate a probability density over many-particle systems. This has been applied to finding meta-stable states of large Lennard-Jones systems. Gebauer, Gastegger, and Schütt[204] introduced G-SchNet that places atoms successively, incorporating rotational and translational symmetries. The model can be fine-tuned to generate molecules with properties in a specified target range.

Future directions: With ML methods affecting every aspect of our ability to explore molecular configurations and compositions, their routine application to facilitate continuous exploration across composition space is not far, which would allow for the variation of the number and type of atoms in the system via **ML-enabled alchemical optimization**. So-called alchemical potentials have long been applied to rational drug design[205,206] and changing of reaction barriers.[207] ML methods, such as NNs, have shown to be capable of modeling alchemical potentials[208,209] as well as to produce smooth paths through alchemical space.[210] We expect a lot of activity in this area in the future with ML methods enabling the continuous variation of elemental composition in materials to optimize their properties.

VI. ML ENABLES CLASSICAL AND QUANTUM DYNAMICS FOR SYSTEMS OF UNPRECEDENTED SCALE AND COMPLEXITY

The dynamical motion of atoms is a central target of a large part of computational research. In molecular simulation, we study the time evolution of electrons and atoms to predict static and dynamic equilibrium properties of molecules and materials at realistic temperature and pressure conditions, but also to understand nonequilibrium dynamics and rare events that govern chemical reactions. Dynamics methods range from classical MD, via mixed quantum-classical dynamics (MQCD) methods (incorporating electronic quantum effects) to quantum dynamics in full quantum or semi-classical formalisms. In all cases, equations of motion need to be integrated over time, which involves numerous evaluations of forces and other properties that govern the dynamics. ML methods can address bottlenecks in such simulations on various levels: Their most prevalent use is to speed up energy, force, and property evaluations in each time step by providing ML-based force fields and interatomic potentials. Other ML approaches directly target MD by supporting coarse-graining and the use of larger time steps, or by replacing MD completely with a direct prediction of dynamical properties, expectation values, and correlation functions.

The most obvious way in which ML can facilitate MD simulations is the **use of ML-based interatomic potentials instead of on-the-fly *ab-initio* MD**. Many early applications of ML in molecular simulation were mostly focused on ML parametrization of electronic structure data for the benefit of MD simulation. ML-based interatomic potentials that replace electronic structure evaluation during dynamics are by now commonly established, see, *e.g.*, Refs. 211, 18, and 212, and have since enabled simulations of unprecedented complexity and scale. For example, a recent breakthrough by Deringer *et al.*[213] showed that Gaussian Approximation Potentials[7,209] could be used to predict phase transitions and electronic properties of systems containing more than 100,000 atoms. Jiang, Li, and Guo[214] have recently reviewed the transformative role that ML-based high-fidelity PESs play in gas-surface dynamics simulations.

In principle, approaches can be distinguished between those that sample molecule deformations around an equilibrium geometry, e.g. for optimizations,[215] or those that consider "reactive" potential energy surfaces.[214,216–220] An alternative approach is to directly predict targeted simulation properties such as reaction yields.[221,222] A key factor in building ML force fields for MD simulations is the efficient and comprehensive sampling of relevant data points. Active learning schemes have been proposed[84,223–227] to efficiently sample the relevant configuration space that a molecule visits during an MD simulation. These schemes are based on an uncertainty measure during ML dynamics, which can be used to detect unexplored or undersampled conformational regions. The uncertainty measure could be for instance the deviation of two NNs or the statistical uncertainty estimate of the inferences made with, *e.g.*, GPR. One way to measure the accuracy and interpolative regime of ML models is to use the previously mentioned adaptive sampling tech-

niques also during the production runs. This allows to detect holes in the potential energy surfaces *on-the-fly*.[84]

By using gradient-domain ML models that are trained on gradients rather than energies, energy-conserving ML force fields can be obtained with high accuracy and little amount of training data required.[156,228,229] Δ-ML models, in the context of MD simulations, have also proven to be very powerful in providing a data-efficient representation of CC accuracy from DFT data[98] or DFT accuracy from mostly DFTB data in the context of QM/MM simulations[90], to name two recent examples. Beyond the use of ML to facilitate accurate force evaluations in MD, ML methods have been used to enable the simulation of rare events that occur on time scales inaccessible to conventional MD. A perspective review that recently arose from a CECAM conference on "Coarse-graining with ML in molecular dynamics" provides a comprehensive overview of ML for free energy sampling, coarse-graining, and long-time MD[17].

ML methods help to identify CVs, which characterize long-time dynamics of molecular systems. This is important to identify long-lived attractor states in phase spaces and to find strategies to efficiently explore dynamics in complex hierarchical energy landscapes, *e.g.*, for the isomerization of alanine dipeptide[230] or for protein folding.[231] ML methods in this domain based on principal component analysis[232] date back to over 20 years ago.[233] More recent approaches include kernel principal component analysis[234–236], diffusion maps,[237–239], the Sketch map method,[240,241] Markov state models[242,243] and various types of autoencoders.[244,245]

Several ML models have been developed that aim to achieve **bottom-up coarse-graining** by representing the potential of mean force or free energy surface as a function of coarse grained variables. This has been done for instance using NNs to infer conformational free energies for oligomers[246] or to construct a coarse-grained liquid water potential[247] or using a Gaussian approximation-based coarse-grained potential for alanine dipeptide[248] and molecular liquids.[249]

MQCD, *i.e.*, classical dynamics of nuclei coupled to the time-dependent quantum mechanical evolution of electrons, are commonly used to simulate light-induced nonadiabatic dynamics of molecules,[250–252] as well as coupled electron-nuclear dynamics in extended systems.[253] While on-the-fly MQCD simulations have become feasible in the last decade, the accessible time scale and the number of non-equilibrium trajectories that can realistically be simulated on-the-fly is too limited to enable comprehensive statistical analysis and ensemble averaging. **ML shows great promise in nonadiabatic excited-state simulations**[20,21] as documented by recent works using NNs to construct excited-state energy landscapes to perform fewest-switches surface hopping MD at longer time scales or with more comprehensive ensemble averaging than would otherwise be possible with on-the-fly dynamics.[84,254,255] Similar progress has been achieved in nonadiabatic dynamics at metal surfaces, where NNs have been used to construct excited-state landscapes[4,256] and continuous representations of the electronic friction tensor[257] used in MD with electronic friction simulations.[258,259]

Even **full quantum dynamics simulations** have recently seen an increasing uptake of ML methodology to push beyond longstanding limitations in the dimensionality of systems that can be simulated. The main bottleneck in quantum dynamics simulations is not the evaluation of the temporal evolution of the electrons, but the temporal evolution of the nuclear wavefunction, which involves computations that (formally) scale exponentially with the number of atoms in the system. Potential energy landscapes in quantum dynamics are typically represented in a diabatic basis rather than the adiabatic representation (directly outputted by electronic structure codes) in a process called (quasi-)diabatization.[260,261] However, quasi-diabatization requires expert knowledge and is highly complex for more than two coupled electronic states. The construction of diabatic representations with deep NNs has recently shown great potential to simplify and automate this laborious task.[87,262–266] Besides the PES generation itself, recent works use GPR to fit the diabatic PESs in reduced dimensions.[267–270] One of the largest ML-enhanced quantum dynamics simulation was recently performed on a 14-dimensional energy landscape for a mycosporine-like amino acid[271].

The computational efficiency of ML models is an important point to consider. MD simulations based on ML models are considerably more efficient than *ab initio* MD, yet still relatively slow compared to empirical force fields. For example, 100 femtosecond MQCD MD of $CH_2NH_2^+$ on a single compute core take 24 seconds with ML potentials compared to 74,224 seconds with the reference method (MR-CISD/aug-cc-pVDZ).[254] The simulation of 100 femtosecond classical MD of the same molecule in the gas phase with an empirical force field takes 0.005 seconds with Amber.[272] The computational efficiency of ML models can become a bottleneck if long time scales or ensemble averages over many thousands of reaction events are required. Similar memory and CPU efficiency bottlenecks can arise during model training of kernel methods and deep neural networks if large training data sets and complex high dimensional models are involved.

Future directions: ML-based interatomic potentials and continuous regression models already play an important role across almost all domains of MD simulations and we expect that the use of ML in MD will further increase in the coming years. As larger and more complex systems are targeted and longer time scales are needed, a future challenge that needs to be tackled is the computational efficiency of ML models, especially for MD simulations. The concept of sparsity in terms of ML methods and data representation can lead to better computational efficiency. Recently, explicit atomic high body order expansions in permutationally invariant polynomials (*e.g.* aPIPs[273], ACE[274]) have emerged as appealing alternative to kernel and deep learning methods as they accurately

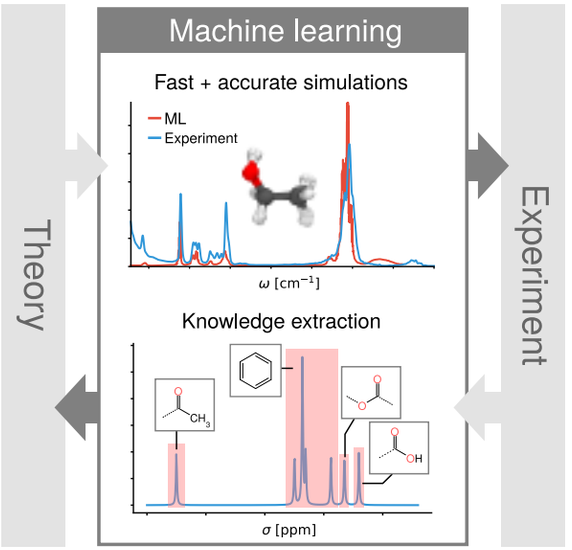


FIG. 5. Depiction of how ML methods can act as a bridge between theory and experiment. ML models trained on theory predict spectra with realistic lineshapes. At the same time, ML models can be used to infer structural information from experimental measurements.

allow high-dimensional parametrization as a function of atomic coordinate spaces and can be trained by linear regression. As a result, both training and evaluation are highly efficient with evaluation times on the order of few milliseconds per atom.[275] While most approaches focus on assisting MD by providing highly-accurate interatomic potentials and force fields, they have also shown great potential in predicting dynamical properties directly and skipping the MD simulation completely or in assessing the validity of different approximations in dynamical simulations. The latter has only recently been shown by Jasinski *et al.*[276] with a Bayesian model to estimate errors due to different approximations in quantum scattering simulations. Going forward, complex dynamical simulation methods will become more accessible to non-expert users with the help of ML and will open avenues to tackle complex systems in solvent environments[71] or dynamics at hybrid organic-inorganic interfaces.[259] It is evident that ML methods will play an important role in extending the range of applications for MQCD methods in the coming years. A recent work by Brieuc *et al.*[277] employing ML methods to achieve converged path-integral MD simulations of reactive molecules in superfluid helium under cryogenic conditions is an exemplary showcase of what the synergy of ML and quantum dynamics methods can achieve.

VII. ML HELPS TO CONNECT THEORY AND EXPERIMENT

The ultimate goal of computational molecular and materials simulation is to connect theory and experiment. This could mean supporting the explanation of experimental outcomes or finding new theoretical rules in observations, in both cases leading to a better understanding of the physical world and its laws. Forming this connection is a hard task. A plethora of different effects need to be considered in even the simplest atomistic systems, making it very difficult to faithfully reproduce experimental conditions in silico. On the other hand, experimental observations can be obscured by a variety of influences or by the sheer complexity of the measured signal. As we have seen in the preceding sections, **ML approaches can increase the accuracy of predictions and the speed with which they can be obtained.** This makes it possible to carry out computational studies which close the gap between theory and experiment by more efficiently incorporating experimental parameters such as finite temperature, measurement conditions, and solvent effects. Moreover, ML techniques can also provide invaluable support in extracting information from experimental observations and uncovering trends that are not directly apparent to the practitioner.

One field which has greatly profited from these developments is **computational spectroscopy**. The prediction of spectroscopic properties is a central aspect of computational modeling, as it provides results which can be directly compared against experiments. Examples of successful ML applications include the prediction of different vibrational spectra, combined with different response properties of the electric field. Gastegger, Behler, and Marquetand[226] have combined a latent charge dipole model with interatomic potentials in order to efficiently simulate infrared spectra (IR) of organic molecules in gas phase without having to resort to electronic structure computations of the molecular dipole. This approach has further been applied to model absorption spectra.[77,144] Raimbault *et al.*[278] introduced a kernel approach for predicting the Raman spectra of organic crystals based on molecular polarizabilities. Using a NN based approach, Sommers *et al.*[279] have demonstrated that ML can also be used to simulate Raman spectra of extended systems such as liquid water, which would be computationally unfeasible when done with DFT. In addition to vibrational spectra, ML models are also capable of modeling response properties, allowing the simulation of electronic excitations using, *e.g.*, MQCD approaches (see Section VI). For example, Zhang *et al.*[144] use NN models to obtain transition dipole moments, which in turn could be used to predict UV and visible light spectra. ML approaches have further been used to predict nuclear magnetic resonance (NMR) spectra from molecular simulations. Paruzzo *et al.*[280], for example, have used the kernel model from Ref. 278 to predict the chemical shifts in molecular solids. Recently, Christensen *et al.* have introduced an electric field dependent descriptor in the FCHL Kernel framework[281]. Based on this, they have derived molecular dipole moments as a general re-

sponse to the electric field, which can be used to simulate IR spectra of small organic molecules. Gastegger, Schütt, and Müller [70] have applied a response theory approach in combination with a deep NN architecture which explicitly depends on electric and magnetic fields. They could show that, in this manner, a single ML model can predict IR, Raman and NMR spectra. Moreover, by introducing the field generated by a molecular environment they were able to model the effect of solvents on the resulting spectra.

Beyond that, ML offers the possibility to **directly extract information from experimental observations** and relate them to fundamental chemical concepts. One example is the use of ML to interpret different types of spectroscopic measurements to determine structural or electronic properties of molecules and materials. Fine *et al.* [282] have recently presented a ML approach to extract data on functional groups from infrared and mass spectroscopy data, while Kiyohara *et al.* [283] have successfully applied a ML scheme to obtain chemical, elemental, and geometric information from the X-ray spectra of materials. Another application where ML shows promise is the automated interpretation of nuclear magnetic resonance spectra with respect to atomic structure, which typically relies heavily on experience.[284]

However, **ML can also be used to leverage information contained in large collections of scientific data**. The majority of chemical knowledge is collected in the form of publications. ML approaches such as natural language processing and image recognition offer the possibility to directly distill functional relationships and chemical insights from the massive body of scientific literature. For instance, Tshitoyan *et al.* [285] have used natural language processing to extract complex materials science concepts, such as structure property relationships, from a large collection of research literature. They could further demonstrate, that their model was able to generalize on the learned concepts and recommend materials for different functional applications. Raccuglia *et al.* [286] recently trained a ML model using information on failed experiments extracted from archived laboratory notebooks to predict the reaction success for the crystallization of templated vanadium selenites. Their model was able to learn general reaction conditions and even revealed new hypotheses regarding the conditions for successful product formation.

Finally, ML offers new ways in which theory can guide experiment. Two fields where ML has played a transformative role are **molecular/materials discovery and computational high-throughput screening**, with several reviews summarizing recent advances.[13,31,33,34,36,287] The combination of high-throughput screening with accurate and efficient ML models has proven to be highly valuable, as it allows to substitute most of the required electronic structure calculations[288]. Examples of what is possible in this space include the objective-free exploration of light-absorbing molecules,[289] drug design,[290] the computational search for highly active transition metal complexes that catalyse C-C cross coupling reactions,[291] or the discovery of new perovskite materials[292] or polymers for organic photovoltaic applications.[293,294]

Still, chemical space is estimated to cover more than 10^{60} molecules[295], hence exhaustive computational screening remains infeasible – even with fast and accurate ML models. In this context, **ML-enabled inverse design** offers a promising alternative by reversing the usual paradigm of obtaining properties from structure[296,297]. Instead, the aim is to create structures exhibiting a range of desired properties. Since such ML models readily provide analytic gradients, an application to property-based structure optimization is straightforward. First steps of applying ML in these areas have recently been achieved. Examples include the optimization of the HOMO-LUMO gap as demonstrated by Schütt *et al.* [75] and relaxation for crystal structure prediction as investigated by Podryabinkin *et al.* [298]. While ML only provides gradient-based local optimization in these examples, it can be combined with genetic algorithms[298] or global optimization methods such as simulated annealing or minima hopping[299].

Future directions: While ML techniques and atomistic ML potentials in particular have contributed greatly to closing the gap between theory and experiment, a range of open issues remains. Problems that have only recently begun to be studied include how to extend ML simulations to efficiently reproduce different experimental conditions, such as solvents or electromagnetic fields. Another frequently encountered issue concerns the data efficiency of ML models, as well as the availability of reliable reference data. For example, most generative models and inverse design approaches to date primarily target simulated properties rather than experimentally measured ones. While calculated quantities (*e.g.* redox potentials, singlet-triplet gaps) can offer invaluable guidance for design endeavors, they ultimately represent approximations to the physical characteristics of a system, which can only be fully captured through experiments (*e.g.* full-cell study for redox kinetics and electrochemical stability). Successful design endeavors therefore often combine theoretical computations with experimental data or calibrate against them[300,301].

VIII. OUTLOOK

We expect that ML methods will soon become an integrated part of electronic structure and molecular simulation software pushing the boundaries of existing techniques toward more computationally efficient simulations. ML methods may for example replace complex integral evaluations in the construction of Hamiltonians and secular equations or they can provide improved initial guesses to iteratively solve integro-differential equations. ML methods can further help to describe nonlocal effects in time and space and provide mechanisms

for on-the-fly uncertainty quantification and accuracy improvements. The beneficial scaling properties of ML algorithms with respect to the size of atomistic systems will play an important role in extending the range of application of existing electronic structure and dynamics simulation methods. The application of ML to MQCD simulations will make it possible to reach currently unfeasible time and length scales beyond few picoseconds and tens of atoms. This will in turn require the improvement of existing molecular simulation methods to capture long time dynamics. As we explore systems of increasing size, we will be able to better study the boundary between quantum effects at the nanoscale as well as collective many-body effects and fluctuations at the meso- and macroscale.[302]

A necessary requirement is the establishment and the distribution of user-friendly and well-maintained **simulation software with tight integration of ML methodology** in chemistry and materials science. Software solutions will need to be modular to allow interfacing with well-established deep learning platforms such as TensorFlow or PyTorch. This should involve the establishment of common data standards to easily communicate atomistic simulation and electronic structure data between chemistry and ML packages. In many ways, this requirement is in line with recent trends of increased modularity of codes via general libraries such as ESL[1] and ELSI[2] (see Fig. 3). A recent initiative toward an integration of ML is the ENTOS quantum chemistry package and ENTOS AI[152].

Another challenge ahead is related to **establishing a culture of openness and willingness to share data and ML models** as the availability of training data is a crucial aspect of driving advances in this field. While data sharing is quite common in material science, it is not yet so common in computational molecular science. Well defined materials data standards as put forward by the Fair Data Infrastructure project (FAIR-DI)[303] and *ab-initio* data repositories such as for example the NOMAD repository[164,165], the Materials Project[304], and the MolSSI QCArchive[166] are needed in all research areas. The need for open access to vast amounts of data will need to be balanced against other needs, such as commercial interests that arise from industrial research or commercial software projects.

Sustainable integration of ML methods into widely-used software will require long-term community effort and might be less glamorous than exciting proof-of-principle applications of ML in chemistry and materials science. Research funding agencies, reviewers, and industrial stakeholders need to acknowledge this and ensure that sustained funding for such efforts is put in place.

If achieved, an integration of ML methodology into electronic structure and molecular simulation software, will induce lasting change in workflows and capabilities for computational molecular scientists. Furthermore, it will offer the opportunity to reconsider many of the underpinning design choices of electronic structure and molecular simulation software packages which, in many cases, historically arose from computational efficiency considerations. For example, Gaussian basis representations have been chosen decades ago in quantum chemistry due to the ease of evaluating multi-centre integrals. If ML methods can vastly facilitate the evaluation of multi-centre integrals, are Gaussian basis functions still the best choice of basis representation?

An integration of ML and molecular simulation will drastically widen participation in the field and uptake of our methods and problem solving approaches. If codes require dramatically fewer computing resources and offer the ability to directly predict experimentally accessible quantities, computational simulation will become more appealing as a complementary tool in synthetic and analytical labs. In many industrial applications, cost-benefit analysis requires that a clear correspondence exists between the cost of delivering predictions and the accuracy and precision that is required for an application. The use of ML methods within such workflows will hopefully also provide a drive toward establishing better measures of uncertainty in atomistic simulation.

Finally, **the method portfolio and skill set of computational molecular scientists will need to adapt** as a consequence of the growing importance of ML methods in electronic structure theory and molecular simulation. In many cases, the presence of some aspects of ML "under the hood" of existing methods and workflows will not change how we apply these methods. For example, a DFT functional parametrized by a ML approach, can be applied as any existing functional (although its range of applicability might be very different). In other cases, the presence of ML methods will fundamentally change basic workflows as we have discussed across the sections of this perspective. In those instances, practitioners need a basic understanding of ML concepts and the different models that they are working with. This involves knowledge of the capabilities and limitations of most standard applications to avoid pitfalls. As such, ML methodology will have to become an integral part of education in computational chemistry and materials science.

ACKNOWLEDGMENTS

This work was funded by the Austrian Science Fund (FWF) [J 4522-N] (J.W.), the Federal Ministry of Education and Research (BMBF) for the Berlin Center for Machine Learning / BIFOLD (01IS18037A) (K.T.S.), and the UKRI Future Leaders Fellowship programme (MR/S016023/1) (R.J.M.). M.G. works at the BASLEARN – TU Berlin/BASF Joint Lab for Machine Learning, co-financed by TU Berlin and BASF SE.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

[1]M. J. T. Oliveira et al., "The cecam electronic structure library and the modular software development paradigm," J. Chem. Phys **153**, 024117 (2020).
[2]V. W.-Z. Yu et al., "ELSI - An open infrastructure for electronic structure solvers," Comput. Phys. Commun. **256**, 107459 (2020).
[3]J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[4]C. Carbogno, J. Behler, A. Groß, and K. Reuter, "Fingerprints for Spin-Selection Rules in the Interaction Dynamics of O_2 at Al(111)," Phys. Rev. Lett. **101**, 096104 (2008).
[5]R. Dawes, D. L. Thompson, A. F. Wagner, and M. Minkoff, "Interpolating moving least-squares methods for fitting potential energy surfaces: A strategy for efficient automatic data point placement in high dimensions," J. Chem. Phys. **128**, 084107 (2008).
[6]S. Manzhos and T. Carrington, "An improved neural network method for solving the Schrödinger equation," Can. J. Chem. **87**, 864–871 (2009).
[7]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons," Phys. Rev. Lett. **104**, 136403 (2010).
[8]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).
[9]H. M. Netzloff, M. A. collins, and M. S. Gordon, "Growing Multiconfigurational Potential Energy Surfaces with Applications to X+H_2 (X=C,N,O) Reactions," J. Chem. Phys. **124**, 154104 (2006).
[10]C. Evenhuis and T. J. Martínez, "A scheme to interpolate potential energy surfaces and derivative coupling vectors without performing a global diabatization," J. Chem. Phys. **135**, 224110 (2011).
[11]B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," Int. Rev. Phys. Chem. **28**, 577–606 (2009).
[12]J. G. Freeze, H. R. Kelly, and V. S. Batista, "Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists," Chem. Rev. **119**, 6595–6612 (2019).
[13]D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, "Deep Learning for Molecular Design – A Review of the State of the Art," Mol. Syst. Des. Eng. **4**, 828–849 (2019).
[14]J. Behler, "First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems," Ang. Chem. Int. Ed. **56**, 12828–12840 (2017).
[15]T. Mueller, A. Hernandez, and C. Wang, "Machine learning for interatomic potential models," J. Chem. Phys. **152**, 50902 (2020).
[16]S. Manzhos and T. Carrington, "Neural Network Potential Energy Surfaces for Small Molecules and Reactions," Chem. Rev. **in press**, doi:10.1021/acs.chemrev.0c00665 (2020).
[17]P. Gkeka et al., "Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems," J. Chem. Theory Comput. **16**, 4757–4775 (2020).
[18]O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," arXiv:2010.07067 (2020).
[19]V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," Adv. Mater. **31**, 1902765 (2019).
[20]J. Westermayr and P. Marquetand, "Machine learning for electronically excited states of molecules," Chem. Rev. **in press**, doi:10.1021/acs.chemrev.0c00749 (2020).
[21]J. Westermayr and P. Marquetand, "Machine learning and excited-state molecular dynamics," Mach. Learn.: Sci. Technol. **1**, 043001 (2020).
[22]P. Dral and M. Barbatti, "Molecular excited states through a machine learning lens," Nat. Rev. Chem. **in press**, doi:10.1038/s41570–021–00278–1 (2021).
[23]P. O. Dral, "Quantum Chemistry in the Age of Machine Learning," J. Phys. Chem. Lett. , 2336–2347 (2020).
[24]O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Exploring chemical compound space with quantum-based machine learning," Nat. Rev. Chem. **4**, 347–358 (2020).
[25]S. Manzhos, "Machine learning for the solution of the Schrödinger equation," Mach. Learn.: Sci. Technol. **1**, 013002 (2020).
[26]M. Ceriotti, "Unsupervised machine learning in atomistic simulations, between predictions and understanding," J. Chem. Phys. **150**, 150901 (2019).
[27]O. A. von Lilienfeld and K. Burke, "Retrospective on a decade of machine learning for chemical discovery," Nat. Commun. **11**, 4895 (2020).
[28]O. A. von Lilienfeld, "Quantum Machine Learning in Chemical Compound Space," Angew. Chem. - Int. Ed. **57**, 4164–4169 (2018).
[29]K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, eds., *Machine Learning Meets Quantum Physics* (Springer, 2020).
[30]A. D. Becke, "Perspective: Fifty years of density-functional theory in chemical physics," J. Chem. Phys. **140**, 18A301 (2014).
[31]B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel, and C. Sutton, "Machine learning for heterogeneous catalyst design and discovery," AIChE J. **64**, 2311–2323 (2018).
[32]G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," Rev. Modern Phys. **91**, 045002 (2019).
[33]X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of artificial intelligence for computer-assisted drug discovery," Chem. Rev. **119**, 10520–10594 (2019).
[34]G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, "From DFT to machine learning: recent approaches to materials science–a review," J. Phys. Mater. **2**, 032001 (2019).
[35]F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," Annu. Rev. Phys. Chem. **71**, 361–390 (2020).
[36]T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K. I. Shimizu, "Machine Learning for Catalysis Informatics: Recent Applications and Prospects," ACS Catal. **10**, 2260–2297 (2020).
[37]J. Behler, "Four generations of high-dimensional neural network potentials," Chemical Reviews **0**, null (0).
[38]T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).
[39]M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," Neural Netw. **6**, 861–867 (1993).
[40]K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and validation of machine learning methods for predicting molecular atomization energies," J. Chem. Theory Comput. **9**, 3404–3419 (2013).
[41]K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," IEEE Trans. Neural Netw. **12**, 181–201 (2001).
[42]G. Tao, "Trajectory-guided sampling for molecular dynamics simulation," Theor. Chem. Acc. **138**, 34 (2019).

[43]Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, "Enhanced sampling in molecular dynamics," J. Chem. Phys. **151**, 070902 (2019).
[44]Q. Lin, Y. Zhang, B. Zhao, and B. Jiang, "Automatically growing global reactive neural network potential energy surfaces: A trajectory-free active learning strategy," J. Chem. Phys. **152**, 154104 (2020).
[45]K. T. Schütt, P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K. R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Advances in Neural Information Processing Systems*, Vol. 2017-Decem (2017) pp. 992–1002.
[46]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big data meets quantum chemistry approximations: the δ-machine learning approach," Journal of chemical theory and computation **11**, 2087–2096 (2015).
[47]D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations* (2014).
[48]I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in Neural Information Processing Systems **27**, 2672–2680 (2014).
[49]Z. Shamsi, K. J. Cheng, and D. Shukla, "Reinforcement learning based adaptive sampling: Reaping rewards by exploring protein conformational landscapes," J. Phys. Chem. B **122**, 8386–8395 (2018).
[50]J. A. Pople, "Nobel Lecture: Nobel Lecture: Quantum chemical models," Rev. Mod. Phys. **71**, 1267 (1999).
[51]L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, "Assessment of gaussian-3 and density functional theories for a larger experimental test set," J. Chem. Phys. **112**, 7374–7383 (2000).
[52]C. Puzzarini, J. Bloino, N. Tasinato, and V. Barone, "Accuracy and interpretability: The devil and the holy grail. new routes across old boundaries in computational spectroscopy," Chem. Rev. **119**, 8131–8191 (2019).
[53]B. Ruscic, "Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and active thermochemical tables," Int. J. Quantum Chem. **114**, 1097–1101 (2014).
[54]A. Chernatynskiy, S. R. Phillpot, and R. LeSar, "Uncertainty quantification in multiscale simulation of materials: A prospective," Ann. Rev. Mater. Res. **43**, 157–182 (2013).
[55]J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, "Density functionals for surface science: Exchange-correlation model development with bayesian error estimation," Phys. Rev. B **85**, 235149 (2012).
[56]M. Aldegunde, J. R. Kermode, and N. Zabaras, "Development of an exchange–correlation functional with uncertainty quantification capabilities for density functional theory," J. Comput. Phys. **311**, 173–195 (2016).
[57]G. N. Simm and M. Reiher, "Systematic error estimation for chemical reaction energies," J. Chem. Theory Comput. **12**, 2762–2773 (2016).
[58]J. Proppe, T. Husch, G. N. Simm, and M. Reiher, "Uncertainty quantification for quantum chemical models of complex reaction networks," Faraday Discuss. **195**, 497–520 (2017).
[59]O. Schütt and J. VandeVondele, "Machine learning adaptive basis sets for efficient large scale density functional theory simulation," J. Chem. Theory Comput. **14**, 4168–4175 (2018).
[60]J. Lüder and S. Manzhos, "Nonparametric local pseudopotentials with machine learning: A tin pseudopotential built using gaussian process regression," J. Phys. Chem. A **124**, 11111–11124 (2020).
[61]C. Duan, F. Liu, A. Nandy, and H. J. Kulik, "Semi-supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost," J. Phys. Chem. Lett. **11**, 6640–6648 (2020).
[62]L. A. Curtiss, K. Raghavachari, G. W. Trucks, and J. A. Pople, "Gaussian-2 theory for molecular energies of first- and second-row compounds," J. Chem. Phys. **94**, 7221–7230 (1991).
[63]L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, and J. A. Pople, "Gaussian-3 (g3) theory for molecules containing first and second-row atoms," J. Chem. Phys. **109**, 7764–7776 (1998).
[64]L. A. Curtiss, P. C. Redfern, and K. Raghavachari, "Gaussian-4 theory," J. Chem. Phys. **126**, 084108 (2007).
[65]P. Zaspel, B. Huang, H. Harbrecht, and O. A. von Lilienfeld, "Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited," J. Chem. Theory Comput. **15**, 1546–1559 (2018).
[66]W. Jeong, S. J. Stoneburner, D. King, R. Li, A. Walker, R. Lindh, and L. Gagliardi, "Automation of active space selection for multireference methods via machine learning on chemical bond dissociation," J. Chem. Theory Comput. **16**, 2389–2399 (2020).
[67]Y.-J. Zhang, A. Khorshidi, G. Kastlunger, and A. A. Peterson, "The potential for machine learning in hybrid QM/MM calculations," J. Chem. Phys. **148**, 241740 (2018).
[68]P. Zhang, L. Shen, and W. Yang, "Solvation free energy calculations with quantum mechanics/molecular mechanics and machine learning models," J. Phys. Chem. B **123**, 901–908 (2018).
[69]L. Böselt, M. Thürlemann, and S. Riniker, "Machine learning in qm/mm molecular dynamics simulations of condensed-phase systems," arXiv:2010.11610 (2020).
[70]M. Gastegger, K. T. Schütt, and K.-R. Müller, "Machine learning of solvent effects on molecular spectra and reactions," arXiv:2010.14942 (2020).
[71]W.-K. Chen, W.-H. Fang, and G. Cui, "Integrating machine learning with the multilayer energy-based fragment method for excited states of large systems," J. Phys. Chem. Lett. **10**, 7836–7841 (2019).
[72]M. Caccin, Z. Li, J. R. Kermode, and A. De Vita, "A framework for machine-learning-augmented multiscale atomistic simulations on parallel supercomputers," Int. J. Quantum Chem. **115**, 1129–1139 (2015).
[73]W. Pronobis, K. R. Schütt, A. Tkatchenko, and K.-R. Müller, "Capturing intensive and extensive dft/tddft molecular properties with machine learning," Eur. Phys. J. B **91**, 178 (2018).
[74]K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke, "Deep learning spectroscopy: Neural networks for molecular excitation spectra," Adv. Sci. **6**, 1801367 (2019).
[75]K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, "Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions," Nat. Commun. **10**, 5024 (2019).
[76]R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. von Lilienfeld, "Electronic spectra from TDDFT and machine learning in chemical space," J. Chem. Phys. **143**, 084111 (2015).
[77]J. Westermayr and P. Marquetand, "Deep learning for uv absorption spectra with schnarc: First steps toward transferability in chemical compound space," J. Chem. Phys. **153**, 154112 (2020).
[78]J. Westermayr and R. J. Maurer, "Physically inspired deep learning of molecular excitations and photoemission spectra," arXiv:2103.09948 (2021).
[79]K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.-R. Müller, and E. K. Gross, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," Phys. Rev. B **89**, 205118 (2014).
[80]Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," J. Phys. Chem. Lett. **9**, 1668–1673 (2018).
[81]J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, "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).

[82]G. Pilania, J. Gubernatis, and T. Lookman, "Multi-fidelity machine learning models for accurate bandgap predictions of solids," Comput. Mat. Sci. **129**, 156–163 (2017).
[83]C. B. Mahmoud, A. Anelli, G. Csányi, and M. Ceriotti, "Learning the electronic density of states in condensed matter," Phys. Rev. B **102**, 235130 (2020).
[84]J. Westermayr, M. Gastegger, M. F. S. J. Menger, S. Mai, L. González, and P. Marquetand, "Machine Learning Enables Long Time Scale Molecular Photodynamics Simulations," Chem. Sci. **10**, 8100–8107 (2019).
[85]C. D. Rankine, M. M. M. Madkhali, and T. J. Penfold, "A deep neural network for the rapid prediction of x-ray absorption spectra," J. Phys. Chem. A **124**, 4263–4270 (2020).
[86]C. D. Rankine and T. J. Penfold, "Progress in the theory of x-ray spectroscopy: From quantum chemistry to machine learning and ultrafast dynamics," J. Phys. Chem. A **in press**, doi: 10.1021/acs.jpca.0c11267 (2021).
[87]Y. Shu and D. G. Truhlar, "Diabatization by machine intelligence," J. Chem. Theory Comput. **16**, 6456–6464 (2020).
[88]S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE T. Know. Data En. **22**, 1345–1359 (2010).
[89]P. O. Dral, A. Owens, A. Dral, and G. Csányi, "Hierarchical machine learning of potential energy surfaces," J. Chem. Phys. **152**, 204110 (2020).
[90]L. Böselt, M. Thürlemann, and S. Riniker, "Machine learning in qm/mm molecular dynamics simulations of condensed-phase systems," J. Chem. Theory Comput. **in press**, doi:10.1021/acs.jctc.0c01112 (2021).
[91]A. Nandi, C. Qu, P. L. Houston, R. Conte, and J. M. Bowman, "Δ-machine learning for potential energy surfaces: A PIP approach to bring a DFT-based PES to CCSD(T) level of theory," J. Chem. Phys. **154**, 051102 (2021).
[92]J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereuz, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning," Nat. Commun. **10** (2019), 10.1038/s41467-019-10827-4.
[93]L. Ward, B. Blaiszik, I. Foster, R. S. Assary, B. Narayanan, and L. Curtiss, "Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations," arXiv **1906.03233** (2019).
[94]S. Käser, D. Koner, A. S. Christensen, O. A. von Lilienfeld, and M. Meuwly, "Machine Learning Models of Vibrating H_2CO: Comparing Reproducing Kernels, FCHL, and PhysNet," J. Phys. Chem. A **124**, 8853–8865 (2020).
[95]S. Käser, E. Boittier, M. Upadhyay, and M. Meuwly, "Mp2 is not good enough: Transfer learning ml models for accurate vpt2 frequencies," arXiv **2103.05491** (2021).
[96]C. Qu, P. Houston, R. Conte, A. Nandi, and J. M. Bowman, "Breaking the CCSD(T) Barrier for Machine Learned Potentials of Large Molecules: Demonstration for Acetylacetone," arXiv **2103.12333** (2021).
[97]J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci. **8**, 3192–3203 (2017).
[98]M. Bogojeski, L. Vogt-Maranto, M. Tuckerman, K.-R. Müller, and K. Burke, "Quantum chemical accuracy from density functional approximations via machine learning," Nat. Commun. **11**, 5223 (2020).
[99]S. Batzner, T. E. Smidt, L. Sun, J. P. Mailoa, M. Kornbluth, N. Molinari, and B. Kozinsky, "Se(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," arXiv:2101.03164 (2021).
[100]K. T. Schütt, O. T. Unke, and M. Gastegger, "Equivariant message passing for the prediction of tensorial properties and molecular spectra," arXiv:2102.03150 (2021).
[101]M. T. Cvitaš, P. Soldán, and J. M. Hutson, "Long range intermolecular forces in triatomic systems: connecting the atom–diatom and atom–atom–atom representations," Mol. Phys. **104**, 23–31 (2006).
[102]H. Li, C. Collins, M. Tanha, G. J. Gordon, and D. J. Yaron, "A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians," J . Chem. Theory Comput. **14**, 5764–5776 (2018).
[103]M. Welborn, L. Cheng, and T. F. Miller, "Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis," J. Chem. Theory Comput. **14**, 4772–4779 (2018).
[104]L. Cheng, M. Welborn, A. S. Christensen, and T. F. Miller, "A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules," J. Chem. Phys. **150**, 131103 (2019).
[105]T. Husch, J. Sun, L. Cheng, S. J. R. Lee, and T. F. M. III, "Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states," arXiv:2010.03626 (2020).
[106]J. Townsend and K. D. Vogiatzis, "Data-driven acceleration of the coupled-cluster singles and doubles iterative solver," J. Phys. Chem. Lett. **10**, 4129–4135 (2019).
[107]S. Dick and M. Fernandez-Serra, "Machine learning accurate exchange and correlation functionals of the electronic density," Nat. Commun. **11**, 3509 (2020).
[108]Y. Chen, L. Zhang, H. Wang, and W. Weinan, "Ground State Energy Functional with Hartree-Fock Efficiency and Chemical Accuracy," J. Phys. Chem. A **124**, 7155–7165 (2020).
[109]Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, and T. F. Miller, "Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features," J. Chem. Phys. **153**, 124111 (2020).
[110]I. Lagaris, A. Likas, and D. Fotiadis, "Artificial neural network methods in quantum mechanics," Comput. Phys. Commun. **104**, 1 – 14 (1997).
[111]I. E. Lagaris, A. Likas, and D. G. Papageorgiou, "Neural network methods for boundary value problems defined in arbitrarily shaped domains," CoRR **cs.NE/9812003** (1998).
[112]M. Sugawara, "Numerical solution of the Schrödinger equation by neural network and genetic algorithm," Comput. Phys. Commun. **140**, 366–380 (2001).
[113]G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks," Science **355**, 602–606 (2017).
[114]H. Saito, "Solving the bose–hubbard model with machine learning," J. Phys. Soc. Jpn. **86**, 093001 (2017).
[115]Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, "Restricted boltzmann machine learning for solving strongly correlated quantum systems," Phys. Rev. B **96**, 205152 (2017).
[116]J. Han, L. Zhang, and W. E, "Solving Many-Electron Schrödinger Equation using Deep Neural Networks," J. Comput. Phys. **399**, 108929 (2019).
[117]D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, "Ab initio solution of the many-electron schrödinger equation with deep neural networks," Phys. Rev. Res. **2**, 033429 (2020).
[118]J. Hermann, Z. Schätzle, and F. Noé, "Deep-neural-network solution of the electronic schrödinger equation," Nat. Chem. **12**, 891–897 (2020).
[119]K. Choo, G. Carleo, N. Regnault, and T. Neupert, "Symmetries and many-body excitations with neural-network quantum states," Phys. Rev. Lett. **121**, 167204 (2018).
[120]F. Zheng, X. Gao, and A. Eisfeld, "Excitonic wave function reconstruction from near-field spectra using machine learning techniques," Phys. Rev. Lett. **123**, 163202 (2019).
[121]K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller, "SchNet - A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).
[122]K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, "Schnetpack: A deep learning toolbox for atomistic systems," J. Chem. Theory Comput. **15**, 448–455 (2019).

[123]M. Gastegger, A. McSloy, M. Luya, K. T. Schütt, and R. J. Maurer, "A deep neural network for molecular wave functions in quasi-atomic minimal basis representation," J. Chem. Phys. **153**, 044123 (2020).
[124]L. Li, S. Hoyer, R. Pederson, R. Sun, E. D. Cubuk, P. Riley, and K. Burke, "Kohn-sham equations as regularizer: building prior knowledge into machine-learned physics," Phys. Rev. Lett. **126**, 036401 (2021).
[125]P. B. Jørgensen and A. Bhowmik, "Deepdft: Neural message passing network for accurate charge density prediction," arXiv:2011.03346 (2020).
[126]A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf, "Electron density learning of non-covalent systems," Chem. Sci. **10**, 9424–9432 (2019).
[127]A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, "Transferable machine-learning model of the electron density," ACS Cent. Sci. **5**, 57–64 (2019).
[128]A. Fabrizio, K. R. Briling, D. D. Girardier, and C. Corminboeuf, "Learning on-top: Regressing the on-top pair density for real-space visualization of electron correlation," J. Chem. Phys. **153**, 204111 (2020).
[129]J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, "Finding density functionals with machine learning," Phys. Rev. Lett. **108**, 253002 (2012).
[130]F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K. R. Müller, "Bypassing the Kohn-Sham equations with machine learning," Nat. Commun. **8**, 872 (2017).
[131]M. Babaei, Y. T. Azar, and A. Sadeghi, "Locality meets machine learning: Excited and ground-state energy surfaces of large systems at the cost of small ones," Phys. Rev. B **101**, 115132 (2020).
[132]J. Schmidt, C. L. Benavides-Riveros, and M. A. L. Marques, "Machine learning the physical nonlocal exchange–correlation functional of density-functional theory," J. Phys. Chem. Lett. **10**, 6425–6431 (2019).
[133]J. Nelson, R. Tiwari, and S. Sanvito, "Machine learning density functional theory for the hubbard model," Phys. Rev. B **99**, 075132 (2019).
[134]X. Lei and A. J. Medford, "Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors," Phys. Rev. Mater. **3**, 063801 (2019).
[135]Y. Suzuki, R. Nagai, and J. Haruyama, "Machine learning exchange-correlation potential in time-dependent density-functional theory," Phys. Rev. A **101**, 050501 (2020).
[136]V. L. Lignères and E. A. Carter, "An introduction to orbital-free density functional theory," in *Handbook of Materials Modeling: Methods*, edited by S. Yip (Springer Netherlands, Dordrecht, 2005) pp. 137–148.
[137]Y. A. Wang, N. Govind, and E. A. Carter, "Orbital-free kinetic-energy density functionals with a density-dependent kernel," Phys. Rev. B **60**, 16350–16358 (1999).
[138]P. Golub and S. Manzhos, "Kinetic energy densities based on the fourth order gradient expansion: performance in different classes of materials and improvement via machine learning," Phys. Chem. Chem. Phys. **21**, 378–395 (2019).
[139]J. Seino, R. Kageyama, M. Fujinami, Y. Ikabata, and H. Nakai, "Semi-local machine-learned kinetic energy density functional demonstrating smooth potential energy curves," Chem. Phys. Lett. **734**, 136732 (2019).
[140]R. Meyer, M. Weichselbaum, and A. W. Hauser, "Machine learning approaches toward orbital-free density functional theory: Simultaneous training on the kinetic energy density functional and its functional derivative," J. Chem. Theory Comput. **16**, 5685–5694 (2020).
[141]T. Zubatyuk, B. Nebgen, N. Lubbers, J. S. Smith, R. Zubatyuk, G. Zhou, C. Koh, K. Barros, O. Isayev, and S. Tretiak, "Machine Learned Hückel Theory: Interfacing Physics and Deep Neural Networks," arXiv:1909.12963 (2019).
[142]A. Farahvash, C.-K. Lee, Q. Sun, L. Shi, and A. P. Willard, "Machine learning frenkel hamiltonian parameters to accelerate simulations of exciton dynamics," J. Chem. Phys. **153**, 074111 (2020).
[143]F. Häse, S. Valleau, E. Pyzer-Knapp, and A. Aspuru-Guzik, "Machine learning exciton dynamics," Chem. Sci. **7**, 5139–5147 (2016).
[144]Y. Zhang, S. Ye, J. Zhang, J. Jiang, and B. Jiang, "Towards efficient and accurate spectroscopic simulations in extended systems with symmetry-preserving neural network models for tensorial properties," arXiv:2004.13605 (2020).
[145]M. Krämer, P. M. Dohmen, W. Xie, D. Holub, A. S. Christensen, and M. Elstner, "Charge and exciton transfer simulations using machine-learned hamiltonians," J. Chem. Theory Comput. **16**, 4061–4070 (2020).
[146]Z. Wang, S. Ye, H. Wang, J. He, Q. Huang, and S. Chang, "Machine learning method for tight-binding Hamiltonian parameterization from ab-initio band structure," npj Comput. Mater. **7**, 11 (2021).
[147]P. O. Dral, O. A. Von Lilienfeld, and W. Thiel, "Machine learning of parameters for accurate semiempirical quantum chemical calculations," J. Chem. Theory Comput. **11**, 2120–2125 (2015).
[148]C.-P. Chou, Y. Nishimura, C.-C. Fan, G. Mazur, S. Irle, and H. A. Witek, "Automatized parameterization of dftb using particle swarm optimization," J. Chem. Theory Comput. **12**, 53–64 (2016).
[149]M. Stöhr, L. Medrano Sandonas, and A. Tkatchenko, "Accurate many-body repulsive potentials for density-functional tight binding from deep tensor neural networks," J. Phys. Chem. Lett. **11**, 6835–6843 (2020).
[150]C. Panosetti, A. Engelmann, L. Nemec, K. Reuter, and J. T. Margraf, "Learning to use the force: Fitting repulsive potentials in density-functional tight-binding with gaussian process regression," J. Chem. Theory Comput. **16**, 2181–2191 (2020).
[151]J. Hoja, L. M. Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr, and A. Tkatchenko, "QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules," Sci. Data **8**, 43 (2021).
[152]F. Manby, T. Miller, P. Bygrave, F. Ding, T. Dresselhaus, F. Batista-Romero, A. Buccheri, C. Bungey, S. Lee, R. Meli, K. Miyamoto, C. Steinmann, T. Tsuchiya, M. Welborn, T. Wiles, and Z. Williams, "entos: A Quantum Molecular Simulation Package," chemrxiv , 10.26434/chemrxiv.7762646.v2 (2019).
[153]B. Hourahine et al., "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations," J. Chem. Phys. **152**, 124101 (2020).
[154]"Quantum Chemistry's Modular Movement," Chem. Eng. News **92**, 26 (2014).
[155]A. Khorshidi and A. A. Peterson, "Amp: A modular approach to machine learning in atomistic simulations," Comput. Phys. Commun. **207**, 310–324 (2016).
[156]S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, "sgdml: Constructing accurate and data efficient molecular force fields using machine learning," Comput. Phys. Commun. **240**, 38 – 45 (2019).
[157]V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "Ab initio molecular simulations with numeric atom-centered orbitals," Comp. Phys. Commun. **180**, 2175–2196 (2009).
[158]D. G. A. Smith, "Psi4 1.4: Open-source software for high-throughput quantum chemistry," J. Chem. Phys. **152**, 184108 (2020).
[159]Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K. Chan, "Pyscf: the python-based simulations of chemistry framework," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **8**, e1340 (2017).
[160]V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M. Wilkins, B. A. Helfrecht, P. Juda, S. P. Bienvenue, W. Fang,

J. Kessler, I. Poltavsky, S. Vandenbrande, J. Wieme, C. Corminboeuf, T. D. Kühne, D. E. Manolopoulos, T. E. Markland, J. O. Richardson, A. Tkatchenko, G. A. Tribello, V. Van Speybroeck, and M. Ceriotti, "i-pi 2.0: A universal force engine for advanced molecular simulations," Comput. Phys. Commun. **236**, 214–223 (2019).

[161]S. Mai, P. Marquetand, and L. González, "Nonadiabatic Dynamics: The SHARC Approach," WIREs Comput. Mol. Sci. **8**, e1370 (2018).

[162]M. Richter, P. Marquetand, J. González-Vázquez, I. Sola, and L. González, "SHARC: Ab initio molecular dynamics with surface hopping in the adiabatic representation including arbitrary couplings," J. Chem. Theory Comput. **7**, 1253–1258 (2011).

[163]A. H. Larsen, "The atomic simulation environment—a python library for working with atoms," J. Phys.: Condens. Matter **29**, 273002 (2017).

[164]C. Draxl and M. Scheffler, "NOMAD: The FAIR concept for big data-driven materials science," MRS Bull. **43**, 676–682 (2018).

[165]C. Draxl and M. Scheffler, "The NOMAD laboratory: from data sharing to artificial intelligence," J. Phys. Mater. **2**, 036001 (2019).

[166]D. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard, and T. D. Crawford, "The molssi qcarchive project: An open-source platform to compute, organize, and share quantum chemistry data," WIREs Computational Molecular Science **11**, e1491 (2021).

[167]"Quantum machine repository," http://quantum-machine.org/datasets/ .

[168]A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, "Structure prediction drives materials discovery," Nat. Rev. Mater. **4**, 331–348 (2019).

[169]G. R. Fleming and P. G. Wolynes, "Chemical dynamics in solution," Phys. Today **43**, 36–43 (1990).

[170]A. Denzel and J. Kästner, "Gaussian Process Regression for Transition State Search," J. Chem. Theory Comput. **14**, 5777–5786 (2018).

[171]R. Meyer and A. W. Hauser, "Geometry optimization using Gaussian process regression in internal coordinate systems," J. Chem. Phys. **152**, 84112 (2020).

[172]G. Raggi, I. F. Galván, C. L. Ritterhoff, M. Vacher, and R. Lindh, "Restricted-Variance Molecular Geometry Optimization Based on Gradient-Enhanced Kriging," J. Chem. Theory Comput. **16**, 3989–4001 (2020).

[173]G. Schmitz and O. Christiansen, "Gaussian process regression to accelerate geometry optimizations relying on numerical differentiation," J. Chem. Phys. **148**, 241704 (2018).

[174]E. Garijo del Río, J. J. Mortensen, and K. W. Jacobsen, "Local bayesian optimizer for atomic structures," Phys. Rev. B **100**, 104103 (2019).

[175]E. Garijo del Río, S. Kaappa, J. A. Garrido Torres, T. Bligaard, and K. W. Jacobsen, "Machine learning with bond information for local structure optimizations in surface science," J. Chem. Phys. **153**, 234116 (2020).

[176]A. Denzel and J. Kästner, "Hessian Matrix Update Scheme for Transition State Search Based on Gaussian Process Regression," J. Chem. Theory Comput. **16**, 5083–5089 (2020).

[177]A. A. Peterson, "Acceleration of saddle-point searches with machine learning," J. Chem. Phys. **145**, 074106 (2016).

[178]O. P. Koistinen, F. B. Dagbjartsdóttir, V. Ásgeirsson, A. Vehtari, and H. Jónsson, "Nudged elastic band calculations accelerated with Gaussian process regression," J. Chem. Phys. **147**, 152720 (2017).

[179]J. A. Garrido Torres, P. C. Jennings, M. H. Hansen, J. R. Boes, and T. Bligaard, "Low-Scaling Algorithm for Nudged Elastic Band Calculations Using a Surrogate Machine Learning Model," Phys. Rev. Lett. **122**, 156001 (2019).

[180]F. Curtis, X. Li, T. Rose, Á. Vázquez-Mayagoitia, S. Bhattacharya, L. M. Ghiringhelli, and N. Marom, "GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction," J. Chem. Theory Comput. **14**, 2246–2264 (2018).

[181]C. J. Pickard and R. J. Needs, "Ab initio random structure searching," J. Phys. Cond. Matter **23**, 053201 (2011).

[182]D. J. Wales and J. P. K. Doye, "Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms," J. Phys. Chem. A **101**, 5111–5116 (1997).

[183]C. Panosetti, K. Krautgasser, D. Palagin, K. Reuter, and R. J. Maurer, "Global materials structure search with chemically-motivated coordinates." Nano Lett. **15**, 8044–8048 (2015).

[184]A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, "Improved protein structure prediction using potentials from deep learning," Nature **577**, 706–710 (2020).

[185]J. J. et al., "High Accuracy Protein Structure Prediction Using Deep Learning," in *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction* (2020).

[186]M. S. Jørgensen, H. L. Mortensen, S. A. Meldgaard, E. L. Kolsbjerg, T. L. Jacobsen, K. H. Sørensen, and B. Hammer, "Atomistic structure learning," J. Chem. Phys. **151**, 054111 (2019).

[187]H. L. Mortensen, S. A. Meldgaard, M. K. Bisbo, M.-P. V. Christiansen, and B. Hammer, "Atomistic structure learning algorithm with surrogate energy model relaxation," Phys. Rev. B **102**, 075427 (2020).

[188]S. R. A. Meldgaard, H. L. Mortensen, M. S. Jorgensen, and B. Hammer, "Structure prediction of surface reconstructions by deep reinforcement learning," J. Condens. Matter Phys. **32**, 404005 (2020).

[189]T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, "Crystal structure prediction accelerated by Bayesian optimization," Phys. Rev. Mater. **2**, 013803 (2018).

[190]V. L. Deringer, D. M. Proserpio, G. Csányi, and C. J. Pickard, "Data-driven learning and prediction of inorganic crystal structures," Faraday Discuss. **211**, 45–59 (2018).

[191]M. K. Bisbo and B. Hammer, "Efficient Global Structure Optimization with a Machine-Learned Surrogate Model," Phys. Rev. Lett. **124**, 086102 (2020).

[192]M. Todorović, M. U. Gutmann, J. Corander, and P. Rinke, "Bayesian inference of atomistic structure in functional materials," npj Comput. Mater. **5**, 35 (2019).

[193]L. Hörmann, A. Jeindl, A. T. Egger, M. Scherbela, and O. T. Hofmann, "SAMPLE: Surface structure search enabled by coarse graining and statistical learning," Comput. Phys. Commun. **244**, 143–155 (2019).

[194]B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning:Generative models for matter engineering," Science **361**, 360–365 (2018).

[195]D. Schwalbe-Koda and R. Gómez-Bombarelli, "Generative models for automatic chemical design," in *Machine Learning Meets Quantum Physics* (Springer, 2020) pp. 445–467.

[196]R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," ACS Cent. Sci. **4**, 268–276 (2018).

[197]Q. Liu, M. Allamanis, M. Brockschmidt, and A. Gaunt, "Constrained graph variational autoencoders for molecule design," in *Advances in Neural Information Processing Systems* (2018) pp. 7795–7804.

[198]E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, and A. Zhavoronkov, "Reinforced Adversarial Neural Computer for de Novo Molecular Design," J. Chem. Inf. Model. **58**, 1194–1204 (2018).

[199]M. Popova, O. Isayev, and A. Tropsha, "Deep reinforcement learning for de novo drug design," Sci. Adv. **4**, eaap7885 (2018).

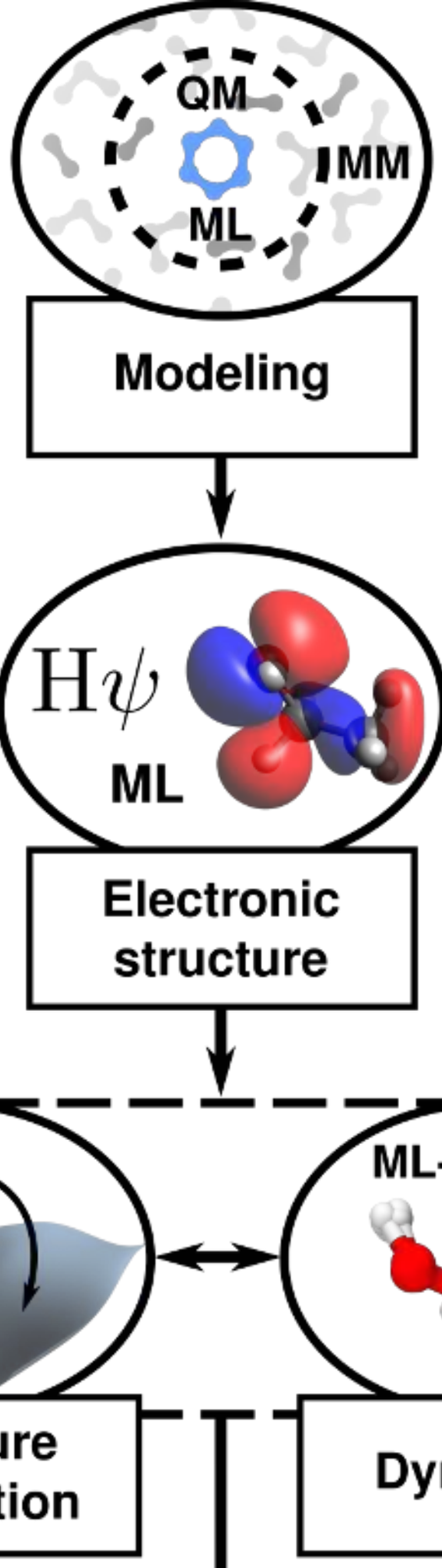
[200]M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," arXiv:1703.01925 (2017).

[201]Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, "Optimization of Molecules via Deep Reinforcement Learning," Sci. Rep. **9**, 10752 (2019).
[202]E. Mansimov, O. Mahmood, S. Kang, and K. Cho, "Molecular geometry prediction using a deep generative graph neural network," Sci. Rep. **9**, 1–13 (2019).
[203]J. Köhler, L. Klein, and F. Noé, "Equivariant flows: sampling configurations for multi-body systems with symmetric energies," in *Proceedings of the 37th International Conference on Machine Learning* (2019).
[204]N. Gebauer, M. Gastegger, and K. Schütt, "Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules," Advances in Neural Information Processing Systems **32**, 7566–7578 (2019).
[205]O. A. von Lilienfeld, R. D. Lins, and U. Rothlisberger, "Variational particle number approach for rational compound design," Phys. Rev. Lett. **95**, 153002 (2005).
[206]O. A. Von Lilienfeld and M. Tuckerman, "Alchemical variations of intermolecular energies according to molecular grand-canonical ensemble density functional theory," J. Chem. Theory Comput. **3**, 1083–1090 (2007).
[207]D. Sheppard, G. Henkelman, and O. A. von Lilienfeld, "Alchemical derivatives of reaction energetics," J. Chem. Phys. **133**, 084104 (2010).
[208]F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld, "Alchemical and structural distribution based representation for universal quantum machine learning," J. Chem. Phys. **148**, 241717 (2018).
[209]S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).
[210]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 1609.08259 (2017).
[211]J. Behler, "Perspective: Machine learning potentials for atomistic simulations," J. Chem. Phys. **145**, 170901 (2016).
[212]V. Botu, R. Batra, J. Chapman, and R. Ramprasad, "Machine learning force fields: Construction, validation, and outlook," J. Phys. Chem. C **121**, 511–522 (2017).
[213]V. L. Deringer, N. Bernstein, G. Csányi, C. B. Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott., "Origins of structural and electronic transitions in disordered silicon," Nature **589**, 59–64 (2021).
[214]B. Jiang, J. Li, and H. Guo, "High-Fidelity Potential Energy Surfaces for Gas Phase and Gas-Surface Scattering Processes from Machine Learning," J. Phys. Chem. Lett. **11**, 5120–5131 (2020).
[215]R. Meyer and A. W. Hauser, "Geometry optimization using gaussian process regression in internal coordinate systems," J. Chem. Phys. **152**, 084112 (2020).
[216]O. T. Unke, D. Koner, S. Patra, S. Käser, and M. Meuwly, "High-dimensional potential energy surfaces for molecular simulations: from empiricism to machine learning," Mach. Learn.: Sci. Technol. **1**, 013001 (2020).
[217]D. Koner, O. T. Unke, K. Boe, R. J. Bemish, and M. Meuwly, "Exhaustive state-to-state cross sections for reactive molecular collisions from importance sampling simulation and a neural network representation," J. Chem. Phys. **150**, 211101 (2019).
[218]J. Danielsson and M. Meuwly, "Atomistic simulation of adiabatic reactive processes based on multi-state potential energy surfaces," J. Chem. Theory Comput. **4**, 1083–1093 (2008).
[219]J. M. Bowman, G. Czakó, and B. Fu, "High-dimensional ab initio potential energy surfaces for reaction dynamics calculations," Phys. Chem. Chem. Phys. **13**, 8094–8111 (2011).
[220]M. Meuwly, "Transformative applications of machine learning for chemical reactions," arXiv **2101.03530** (2021).
[221]F. Häse, I. Fdez. Galván, A. Aspuru-Guzik, R. Lindh, and M. Vacher, "How machine learning can assist the interpretation of *Ab Initio* molecular dynamics simulations and conceptual understanding of chemistry," Chem. Sci. **10**, 2298–2307 (2019).
[222]P. L. Houston, A. Nandi, and J. M. Bowman, "A machine learning approach for prediction of rate constants," J. Phys. Chem. Lett. **10**, 5250–5258 (2019).
[223]Z. Li, J. R. Kermode, and A. De Vita, "Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces," Phys. Rev. Lett. **114**, 096405 (2015).
[224]V. Botu and R. Ramprasad, "Learning scheme to predict atomic forces and accelerate materials simulations," Phys. Rev. B **92**, 094306 (2015).
[225]J. Behler, "Constructing high-dimensional neural network potentials: A tutorial review," Int. J. Quantum Chem. **115**, 1032–1050 (2015).
[226]M. Gastegger, J. Behler, and P. Marquetand, "Machine learning molecular dynamics for the simulation of infrared spectra," Chem. Sci. **8**, 6924–6935 (2017).
[227]A. V. Akimov, "A simple phase correction makes a big difference in nonadiabatic molecular dynamics," J. Phys. Chem. Lett. **9**, 6096–6102 (2018).
[228]S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).
[229]S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," Nat. Commun. **9**, 3887 (2018).
[230]A. Ma and A. R. Dinner, "Automatic method for identifying reaction coordinates in complex systems," J. Phys. Chem. B **109**, 6769–6779 (2005), pMID: 16851762.
[231]F. Noé, G. De Fabritiis, and C. Clementi, "Machine learning for protein folding and dynamics," Curr. Opin. Struct. Biol. **60**, 77 – 84 (2020).
[232]K. Pearson, "On lines of closes fit to system of points in space, london, e dinb," Dublin Philos. Mag. J. Sci **2**, 559–572 (1901).
[233]M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, "Principal component analysis and long time protein dynamics," J. Phys. Chem. **100**, 2567–2572 (1996).
[234]B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comput. **10**, 1299–1319 (1998).
[235]W. W. Zhang Z., "Coarse-graining protein structures with local multivariate features from molecular dynamics," J. Phys. Chem. B **112**, 44 (2008).
[236]O. F. Lange and H. Grubmüller, "Full correlation analysis of conformational protein dynamics," Proteins: Struct., Funct., Bioinf. **70**, 1294–1312 (2008).
[237]R. R. Coifman and S. Lafon, "Diffusion maps," Appl. Comput. Harmon. Anal. **21**, 5–30 (2006).
[238]J. Preto and C. Clementi, "Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics," Phys. Chem. Chem. Phys. **16**, 19181–19191 (2014).
[239]W. Zheng, M. A. Rohrdanz, and C. Clementi, "Rapid exploration of configuration space with diffusion-map-directed molecular dynamics," J. Phys. Chem. B **117**, 12769–12776 (2013).
[240]G. A. Tribello, M. Ceriotti, and M. Parrinello, "Using sketch-map coordinates to analyze and bias molecular dynamics simulations," Proc. Natl. Acad. Sci **109**, 5196–5201 (2012).
[241]M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," Proc. Natl. Acad. Sci **108**, 13023–13028 (2011).
[242]A. Mardt, L. Pasquali, H. Wu, and F. Noé, "VAMPnets for deep learning of molecular kinetics," Nat. Commun. **9**, 5 (2018).
[243]F. Noé and E. Rosta, "Markov Models of Molecular Kinetics," J. Chem. Phys. **151**, 190401 (2019).
[244]W. Chen, A. R. Tan, and A. L. Ferguson, "Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design," J. Chem. Phys. **149**, 072312 (2018).
[245]J. M. L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary, "Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)," J. Chem. Phys. **149**, 072301 (2018).

[246]T. Lemke and C. Peter, "Neural Network Based Prediction of Conformational Free Energies - A New Route toward Coarse-Grained Simulation Models," J. Chem. Theory Comput. **13**, 6213–6221 (2017).
[247]L. Zhang, J. Han, H. Wang, R. Car, and W. E. Weinan, "DeePCG: Constructing coarse-grained models via deep neural networks," J. Chem. Phys. **149**, 034101 (2018).
[248]J. Wang, S. Chmiela, K.-R. Müller, and C. C. Frank Noé, "Ensemble learning of coarse-grained molecular dynamics force fields with a kernel approach," J. Chem. Phys. **152**, 194106 (2020).
[249]S. T. John and G. Csányi, "Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials," J. Phys. Chem. B **121**, 10934–10949 (2017).
[250]M. Barbatti, "Nonadiabatic dynamics with trajectory surface hopping method," Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**, 620–633 (2011).
[251]L. González and R. Lindh, eds., *Quantum Chemistry and Dynamics of Excited States: Methods and Applications* (John Wiley & Sons, 2020).
[252]S. Mai and L. González, "Molecular photochemistry: Recent developments in theory," Ang. Chem. Int. Ed. **59**, 16832–16846 (2020).
[253]B. Smith and A. V. Akimov, "Modeling nonadiabatic dynamics in condensed matter materials: Some recent advances and applications," J. Phys. Cond. Matter **32**, 073001 (2020).
[254]J. Westermayr, M. Gastegger, and P. Marquetand, "Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics," J. Phys. Chem. Lett. **11**, 3828–3834 (2020).
[255]J. Li, P. Reiser, A. Eberhard, P. Friederich, and S. Lopez, "Nanosecond photodynamics simulations of a cis-trans isomerization are enabled by machine learning," ChemRxiv , DOI:10.26434/chemrxiv.13047863.v1 (2020).
[256]C. Carbogno, J. Behler, K. Reuter, and A. Groß, "Signatures of nonadiabatic O_2 dissociation at Al(111): First-principles fewest-switches study," Phys. Rev. B **81**, 035410 (2010).
[257]Y. Zhang, R. J. Maurer, and B. Jiang, "Symmetry-Adapted High Dimensional Neural Network Representation of Electronic Friction Tensor of Adsorbates on Metals," J. Phys. Chem. C **124**, 186–195 (2020).
[258]Y. Zhang, R. J. Maurer, H. Guo, and B. Jiang, "Hot-electron effects during reactive scattering of H_2 from Ag(111): the interplay between mode-specific electronic friction and the potential energy landscape," Chem. Sci. **10**, 1089–1097 (2019).
[259]C. L. Box, Y. Zhang, R. Yin, B. Jiang, and R. J. Maurer, "Determining the effect of hot electron dissipation on molecular scattering experiments at metal surfaces," JACS Au **in press** (2020), 10.1021/jacsau.0c00066.
[260]D. R. Yarkony, "Nonadiabatic quantum chemistry - past, present, and future," Chem. Rev. **112**, 481–498 (2012).
[261]H. Köppel, W. Domcke, and L. S. Cederbaum, *in: Conical Intersections (W. Domcke, D. R. Yarkony, H. Köppel, Eds.)* (World Scientific, New York, 2004).
[262]B. Jiang, J. Li, and H. Guo, "Potential energy surfaces from high fidelity fitting of *Ab Initio* points: The permutation invariant polynomial - neural network approach," Int. Rev. Phys. Chem. **35**, 479–506 (2016).
[263]T. Lenzen and U. Manthe, "Neural network based coupled diabatic potential energy surfaces for reactive scattering," J. Chem. Phys. **147**, 084105 (2017).
[264]D. M. G. Williams and W. Eisfeld, "Neural Network Diabatization: A New Ansatz for Accurate High-Dimensional Coupled Potential Energy Surfaces," J. Chem. Phys. **149**, 204106 (2018).
[265]C. Xie, X. Zhu, D. R. Yarkony, and H. Guo, "Permutation invariant polynomial neural network approach to fitting potential energy surfaces. IV. coupled diabatic potential energy matrices," J. Chem. Phys. **149**, 144107 (2018).
[266]D. M. G. Williams and W. Eisfeld, "Complete nuclear permutation inversion invariant artificial neural network (cnpi-ann) diabatization for the accurate treatment of vibronic coupling problems," J. Phys. Chem. A **in press**, DOI:10.1021/acs.jpca.0c05991 (2020).
[267]G. W. Richings and S. Habershon, "Direct grid-based quantum dynamics on propagated diabatic potential energy surfaces," Chem. Phys. Lett. **683**, 228 – 233 (2017).
[268]G. W. Richings and S. Habershon, "MCTDH on-the-fly: Efficient grid-based quantum dynamics without pre-computed potential energy surfaces," J. Chem. Phys. **148**, 134116 (2018).
[269]G. W. Richings, C. Robertson, and S. Habershon, "Improved on-the-fly MCTDH simulations with many-body-potential tensor decomposition and projection diabatization," J. Chem. Theory Comput. **15**, 857–870 (2019).
[270]G. W. Richings and S. Habershon, "A new diabatization scheme for direct quantum dynamics: Procrustes diabatization," J. Chem. Phys. **152**, 154108 (2020).
[271]G. W. Richings, C. Robertson, and S. Habershon, "Can we use on-the-fly quantum simulations to connect molecular structure and sunscreen action?" Faraday Discuss. **216**, 476–493 (2019).
[272]R. Salomon-Ferrer, D. A. Case, and R. C. Walker, "An overview of the amber biomolecular simulation package," WIREs: Comput. Mol. Sci. **3**, 198–210 (2013).
[273]C. van der Oord, G. Dusson, G. Csányi, and C. Ortner, "Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials," Mach. Learn.: Sci. Technol. **1**, 015004 (2020).
[274]R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).
[275]A. E. A. Allen, G. Dusson, C. Ortner, and G. Csányi, "Atomic permutationally invariant polynomials for fitting molecular force fields," Mach. Learn.: Sci. Technol. **2**, 025017 (2021).
[276]A. Jasinski, J. Montaner, R. C. Forrey, B. H. Yang, P. C. Stancil, N. Balakrishnan, J. Dai, R. A. Vargas-Hernández, and R. V. Krems, "Machine learning corrected quantum dynamics calculations," Phys. Rev. Research **2**, 3 (2020).
[277]F. Brieuc, C. Schran, F. Uhl, H. Forbert, and D. Marx, "Converged quantum simulations of reactive solutes in superfluid helium: The bochum perspective," J. Chem. Phys. **152**, 210901 (2020).
[278]N. Raimbault, A. Grisafi, M. Ceriotti, and M. Rossi, "Using gaussian process regression to simulate the vibrational raman spectra of molecular crystals," New J. Phys. **21**, 105001 (2019).
[279]G. M. Sommers, M. F. C. Andrade, L. Zhang, H. Wang, and R. Car, "Raman spectrum and polarizability of liquid water from deep neural networks," Phys. Chem. Chem. Phys. **22**, 10592–10602 (2020).
[280]F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, "Chemical shifts in molecular solids by machine learning," Nat. Commun. **9**, 1–10 (2018).
[281]A. S. Christensen, F. A. Faber, and O. A. Von Lilienfeld, "Operators in quantum machine learning: Response properties in chemical space," J. Chem. Phys. **150**, 064105 (2019).
[282]J. A. Fine, A. A. Rajasekar, K. P. Jethava, and G. Chopra, "Spectral deep learning for prediction and prospective validation of functional groups," Chem. Sci. **11**, 4618–4630 (2020).
[283]S. Kiyohara, T. Miyata, K. Tsuda, and T. Mizoguchi, "Data-driven approach for the prediction and interpretation of core-electron loss spectroscopy," Sci. Rep. **8**, 1–12 (2018).
[284]C. Cobas, "Nmr signal processing, prediction, and structure verification with machine learning techniques," Magn. Reson. Chem. **58**, 512–519 (2020).
[285]V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," Nature **571**, 95–98 (2019).
[286]P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, "Machine-learning-assisted materials discovery using failed experiments," Nature **533**, 73–76 (2016).

[287]K. McCullough, T. Williams, K. Mingle, P. Jamshidi, and J. Lauterbach, "High-throughput experimentation meets artificial intelligence: A new pathway to catalyst discovery," Phys. Chem. Chem. Phys. **22**, 11174–11196 (2020).
[288]J. L. Melville, E. K. Burke, and J. D. Hirst, "Machine learning in virtual screening," Comb. Chem. High Throughput Screening **12**, 332–343 (2009).
[289]K. Terayama, K. Terayama, K. Terayama, K. Terayama, M. Sumita, M. Sumita, R. Tamura, R. Tamura, R. Tamura, D. T. Payne, M. K. Chahal, S. Ishihara, K. Tsuda, K. Tsuda, and K. Tsuda, "Pushing property limits in materials discovery: Via boundless objective-free exploration," Chem. Sci. **11**, 5959–5968 (2020).
[290]S. Ekins, A. C. Puhl, K. M. Zorn, T. R. Lane, D. P. Russo, J. J. Klein, A. J. Hickey, and A. M. Clark, "Exploiting machine learning for end-to-end drug discovery and development," Nat. Mater. **18**, 435 (2019).
[291]B. Meyer, B. Sawatlon, S. Heinen, O. A. Von Lilienfeld, and C. Corminboeuf, "Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts," Chem. Sci. **9**, 7069–7077 (2018).
[292]J. I. Gómez-Peralta and X. Bokhimi, "Discovering new perovskites with artificial intelligence," J. Solid State Chem. **285**, 121253 (2020).
[293]P. B. Jørgensen, M. Mesta, S. Shil, J. M. García Lastra, K. W. Jacobsen, K. S. Thygesen, and M. N. Schmidt, "Machine learning-based screening of complex molecules for polymer solar cells," J. Chem. Phys. **148**, 241735 (2018).
[294]P. C. St John, C. Phillips, T. W. Kemper, A. N. Wilson, Y. Guan, M. F. Crowley, M. R. Nimlos, and R. E. Larsen, "Message-passing neural networks for high-throughput polymer screening," J. Chem. Phys. **150**, 234111 (2019).
[295]C. M. Dobson, "Chemical space and biology," Nature **432**, 824–828 (2004).
[296]T. Weymuth and M. Reiher, "Inverse quantum chemistry: Concepts and strategies for rational compound design," Int. J. Quantum Chem. **114**, 823–837 (2014).
[297]A. Zunger, "Inverse design in search of materials with target functionalities," Nat. Rev. Chem. **2**, 1–16 (2018).
[298]E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, "Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning," Phys. Rev. B **99**, 064114 (2019).
[299]J. Noh, G. H. Gu, S. Kim, and Y. Jung, "Machine-enabled inverse design of inorganic solid materials: promises and challenges," Chem. Sci. **11**, 4871–4881 (2020).
[300]R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, *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).
[301]K. Lin, R. Gómez-Bombarelli, E. S. Beh, L. Tong, Q. Chen, A. Valle, A. Aspuru-Guzik, M. J. Aziz, and R. G. Gordon, "A redox-flow battery with an alloxazine-based organic electrolyte," Nat. Energy **1**, 1–8 (2016).
[302]A. Ambrosetti, N. Ferri, R. A. DiStasio, and A. Tkatchenko, "Wavelike charge density fluctuations and van der waals interactions at the nanoscale," Science **351**, 1171–1176 (2016).
[303]M. Wilkinson et al., "The fair guiding principles for scientific data management and stewardship," Sci. Data **3**, 160018 (2016).
[304]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "The Materials Project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

Model selection
Bayesian error estimation
adaptive basis sets
pseudo-potentials



Model building
active space selection
solvation models

ML integration
MO & density prediction
NN wavefunctions

Semi-empirical methods
ML-based Hückel,
TB Hamiltonians
& repulsive energy

Physically informed ML
atomic orbital features
symmetries
2-electron integrals

Mixed levels of theory
Δ- learning
transfer learning

Local exploration
equilibrium search
transition states

Global exploration
structure prediction
generative models
reinforcement learning

Simulation
interatomic potentials
nonadiabatic excited states
quantum dynamics

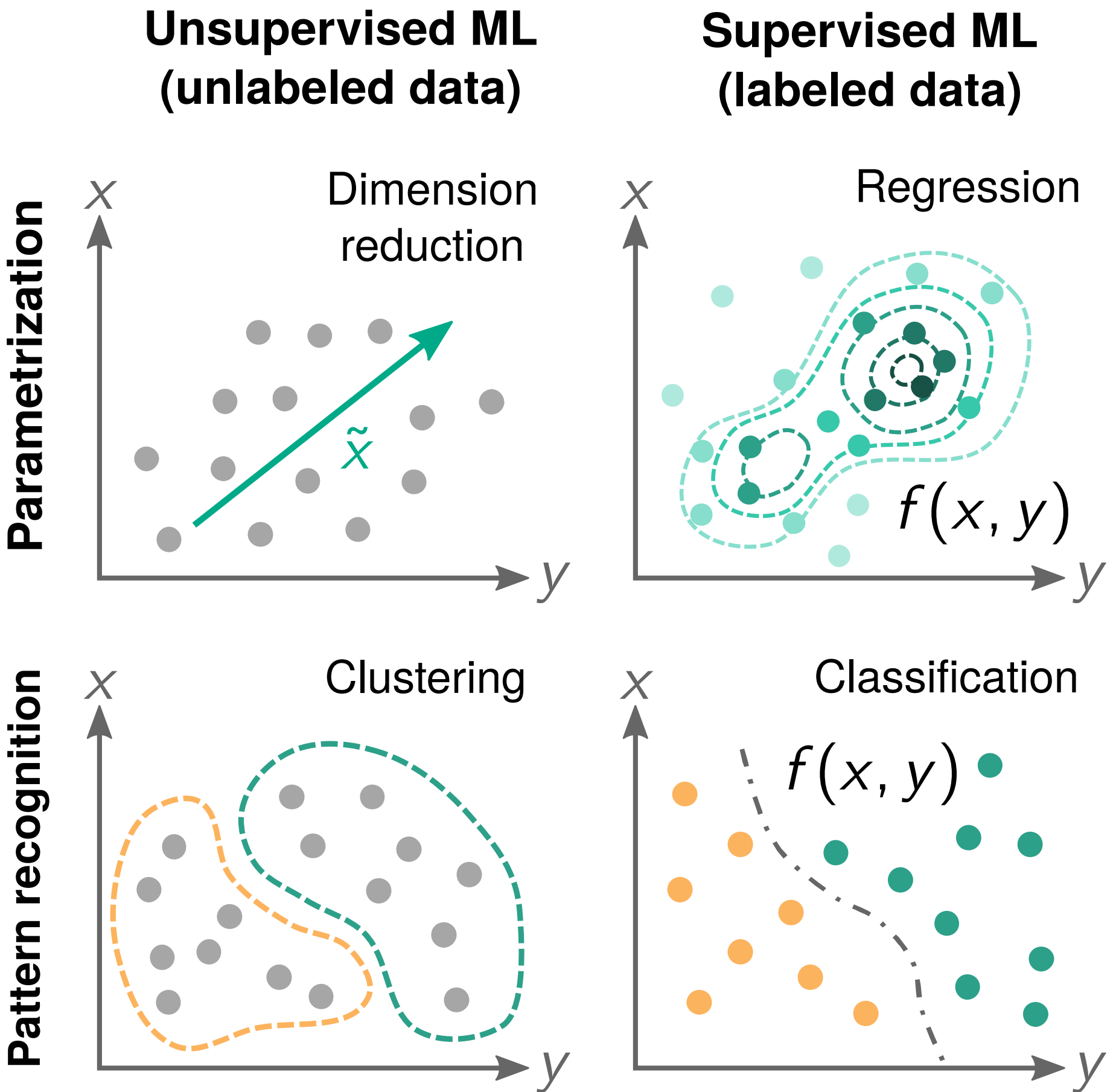
Rare events
long-time dynamics
coarse-graining

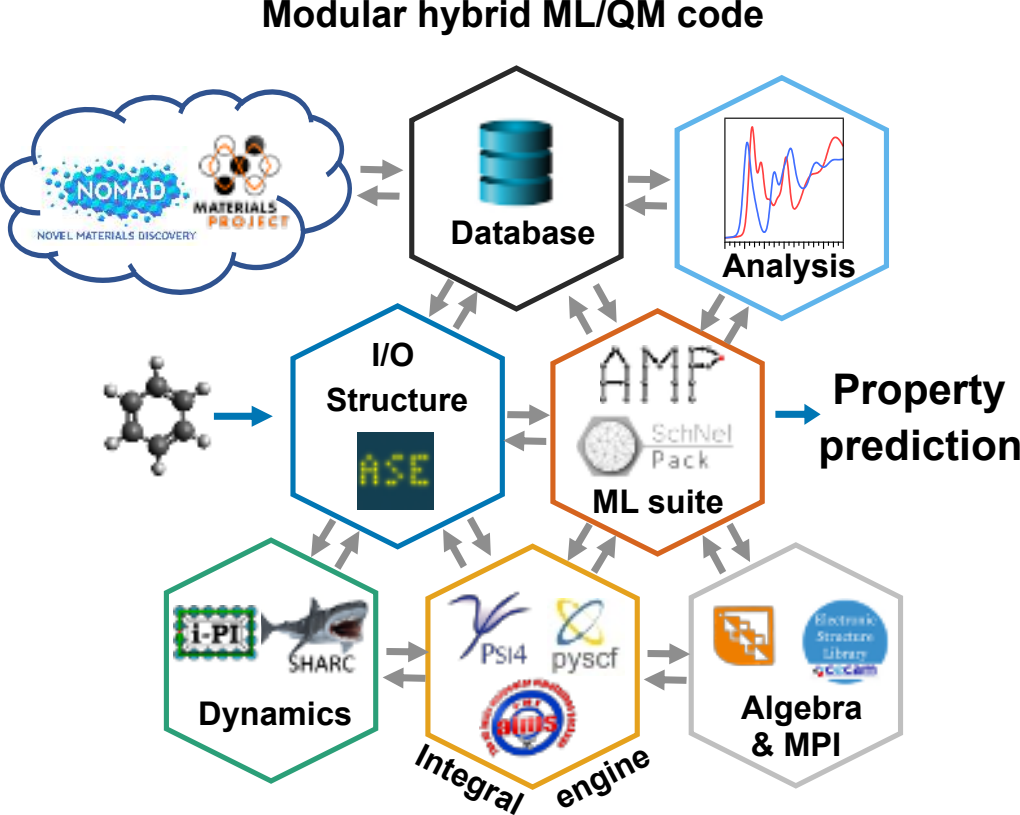
Computational spectroscopy
interatomic potentials
response properties
direct prediction

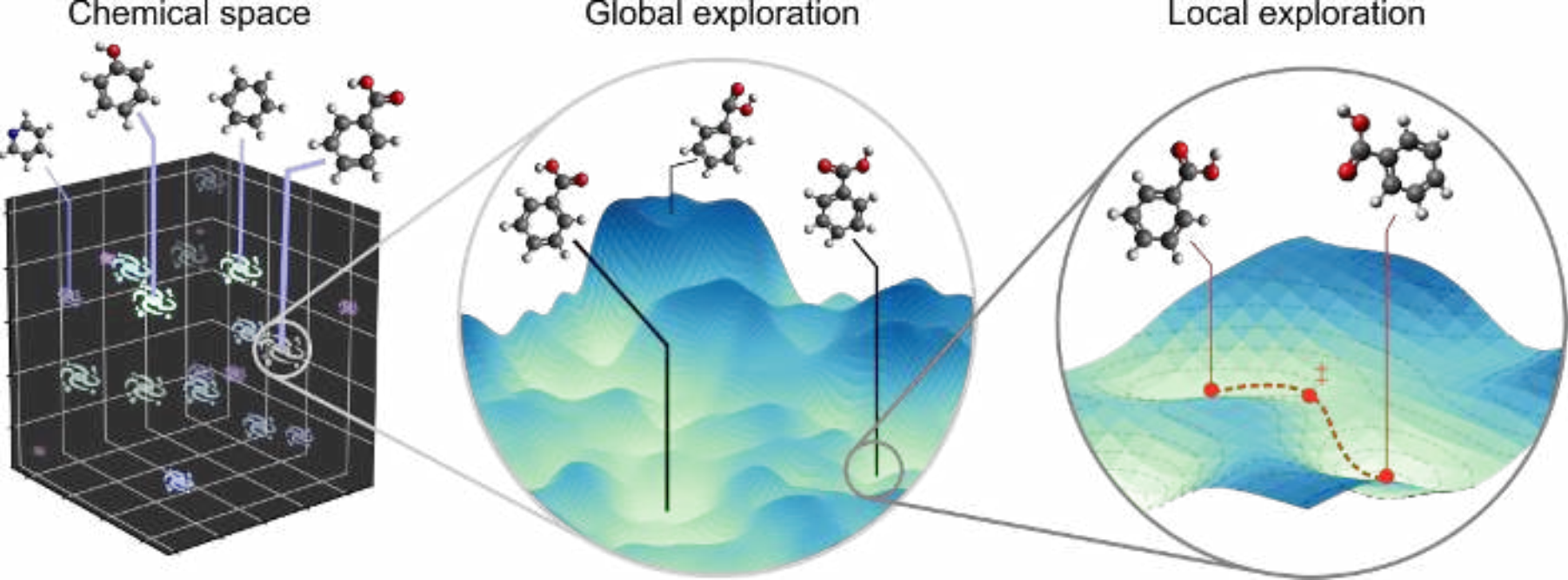
Data analysis
X-ray spectroscopy
NMR spectroscopy

High-troughput screening
ML surrogate models
drug & catalyst design

Inverse design
generative models,
reinforcement learning

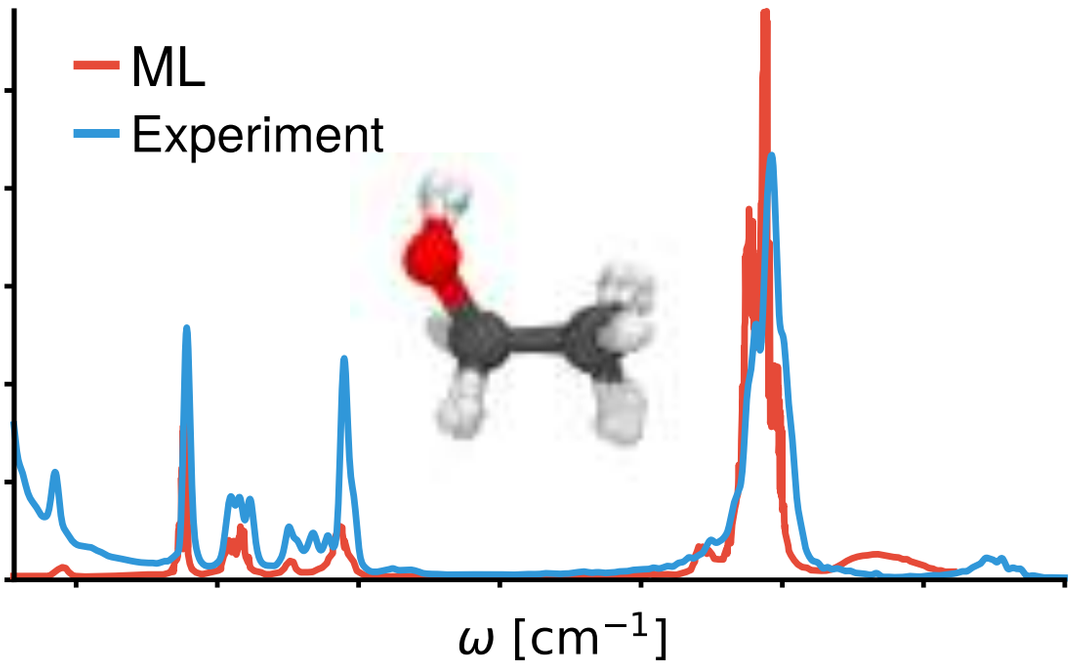

Unsupervised ML
(unlabeled data)
Supervised ML
(labeled data)
Parametrization
Pattern recognition
Dimension
reduction
Regression
Clustering
Classification
x
y
$\tilde{x}$
$f(x, y)$


Modular hybrid ML/QM code
NOMAD
NOVEL MATERIALS DISCOVERY
MATERIALS PROJECT
Database
Analysis
I/O
Structure
ASE
AMP
SchNet Pack
ML suite
Property prediction
i-PI
SHARC
Dynamics
Psi4
pyscf
Integral engine
Algebra & MPI

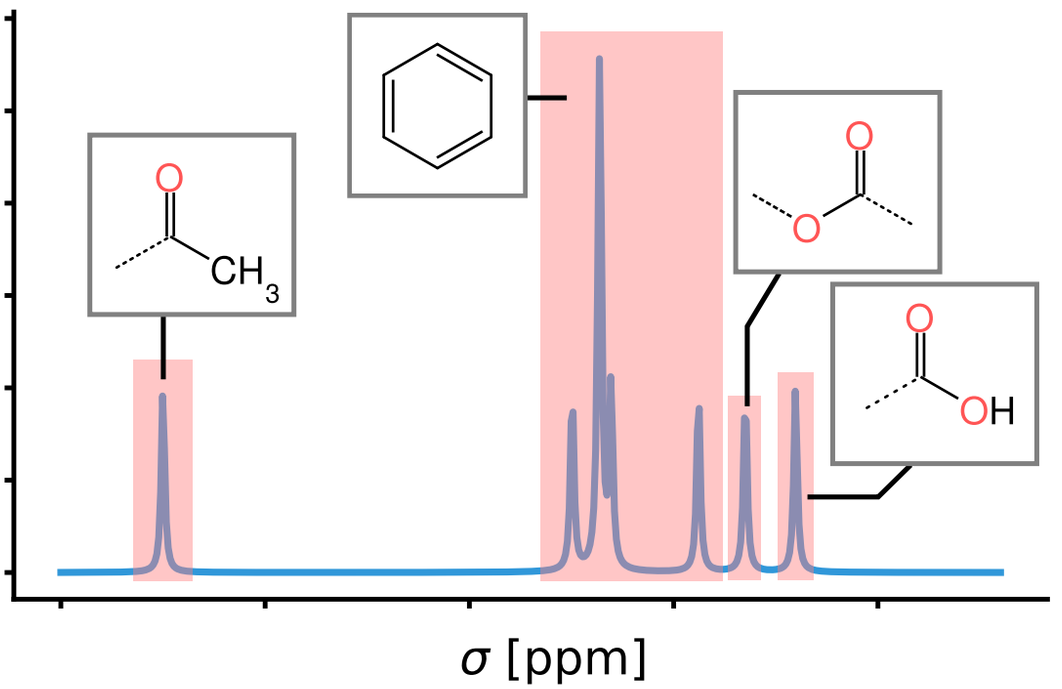

Chemical space
Global exploration
Local exploration

Machine learning

Fast + accurate simulations



Knowledge extraction



Theory

Experiment