



Machine learning potentials for extended systems: a perspective

Jörg Behler^{1,a} and Gábor Csányi²

¹ Universität Göttingen, Institut für Physikalische Chemie, Theoretische Chemie, Tammannstr. 6, 37077 Göttingen, Germany

² Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, UK

Received 20 April 2021 / Accepted 5 July 2021
© The Author(s) 2021

Abstract. In the past two and a half decades machine learning potentials have evolved from a special purpose solution to a broadly applicable tool for large-scale atomistic simulations. By combining the efficiency of empirical potentials and force fields with an accuracy close to first-principles calculations they now enable computer simulations of a wide range of molecules and materials. In this perspective, we summarize the present status of these new types of models for extended systems, which are increasingly used for materials modelling. There are several approaches, but they all have in common that they exploit the locality of atomic properties in some form. Long-range interactions, most prominently electrostatic interactions, can also be included even for systems in which non-local charge transfer leads to an electronic structure that depends globally on all atomic positions. Remaining challenges and limitations of current approaches are discussed.

1 Modelling extended systems

It was recognised early on in the development of quantum mechanics that the time scale separation between the motion of electrons and nuclei allows for the Born–Oppenheimer approximation, valid for a large part of chemistry, solid-state physics and materials science [1]. This gave rise to the quest for accurate analytic models of the resulting potential energy associated with a set of nuclear positions to avoid the prohibitive costs of repeatedly performing explicit electronic structure calculations for systems containing more than a few hundred atoms, e.g. in molecular dynamics (MD) and high throughput screening. It quickly became obvious that apart from low-dimensional cases, e.g. diatomics, or an infinitesimally small configuration space, like the vibrational modes of a molecule and the phonon spectrum of a crystal at zero temperature, systematically convergent fits of the potential energy surface (PES) were not feasible. Thus, for a long time “force fields” in chemistry and “interatomic potentials” in physics and materials science have been largely dominated by empirical modelling, where physics and chemistry-based principles are coupled with some form of nonlinear model fitting (e.g. Morse potential for chemical bonds [2], Lennard–Jones for dispersion interactions [3]), having just about enough explanatory and predictive power

for a qualitative understanding of a specific problem at hand.

In many areas, remarkable progress has been made using this approach, e.g. organic force fields enabled the simulation of large biopolymers, often in aqueous solution. The price has been the necessity of drastic simplifications, such as the use of low-dimensional additive terms with simple functional forms and, with a few exceptions [4], the inability to describe the making and breaking of chemical bonds. In materials science, on the other hand, reactivity has been an essential feature of even basic potentials, but the resulting functional forms turned out to be difficult to use for systems containing many different elements.

In parallel to these developments, the direct solution of the electronic structure problem in some approximate form, typically density functional theory (DFT), reached a level of usefulness that revolutionised atomistic modelling of extended systems putting it, for the last couple of decades, right at the center of our understanding of materials on the electronic scale – as long as the questions could be addressed by calculating the total energy (or short dynamical paths) of a few hundred atoms. While understanding the electronic structure of materials and its implications for material properties is clearly an important and often necessary endeavour, the situation has arisen by the turn of this century that much of the world’s academic computing power was used to carry out *ab initio molecular dynamics* and similar tasks solving the electronic structure problem at each time step simply because there

^ae-mail: joerg.behler@uni-goettingen.de (corresponding author)

has been no alternative access to energies and forces of comparable quality while the research questions rarely require detailed knowledge about the electronic structure, like band gaps or orbital energies.

2 The challenge of dimensionality

Progress in high throughput electronic structure calculations regarding the robustness of software, advances in algorithms, and wide availability of the necessary computational capacity on the one hand, and rapid conceptual advances of “machine learning approaches” in data science, statistics and artificial intelligence research on the other hand resulted in a new kind of numerical solution, using *machine learning potentials* (MLP), to the age-old problem of materials modelling, the high-dimensional and accurate representation of the Born–Oppenheimer PES for extended systems [5–9].

There is a multitude of intellectual roots for this development in many different fields. The representation of potential energy surfaces for small molecules has been a very active field in quantum chemistry for a long time [10–15]. Many pertinent issues were explored, such as the symmetry of the representations and basis sets, and systematic convergence, but all schemes scaled badly, essentially exponentially, with the number of degrees of freedom, restricting these approaches to small systems. Thus, the application area of the resulting potentials was mostly molecular spectroscopy, which placed very stringent demands on accuracy. Elsewhere, empirical force fields—or interatomic potentials, such as the Embedded Atom Method [16]—were widely used for modelling extended materials. These scale very favourably with system size, but the general sentiment in the community developing these models was that a phenomenologically correct prediction of macroscopic properties was more important than quantitative convergence towards the first-principles potential energy surface, the latter generally held to be impossible.

The first generation of MLPs was introduced in the seminal work of Doren et al. [17]. They employed a feed-forward neural network to represent the DFT PES for a H_2 molecule interacting with a silicon surface. Many other neural network potentials have been proposed in the following decade employing one or a set of neural networks [18–20], but they all had in common that only a few degrees of freedom were taken into account explicitly. The main reason for this limitation of first-generation MLPs was the lack of descriptors that had the physically mandatory invariances, which could only be written down for special cases of low-dimensional problems [21, 22] and did not scale to many-atom systems.

A solution to the above problem was the introduction of a *many-body descriptor* of the local atomic neighbourhood environment by Behler and Parrinello [23] that had all the required symmetries. They used it to model the *atomic energy*, E_i , of each atom i , using element-

specific atomic feed-forward neural networks, fitted to reproduce the total energy, as

$$E_{\text{total}} = \sum_{i=1}^{N_{\text{atom}}} E_i$$

$$E_i = E(r_{ij_1} \dots r_{ij_k} \dots) \quad \text{with } r_{ij} = |\mathbf{r}_i - \mathbf{r}_j| < R_{\text{cut}}, \quad (1)$$

where $\mathbf{r}_i$ is the position vector of atom i and r_{ij} is the distance between atom i and its neighbor j . The key in this ansatz of second-generation MLPs is the restriction of arguments of the atomic energy to the coordinates of atoms within a distance cutoff R_{cut} , which makes the atomic energy function moderate dimensional, approximately three times the number of atoms within the cutoff (which in practice ranges between 10 and 50 atoms, depending on the material and cutoff radius). The descriptors of Behler and Parrinello (often called Atom-Centered Symmetry Functions [24]) parametrise this function, using a fixed length array, in a way that it remains a smooth function of the atomic positions, but allowing for atoms to cross in and out of the cutoff sphere which is necessary for performing atomistic simulations. Note that most conventional empirical potentials can also be written in or reformulated into the above form, but the representation of the atomic neighbourhood environment they use (effectively the unordered set of distances for bond terms, angles for three-body terms) are low-dimensional and thus cannot be used to model an arbitrary function of the neighbourhood environment.

Using the distance-restricted atomic energy is a controlled approximation. The larger the cutoff, the smaller the contribution of the long-range interactions that are ignored, but the space in which the atomic energy function has to be fitted is larger, necessitating more training data. There are plenty of examples, and not coincidentally these constitute the earliest successes of MLPs, where long-range terms beyond 5–6 Å or so can be essentially ignored, and the resulting potentials are still extremely useful [8, 25, 26]. For example in elemental systems, and many bulk materials, the degree of charge transfer is either minimal or screening lengths are very short. There have been recent successes in materials where *a priori* one perhaps would not expect that, e.g. some bulk oxides [27, 28] and even bulk liquid water [29, 30].

Fitting interatomic potentials to electronic structure data in this manner was a new type of “task” for computational materials science, and given the developments of the past decades, we can now confidently say that this task has been solved eminently by many types of modern MLPs like high-dimensional neural network potentials (HDNNP) [23], Gaussian approximation potentials (GAP) [31], moment tensor potentials [32], spectral neighbor analysis potentials [33], atomic cluster expansion [34], DeepMD [35], and ANI [36], capable of reaching accuracies of 0.001 eV/atom for energies and 0.05 eV/Å for forces for many systems. This accuracy is near the convergence threshold of typical DFT

Table 1 The three main approaches to parametrising atomic properties (just energies and forces directly, or also including electrostatic properties such as charges and dipoles)

Descriptors + nonlinear fitting

A compact symmetry preserving representation of the (usually local) atomic geometry is defined, and used as input to nonlinear regression schemes such as neural networks or kernel regression

Basis + linear fitting

A basis of many symmetric functions of local atomic geometries is constructed, and coefficients are determined using regularised least squares fitting

Message passing networks

The representation and regression problems are solved simultaneously, starting with just atomic identities, which are iteratively passed along the neighbour graph of the system and combined into a symmetric nonlinear function

calculations when performing first principles molecular dynamics.

3 Representation and regression

The total potential energy of a system must be invariant to many symmetries: global rotation and translation of the system, and the permutation of atoms of the same element. Having made the approximation that the total energy is the sum of local (atomic or site) energy terms, it is natural to want to transfer these symmetries or invariances to the site energy in Eq. 1, but that is hard to do without sacrificing the ability of the model to fit an arbitrary (symmetric) function. For example, empirical force fields, which have a venerable history, split the potential energy into terms that each depend on the positions of only one, two, three, etc. atoms, and parametrise them using invariant internal coordinates (interatomic distance, angle, etc), see top panels of Fig. 1. Force field energies fulfill these invariances because they consist of additive terms, each of which are invariant to rotations and translations, and which are evaluated independently. While a large fraction of the intramolecular potential energy of organic molecules can be captured in this way, when one continues this sequence systematically to high body order (e.g. in the permutationally invariant polynomial scheme of Bowman and Braams [14]), the computational expense explodes exponentially. For extended materials, such as metals, oxides, semiconductors, this strategy runs into difficulties very quickly already for more than a few atoms. On the other hand when the body order is truncated at 3 or 4-body level, the potentials are not accurate enough in general.

Table 1 summarises the three main approaches for solving this problem that have emerged over the past decade, and Fig. 1 gives a graphical representation of

each in comparison to classical force fields. In the first approach, presented in the second panel on the right of Fig. 1, no body ordered splitting of the potential energy is imposed, but descriptors consisting of low body order terms are used in a fully symmetric *representation* of the local geometry, which is then subsequently used in a nonlinear regression scheme to fit the potential energy (e.g. using artificial neural networks [23] or kernel regression [31]). The set of Atom Centered Symmetry Functions (ACSF) [23,24] is an example, where a compact representation of all neighboring atoms inside the atomic environment is built using two- and three-body functions. The power spectrum, as well as its commonly used incarnation, the Smooth Overlap of Atomic Positions (SOAP) [37], is a similar three-body representation, while the bispectrum is four-body [37]. These are built up as products of the single-particle density using the “density trick” and therefore the cost of their evaluation scales linearly with the number of neighbours at arbitrarily high body order [38].

The success of MLPs, built using compact low body order representations and nonlinear regressors (such as HDNNPs [23] or GAPs [31]), in correctly fitting finely nuanced *ab initio* potential energy surfaces [25,26] has moved the field forward significantly, and brought with it renewed interest in designing symmetric representations for materials. In particular, Moment Tensor Potentials [32] and later the Atomic Cluster Expansion [34] (see third panel on the right of Fig. 1) bring together some features of the descriptor approach and empirical force fields. The potential energy is again split up into body ordered contributions, but each of these are represented by symmetric basis functions that are *complete*, i.e. arbitrary smooth functions can be constructed using simple *linear* combinations. The same “density trick” that allowed the efficient evaluation of the power- and bispectrum is used to keep the evaluation cost independent of the body order and number of neighbours.

Very recently, a third approach to representation and regression has been introduced based on learning the atomic descriptors directly from the geometric structure, see bottom panel of Fig. 1. This class of neural network potentials employs message passing networks [39]. The central idea of these methods is to map the chemical elements into a high dimensional atomic feature vector, which is then iteratively refined for each atom by exchanging information about the features of neighboring atoms and local geometry. First implementations used only interatomic distances, but more recently the full many-body description is utilized e.g. in “equivariant message passing networks” [40]. For high-dimensional systems, a cutoff radius for the information exchange is introduced making the methods feasible for larger systems, and for each message passing step (typically between two and six are employed) the effective range of the description of the atomic environments is increased. Thus, no representation is constructed independently of the regression, but both tasks are accomplished simultaneously. Several different methods have been proposed to date and applied to

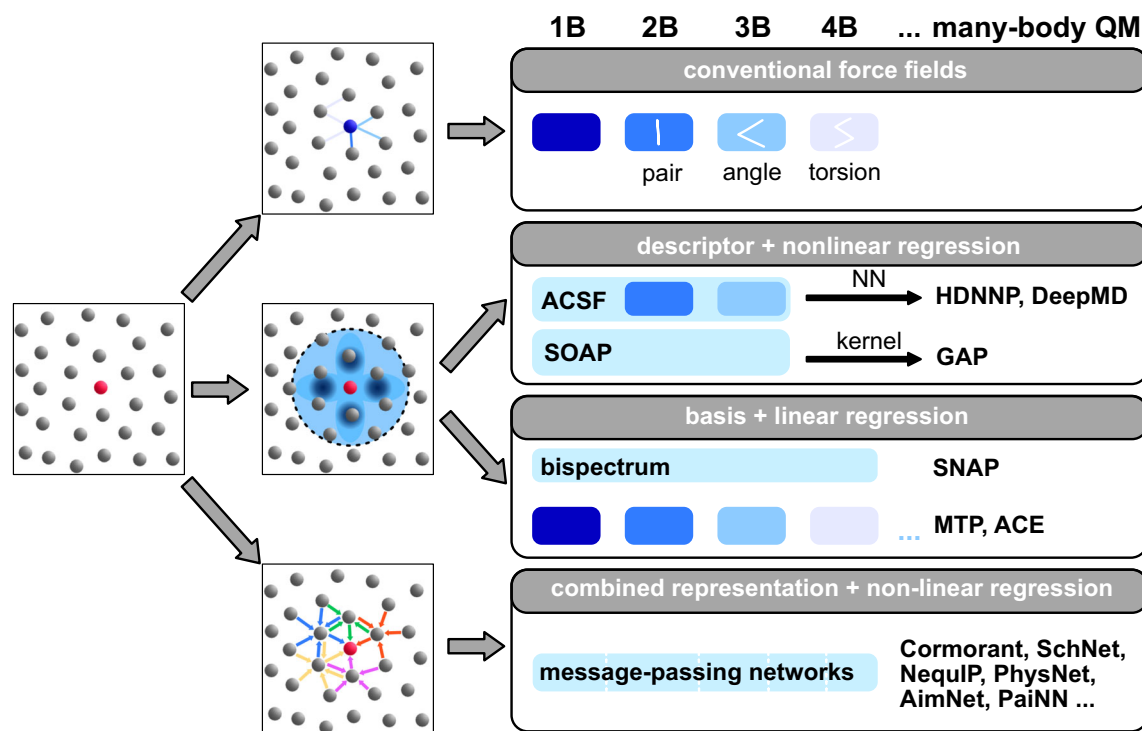


Fig. 1 Illustration of the different strategies for representing local atomic environments and regressing the potential energy

benchmark data sets, in particular for molecules, yielding highly accurate energies and forces [40–46].

4 Beyond locality—long-ranged MLPs

In spite of many successful applications of second-generation MLPs based on Eq. 1, applying a cut-off truncating the interactions necessarily represents an approximation. This is well justified and introduces only very small errors e.g. for systems without significant charge transfer like in elemental compounds [47, 48] or in the presence of efficient screening, like in bulk liquid water [29, 30]. Still, any interaction present beyond the cutoff limits the accuracy that can be achieved, because for the ML algorithm such interactions result in inconsistent data that cannot be represented with the available information about the local atomic environments. Consequently, interactions beyond the cutoff will appear like noise, and to minimize the overall error an averaged interaction is obtained for interatomic distances larger than the cut-off radius. Apart from electrostatics, e.g. also dispersion interactions and delocalized electrons in aromatic compounds or conjugated π -systems [48, 49] can give rise to rather long-ranged interactions.

While in principle it is possible to systematically extend the cutoff radius to include an increasing part of the atomic interactions, which is very similar in spirit to increasing the number of passing steps in message-passing methods, there are practical limitations in both

approaches. These concern the complexity in describing the rapidly growing configuration space in the atomic environments by a set of descriptors, but also the unfavorable scaling of the performance regarding the number of neighbors in the cutoff spheres, or the number of passing steps in message passing networks, respectively.

Assessing the relevance of long-range interactions is thus an important task and has been done e.g. for the cases of amorphous carbon [48], carbon clusters [50], water clusters [29], organic molecules [49, 51] and the adsorption of metal clusters at oxides [51]. For instance, it has been shown in a locality test for carbon [48] how long-range interactions can be quantified by sampling different atomic positions outside the cutoff sphere as illustrated in Fig. 2. In essence, for fixed atomic positions inside the cutoff sphere the force acting on the central atom is monitored while moving the atoms beyond the cutoff. For mainly local interactions, only small changes of the force will be observed. Similarly, the dependence of atomic charges on structural features outside the local atomic environments can be studied, e.g. by varying distant functional groups [51]. Next to an overall reduced accuracy of the potential, too small cutoffs can also induce artificial oscillations in the energy and forces close to the cutoff radius if energies and forces are used for training [24]. Note that such locality tests are very general, and not only applicable for assessing the level of nonlocality, but can be used to measure the effectiveness of a long-range model: instead of measuring the standard deviation of the total force at the center of the fixed sphere (see Fig. 2), it is possible to consider the force difference between the ref-

erence electronic structure method and the proposed long-range model.

In particular the inclusion of electrostatic interactions in MLPs has received a lot of attention in recent years. Earliest attempts have used neural networks to express electrostatic multipoles with the aim to improve classical force fields [52]. Further, it has been suggested to augment MLPs by long-range electrostatic energies based on fixed element-specific charges e.g. for GAP [31] and SNAP [53], which is an efficient solution for specific systems but is limited to compounds with very similar charges for all atoms of a given species.

Another path is followed in third-generation MLPs, in which both the short-range and the long-range energy are expressed by a ML fit. For this purpose, the long-range electrostatic energy is computed using flexible charges, which are represented as local environment-dependent atomic properties by machine learning algorithms. An example are third-generation HDNNPs [54, 55], in which the charges are expressed by element-specific atomic feed-forward neural networks and in which the resulting electrostatic energy is combined with a short-range energy according to Eq. 1 to account for local bonding. Double counting of electrostatic energy contributions, which within the cutoff could equally be included in the short-range part, is avoided by separating the reference total energies and forces into an electrostatic and a remaining short-range part before training. To prevent an increased energy range to be covered by the short-range energies due to the singularity of the Coulomb potential for short interatomic distances, it has been found useful to screen the electrostatic part inside the cutoff radius [55] similarly to what is done in dispersion correction schemes [56].

Many other approaches have been proposed to express atomic charges by machine learning [42, 57–62]. Often, reference charges obtained in electronic structure calculations are used for training. Such charges are not physical observables and thus there is no unique choice, but still there is a range of methods which are mathematically well defined. Consequently, quite different reference atomic charges can be obtained depending on the chosen method [62]. Alternatively, charges can also be determined using dipole moments as Refs. [57, 63], which offers the advantage of an observable target property, but this method is restricted to small molecular systems. In principle, tests could be performed for a given system which charge partitioning scheme provides the best representation of long-range electrostatic interactions by performing locality tests after removing the corresponding long-range electrostatic forces. In addition to the local atomic descriptors and message passing networks, it has also been demonstrated that the electrostatic energy can be included using long-ranged representations like the long-distance equivariant representation [64].

An obvious problem when learning individual atomic charges is to ensure the correct overall charge of the system because of small fitting errors, and in particular for periodic systems any net charge must be avoided. A straightforward solution is the rescaling of

the charges [54], but also using a compensating background charge has been proposed [57].

While the incorporation of long-range electrostatic interactions relying on flexible environment-dependent charges in third-generation MLPs has been a step forward, the underlying assumption of only locally dependent charges still limits the applicability of these methods. The reason is that in numerous systems the local electronic structure is influenced by very distant structural features giving rise to non-local charge transfer. While this is naturally incorporated in electronic structure calculations, i.e. in the reference data, in such a situation charges cannot be accurately learned as a function of the local environment only, resulting in ambiguous or even contradictory information in the training data.

Consequently, when discussing interatomic potentials, it is important to clearly distinguish on the one hand between potentials that include *long-range interactions* in general, e.g. in form of an explicit Coulomb term like in third-generation MLPs, and on the other hand potentials considering *non-local interactions* resulting from dependencies extending over distances beyond the cutoff radius, or beyond the environment covered by the message-passing steps. This non-locality cannot be captured by third-generation potentials relying on local charges only.

Atomic partial charges can be considered as a coarse-grained qualitative fingerprint of the electronic structure, and thus they can be used to detect non-local dependencies of the electronic structure in a simple way [51]. These are omnipresent in many systems, from organic molecules to bulk materials and interfaces [49, 51]. Apart from these dependencies, another possible origin of global changes in the electronic structure is a change of the total charge, which may not only be the attachment or removal of an electron, but also chemical reactions like (de)protonation. MLPs up to the third generation do not depend on the total charge of the system and are thus restricted to a single charge state.

MLPs, which are able to include long-range electrostatic interactions based on atomic charges depending on the structure of the full system, define fourth-generation MLPs. The first potential of this generation has been the charge equilibration neural network technique (CENT) [65]. In this method, the atomic partial charges are determined in a global charge equilibration step [66] employing atomic electronegativities expressed by individual atomic neural networks as a function of the local chemical environments. The method employs a charge density-based total energy expression making it particularly suitable for systems with primarily ionic bonding. A generalization of this method has recently been proposed in form of fourth-generation high-dimensional neural network potentials (4G-HDNNP) [51], which combine the properties of CENT and second-generation HDNNPs of the Behler-Parrinello type. The resulting method includes long-range electrostatic interactions based on globally-dependent charges and atomic short-

range energies accounting for local bonding. Due to the charge equilibration step, the naive implementation is cubically scaling with system size, but using iterative schemes close to linear scaling can be achieved. Another related method is the Becke population neural network (BpopNN) [67]. This method relies on atomic neural networks using modified SOAP descriptors including also the atomic populations, i.e. information about the atomic charges, which are determined in an iterative process. All fourth-generation MLPs can be constructed to simultaneously describe several charge states of a system.

It is interesting to note that message passing networks can form a bridge between second, third and fourth-generation potentials. For instance, apart from the total energy they have been suggested to predict charges [42, 58, 59], although to date only rarely explicit Coulomb terms are used to include long-range interactions without truncation [42] as needed for a classification as a third or fourth generation potential. Still, by increasing the number of passing steps in principle they allow to describe more and more interactions irrespective of the physical nature including electrostatics between close atoms, like second-generation MLPs based on Eq. 1. They would become equivalent to fourth-generation potentials for an infinite number of passing steps, or for small molecules in vacuum that can be treated globally. Recently, several message passing network have been proposed which are also capable to describe systems in different total charge states [58, 59].

Apart from electrostatics, also dispersion interactions can be rather long-ranged and thus relevant in the cutoff region. They are much weaker than electrostatic interactions, but depending on the system, still important parts of the dispersion energy might be missed by second-generation MLPs. Like for electrostatics, a hierarchy of method can be used to included dispersion interactions, but in comparison to electrostatics many details remain to be explored. A straightforward possibility to include dispersion is to simply add a correction like those also frequently used in electronic structure calculations, e.g. the family of methods suggested by Grimme [56], which is used in Phys-Net [42] and tensormol [57]. In chemically simple systems, such as fluid methane, the long-range part of the interaction can be represented by a $1/r^6$ potential with a fixed coefficient, such that on top that a second-generation MLP is sufficient [68]. More advanced methods could be based on environment-dependent C_6 coefficients, qualifying as third-generation methods, to globally-dependent parameters of the fourth generation. Such methods are not yet available and the importance of a high-level treatment of dispersion remains to be investigated in future work in view of the importance of many-body dispersion in some systems [69].

The inclusion of long-range energy contributions in the different types of MLPs is summarized in Fig. 3. Second-generation MLPs based on environment-dependent atomic energies typically include all short-ranged interactions arising e.g. from covalent bonding, and all electrostatic and dispersion interactions up to

the cutoff radius, since the atomic energies in Eq. 1 can include all types of interactions. On the other hand, all interactions beyond the cutoff are neglected. In third-generation MLPs the atomic energies are used to represent the short-range energies only, which are extracted for training by first removing the electrostatic and possibly the dispersion energy. These energy terms are then explicitly added without truncation using Coulomb's law in case of electrostatics. Any inaccuracies in the electrostatic and dispersion energies can be corrected inside the cutoff spheres by adapting the atomic short-range energies accordingly. Still, depending on the specific system a part of the long-range interactions might be missed beyond the cutoff, e.g. in case of charge transfer over long distances that cannot be captured by environment-dependent charges. Fourth-generation MLPs are based on charges depending on the global structure of the system. Therefore, they in principle include the full long-range energies, and the accuracy of the long-range interactions beyond the cutoff is only limited by the accuracy obtained in the fitting process and the chosen charge partitioning method. Message passing networks, like second-generation MLPs, can describe all types of interactions with the number of passing steps taking the role of the cutoff radius, which can be systematically increased.

5 Discussion and outlook

When considering any task, especially a newly defined task such as the regression of the Born–Oppenheimer potential energy surface, which we discussed in the introduction, it is important to also define the measures of success. What is a good potential? It is surprisingly hard to give a general and quantitative answer. In the world of empirical potentials, this question seemed easy. The potentials used very few parameters, and once those were fixed (either by fitting them to a few key electronic structure calculations or by matching the results of molecular dynamics simulations to experiments), the performance of the empirical potential can be tested in a wide variety of situations, all well away from those that went into “training” the potential. This has been possible due to the physics-based functional forms ensuring at least a qualitatively reasonable transferability. In contrast, the MLPs work differently. Here, the chosen functional forms have no physical basis but need to be flexible to reach a high numerical accuracy, and extrapolation in high dimensional spaces is necessarily limited, so the training database has to include local environments that are similar to those where the potential will be used. But then how does one test the potential? On configurations that are far away from the training set, errors will be large, but that is part and parcel of using an MLP. The question then is turned around and we need to ask: is our training database diverse enough that it covers all the local environments where our potential is ever likely to be evaluated in the course of the simulations? Simulations with MLPs can

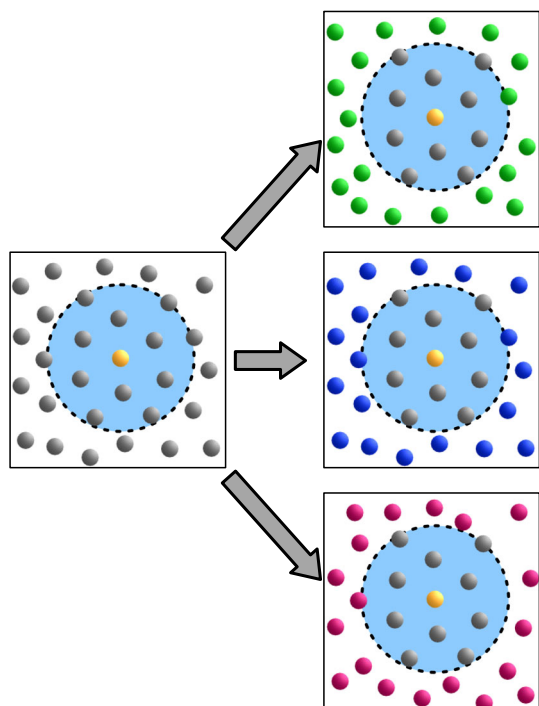


Fig. 2 Illustration of the “locality test”. An atom (yellow) is selected from a larger configuration, and the positions of all grey atoms within a sphere of a certain radius of it are fixed. Many new configurations are generated which all share this fixed sphere, but differ elsewhere. The forces are computed for all structures, and the standard deviation of the force on the central atom is recorded. Any local interatomic potential whose cutoff is equal to half of the sphere radius cannot predict the force on the central atom with a smaller standard error than that from the locality test, simply because the structural information required is not available to it. This is independent of the representation and regression method, and is a consequence of the underlying locality approximation

go *catastrophically* wrong if the simulation leaves the range of validity. We illustrate the asymmetry of consequences of positive and negative errors in the PES in Fig. 4, where the effect is exaggerated for clarity. An error in the positive direction will lead to a lowering of the probability in visiting that state in a finite temperature simulation, but this effect is usually very local leaving most of the probability distribution unchanged. A similar-sized error in the negative direction acting as an attractive well however could lead to a dramatic change in the overall probability distribution and thus the values of many observables that are computed from the simulation. The key point is that the MLP is not merely used to predict properties of structures drawn from some known distribution, but the MLP is used to generating the distribution itself.

So while testing MLPs necessarily involves computing root mean square errors (RMSEs), careful judgement has to be made over what distribution of configurations these RMSEs are evaluated. A well-fitted potential will always show good RMSEs over the ref-

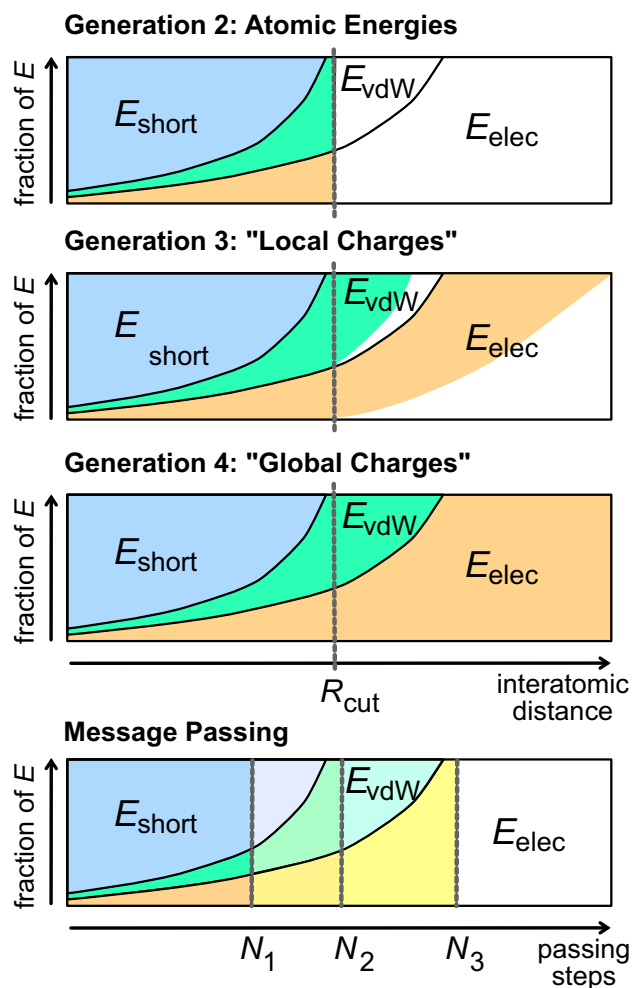


Fig. 3 Relative range and coverage of energy contributions in different types of MLPs. In second-generation MLPs, all interactions are completely included in the atomic energies inside the cutoff radii. In third-generation MLPs electrostatic and dispersion interactions are separated and computed analytically outside the cutoff radius based on local properties like environment-dependent charges. Any error can be corrected inside the cutoff radii by the atomic energies. Fourth-generation MLPs are based on atomic properties depending on the global structure and total charge of the systems resulting, e.g., in improved charges for systems with non-local charge transfer. In message passing MLPs the situation is similar to second-generation MLPs, with the range determined by the number of passing steps N_i

erence configurations, including test sets split out for validation purposes. The far more important question is how well *properties* are predicted. This opens up a whole set of new, as yet unanswered questions: what is the relationship between the pointwise (RMSE) errors of the potential energy surface, and the accuracy of predicting material properties?

There appears to be two strategies emerging with respect to how wide ranging the MLP training databases are. On the one hand, one can make “special purpose” potentials for a particular project, which can be vali-

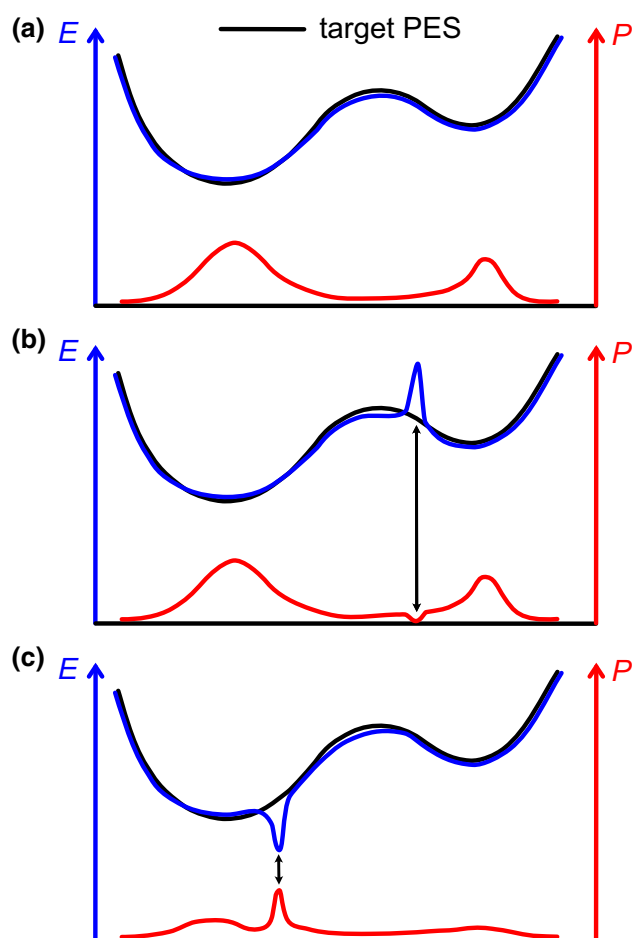


Fig. 4 Impact of positive and negative energy errors on sampling the Boltzmann distribution. In **a** the energy E is close to the target PES and the probability of the system to be in the global or the local minimum is high. In **b** the energy in a small part of the PES is overestimated resulting in a local decrease of the population P while the population of the other configurations is hardly changed. In case of a severe underestimation of the energy in **c**), the population of the artificial minimum is strongly increased resulting in a depletion of the population elsewhere

dated for the configurations relevant to the simulations that are planned in the context of that project. There is no expectation of transferability beyond the intended application. Alternatively, and much in common with the spirit of the empirical potentials, one can attempt to create an MLP without knowing in advance what scientific questions it will be used to address, what simulations will be carried out with it and (ultimately) what local chemical environments will be present in those simulations. This is a harder task, and limits still have to be placed (and explicitly communicated!) on the range of chemical environments that are “covered”. There have been some successes, notably for elemental systems, such as carbon [48,70] and silicon [47,71], as well as also organic molecules [36] and liquids like bulk [30,72] and interfacial [73] water.

The key for successful applications is the inclusion of the relevant atomic environments in the reference data set. Since a systematic mapping of high-dimensional configurations on a regular grid is a hopeless endeavor, strategies yielding the chemically most important (i.e., accessible) and diverse structures are needed. Often, ab initio molecular dynamics simulations are used as starting point, but this is usually insufficient as including rarely visited structures can be crucial for obtaining stable potentials. Hence, nowadays, active learning strategies, as used in the ML community for quite some time [74], have become a standard approach for the construction of MLPs [75–79]. This search can be complemented by selecting non-redundant structural features from a large number of structures [80–82]. In both approaches the identification of relevant structures is even possible before carrying out expensive electronic structure calculations, allowing to focus the available computing time on the most important information.

Another aspect to be considered when training MLPs is the calculation of the reference data by electronic structure methods. At the moment, in particular in the field of materials science, DFT is the dominant reference method since it offers a good compromise between accuracy and efficiency for many systems. Generalized Gradient Approximation (GGA) functionals still represent today’s workhorse methods, but more and more studies consider hybrid functionals, which are often much more expensive; on the other hand very accurate wave function methods have been used, but to date only for small molecular systems. This has two reasons: the high costs, but also the need for periodic boundary conditions in the fields of materials science and chemistry in solution, which are not generally available for many wave function methods. Hence, there is currently an accuracy gap for the reference methods between DFT and wave function methods, which in the future might be filled by methods like quantum Monte Carlo (QMC) [83], RPA [84] or coupled-cluster methods [85]. A challenge still to be solved for a routine use of QMC is the treatment of noisy data, as noise reduction for QMC is possible but comes with significantly increased computational costs.

An important remaining challenge is the construction of MLPs for systems containing many different elements. The configuration space inside the atomic environments grows rapidly with the number of elements, which is challenging for both, the sampling of these configurations by electronic structure calculations, but also for dealing with the increasing number of descriptors growing combinatorially with the number of elements. To address this problem, some solutions have been proposed. For instance, the configuration space can be reduced by decreasing the cutoff radius defining the atomic environments, like in the ANI-1 potential [36]. Here, the aim is to construct a MLP being transferable over a wide range of organic molecules necessitating a compromise in the accuracy that can be achieved. Alternatively, for specific problems in chemistry and materials science it may not be necessary to obtain very transferable potentials as for physical reasons many in

principle possible atomic configurations do not occur, thus reducing the configuration space that needs to be accurately represented by the potential. Consequently, it is close to impossible to give a generally valid number of elements that can be included in current MLPs, but as a rule of thumb constructing potentials containing three to four elements can be routinely done.

Concerning the descriptor, several suggestions have been made to break the combinatorial growth, e.g. by including the nuclear charge into the descriptor in weighted atom-centered symmetry functions (wACSF) [86] or by using combined structural and compositional descriptors [87]. In spite of some proof-of-principle benchmarks, to date the resulting sparse representations have not been frequently used and more work is required to investigate to what extent the number of descriptors can be reduced, because an accurate distinction of different structures is essential for the successful construction of MLPs. Interestingly, even for established descriptors like ACSFs the number of descriptors is typically smaller than the formal dimensionality of the atomic environments. Hence, a certain level of coarse graining the structural information seems to be tolerable, but the essential information must be retained. To identify the important descriptors, also automatic procedures for selecting descriptors from a large pool based on CUR [88] or the correlation of the descriptor values have been proposed [81]. Further, it has been suggested to determine the descriptor parameters using genetic algorithms [86].

There are essentially two different selection strategies for descriptors, as they could either be chosen based on their spatial shape characterizing the atomic environments or to achieve the best-possible description of a given specific reference data set. In early stages of the generation of data sets, unbiased descriptors should generally be preferred, while for large converged data sets a focus might shift on using the smallest possible number of descriptors needed for this data.

In summary, the construction of atomistic potentials has a long history, both in chemistry as well as in materials science, with traditionally different approaches to meet the requirements of the respective questions to be answered. With MLPs becoming more and more established, many groups have entered the field resulting in further rapid advances bringing these fields closer together. Although being less efficient than simple empirical force fields, MLPs allow to extend first-principles methods to problems where extensive sampling is required in particular for those systems in which similar structural patterns are repeatedly encountered.

Author contributions

Both authors contributed equally to this work.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: This is a review. There is no original research data to deposit. There is no other change to be made to the manuscript.]

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. M. Born, R. Oppenheimer, *Ann. Phys.* **389**, 457 (1927)
2. P.M. Morse, *Phys. Rev.* **34**, 57 (1929)
3. J.E. Jones, *Proc. R. Soc. Lond. A* **106**, 463 (1924)
4. A.C.T. van Duin, S. Dasgupta, F. Lorant, W.A. Goddard III, *J. Phys. Chem. A* **105**, 9396 (2001)
5. J. Behler, *Phys. Chem. Chem. Phys.* **13**, 17930 (2011)
6. J. Behler, *J. Chem. Phys.* **145**, 170901 (2016)
7. F. Noé, A. Tkatchenko, K.R. Müller, C. Clementi, *Ann. Rev. Phys. Chem.* **71**, 361 (2020)
8. V.L. Deringer, M.A. Caro, G. Csányi, *Adv. Mater.* **31**, 1902765 (2019)
9. V. Botu, R. Batra, J. Chapman, R. Ramprasad, *J. Phys. Chem. C* **121**, 511 (2017)
10. T. Carrington Jr., W.H. Miller, *J. Chem. Phys.* **84**, 4364 (1986)
11. S.N. Yurchenko, J. Zheng, H. Lin, P. Jensen, W. Thiel, *J. Chem. Phys.* **123**, 134308 (2005)
12. G.C. Schatz, *Rev. Mod. Phys.* **61**, 669 (1989)
13. J.M. Bowman, G. Czako, B. Fu, *Phys. Chem. Chem. Phys.* **13**, 8094 (2011)
14. B.J. Braams, J.M. Bowman, *Int. Rev. Phys. Chem.* **28**, 577 (2009)
15. J.M. Bowman, B.J. Braams, S. Carter, C. Chen, G. Czako, B. Fu, X. Huang, E. Kamarchik, A.R. Sharma, B.C. Shepler et al., *J. Phys. Chem. Lett.* **1**, 1866 (2010)
16. M.S. Daw, S.M. Foiles, M.I. Baskes, *Mater. Sci. Rep.* **9**, 251 (1993)
17. T.B. Blank, S.D. Brown, A.W. Calhoun, D.J. Doren, *J. Chem. Phys.* **103**, 4129 (1995)
18. S. Lorenz, A. Groß, M. Scheffler, *Chem. Phys. Lett.* **395**, 210 (2004)
19. M. Malshe, R. Narulkar, L.M. Raff, M. Hagan, S. Bukkapatnam, P.M. Agrawal, R. Komanduri, *J. Chem. Phys.* **130**, 184102 (2009)
20. S. Manzhos, T. Carrington Jr., *J. Chem. Phys.* **127**, 014103 (2007)
21. H. Gassner, M. Probst, A. Lauenstein, K. Hermansson, *J. Phys. Chem. A* **102**, 4596 (1998)

22. J. Behler, S. Lorenz, K. Reuter, J. Chem. Phys. **127**, 014705 (2007)
23. J. Behler, M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007)
24. J. Behler, J. Chem. Phys. **134**, 074106 (2011)
25. J. Behler, Angew. Chem. Int. Ed. **56**, 12828 (2017)
26. V.L. Deringer, A.P. Bartók, N. Bernstein, D.M. Wilkins, M. Ceriotti, G. Csányi, Chem. Rev. (2021) (accepted)
27. G. Sivaraman, A.N. Krishnamoorthy, M. Baur, C. Holm, M. Stan, G. Csányi, C. Benmore, A. Vazquez-Mayagoitia, NPJ Comput. Mater. **6**, 1 (2020)
28. Y.B. Liu, J.Y. Yang, G.M. Xin, L.H. Liu, G. Csányi, B.Y. Cao, J. Chem. Phys. **153**, 144501 (2020)
29. S. Yue, M.C. Muniz, M.F.C. Andrade, L. Zhang, R. Car, A.Z. Panagiotopoulos, J. Chem. Phys. **154**, 034111 (2021)
30. T. Morawietz, A. Singraber, C. Dellago, J. Behler, PNAS **113**, 8368 (2016)
31. A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010)
32. A.V. Shapeev, Multiscale Model. Simul. **14**, 1153 (2016)
33. A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, G.J. Tucker, J. Comp. Phys. **285**, 316 (2015)
34. R. Drautz, Phys. Rev. B **99**, 014104 (2019)
35. H. Wang, L. Zhang, J. Han, W. E, Comput. Phys. Commun. **228**, 178 (2018)
36. J.S. Smith, O. Isayev, A.E. Roitberg, Chem. Sci. **8**, 3192 (2017)
37. A.P. Bartók, R. Kondor, G. Csányi, Phys. Rev. B **87**, 184115 (2013)
38. F. Musil, A. Grisafi, A.P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Physics-inspired structural representations for molecules and materials* (2021). [arXiv:2101.04673](https://arxiv.org/abs/2101.04673)
39. J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, *Neural Message Passing for Quantum Chemistry*, in *Proceedings of the 34th International Conference on Machine Learning*, edited by D. Precup, Y.W. Teh (PMLR, 2017), Vol. 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272
40. S. Batzner, T.E. Smidt, L. Sun, J.P. Mailoa, M. Kornbluth, N. Molinari, B. Kozinsky (2021). [arXiv:2101.03164](https://arxiv.org/abs/2101.03164)
41. K.T. Schütt, H.E. Sauceda, P.J. Kindermans, A. Tkatchenko, K.R. Müller, J. Chem. Phys. **148**, 241722 (2018)
42. O.T. Unke, M. Meuwly, J. Chem. Theory Comput. **15**, 3678 (2019)
43. R. Zubatyuk, J.S. Smith, J. Leszczynski, O. Isayev, Sci. Adv. **5**, eaav6490 (2019)
44. B. Anderson, T.S. Hy, R. Kondor (2019). [arXiv:1906.04015v3](https://arxiv.org/abs/1906.04015v3)
45. K.T. Schütt, O.T. Unke, M. Gastegger (2021). [arXiv:2102.03150v2](https://arxiv.org/abs/2102.03150v2)
46. V.G. Satorras, E. Hoogeboom, M. Welling, *E(n) equivariant graph neural networks* (2021). [arXiv:2102.09844](https://arxiv.org/abs/2102.09844)
47. J. Behler, R. Martoňák, D. Donadio, M. Parrinello, Phys. Rev. Lett. **100**, 185501 (2008)
48. V.L. Deringer, G. Csányi, Phys. Rev. B **95**, 094203 (2017)
49. T.W. Ko, J.A. Finkler, S. Goedecker, J. Behler, Acc. Chem. Res. **54**, 808 (2021)
50. B. Parsaeifard, J.A. Finkler, S. Goedecker (2020). [arXiv:2008.11277](https://arxiv.org/abs/2008.11277)
51. T.W. Ko, J.A. Finkler, S. Goedecker, J. Behler, Nat. Commun. **12**, 398 (2021)
52. S. Houlding, S.Y. Liem, P.L.A. Popelier, Int. J. Quantum Chem. **107**, 2817 (2007)
53. Z. Deng, C. Chen, X.G. Li, S.P. Ong, NPJ Comput. Mater. **5**, 75 (2019)
54. N. Artrith, T. Morawietz, J. Behler, Phys. Rev. B **83**, 153101 (2011)
55. T. Morawietz, V. Sharma, J. Behler, J. Chem. Phys. **136**, 064103 (2012)
56. S. Grimme, J. Antony, S. Ehrlich, H. Krieg, J. Chem. Phys. **132**, 154104 (2010)
57. K. Yao, J.E. Herr, D.W. Toth, R. Mckintyre, J. Parkhill, Chem. Sci. **9**, 2261 (2018)
58. D.P. Metcalf, A. Jiang, S.A. Spronk, D.L. Cheney, C.D. Sherrill, J. Chem. Inf. Model. **61**, 115 (2021)
59. R. Zubatyuk, J.S. Smith, B.T. Nebgen, S. Tretiak, O. Isayev, ChemRxiv (2020). <https://doi.org/10.26434/chemrxiv.12725276.v2>
60. B.K. Rai, G.A. Bakken, J. Comp. Chem. **34**, 1661 (2013)
61. T. Bereau, D. Andrienko, O.A. von Lilienfeld, J. Chem. Theory Comput. **11**, 3225 (2015)
62. A.E. Sifain, N. Lubbers, B.T. Nebgen, J.S. Smith, A.Y. Lokhov, O. Isayev, A.E. Roitberg, K. Barros, S. Tretiak, J. Phys. Chem. Lett. **9**, 4495 (2018)
63. M. Gastegger, J. Behler, P. Marquetand, Chem. Sci. **8**, 6924 (2017)
64. A. Grisafi, M. Ceriotti, J. Chem. Phys. **151**, 204105 (2019)
65. S.A. Ghasemi, A. Hofstetter, S. Saha, S. Goedecker, Phys. Rev. B **92**, 045131 (2015)
66. A.K. Rappe, W.A. Goddard III, J. Phys. Chem. **95**, 3358 (1991)
67. X. Xie, K.A. Persson, D.W. Small, J. Chem. Theory Comput. **16**, 4256 (2020)
68. M. Veit, S.K. Jain, S. Bonakala, I. Rudra, D. Hohl, G. Csányi, J. Chem. Theory Comput. **15**, 2574 (2019)
69. A.M. Reilly, A. Tkatchenko, J. Phys. Chem. Lett. **4**, 1028 (2013)
70. P. Rowe, V.L. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, J. Chem. Phys. **153**, 034702 (2020)
71. A.P. Bartók, J. Kermode, N. Bernstein, G. Csányi, Phys. Rev. X **8**, 041048 (2018)
72. B. Cheng, E.A. Engel, J. Behler, C. Dellago, M. Ceriotti, PNAS **116**, 1110 (2019)
73. V. Quaranta, M. Hellström, J. Behler, J. Phys. Chem. Lett. **8**, 1476 (2017)
74. H.S. Seung, M. Opper, H. Sompolinsky, In *Proceedings of the fifth annual workshop on computational learning theory*, pp. 287–294 (1992)
75. N. Artrith, J. Behler, Phys. Rev. B **85**, 045439 (2012)
76. E.V. Podryabinkin, A.V. Shapeev, Comp. Mater. Sci. **140**, 171 (2017)
77. L. Zhang, D.Y. Lin, H. Wang, R. Car, W. E, Phys. Rev. Mater. **3**, 023804 (2019)
78. C. Schran, J. Behler, D. Marx, J. Chem. Theory Comput. **16**, 88 (2020)
79. N. Bernstein, G. Csányi, V.L. Deringer, NPJ Comput. Mater. **5**, 1 (2019)
80. A.P. Bartók, S. De, C. Poelking, N. Bernstein, J.R. Kermode, G. Csányi, M. Ceriotti, Sci. Adv. **3**, e1701816 (2017)

81. G. Imbalzano, A. Anelli, D. Giofre, S. Klees, J. Behler, M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018)
82. S. Stocker, G. Csányi, K. Reuter, J.T. Margraf, Nat. Commun. **11**, 1 (2020)
83. W.M.C. Foulkes, L. Mitas, R.J. Needs, R. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001)
84. V.L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D.A. Drabold, S.R. Elliott, Nature **589**, 59 (2021)
85. I.Y. Zhang, A. Grüneis, Front. Mater. **6**, 123 (2019)
86. M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsényi, P. Marquetand, J. Chem. Phys. **148**, 241709 (2018)
87. N. Artrith, A. Urban, G. Ceder, Phys. Rev. B **96**, 014112 (2017)
88. M.W. Mahoney, P. Drineas, PNAS **106**, 697 (2009)