REVIEW ARTICLE | INSIGHT
https://doi.org/10.1038/s41563-020-0777-6
nature materials

Machine-learned potentials for next-generation matter simulations

Pascal Friederich[1,2,3,7], Florian Häse[1,2,4,5], Jonny Proppe[1,2,8] and Alán Aspuru-Guzik[1,2,4,5,6] ✉

The choice of simulation methods in computational materials science is driven by a fundamental trade-off: bridging large time- and length-scales with highly accurate simulations at an affordable computational cost. Venturing the investigation of complex phenomena on large scales requires fast yet accurate computational methods. We review the emerging field of machine-learned potentials, which promises to reach the accuracy of quantum mechanical computations at a substantially reduced computational cost. This Review will summarize the basic principles of the underlying machine learning methods, the data acquisition process and active learning procedures. We highlight multiple recent applications of machine-learned potentials in various fields, ranging from organic chemistry and biomolecules to inorganic crystal structure predictions and surface science. We furthermore discuss the developments required to promote a broader use of ML potentials, and the possibility of using them to help solve open questions in materials science and facilitate fully computational materials design.

Virtualizing the design process of new materials offers great prospects to accelerate technological innovation[1–3]. However, the process of virtual materials screening and design depends on the availability of fast and accurate modelling techniques that reliably predict materials properties at realistic operating conditions. Widely used materials simulation methods include highly accurate but computationally demanding ab initio methods[4,5], and fast classical force field-based methods[6–8] that are often limited in their applicability (notably due to fixed covalent topology) and accuracy[9,10] as illustrated in Fig. 1. Machine-learned (ML) potentials constitute a promising approach to solve computationally challenging problems in materials sciences (for example, the simulation of enzymes, chemical reactions or complex materials and interfaces as depicted in the inset in Fig. 1) that were previously neither accessible by classical force fields nor by computationally demanding ab initio molecular dynamics (MD) methods. Behler and Parrinello pioneered the systematic development of ML potentials[11,12], while the idea of fitting force field parameters using ab initio data dates back to earlier work[13–17]. More recent breakthroughs in constructing highly accurate ML potentials for organic molecules[18–20] as well as for disordered and crystalline materials[21–23] demonstrate that ML potentials have the capability to initiate a paradigm shift in materials modelling and design.

In this Review, we outline some commonly used approaches to train ML potentials, including promising recent developments and successful examples of applications. We then propose and discuss auspicious avenues for future developments that have the potential to enable predictive large-scale materials simulations far beyond the scope of currently available classical and ab initio approaches. We will in particular discuss how hybrid machine learning/molecular mechanics (ML/MM) methods will help to approach substantially larger systems[24–26] and longer time scales[27] than currently accessible with state-of-the-art methods (Fig. 1). We will furthermore outline developments required to make ML potentials robust and widely applicable tools, similar to today's classical force fields or density functional theory-based MD simulations. Finally, we will provide our perspective for ML potentials to become a method capable of reliably describing complex systems on an atomistic level, including realistic environmental conditions. As such, ML potentials will be an essential step towards fully computational materials design.

Overview of the history of and recent progress in ML potentials

The origin of ML potentials is based on the idea to use ab initio data to fit free parameters in classical force fields and analytical interatomic interaction potentials, dating back to work by Ercolessi and Adam in 1992 who found that an "extensive data set overcomes the difficulties encountered by traditional fitting approaches when using rich and complex analytic forms, allowing to construct potentials with a degree of accuracy comparable to that obtained by ab initio methods"[15,28]. In 1995, Blank et al. proposed the use of generally applicable neural networks instead of predefined analytical functions[14]. Lorenz et al. continued to use neural networks to fit intermolecular interaction energies[29]. In 2007, Behler and Parrinello introduced high-dimensional symmetry functions as a generalized, system-size-independent approach to represent atomistic configurations: "a NN [neural network] optimized for a certain number of degrees of freedom, i.e., number of atoms, cannot be used to predict energies for a different system size". Thus, they suggested the use of symmetry functions to "describe the energetically relevant local environment of each atom and are subsequently used as input for the NN"[11].

While computational power[30–32], ML models[19,33] and architectures[34–36], as well as the amount of available ab initio training data[37–39], have substantially increased during the past 25 years, the challenges addressed in these early works remain part of current research[19,40]. The need for large amounts of accurate training data, informative representations that are independent of the system size, and accurate ML models remain topics of high interest and active research. Even topics that recently gained increased interest—such as uncertainty quantification[41], as well as active learning approaches[42] to dynamically generate training data—often date back to early work,

---

[1]Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, Ontario, Canada. [2]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [3]Institute of Nanotechnology, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. [4]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. [5]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. [6]Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario, Canada. [7]Present address: Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany. [8]Present address: Institute of Physical Chemistry, Georg-August University, Göttingen, Germany.
✉e-mail: alan@aspuru.com

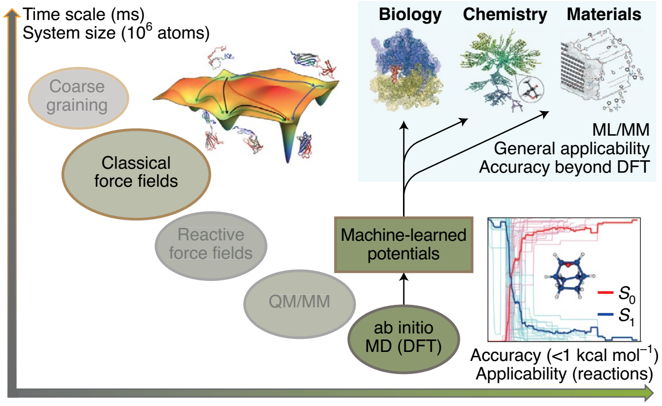


Fig. 1 | Simulation methods. ML potentials as a potential solution to the trade-off between cost and accuracy of conventional atomistic simulations. Potential future developments include hybrid machine learning/molecular mechanics (ML/MM) methods, more efficient representations to decrease simulation times and more accurate training data (proposed by an active learning algorithm) to improve the model accuracy beyond density functional theory. Potential future applications are shown in the blue box (only approximately positioned according to their system-size and accuracy requirements). They include the simulation of enzymes and biomolecules such as ribosomes, the quantitative simulation of chemical reactions and reaction networks as well as the atomistic simulation of complex reactive materials as found in, for example, Li-ion batteries. The inset on the top left shows the energy landscape of a protein folding simulation, which is a prototypical example of a classical force field calculation. No covalent bonds are formed or broken during the simulation. The inset on the right shows an excited state dynamics simulation of a S_1 to S_0 transition, which requires ab initio methods to compute the excited state properties. The 'Coarse graining', 'Reactive force fields', and 'QM/MM' boxes are faded out as these methods are not discussed in depth. Figure adapted with permission from: Marat Yusupov, Roland Beckmann and Anthony Schuller (biology image); ref. [24], American Chemical Society (chemistry image); ref. [25], Elsevier (materials image); ref. [26], PNAS (left inset); ref. [27], AIP (right inset).

for example, by Brown et al. in 1996: "The calculation of error surfaces for interpolants is also useful since it provides an easily interpretable guide to where in the input space the network is uncertain about the interpolant. This allows the choice of subsequent training data to be altered accordingly"[43].

Although many of the fundamental concepts in the field of ML potentials were conceived and described in the literature more than 15 years ago, the computational tools and power for extensive data generation and analysis, as well as the efficient training and application of machine learning models, only recently enabled ground-breaking successes with mature technologies. This initiated the development of a large variety of promising methods to train and apply ML potentials. In this work, we will introduce basic concepts of popular methods to construct ML potentials, notably neural networks and kernel-based regression models, and illustrate how they can be used to predict energies and forces in atomically resolved systems (see the 'Machine learning models and system representations' section). All machine learning methods require data to be trained, with deep neural networks typically requiring large amounts[19]. We therefore discuss promising recent approaches of large scale data acquisition (see the 'Data acquisition' section). Many ML methods do not necessarily generalize well to data points far beyond the training data, which can limit the reliability of machine learning-based simulations. Uncertainty quantification and active learning approaches (see the 'Uncertainty quantification and active learning' section) alleviate this problem by enabling automated model correction and prediction accuracy improvement via selective training data acquisition. We will finally highlight applications of ML potentials to study both organic and inorganic materials (see the 'Applications' section).

Machine learning models and system representations. In the following, we will introduce the main challenges of applying machine learning models to atomistically resolved systems, and present solutions in the form of neural network- and kernel-based machine learning models, along with widely used system representations.

The challenge. Machine learning models are capable of building mathematical models for pattern identification and inference on numerically represented data. ML potentials are specifically designed to find a mapping between a three-dimensional configuration of an arbitrary number of atoms representing a material of interest and its corresponding conformational energy. One of the biggest challenges in designing ML potentials is the identification of a suitable numerical representation of the material. This challenge is two-fold: materials systems of arbitrary size (number of atoms N) have to be represented by vectors of fixed dimension[44,45] while the complete information about the three-dimensional configuration (generally $3N - 6$ rotationally and translationally invariant internal degrees of freedom) has to be encoded in this vector. Further constraints on the representation can be posed by additional symmetries of the system such as periodicity in the case of crystalline materials[46,47]. The choice of regression model used for the prediction of properties (energies and gradients) is typically tightly connected to the choice of representation.

Many successful examples of ML methods applied to questions in materials sciences and chemistry only require a solution to one of the two challenges. Properties of molecular equilibrium structures[48,49] can be accurately predicted without accounting for all conformational degrees of freedom while molecule-specific approaches[39,50–52] do not need to account for variations in the system size. In the following section, we will give a non-exhaustive overview of a selection of representations and corresponding regression models that address both of these challenges and thus constitute promising candidates for future materials modelling and virtual materials design. Box 1 explains the basic principles of neural network- and kernel-based ML potentials and Box 2 summarizes state-of-the-art ML potentials and their respective representations and model architectures.

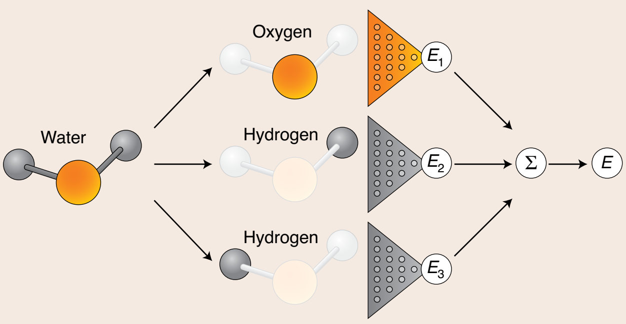
ML potentials based on neural networks. Neural networks constitute a different class of machine learning algorithms that are loosely inspired by the human brain. Organized as a directed graph of trainable units (neurons), neural networks can learn highly nonlinear relations between input data and output data. The machine learning models and system representations shown in Fig. 2a,b are designed to describe local contributions to the total energy of a system.

One of the most widely used approaches to generate input data for the training of neural network potentials is to define an informative representation of an atomic system of arbitrary size by using symmetry functions initially developed by Behler and Parrinello[11,36]. Instead of finding representations encoding molecules on a global scale, the symmetry functions describe the local chemical and geometrical environment of every single atom independently (see Fig. 2a and Box 1). While being inherently translationally invariant, the symmetry functions are defined in a rotationally invariant manner[53]. The total energy of an N-atom system is predicted by summing up the energy contributions of every single atom predicted by element-type specific machine learning models (Fig. 2b)[11,18,40,54], leading to a linear scaling of the computational cost with the system size[55].

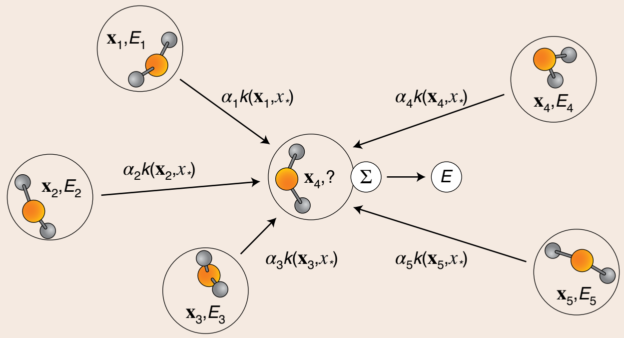
Box 1 | Neural network potentials and kernel methods

Neural network-based potentials predict the total energies of molecules using a summation of the predictions of multiple neural networks for each atom in the molecule. Each neural network is trained on representations of the local environment of an atom, called symmetry functions[11,18,36,40]. Separate element-specific neural networks are trained, each consisting of multiple fully connected layers that typically decrease in size from the input layer to the final prediction.

Kernel-based models such as Gaussian processes predict the expected value of the energy of a new data point x_*, that is, a new conformation or molecule as a linear combination of coefficients $\alpha = (\mathrm{K} + \sigma_n^2 I)^{-1}\mathbf{y}$, with covariance matrix K with $K_{ij} = k(x_i, x_j)$, training points x_i, target vector $\mathbf{y}$ (here, its elements are energies, E_i) and kernel function $k(x, x_*)$[152]. The SOAP kernel[74] provides a computationally efficient similarity measure between local atomic environments. It integrates over the local environment of two different atoms, where each atom in the environment is represented as a Gaussian. Two integrations are performed, one over space to obtain the overlap between the two environments and one over all rotations of the local environment to make the kernel rotationally invariant.



Neural network-based potentials. The system is split into single atom environments for which element-specific neural networks predict energies that are summed up to the total energy of the system.



Kernel-based potentials. The energy is predicted by using a kernel function to compare to all other training samples and to interpolate the energy as a weighted sum of contributions of all known data points.

Even though most neural network potentials focus on short-range interactions, long-range interactions pose a major computational challenge in classical MD simulations, where long-range electrostatic interactions are calculated using static point charges that are assigned pre-specified points within molecules, typically to each atom position. To improve this description[56], machine learning can be used to predict environment-dependent point charges or molecular dipole/multipole moments[57–59]. The neural network-based TensorMol model[34] combines local atomic energy contributions with long-range electrostatics as well as van der Waals energies[60] to improve the prediction of potential energy surfaces and enable the simulation of large systems (Box 2). Point charges and multipole moments can furthermore be used to computationally predict infrared spectra and weak intermolecular interaction energies in organic crystals[61–63].

A rather novel route to machine learning-based atomistic potentials in general and long range-corrected potentials in particular has been proposed by Bereau and colleagues[64]. Instead of employing neural networks (or other machine learning models) to construct atomistic potentials, they adhere to the idea of using physics-based potentials. All local, environment-dependent atomic properties of these potentials (for example, electrostatic multipole coefficients or polarizabilities), however, are machine-learned. These data-driven intermolecular potentials from a combination of physics- and machine learning-based models (IPMLs) enable the accurate calculation of noncovalent interactions including many-body dispersion. They were shown to be transferable across conformational and chemical space (hydrogen, carbon, nitrogen and oxygen atoms) without explicit prior parametrization.

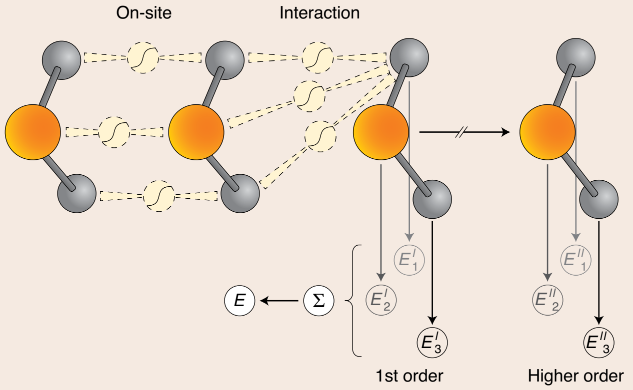
ML potentials based on kernel methods. One widely used group of regression models for ML potentials are kernel-based methods[47,65,66], in particular Gaussian processes (GPs; see Fig. 2c and Box 1)[67–71]. Kernel methods constitute machine learning algorithms typically used for pattern recognition. They use the 'kernel trick' to express similarity in potentially higher-dimensional spaces (using inner products) derived from the set of input features to identify nonlinear decision boundaries with linear learning algorithms. They are typically easier to train and have fewer hyperparameters than (deep) neural networks. However, training time and prediction time scale with M^3 and M^2, respectively, with M being the number of training samples, as opposed to the more favourable scaling of neural network models that were explained in more detail earlier. This adverse computational scaling presents an obstacle to covering large amounts of training data and diverse parts of chemical space with a single GP. Sparse GP models reduce this unfavourable scaling to MX^2, with X being the number of inducing points that have to be selected in advance, balancing the reduction in computational demand with a loss in accuracy[72,73]. A widely used regression model, in particular for inorganic materials, is the smooth overlap of atomic positions (SOAP) kernel[74] to train the Gaussian approximation potential (GAP)[68,75]. The SOAP-GAP model and similar methods[76] can furthermore be used to compare environment-embedded atoms in molecules and solids[76,77] as well as to predict molecular properties[78].

Discussion. Symmetry functions-based representations are typically high-dimensional, which increases the number of parameters as well as the training and prediction times of the associated regression models. Dimensionality reduction of the representation by systematic selection of features (individual dimensions of the representation) can reduce training and simulation time substantially. Feature selection methods range from simple approaches employing the Pearson correlation coefficient as selection criterion[79] to more advanced techniques such as CUR decomposition[80], a modified version of farthest-point sampling, or automatic relevance determination[81], a Bayesian method that estimates the relevance of features by optimizing the marginal likelihood of the underlying model. Some of these methods can also be used to reduce the number of training points, which is of crucial importance to accelerate the training and evaluation of kernel-based methods. Further methods to systematically reduce the required training set size are based on

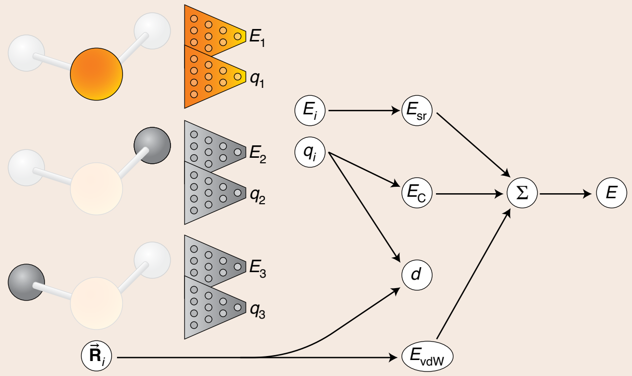
Box 2 | State of the art ML potentials

Three recently developed approaches to predict total energies of molecules with arbitrary size and geometry are the deep tensor neural network (DTNN)[38,153], the closely related SchNet approach[151] and the hierarchically interacting particle neural network (HIP-NN) model[35,154]. These models represent each atom as a vector of fixed size, initially encoding only its atomic number. The representation vectors are propagated through the model in a sequence of on-site and interaction layers. On-site layers are fully connected neural network layers, while the interaction layers combine the representations of all atoms using distance-dependent, nonlinear transformations. The total energy is obtained by summing over atomic energy contributions, each of which is computed from the respective atom representation based on linear regression. In the case of HIP-NN, this procedure is repeated several times at different depths of the entire model, leading to a hierarchy of energy contributions at different effective interaction distances.

Long-range electrostatics and van der Waals corrections can be modelled explicitly and added to the energy. In the TensorMol model[34], for example, neural networks are trained to predict point charges q_i of molecules that reproduce the dipole moment d of the molecules. In a second step, the Coulomb interaction E_C between these point charges and van der Waals energy E_{vdW} are calculated and neural networks are trained that predict the short-range energy contribution E_{sr} in a way that all three energy terms sum up to the total energy E of the system.



Interaction networks. Representations are propagated through the neural network and energies are predicted as sums of single-atom energies at multiple depths of the model.



Long-range interactions. Local interaction energies are predicted by the neural network directly, while long-range interactions are modelled using physics-based Coulomb interactions based on partial charges predicted by the neural network as well as parameterized van der Waals interactions.

active learning strategies and will be discussed in the 'Uncertainty quantification and active learning' section.

It was shown that the incorporation of forces acting on each atom (gradients of the potential energy surface) in addition to energies in the learning process can increase the accuracy of models by reducing artefacts due to inconsistencies of energies and gradients[34]. One example of an intrinsically energy-conserving method is the (symmetric) gradient domain machine learning ((s)GDML) technique[39,82]. GDML is a kernel-based method that directly predicts the forces acting on each atom. Using a molecule-specific representation, it has to be re-parameterized for each new molecular species. However, it already reaches high accuracy when trained on a relatively small number of data points.

Data acquisition. Supervised machine learning models for ML potentials require training data that generally consists of the coordinates and elements of all atoms in a molecule or unit cell and their corresponding energies and forces. ML potentials target the accurate prediction of energies and forces for arbitrary conformations of atoms that are sampled during an MD simulation. In contrast to many other machine learning tasks in materials and molecular sciences[49,83–85], off-equilibrium conformations are therefore of crucial importance when training ML potentials. Multiple approaches to systematically generate a diverse set of off-equilibrium conformations are described in the literature, some of which are outlined in detail in this section. Afterwards, attempts to systematically reduce the amount of data required to train models to a given target accuracy will be discussed.

Apart from a random sampling of the configuration space (Fig. 3a), one of the most straightforward but not necessarily the most efficient approach is the extraction of data from existing ab initio MD trajectories (Fig. 3b)[45]. Structural resemblance of consecutive timeframes, as well as long-time correlations, will yield an unbalanced dataset that primarily samples from selected parts of the conformational space while many (high-energy) conformations and disconnected local minima are never observed. However, the procedure can be partially improved by sampling from trajectories at higher temperatures to enhance coverage of high-energy conformations during the simulation[45].

The approach of data acquisition described by Smith et al. for the training of the ANI1[18,37] model consists of a three-step process. In the first step, equilibrium geometries of a large number of small molecules are obtained using density functional theory (DFT)-based geometry optimizations. In the second step, eigenmodes are calculated and the equilibrium conformations are distorted along linear combinations of the eigenmodes, scaled with temperature- and frequency-dependent factors. In the last step, the energies of all conformers are evaluated using DFT calculations (Fig. 3c). Using this procedure, about 22 million conformers of 57,000 molecules were generated and used to train the ANI1 potential[34,37].

Kernel-based methods—often used to model inorganic systems with one or a few different elements[68,75,86]—can achieve high accuracies when trained on only a few thousand data points. Deringer et al. and other groups developed schemes that start without any prior knowledge of the system by running ab initio calculations on simulation boxes with randomly distributed atoms. Active learning

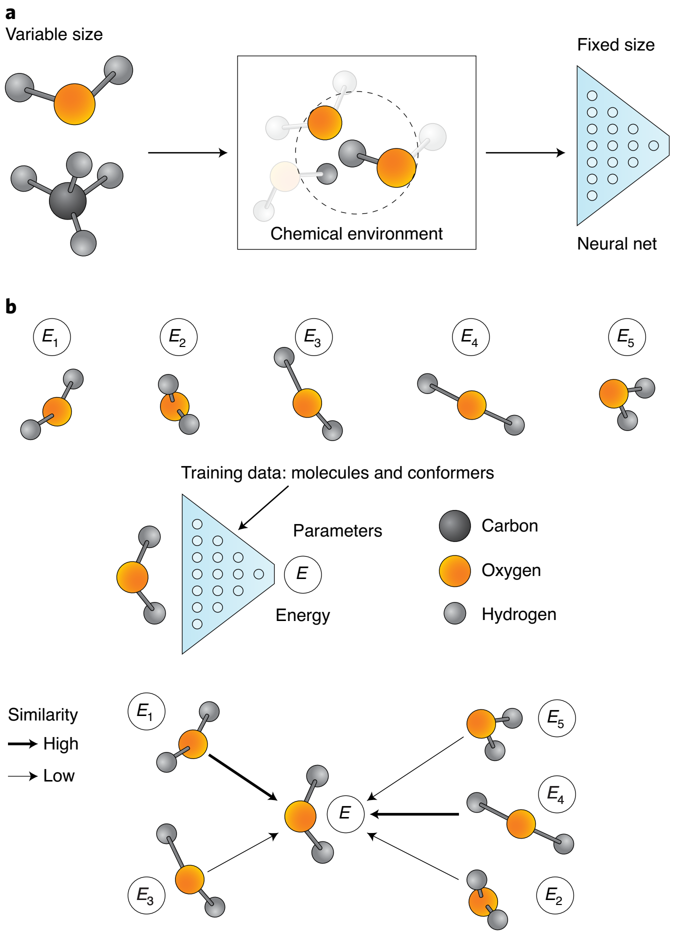


Fig. 2 | Representation and models. a, The challenge of finding a fixed-size representation of a system of arbitrary size can be solved using representations of atom centred local chemical environments. **b**, Neural network-based ML potentials are trained by optimizing the parameters given the training data (multiple conformations of multiple molecules and corresponding energies *E*). Predictions are done using a fixed set of parameters. In Behler et al.[11] and Smith et al.[18], different neural networks are trained for the chemical elements. The total energy *E* of a system is computed as the sum of the neural network predictions for each atom. **c**, Kernel-based methods predict the properties of a molecule by comparing its representation with the representations of all molecules in the training set. The kernel estimates the similarity between the molecules and uses it to weight the training labels (energies).

procedures as described in the next section are used to refine the dataset by iterating between model training, MD simulation and systematic acquisition of further ab initio data points[21,87–89].

Uncertainty quantification and active learning. Rendering materials design a successful endeavour requires transferable machine learning models. To achieve this goal, the training set must represent the important features of the problem-specific configuration/compound space (application domain), which usually is unknown at the outset. Hence, in an optimal learning process, training data would be generated successively (Fig. 4a). While the accuracy of data-driven models can be (most often) improved by increasing the amount of training data, their complex, time-consuming and/or expensive generation constitutes a major bottleneck. This limitation incentivizes the generation of informative training sets that contain a minimum number of examples to achieve a predefined accuracy on a given application domain. In active learning approaches[42], predictions inferred from a given training set are harnessed to determine a minimal set of new systems (for example, molecular configurations) for which target values (labels) are required. The training set will be extended by these new labelled data points on the fly to optimally improve the underlying prediction model for those systems that are not yet contained in the training set. Active learning is particularly applicable to data-extensive scenarios such as chemical-space exploration or MD simulations, which yield constant data streams.

A simple but effective data selection algorithm was proposed by Botu and Ramprasad[90]. They used the Euclidean distance between configurational fingerprints to actively learn energies and forces for MD simulations of bulk aluminium and aluminium surfaces. If the inter-configurational distance was too large, electronic-structure calculations were performed to update the ML force field. The more sophisticated 'query by committee' algorithm[42] (Fig. 4b) was applied first in the context of ML potentials by Behler[36], although Smith et al.[91] were the first to introduce the term to the community. Herein, a committee (ensemble) of neural-network models with different architectures but comparable average performance is selected. One member of the committee is employed as a force field in an MD simulation. While the simulation is executed, each committee member predicts the potential energy of every configuration along the resulting trajectory. If a large disagreement between the committee members is detected—indicated by a high prediction variance—the corresponding configuration will be subjected to an electronic-structure calculation. The resulting labelled data point is then used to update the committee models.

While query by committee methods require training of multiple models to estimate the prediction variance for a molecular configuration, the same can be achieved by a single model with built-in uncertainty quantification (such as Bayesian neural networks or GPs), possibly leading to substantial computational savings (Fig. 4c). In contrast to conventional neural networks, Bayesian neural networks incorporate distributions (instead of single values) of their weights and biases, which are propagated through the network to yield target values equipped with uncertainty measures (square root of the prediction variance)[92]. GPs go even beyond this idea and infer uncertainty from training data alone, that is, without the need to define any functions explicitly. Simm and Reiher[24] used the prediction variance of a GP to learn complex chemical reaction mechanisms. Their active learning workflow is sequential, allowing for only one label (for example, an ab initio electronic energy) at a time to be generated. Proppe et al.[93] extended this approach to enable the generation of multiple labels simultaneously, which offers a striking acceleration of matter simulations.

Instead of providing the active learner with a stream of unlabelled data, it can be made even more active so as to synthesize/propose unlabelled data itself. This approach is useful, for example, to accelerate the exploration of potential energy surfaces by synthesizing representations of unseen configurations. To be useful for materials science, one needs to ensure that the synthesized representations correspond to reasonable conformations, unit cells and so on. Variational autoencoders were developed in our lab for this purpose[94]. Similarity measures used in kernel-based regression models can further help to confine the search for new configurations to regions that are vicinal to those that have already been sampled. Synthesizing unlabelled data is also appealing for optimization tasks. Instead of aiming at exploring the application domain evenly, the active learner can be made selective so as to focus on particularly interesting subdomains. For instance, identifying the local minima of a potential energy surface is of crucial importance. In this case, one could introduce a heuristic that rewards synthesizing representations that correspond to supposedly low-energy configurations. Bayesian optimization is closely related to this approach

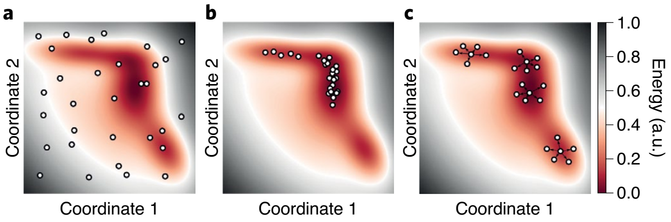


Fig. 3 | Data acquisition. **a**, Random sampling of the high-dimensional configuration space will overestimate the importance of high-energy points and might miss local optima. **b**, Sampling along (low-energy) MD trajectories samples with a high density around local optima but might miss important parts of the configuration space. **c**, Sampling around local optima (for example, ANI1 data acquisition scheme) achieves homogeneous sampling around all local optima.

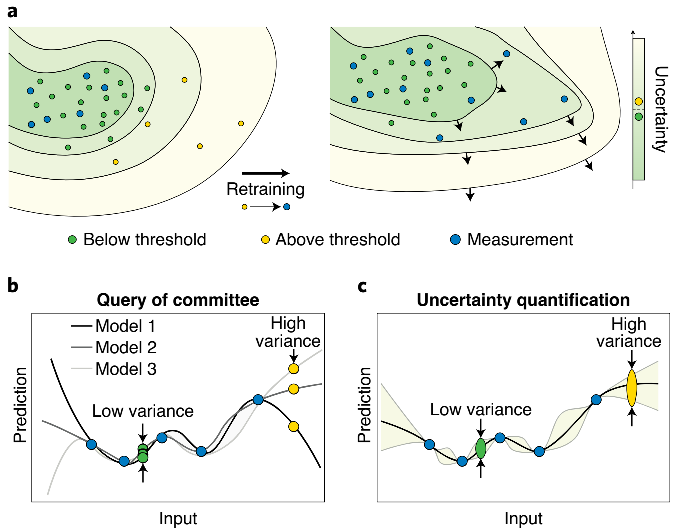


Fig. 4 | Active learning. **a**, Active learning scheme where the uncertainty of predictions is used to decide if the new reference calculations and a retraining of the ML potential is required. The points denote configurations of atoms that are encountered during a MD simulation. Green points are within acceptable uncertainty whereas yellow points are chosen for retraining of the model (denoted by blue points), which will increase the configuration space of low model uncertainty, illustrated by the expansion of the isosurfaces of model uncertainty. **b**, Query by committee method to estimate the uncertainty of a model without implicit uncertainty estimations. Multiple models (lines) are trained and their (dis)agreement is used as a measure of uncertainty. **c**, Models with intrinsic uncertainty quantification (for example, Bayesian neural networks or GP models) can directly be used to probe the reliability of predictions and in case of high uncertainty retrain the model.

and has been successfully demonstrated for problems in chemistry including multi-objective formulations[95,96].

Active learning approaches have furthermore been applied to accelerate chemical-space exploration (based on Thompson sampling[97]), to identify suitable collective variables (based on reinforcement learning[98]), as well as for the prediction of molecular properties[99] (for example, atomization energies, polarizabilities) and interatomic potentials[100] (based on the D-optimality criterion).

Applications. The very generic mapping from an arbitrary three-dimensional configuration of atoms to energies and forces acting on each atom makes classical force fields, ab initio calculations and also ML potentials a versatile method to drive dynamics simulations of practically any material, ranging from small organic molecules and large biomolecules over disordered (gaseous, liquid or amorphous solid) materials to (complex) crystalline systems such as metal–organic frameworks. The development of ML potentials that are as accurate and reliable but much faster than the ab initio calculations on which they were trained (Fig. 1), will enable atomistic simulations of large systems over large time scales and, thus, lead to a further boost in the endeavour of virtualizing materials design. Multiple promising examples of successful applications of ML potentials to a variety of material classes exist, some of which are illustrated in Fig. 5. Most of them are based on the models and representations presented earlier.

Inorganic materials, both disordered and crystalline, offer a wide variety of open research questions, where ML potentials can be readily applied to better understand the mechanical properties of the materials, predict surface reconstruction and surface properties (Fig. 5a)[101–103], simulate phase diagrams[104], find new polymorphs or crystal structures (Fig. 5b)[22,44,45,68,105], simulate the impact and propagation of lattice defects with atomistic resolution[75,106], simulate nanoparticles (Fig. 5c)[107], liquids[52], or study properties of complex (for example, mixed organic–inorganic) systems such as metal–organic frameworks (Fig. 5d)[108]. The accuracy of ML potentials combined with active learning approaches, in particular for high-energy configurations, enables the reliable simulation of various phases of single-component materials such as carbon[68,109], silicon[78,87,104] or boron[89], as well as more complex materials such as GeTe (ref. [110]), GeSbTe (ref. [111]) or CuAu (ref. [112]). Multiple research groups in particular showed that the use of ML potentials for inorganic materials not only helps to better understand microscopic processes (for example, crack propagation, formation of polymorphs under certain conditions, and so on) but can enable the in silico discovery of unknown microstructures and polymorphs of materials, interfaces and nanoparticles.

While for many inorganic applications, material-specific machine learning models only need to describe the interaction between a small number of chemical elements, even the most basic ML potentials for organic materials need to cover carbon, oxygen, nitrogen, fluorine and hydrogen simultaneously in one model (Fig. 5e)[18,113]. In addition, sulfur and phosphorus have to be included to cover the space of amino acids, nucleotides and other biomolecules (Fig. 5f)[114], while most simulations of chemical reactions will include halogens and metal/heavy atom-based catalysts. This poses a particular challenge to many of the previously described approaches because the amount of required training data, as well as the size of the input representation and thus the size of the machine learning models, often scale with the number of chemical elements[40]. However, the simulation of reaction dynamics and the simulation of large biological systems are two of the most promising applications of ML potentials, where the high accuracy of ab initio methods is required but not affordable at needed time and length scales[115].

The simulation of chemical reactions and the search for transition states is a prototypical task that is not solvable with classical force fields due to their missing ability to break or form bonds. A variety of reactive generalizations of classical force fields are being developed but it is beyond the scope of this Review to discuss their advantages and weaknesses[10,116]. Ab initio-based simulations of reactions have been demonstrated on selected systems but can become computationally intractable, notably for large system sizes, long time scales and, particularly relevant to materials design, when a large number of systems needs to be analysed[117,118]. Computationally less costly semi-empirical and tight-binding methods[119–121], as well as quantum mechanics/molecular mechanics (QM/MM) models[115,122–124], are a promising approach to speed up calculations that we will discuss later on in the context of ML potentials.

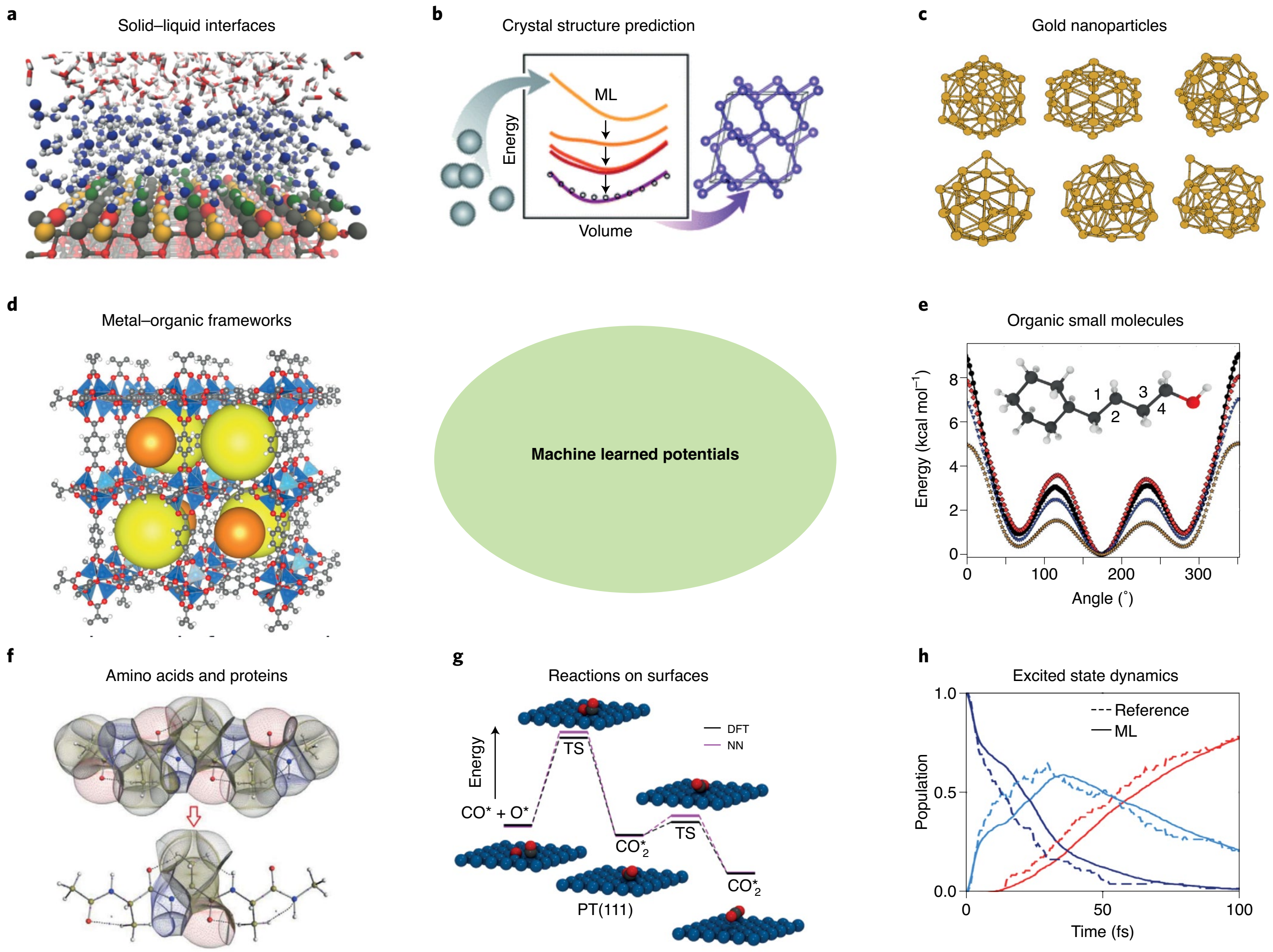


Fig. 5 | Applications. a–h, Application of ML potentials to ZnO–water interfaces (**a**); crystal structure prediction (**b**); gold nanoparticles (**c**); metal–organic frameworks (**d**); MD simulations of small organic molecules (**e**); amino acids, polypeptides and proteins (**f**); reactions on surfaces (**g**); and excited state dynamics of small organic molecules (**h**). Figure adapted with permission from: ref. [102], American Chemical Society (**a**); ref. [22], RSC (**b**); ref. [107], AIP (**c**); ref. [108], American Chemical Society (**d**); ref. [18], RSC (**e**); ref. [114], Wiley (**f**); ref. [132], American Chemical Society (**g**); ref. [51], RSC (**h**).

Contrary to classical force fields, ML potentials do not distinguish between bonded and non-bonded interactions and thus are intrinsically capable of simulating chemical reactions. Therefore, ML potentials were successfully applied in transition state search[125], reactions between molecules[126–129], organic reactions[130], simulation of surface scattering[101,131], as well as heterogeneous catalysis and surface reactions (Fig. 5g)[102,103,132,133]. The high-dimensional conformation space combined with potentially high model uncertainties[134] when extrapolating beyond the training space require particular precaution when selecting training data (see the active learning approaches introduced in the previous section).

Fast and accurate ML potentials can be used to implement reliable and fast transition state search algorithms[71] that in the future might enable automated workflows for the design of novel catalysts for heterogeneous and homogeneous catalysis, including the optimization of substrate-dependent catalysts for industrially relevant reactions as well as the development of cheaper metal–organic catalysts or organocatalysts with similar performance as state-of-the-art second-and third-row transition metal-based catalysts. At the same time, ML potentials for the simulation of biological processes involving proteins and enzymes can potentially act as a computational microscope to better understand the function of biomolecules (for example, metabolic processes or trans-membrane transport) as well as their interaction with cofactors.

Figure 5h finally shows one example of using ML potentials to run (long time scale) excited state dynamics or photodynamics simulations[135] using a surface hopping approach. Several research groups have shown how multiple potential energy surfaces of a small organic molecule can be predicted using molecule specific ML potentials[51,136,137]. While Westermayr et al. predict non-adiabatic couplings and spin–orbit couplings close to conical intersections[51], Hu et al. switch from ML potentials to ab initio methods in the vicinity of conical intersections[137]. To our knowledge, a more universal, beyond molecule-specific approach to predict excited-state potential energy surfaces is yet to be demonstrated. Further machine learning-based approaches were applied to the modelling of excitation energy transfer[138] and proton-transfer reactions[139].

Perspective on how ML potentials can enable reliable in silico experiments and virtualize materials design

Despite promising developments in the field of ML potentials during the past few years, they are not meant to replace classical force fields and ab initio methods, but to complement them in order

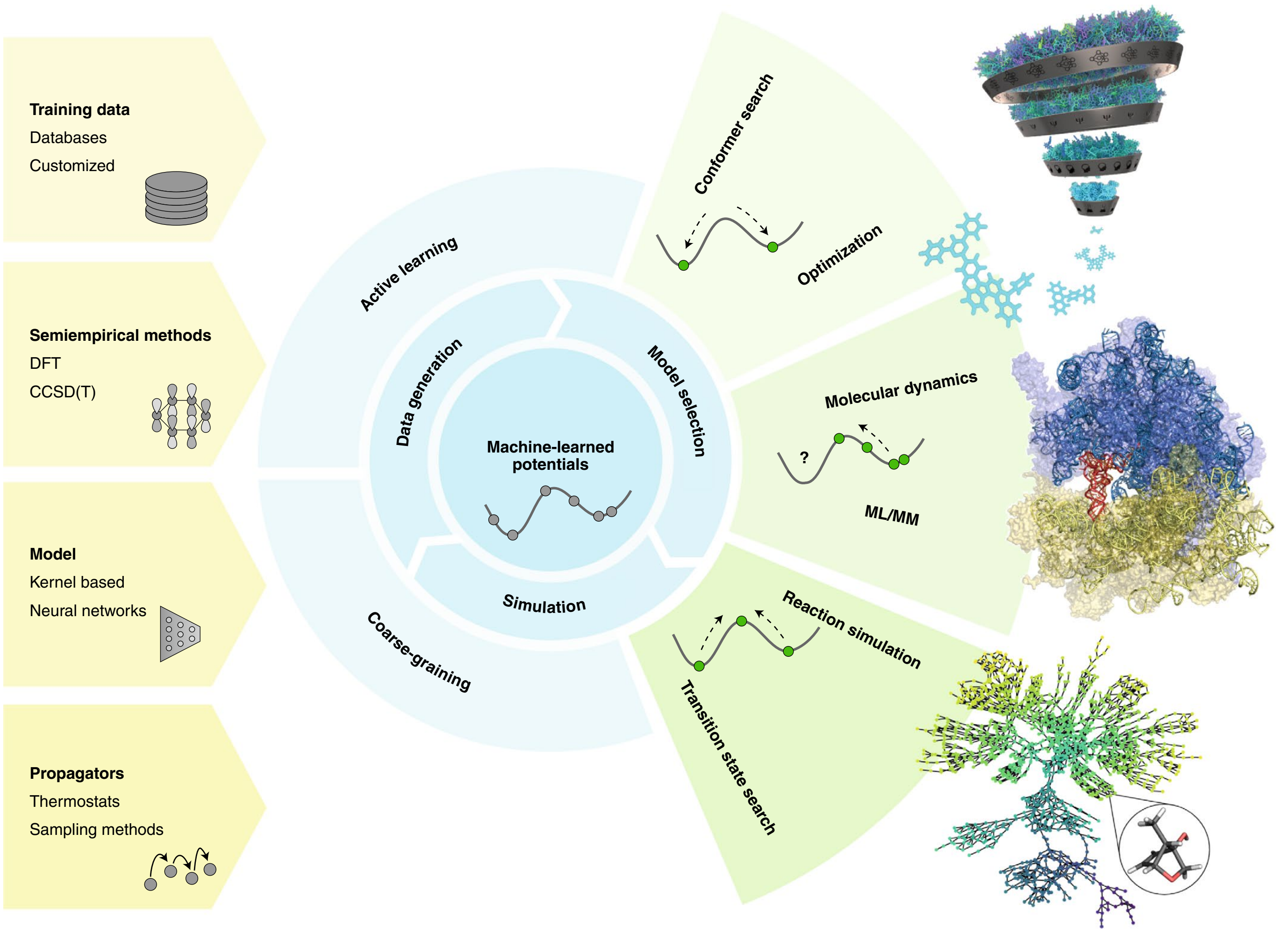


Fig. 6 | ML potentials as a general-purpose tool. We envision the development of ML potential packages that incorporate the entire workflow of data selection or generation of customized training data, model selection or training, simulation and uncertainty quantification, active learning approaches and application of the ML potential to a variety of tasks. Potential applications include high-throughput screening approaches for molecule design, MD simulations (for example, ribosomes), and automated exploration of chemical reaction networks. Figure adapted with permission from: Lauren A. Kate (top right); Marat Yusupov, Roland Beckmann and Anthony Schuller (middle right); ref. [24], American Chemical Society (bottom right).

to build faster and more accurate models to understand and design the materials of the future. In the following section, we will discuss ways to further improve the availability and usability of ML potentials to make them more accessible to researchers working in the area of materials simulations (see the 'From an expert method to a robust and widely applicable tool' section). We will furthermore discuss potential pathways to accelerate simulations with ML potentials to make them applicable to larger system sizes and simulation times (see the 'From microscopic (nm/ns) to macroscopic systems (μm/μs)' section). Finally, we will outline the substantial impact that ML potentials can have on the simulation of chemical reactions and catalytic systems. We will discuss recent successes in training ML potentials with accuracies beyond DFT and how to combine them with active learning approaches to address the large configuration space that has to be described when searching for transition states and simulating chemical reactions in general (see the 'ML potentials beyond DFT accuracy for reaction simulations' section).

From an expert method to a robust and widely applicable tool. Computational materials design inherently requires methods and software tools that can explore the chemical space of molecules and materials in a robust and reliable way. Robustness refers to the practicality and user-friendliness of the software tools independent of the given hardware and software, while reliability refers to the accuracy of the results for a wide range of materials and simulation workflows. Thus, robustness and reliability are two important requirements that need to be satisfied by all simulation tools, including ML potentials, that allow their broad application to answer scientific questions and eventually enable computational materials design. Active learning approaches that incorporate variance estimation[140] and automatically generate training data while exploring chemical space[95] are promising for this purpose.

Many state-of-the-art ML potentials are developed for specific material systems and thus limited in their applicability. The generalization and automation of training protocols will enhance the applicability of ML potentials by providing automated out-of-the-box methods to train and use ML potentials for arbitrary systems[141]. There are two main pathways to help ML potentials become tools that are generally applicable to many classes of materials and molecules. One option is the development of datasets and models that intrinsically cover a wide range of application areas (for example, the ANI-1 potential applicable to organic molecules, similar to many widely applicable classical force fields). Another option is the development of methods that generate (initial) training data in an automated way, combined with easy-to-train expert models that involve active learning procedures to reliably explore chemical space.

We suggest the development of programs similar to current DFT and electronic structure packages that include user-friendly interfaces to implement the following key features (Fig. 6): First, the packages should include pre-trained models such as ANI1 or TensorMol 0.1 as well as kernel-based models including their training data, similar to electronic structure packages that implement a variety of DFT functionals or post-Hartree–Fock methods. The package has to be able to read a standardized input format for the atomistic system that the user wants to simulate and a number of simulation types (geometry optimization, MD simulation and so on), similar to state-of-the-art MD packages. Beyond that, it would be possible to interface such a package with electronic structure methods as well as ML frameworks for (re)training of the implemented models. This will enable the user-friendly execution of active learning procedures as well as fully automated generation of models for new classes of materials, which is of particular interest for kernel-based methods due to the low data requirements. The full package is illustrated in Fig. 6. All required interfaces to databases, external electronic structure packages, machine learning models as well as MD protocols are shown on the left; the core of the ML potential package including workflows for data selection/generation model selection/training and simulation including uncertainty quantification are shown in the middle, potentially extensible with active learning models and coarse graining protocols; potential applications of the ML potential package are illustrated on the right, including high-throughput virtual screening and materials design approaches that require conformer search and geometry optimization of millions of molecules, 'classical' MD and ML/MM simulations and the simulation of chemical simulations and transition state search, both for organic and inorganic chemistry and catalysis.

Seminal examples of such efforts include the integration of the high-dimensional neural network potentials in LAMMPS[142], the TensorMol 0.1 model available as a Python package[34]. The ANI1[18] model as well as SchNetPack[143] and sGDML (http://quantum-machine.org/gdml) provide interfaces to the atomic simulation environment (ASE). The libAtoms website (http://libatoms.org) furthermore provides a variety of datasets and models for the simulation of inorganic materials that can be extended to be used for materials modelling and design.

From microscopic (nm/ns) to macroscopic systems (µm/µs). It has been shown that ML potentials offer the possibility to substantially increase system size and time scales that can be reached with ab initio MD methods. However, currently available representations and models are still computationally more demanding than typical classical MD simulations. To further extend current limitations, we propose the development of hybrid ML/MM methods. In contrast to QM/MM simulations, the similarity of ML and MM potentials allows for a tight integration of both methods with smooth and variable transition regions. Initial efforts in combining classical force fields and machine learning methods[144] and ML/MM simulations[145], as well as approaches to use machine learning to parameterize coarse-grained representations[146] were shown in literature.

ML/MM schemes have the potential to substantially accelerate state-of-the-art QM/MM simulations of, for example, large biomolecules and enzymatic reactions that require accuracy beyond classical force fields only in small parts of the system. While the interface of QM and MM regions is challenging to define in a fully automated way, the description of the interface between machine learning and MM regions is more straightforward. Most ML potentials predict single-atom energy contributions for each atom based on its local chemical environment. Even gradual transitions from the machine learning to the MM region is possible, which only requires to map bonded interactions of the classical force field to single atoms and mix them with the ML energy[145]. Non-bonded interactions, in particular the attractive part of the Lennard–Jones potential, as well as long-range electrostatic interactions are not intrinsically part of the ML potential (Box 1) and remain part of the MM scheme.

Further acceleration of ML potentials-based MD simulations can be achieved by lowering the dimensionality of the atomic representations of the local chemical environment and overcoming the limitations that arise due to the adverse scaling with the number of elements[40] described by the ML potentials[35,147].

Finally, machine learning models can also be used to enhance the sampling of MD simulations by learning bias potentials[147] and thus enable more efficient exploration of the configuration space based on data analysis[148].

ML potentials beyond DFT accuracy for reaction simulations. Reaction simulations and transition state searches[71,149], as well as the simulation of excited-state potential energy surfaces, are two research areas where ab initio methods such as DFT, time-dependent DFT and beyond are required to achieve desired accuracy[82]. The underlying challenge is many-fold: the configuration space that has to be accurately described in reaction simulations or excited states simulations is substantially larger than the configuration space encountered at close-to-equilibrium simulations, which requires predictive models that extrapolate well into unknown regions of the potential energy surface, a task well-suited for an active learning algorithm. At the same time, the accuracy on which these configurations have to be described often goes beyond the accuracy of widely used density functionals. Both of these challenges make the development of ML potentials for reaction simulations and excited-state potential energy surfaces a challenging task. However, the prospect of reducing the computational cost of such simulations in enormous. Having reliable ML potentials for reaction simulations will enable the in silico design and optimization of chemical reactions, including heterogeneous and homogeneous catalysis as well as enzymatic reactions. A potential solution to the reliability challenge are hybrid approaches based on active learning schemes where training data is generated on the fly whenever needed (see the 'Uncertainty quantification and active learning' section). Furthermore, it has been shown that ML potentials can be trained on reference data beyond DFT accuracy in a transfer learning approach where a first model is trained on a large amount of comparably inexpensive data followed by a (partial) refinement of the model using a smaller amount of more accurate data. One example is the ANI-1ccx model, which approaches CCSD(T)/CBS accuracy on benchmarks for thermochemistry, isomerization reactions and torsion potentials of small drug-like molecules[150].

Outlook

Design and optimization of new molecules and materials increasingly depends on the availability of fast and reliable computational tools. ML potentials might revolutionize materials simulations by substantially decreasing the computational cost of MD simulations while reaching the accuracy and wide applicability of ab initio computations. Recent developments in model architectures and training data generation including active learning techniques enabled successful applications of ML potentials to a variety of materials classes. We outlined a roadmap to further increase the usability and applicability of ML potentials to mature into a widely used tool for materials simulation and design.

Received: 21 November 2019; Accepted: 17 July 2020;
Published online: 27 May 2021

References

1. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
2. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
3. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
4. Car, R. & Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55**, 2471–2474 (1985).
5. Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A. & Joannopoulos, J. D. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.* **64**, 1045–1097 (1992).
6. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.* **9**, 646–652 (2002).
7. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
8. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
9. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–49 (2011).
10. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A. ReaxFF: a reactive force field for hydrocarbons. *J. Phys. Chem. A* **105**, 9396–9409 (2001).
11. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
12. Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).
13. Izvekov, S., Parrinello, M., Burnham, C. J. & Voth, G. A. Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: a new method for force-matching. *J. Chem. Phys.* **120**, 10896–10913 (2004).
14. Blank, T. B., Brown, S. D., Calhoun, A. W. & Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **103**, 4129–4137 (1995).
15. Ercolessi, F. & Adams, J. B. Interatomic potentials from first-principles calculations. *MRS Proc.* **291**, 31 (1992).
16. Ischtwan, J. & Collins, M. A. Molecular potential energy surfaces by interpolation. *J. Chem. Phys.* **100**, 8080–8088 (1994).
17. Ho, T. & Rabitz, H. A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations. *J. Chem. Phys.* **104**, 2584–2597 (1996).
18. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
19. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
20. Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys. Condens. Matter* **26**, 183001 (2014).
21. Bernstein, N. et al. Quantifying chemical structure and machine-learned atomic energies in amorphous and liquid silicon. *Angew. Chem. Int. Ed. Engl.* **131**, 7131–7135 (2019).
22. Deringer, V. L., Proserpio, D. M., Csányi, G. & Pickard, C. J. Data-driven learning and prediction of inorganic crystal structures. *Faraday Discuss.* **211**, 45–59 (2018).
23. Geiger, P. & Dellago, C. Neural networks for local structure detection in polymorphic systems. *J. Chem. Phys.* **139**, 164105 (2013).
24. Simm, G. N. & Reiher, M. Error-controlled exploration of chemical reaction networks with Gaussian processes. *J. Chem. Theory Comput.* **14**, 5238–5248 (2018).
25. An, S. J. et al. The state of understanding of the lithium-ion-battery graphite solid electrolyte interphase (SEI) and its relationship to formation cycling. *Carbon* **105**, 52–76 (2016).
26. Reddy, G., Liu, Z. & Thirumalai, D. Denaturant-dependent folding of GFP. *Proc. Natl Acad. Sci. USA* **109**, 17832–17838 (2012).
27. Shu, Y. & Levine, B. G. Communication: non-radiative recombination via conical intersection at a semiconductor defect. *J. Chem. Phys.* **139**, 081102 (2013).
28. Ercolessi, F. & Adams, J. B. Interatomic potentials from first-principles calculations: the force-matching method. *Europhys. Lett.* **26**, 583–588 (1994).
29. Lorenz, S., Groß, A. & Scheffler, M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem. Phys. Lett.* **395**, 210–215 (2004).
30. Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M. & Schmidhuber, J. Flexible, high performance convolutional neural networks for image classification. In *Twenty-Second International Joint Conference on Artificial Intelligence* (IJCAI, 2011).
31. Wang, L., Wu, W., Xiao, J. & Yi, Y. Large scale artificial neural network training using multi-GPUs. Preprint at https://arxiv.org/abs/1511.04348 (2015).
32. Steinkraus, D., Buck, I. & Simard, P. Y. Using GPUs for machine learning algorithms. In *Eighth International Conference on Document Analysis and Recognition* Vol. 2 1115–1120 (IEEE, 2005).
33. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
34. Yao, K., Herr, J. E., Toth, D. W., Mckintyre, R. & Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **9**, 2261–2269 (2018).
35. Lubbers, N., Smith, J. S. & Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **148**, 241715 (2018).
36. Behler, J. Constructing high-dimensional neural network potentials: a tutorial review. *Int. J. Quantum Chem.* **115**, 1032–1050 (2015).
37. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4**, 170193 (2017).
38. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
39. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
40. Behler, J. Perspective: machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
41. Simm, G. N., Proppe, J. & Reiher, M. Error assessment of computational models in chemistry. *Chimia* **71**, 202–208 (2017).
42. Settles, B. *Active Learning* (Morgan & Claypool, 2012).
43. Brown, D. F. R., Gibbs, M. N. & Clary, D. C. Combining ab initio computations, neural networks, and diffusion Monte Carlo: an efficient method to treat weakly bound molecules. *J. Chem. Phys.* **105**, 7597–7604 (1996).
44. Collins, C. R., Gordon, G. J., von Lilienfeld, O. A. & Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **148**, 241718 (2018).
45. Cubuk, E. D., Malone, B. D., Onat, B., Waterland, A. & Kaxiras, E. Representations in neural network based empirical potentials. *J. Chem. Phys.* **147**, 024104 (2017).
46. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quant. Chem.* **115**, 1094–1101 (2015).
47. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
48. Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems* Vol. 28 2224–2232 (NeurIPS, 2015).
49. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
50. Friederich, P., Konrad, M., Strunk, T. & Wenzel, W. Machine learning of correlated dihedral potentials for atomistic molecular force fields. *Sci. Rep.* **8**, 2559 (2018).
51. Westermayr, J. et al. Machine learning enables long time scale molecular photodynamics simulations. *Chem. Sci.* **10**, 8100–8107 (2019).
52. Bartók, A. P., Gillan, M. J., Manby, F. R. & Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: applications to molecular and condensed water. *Phys. Rev. B Condens. Matter* **88**, 054104 (2013).
53. Huo, H. & Rupp, M. Unified representation of molecules and crystals for machine learning. Preprint at https://arxiv.org/abs/1704.06439 (2017).
54. Wang, H., Zhang, L., Han, J. & Weinan, E. DeePMD-kit: a deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).
55. Yoo, D. et al. Atomic energy mapping of neural network potential. *Phys. Rev. Mater.* **3**, 093802 (2019).
56. Chapman, D. E., Steck, J. K. & Nerenberg, P. S. Optimizing protein-protein van der Waals interactions for the AMBER ff9x/ff12 force field. *J. Chem. Theory Comput.* **10**, 273–281 (2014).
57. Handley, C. M. & Popelier, P. L. A. Dynamically polarizable water potential based on multipole moments trained by machine learning. *J. Chem. Theory Comput.* **5**, 1474–1489 (2009).
58. Mills, M. J. L. & Popelier, P. L. A. Electrostatic forces: formulas for the first derivatives of a polarizable, anisotropic electrostatic potential energy function based on machine learning. *J. Chem. Theory Comput.* **10**, 3840–3856 (2014).

59. Unke, O. T. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
60. Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27**, 1787–1799 (2006).
61. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
62. Bereau, T., Andrienko, D. & von Lilienfeld, O. A. Transferable atomic multipole machine learning models for small organic molecules. *J. Chem. Theory Comput.* **11**, 3225–3233 (2015).
63. Sifain, A. E. et al. Discovering a transferable charge assignment model using machine learning. *J. Phys. Chem. Lett.* **9**, 4495–4501 (2018).
64. Bereau, T., DiStasio, R. A. Jr, Tkatchenko, A. & von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **148**, 241706 (2018).
65. Ramakrishnan, R. & von Lilienfeld, O. A. Many molecular properties from one kernel in chemical space. *Chimia* **69**, 182–186 (2015).
66. Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **115**, 1058–1073 (2015).
67. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
68. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B Condens. Matter* **95**, 094203 (2017).
69. Thacker, J. C. R. et al. Towards the simulation of biomolecules: optimisation of peptide-capped glycine using FFLUX. *Mol. Simula.* **44**, 881–890 (2018).
70. Uteva, E., Graham, R. S., Wilkinson, R. D. & Wheatley, R. J. Interpolation of intermolecular potentials using Gaussian processes. *J. Chem. Phys.* **147**, 161706 (2017).
71. Denzel, A. & Kästner, J. Gaussian process regression for transition state search. *J. Chem. Theory Comput.* **14**, 5777–5786 (2018).
72. Weiss, Y., Schölkopf, B. & Platt, J. C. (eds) *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference* (Bradford Books, 2006).
73. Seeger, M., Williams, C. & Lawrence, N. Fast forward selection to speed up sparse Gaussian process regression. In *Artificial Intelligence and Statistics 9* (EPFL, 2003).
74. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
75. Szlachta, W. J., Bartók, A. P. & Csányi, G. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B Condens. Matter* **90**, 104108 (2014).
76. Yang, L., Dacek, S. & Ceder, G. Proposed definition of crystal substructure and substructural similarity. *Phys. Rev. B* **90**, 054102 (2014).
77. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
78. Bartók, A. P. et al. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
79. Imbalzano, G. et al. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **148**, 241730 (2018).
80. Mahoney, M. W. & Drineas, P. CUR matrix decompositions for improved data analysis. *Proc. Natl Acad. Sci. USA* **106**, 697–702 (2009).
81. MacKay, D. J. C. in *Models of Neural Networks III* (eds Domany, E., van Hemmen, J. L. & Schulten, K.) 211–254 (1996).
82. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
83. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard clean energy project: the use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
84. Faber, F. A., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (ABC2D6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
85. Yao, K., Herr, J. E., Brown, S. N. & Parkhill, J. Intrinsic bond energies from a bonds-in-molecules neural network. *J. Phys. Chem. Lett.* **8**, 2689–2694 (2017).
86. Csányi, G., Albaret, T., Payne, M. C. & De Vita, A. 'Learn on the fly': a hybrid classical and quantum-mechanical molecular dynamics simulation. *Phys. Rev. Lett.* **93**, 175503 (2004).
87. Deringer, V. L. et al. Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. *J. Phys. Chem. Lett.* **9**, 2879–2885 (2018).
88. Sosso, G. C., Deringer, V. L., Elliott, S. R. & Csányi, G. Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials. *Mol. Simula.* **44**, 866–880 (2018).
89. Deringer, V. L., Pickard, C. J. & Csányi, G. Data-driven learning of total and local energies in elemental boron. *Phys. Rev. Lett.* **120**, 156001 (2018).
90. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quant. Chem.* **115**, 1074–1083 (2015).
91. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
92. MacKay, D. J. C. Bayesian methods for neural networks: theory and applications. *Netw. Comput. Neural Sys.* **6**, 469–505 (1995).
93. Proppe, J., Gugler, S. & Reiher, M. Gaussian process-based refinement of dispersion corrections. *J. Chem. Theory Comput.* **15**, 6046–6060 (2019).
94. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
95. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: a Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
96. Häse, F., Roch, L. M. & Aspuru-Guzik, A. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem. Sci.* **9**, 7642–7655 (2018).
97. Hernández-Lobato, J. M., Requeima, J., Pyzer-Knapp, E. O. & Aspuru-Guzik, A. Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. In *Proc. 34th Int. Conf. Machine Learning* Vol. 70 1470–1479 (JMLR, 2017).
98. Zhang, L., Wang, H. & E, W. Reinforced dynamics for enhanced sampling in large atomic and molecular systems. *J. Chem. Phys.* **148**, 124113 (2018).
99. Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. Machine learning of molecular properties: locality and active learning. *J. Chem. Phys.* **148**, 241727 (2018).
100. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
101. Kolb, B., Luo, X., Zhou, X., Jiang, B. & Guo, H. High-dimensional atomistic neural network potentials for molecule-surface interactions: HCl scattering from Au(111). *J. Phys. Chem. Lett.* **8**, 666–672 (2017).
102. Quaranta, V., Hellström, M. & Behler, J. Proton-transfer mechanisms at the water-ZnO interface: the role of presolvation. *J. Phys. Chem. Lett.* **8**, 1476–1483 (2017).
103. Boes, J. R. & Kitchin, J. R. Neural network predictions of oxygen interactions on a dynamic Pd surface. *Mol. Simula.* **43**, 346–354 (2017).
104. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
105. Artrith, N., Morawietz, T. & Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **83**, 079114 (2011).
106. Mills, K. et al. Extensive deep neural networks for transferring small scale learning to large scale systems. *Chem. Sci.* **10**, 4129–4140 (2019).
107. Chiriki, S., Jindal, S. & Bulusu, S. S. Neural network potentials for dynamics and thermodynamics of gold nanoparticles. *J. Chem. Phys.* **146**, 084314 (2017).
108. Eckhoff, M. & Behler, J. From molecular fragments to the bulk: development of a neural network potential for MOF-5. *J. Chem. Theory Comput.* **15**, 3793–3809 (2019).
109. Caro, M. A., Deringer, V. L., Koskinen, J., Laurila, T. & Csányi, G. Growth mechanism and origin of high sp^3 content in tetrahedral amorphous carbon. *Phys. Rev. Lett.* **120**, 166101 (2018).
110. Gabardi, S., Sosso, G. G., Behler, J. & Bernasconi, M. Priming effects in the crystallization of the phase change compound GeTe from atomistic simulations. *Faraday Discuss.* **213**, 287–301 (2019).
111. Mocanu, F. C. et al. Modeling the phase-change memory material, GeSbTe, with machine-learned interatomic potential. *J. Phys. Chem. B* **122**, 8998–9006 (2018).
112. Artrith, N. & Kolpak, A. M. Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: a combination of DFT and accurate neural network potentials. *Nano Lett.* **14**, 2670–2676 (2014).
113. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
114. Fletcher, T. L. & Popelier, P. L. A. Toward amino acid typing for proteins in FFLUX. *J. Comput. Chem.* **38**, 336–345 (2017).
115. Blau, S. M., Bennett, D. I. G., Kreisbeck, C., Scholes, G. D. & Aspuru-Guzik, A. Local protein solvation drives direct down-conversion in phycobiliprotein PC645 via incoherent vibronic transport. *Proc. Natl Acad. Sci. USA* **115**, E3342–E3350 (2018).
116. Senftle, T. P. et al. The ReaxFF reactive force-field: development, applications and future directions. *npj Comput. Mater.* **2**, 15011 (2016).
117. Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **53**, 291–318 (2002).

118. Reiher, M. (ed.) *Topics in Current Chemistry: From Quantum Chemistry to Molecular Simulation* (Springer, 2007).
119. Gaus, M., Cui, Q. & Elstner, M. DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* **7**, 931–948 (2012).
120. Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput. Aided Mol. Des.* **4**, 1–103 (1990).
121. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB-an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
122. Lin, H. & Truhlar, D. G. QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **117**, 185–199 (2007).
123. Senn, H. M. & Thiel, W. QM/MM studies of enzymes. *Curr. Opin. Chem. Biol.* **11**, 182–187 (2007).
124. Cui, Q. Perspective: quantum mechanical methods in biochemistry and biophysics. *J. Chem. Phys.* **145**, 140901 (2016).
125. Pozun, Z. D. et al. Optimizing transition states via kernel-based machine learning. *J. Chem. Phys.* **136**, 174101 (2012).
126. Li, J., Song, K. & Behler, J. A critical comparison of neural network potentials for molecular reaction dynamics with exact permutation symmetry. *Phys. Chem. Chem. Phys.* **21**, 9672–9682 (2019).
127. Lu, D. et al. Mode specific dynamics in the $H_2 + SH \rightarrow H + H_2S$ reaction. *Phys. Chem. Chem. Phys.* **18**, 29113–29121 (2016).
128. Kolb, B., Zhao, B., Li, J., Jiang, B. & Guo, H. Permutation invariant potential energy surfaces for polyatomic reactions using atomistic neural networks. *J. Chem. Phys.* **144**, 224103 (2016).
129. Li, J. & Guo, H. Communication: an accurate full 15 dimensional permutationally invariant potential energy surface for the OH + CH4 → H2O + CH3 reaction. *J. Chem. Phys.* **143**, 221103 (2015).
130. Gastegger, M. & Marquetand, P. High-dimensional neural network potentials for organic reactions and an improved training algorithm. *J. Chem. Theory Comput.* **11**, 2187–2198 (2015).
131. Liu, Q. et al. Constructing high-dimensional neural network potential energy surfaces for gas–surface scattering and reactions. *J. Phys. Chem. C.* **122**, 1761–1769 (2018).
132. Del Cueto, M. et al. New perspectives on CO_2–Pt(111) interaction with a high-dimensional neural network potential energy surface. *J. Phys. Chem. C.* **124**, 5174–5181 (2020).
133. Shakouri, K., Behler, J., Meyer, J. & Kroes, G.-J. Accurate neural network description of surface phonons in reactive gas-surface dynamics: N_2 + Ru(0001). *J. Phys. Chem. Lett.* **8**, 2131–2136 (2017).
134. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed. Engl.* **56**, 12828–12840 (2017).
135. Crespo-Otero, R. & Barbatti, M. Recent advances and perspectives on nonadiabatic mixed quantum-classical dynamics. *Chem. Rev.* **118**, 7026–7068 (2018).
136. Chen, W.-K., Liu, X.-Y., Fang, W.-H., Dral, P. O. & Cui, G. Deep Learning for nonadiabatic excited-state dynamics. *J. Phys. Chem. Lett.* **9**, 6702–6708 (2018).
137. Hu, D., Xie, Y., Li, X., Li, L. & Lan, Z. Inclusion of machine learning kernel ridge regression potential energy surfaces in on-the-fly nonadiabatic molecular dynamics simulation. *J. Phys. Chem. Lett.* **9**, 2725–2732 (2018).
138. Häse, F., Kreisbeck, C. & Aspuru-Guzik, A. Machine learning for quantum dynamics: deep learning of excitation energy transfer properties. *Chem. Sci.* **8**, 8419–8426 (2017).
139. Richings, G. W. & Habershon, S. Direct quantum dynamics using grid-based wave function propagation and machine-learned potential energy surfaces. *J. Chem. Theory Comput.* **13**, 4012–4024 (2017).
140. Musil, F., Willatt, M., Langovoy, M. A. & Ceriotti, M. Fast and accurate uncertainty estimation in chemical machine learning. *J. Chem. Theory Comput.* **15**, 906–915 (2019).
141. Huan, T. D. et al. Iterative-learning strategy for the development of application-specific atomistic force fields. *J. Phys. Chem. C.* **123**, 20715–20722 (2019).
142. Singraber, A., Behler, J. & Dellago, C. Library-based LAMMPS implementation of high-dimensional neural network potentials. *J. Chem. Theory Comput.* **15**, 1827–1840 (2019).
143. Schütt, K. T. et al. SchNetPack: a deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.* **15**, 448–455 (2019).
144. Häse, F., Valleau, S., Pyzer-Knapp, E. & Aspuru-Guzik, A. Machine learning exciton dynamics. *Chem. Sci.* **7**, 5139–5147 (2016).
145. Zhang, L., Wang, H. & E, W. Adaptive coupling of a deep neural network potential to a classical force field. *J. Chem. Phys.* **149**, 154107 (2018).
146. Zhang, L., Han, J., Wang, H., Car, R. & E, W. DeePCG: constructing coarse-grained models via deep neural networks. *J. Chem. Phys.* **149**, 034101 (2018).
147. Galvelis, R. & Sugita, Y. Neural network and nearest neighbor algorithms for enhancing sampling of molecular dynamics. *J. Chem. Theory Comput.* **13**, 2489–2500 (2017).
148. Chiavazzo, E. et al. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl Acad. Sci. USA* **114**, E5494–E5503 (2017).
149. Meyer, B., Sawatlon, B., Heinen, S., von Lilienfeld, O. A. & Corminboeuf, C. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.* **9**, 7069–7077 (2018).
150. S Smith, J. *et al.* Outsmarting quantum chemistry through transfer learning. Preprint at https://doi.org/10.26434/chemrxiv.6744440 (2018).
151. Schütt, K. T., Sauceda, H. E.-J., Kindermans, P., Tkatchenko, A.-R. & Müller, K. SchNet – a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
152. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* 10th edn (MIT Press, 2006).
153. Lu, J., Wang, C. & Zhang, Y. Predicting molecular energy using force-field optimized geometries and atomic vector representations learned from an improved deep tensor neural network. *J. Chem. Theory Comput.* **15**, 4113–4121 (2019).
154. Nebgen, B. et al. Transferable dynamic molecular charge assignment using deep neural networks. *J. Chem. Theory Comput.* **14**, 4687–4698 (2018).

Acknowledgements

The authors thank G. dos Passos Gomes and S. Y. Guo for helpful discussions about the literature in the field of ML potentials. P.F. acknowledges funding by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 795206 (MolDesign). J.P. appreciates funding through an Early Postdoc.Mobility fellowship by the Swiss National Science Foundation (project no. 178463). F.H. acknowledges support from the Herchel Smith Graduate Fellowship and the Jacques-Emile Dubois Student Dissertation Fellowship. The authors acknowledge support by the Canadian Institute for Advanced Research and the Canada 150 Research Chair Program, as well as the generous support of A. G. Frøseth.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to A.A.-G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2021