Machine Learning Interatomic Force Fields for Carbon Allotropic Materials

Xiangjun Liu, Quanjie Wang, and Jie Zhang

1 Introduction

Carbon is capable of forming many allotropes due to its valency. The well-known forms of carbon include diamond and graphene. In the past decades, many more allotropes have been further discovered and researched, such as carbon nanotube, nanotubes, and buckminsterfullerene. At the present time, around 500 hypothetical 3-periodic carbon allotropes are known; each of them exhibits significantly different properties and wide potentials of applications. Recently, due to the enhancement of high performance computing (HPC) power, and algorithmic improvements, computational materials science has gradually become an important supplement to traditional theory and experiment for the study of carbon allotropes, as well as a crucial bridge between micro and macro, theory and experiment [1–3]. On one hand, it can help us to understanding the microstructure and behavior of carbon materials in atomic level, and on the other hand, it can predict the properties and formation mechanism of carbon materials without doing experiments.

2 Traditional Force Fields

Molecular dynamic (MD) simulation is most widely used for studying atomistic systems, which can monitor the atomic-level time-evolution of physical and chemical processes and predict macroscopic properties from microscopic details. Starting with an initial atomic locations and velocities, MD simulations require the atomic

X. Liu (✉) · Q. Wang · J. Zhang
School of Mechanical Engineering, Donghua University, Shanghai, China
e-mail: xjliu@dhu.edu.cn; mezhangjie@dhu.edu.cn

Y. Cheng et al. (eds.), *Artificial Intelligence for Materials Science*, Springer Series in Materials Science 312, https://doi.org/10.1007/978-3-030-68310-8_4

forces as input to propagate the atoms locations and their velocities to the next timestep (at which point, the atomic forces are reevaluated), the cycle continues, thus allowing for an iterative time-evolution of the system. The atomic forces at each timestep may be obtained either using quantum mechanics (QM) based methods, such as density functional theory (DFT), or parameterized classical semiempirical analytical interatomic force fields, such as Stillinger-Weber potentials, Tersoff potential, and so forth [4–6]. Choosing between the two approaches depends on which side of the cost-accuracy trade-off ones wishes to be at. QM methods (also referred to as ab initio or first-principles methods) are versatile and offer the capability to accurately model a range of chemistries and chemical environments by solving for the Schrodinger equation. However, the computational complexity of QM methods is at least cubic in the number of electrons; consequently, practical applications of these methods at present are limited to studies of phenomena whose typical length and time scales are of the order of nanometers and picoseconds, respectively [1]. Parameterized classical force fields can be used to access truly large-length and long-time scales, which typically are 6–10 orders of magnitude faster than DFT, because the influence of electrons is not taken into account in the calculation of atomic force [7]. However, these approaches are also problematic, as such force fields cannot precisely reproduce QM forces and have limited transferability; for instance, they are not transferable to situations that were not originally used in the parameterization [8–10]. Facing this scenario, it is necessary to develop novel and efficient force field. The advent of big-data analytics and easy to access to HPC resources has brought powerful machine learning (ML) techniques to the forefront. Meanwhile, ML methods hold promise in resolving the disconnect between force field developers and the end-users, which is common in classical potential function development, in other words, empowering the users to develop new or tailor existing force fields to meet their needs [11].

3 Machine Learning Force Field

Since US President Barack Obama proposed the genome project in 2010, the application cases of ML in material development have been emerging. For example, rapid search for high thermal conductivity materials [12], design of low interface thermal resistance superlattice structure [13], utilizing neural network assisted drug development [14], screening of high-throughput materials [15], prediction of material structure [16], design of ultra-high hard materials [17], and so forth.

ML mainly uses a trained model (such as neural network algorithm, Bayesian optimization algorithm, and random forest) to extracting information from large historical datasets (from experiments, simulations, online database, etc.), and then accurately capture the relationship between structures and properties by data mining techniques for materials discovery and properties prediction [4]. Recently, a data-driven and ML-based atomic force field development research has attracted wide concern due to its flexibility and adaptability. In contrast to conventional interatomic

potentials and QM-based methods, the ML-based paradigm has been verified by many groups is a feasible pathway in the creation of interatomic force field that both has the accuracy and versatility of QM methods and the low computing cost of parameterized semiempirical interatomic potentials. The ML-based force field was first proposed independently by Botu [5] and Li [18]; they used vector structural descriptors as fingerprints of atomic environments, and separately learned individual force components using kernel regression and Gaussian process regression. Recently, Glielmo et al. [4] proposed a novel scheme, which predicts the forces as vector quantities using Gaussian process regression. Additionally, they added the many-body kernel to represent the dependence of force not only on interatomic distance but also bone angle. Because many more force components as a training dataset can be obtained from DFT calculation directly, the construction of the ML force field is easier than that of ML potentials. ML force field has been successfully constructed for many elemental materials, such as Al [4, 5, 19, 20], Si [21], and Cu [12], and a few multicomponent materials, such as SiO_2 [22]. Moreover, the feasibility of ML force field has been verified by several static and dynamic applications, including melting, stress-strain behavior, point defect diffusion in bulk, proper description of dislocation core regions, metal phase transition and adatom organization as surface, and so forth [4, 22].

Traditionally, the parameters in classical semiempirical interatomic potential are obtained by fitting to QM calculations or experimental data under equilibrium state, as the red dotted line region shown in Fig. 1a, therefore, overemphasis on equilibrium configurations often result in performs poorly in predicting the relative energies in transition state or far-from-equilibrium position. However, ML-based force field can overcome this issue by learning from reference datasets in the whole potential energy space, as shown in Fig. 1b.

The application of ML in molecular force field development largely falls into two broad categories: one approach is based on classical semiempirical analytical interatomic potentials, employing ML algorithms to optimize the potential parameters, namely, ML-based optimized force field. The other approach is to establish nonlinear mapping by ML models between atomic configurations and potential energies or force, which has no fixed mathematical functional form, namely,

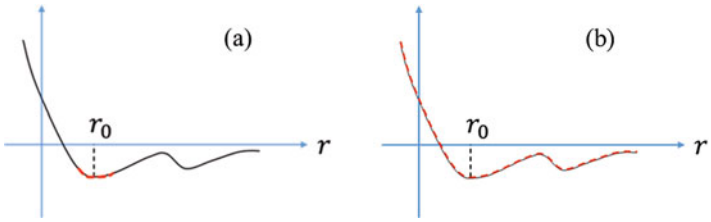

Fig. 1 The data obtained used for constructing atomic force fields: (**a**) classical semiempirical force field, (**b**) ML-based force field

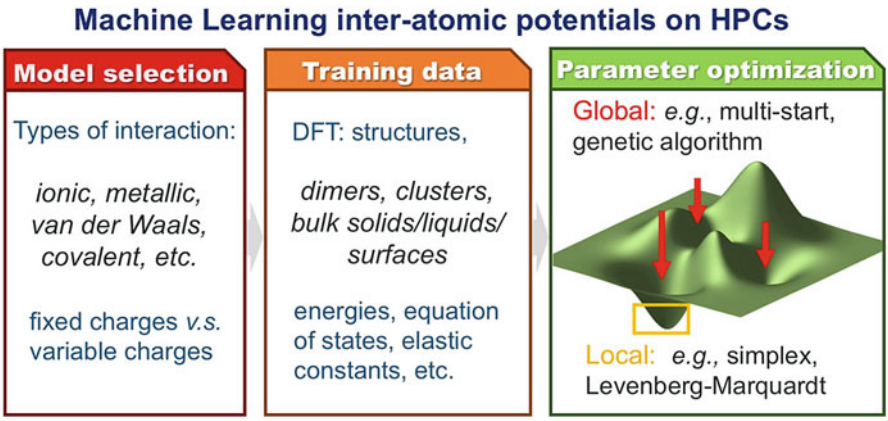

Fig. 2 An overview of ML-based potential function parameters optimization framework. (Adapted with permission from ref. [7], copyright 2019 Physical Chemistry publishing)

ML-based force field. The train datasets of these two approaches are both obtained from DFT calculation, which can ensure the accuracy of prediction [23–27].

Chan et al. [7] proposed a ML-based optimized force field to accurately simulate the dynamical process at reactive interfaces and low dimensional system, such as clusters and molecules. The procedure as shown in Fig. 2, which involves (1) defining or selecting a functional form, the functional from selection apart from the material being studied but also strongly dependent on the phenomenon being explored, (2) constructing an extensive training data set from electronic structure calculations, and the training as far as possible to encompassing all possible atomic environments and coordinates likely to be encountered in dynamic simulations, (3) optimizing force-field parameters using ML algorithms, such as genetic algorithms (GA), formulating a fitting procedure and implementing these algorithms on HPCs.

For ML-based force field, the force field is transferable and adaptive due to overcoming the limitations result from the predefined mathematical functional form. For instance, new reference configurations can be added to enhance the versatility of the force field as required [28–31]. A typical ML-based force field development workflow mainly consists of four key steps, which are: (1) generation of reference data, such as, using DFT; (2) fingerprinting or quantifying the atomic environments, in a manner that will allow the fingerprint as input in regression model; (3) choosing a subset from the reference data set, using clustering and sampling techniques to reduce the learning cost while ensuring that the dataset retain the diversity of the original reference data set; (4) learning from the training set, thus construct a nonlinear mapping between the atomic configurations and the forces, followed by testing the learned model on the remainder of the data set. The entire framework involved in the construction of force field is portrayed schematically in Fig. 3.

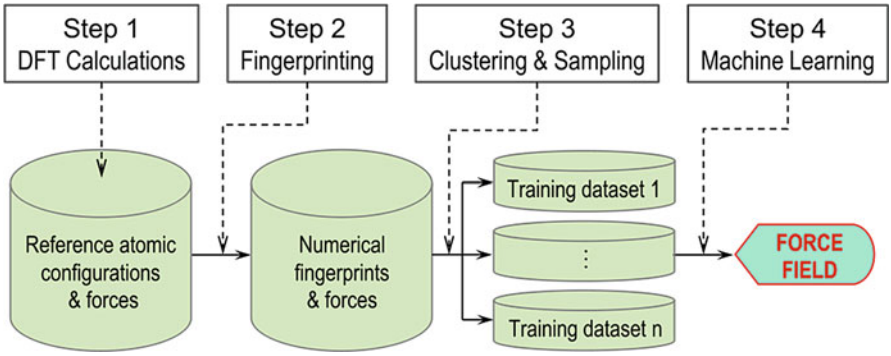

Fig. 3 Workflow for the creation of machine learning force field. (Adapted with permission from ref. [12], copyright 2017 NPJ Computational Materials publishing)

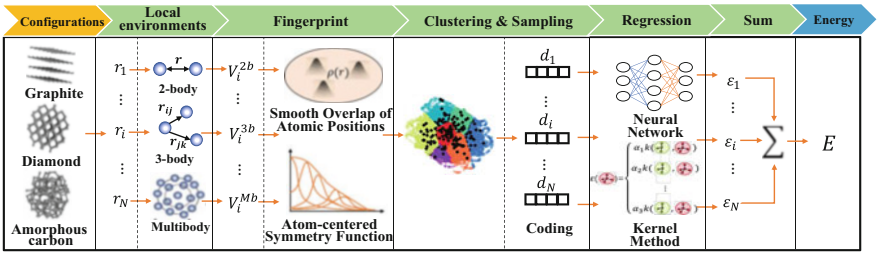

Fig. 4 ML-based force field framework for carbon allotropic

4 Procedure to Develop ML-Based Force Field

A ML-based force field framework for carbon allotropic is given in Fig. 4. First, constructing a comprehensive reference database as possible, in addition to commonly encountered crystalline phases, diamond and graphite, other relevant phase of carbon should also be taken in account, such as amorphous carbon. In order to facilitate the input of regression model, a proper fingerprint or descriptor should be selected to quantifying the local environment of an atom, the local environment of an atom typical divide into 2-body, 3-body, and many-body. Next, a cluster and sample algorithm employed to identify the redundant and noncontributing data to reduce computing costs. Finally, selecting a regression model to establish the fingerprint-force mapping, the energy E is further approximated as a sum of the atomic energies

$$E = \sum_i \varepsilon\left(d_i\right) \tag{1}$$

where d_i is the feature vector of atom i, which accounts for the chemical environment of atom i that depends on positions and chemical identities of its neighboring atoms up to a given cutoff radius. The ε represents the atomic energy as a function of descriptors. The details of each key step will be described next.

4.1 Reference Database

The reference data used for creating force field must be as accurate and comprehensive as possible, ensuring sufficiently low intrinsic errors. Generally, calculations (such as first-principles, molecular dynamics, lattice dynamics, and so forth) [32, 33], experiments [34–36] and online libraries [37, 38] have been used to collect these data. Among these approaches, first-principles is most convenient and quick approach to sample reference data for ML force field construction because abundant reference structures and corresponding quantum mechanical properties (i.e., energy and atomic force) can be directly obtained from one piece of an ab initio MD trajectory. For example, Li et al. [22] obtained abundant atomic configurations of Cu and SiO_2 by fast ab initio MD, such as face-centered cubic supercells, surface (111) supercells, surface (100) supercells, and amorphous supercells. Huan et al. further expanded the atomic configurations and corresponding force by rotating the collected atomic configurations, and providing more force components than that in the original dataset. To mimic the diverse environments an atom could exist in, Botu et al. [4] built several periodical and non-periodical equilibrium configurations, such as (a) defect free bulk, (b) surfaces, (c) point defects, vacancies and adatoms, (d) isolated clusters, (e) grain boundaries, (f) lattice expansion and compression, and (g) edge type dislocations, as shown in Fig. 5. To correctly describe the nonequilibrium behavior of an atom, initially atoms are randomly perturbed to coerce the dynamics into sampling nonequilibrium environments. The combination of ab initio MD and random perturbations resulted in a diverse set of reference atomic environments and forces.

4.2 Structural Fingerprints

Each atomic energy contribution depends only on its local environment, as shown is Fig. 6, which is represented by a feature space vector or fingerprint so as to numerically represent atomic configurations [31]. The accuracy of the ML force field will strongly depend on the selection of the fingerprint; a good fingerprint should possess the following properties: (1) it can be encoded as a fixed-length vector so as to facilitate regression, (2) it should retain basic physical symmetry invariant, such as translation, rotation, or permutation, (3) it is complete, i.e., different atomistic neighborhood configurations lead to different fingerprints and vice versa, and the "distance" between the fingerprints should be proportional

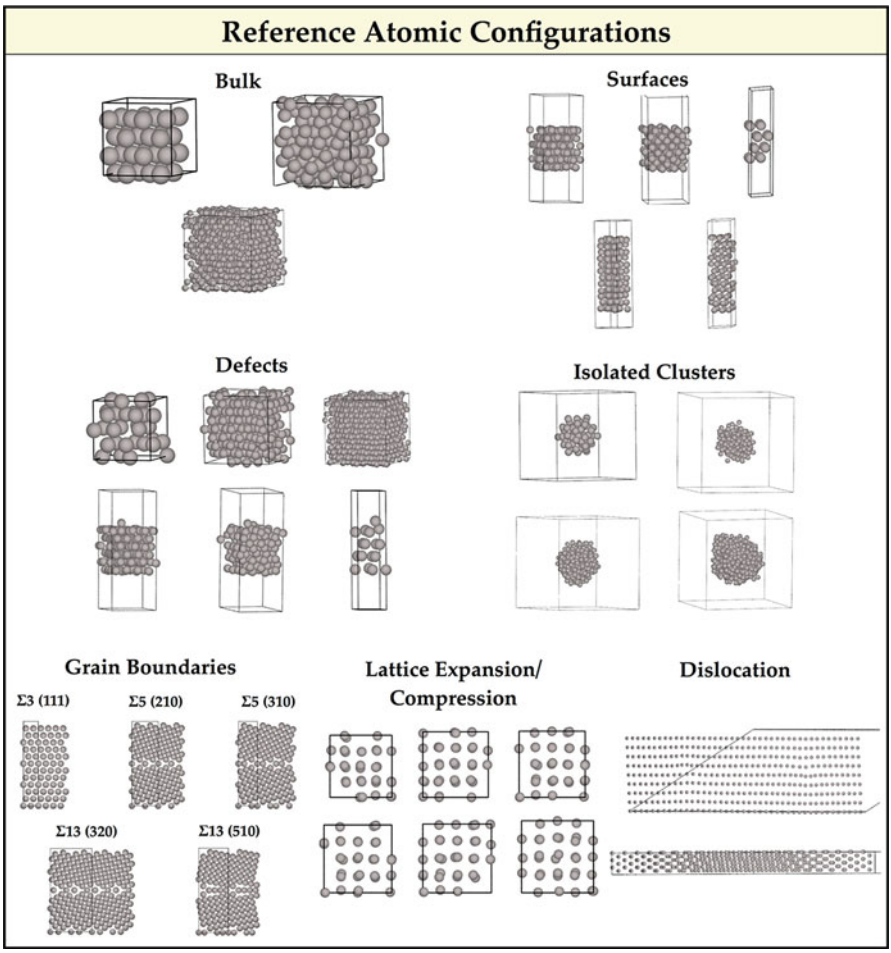

Fig. 5 Reference configurations used to sample atomic environments for training and testing of force field. (Adapted with permission from ref. [4], copyright 2017 Physical Chemistry C publishing)

to the intrinsic difference between the atomistic neighborhood configurations. A number of structural descriptors have been proposed to represent the local atomic environment, such as atom-centered symmetry functions [39], bispectrum [40], and Smooth Overlap of Atomic Positions (SOAP) kernel [41]. The most widely used descriptor is the vector atomic fingerprint function, which was proposed by Botu and Ramprasad, and has been proven to be an effective structural descriptor in the prediction of vectorial atomic properties [4]. Using the fingerprint, the atomic environment of the *i*th atom in a specific atomic configuration can be represented by

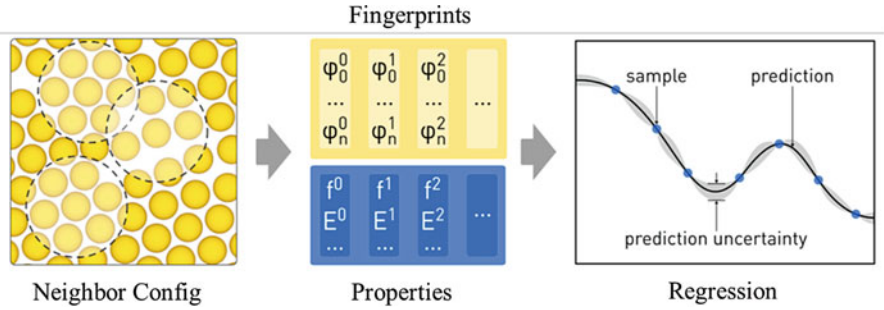

Fig. 6 Transform atomistic neighborhood configurations into feature vectors and train non-liner regression models. (Adapted with permission from ref. [1], copyright 2018 Chemical Physics publishing)

$$V_i^{\alpha} = \sum_j \frac{r_{ij}^{\alpha}}{r_{ij}} \mathrm{e}^{-\left(\frac{R_{ij}}{\eta}\right)^2} \times f_{\mathrm{c}}\left(r_{ij}\right) \tag{2}$$

where r_{ij} signifies the distance between atoms i and j, and r_{ij}^{α} is a scalar projection of this distance along the α direction ($\alpha = x$, y or z). η is a parameter that controls the decay rate, and f_{c} is the cutoff function that gradually reduces the contribution of distant atoms and truncates the interatomic interaction when r_{ij} is larger than cutoff distance R_{c}.

Although Eq. (2) has been proved to be very effective in various materials, it ignores bond angle information, which might be insufficient for complex covalent materials. For this, Li et al. [22] have modified Eq. (2) and proposed another structural fingerprint that takes the bond angle into consideration. The formulas of the two structural fingerprints are:

$$V_i^{1,\alpha} = \sum_j \frac{r_{ij}^{\alpha}}{r_{ij}} \mathrm{e}^{-\eta\left(r_{ij}-R_{\mathrm{s}}\right)^2} \times f_{\mathrm{c}}\left(r_{ij}\right) \tag{3}$$

$$V_i^{2,\alpha} = 2^{1-\zeta} \sum_j \sum_k \left(\overrightarrow{r_{ij}} + \overrightarrow{r_{ik}}\right)^{\alpha} \left(1 + \cos\left(\theta_{ijk} - \theta_{\mathrm{s}}\right)\right)^{\zeta} \times \mathrm{e}^{-\eta\left(\frac{r_{ij}+r_{ik}}{2}-R_{\mathrm{s}}\right)^2} \times f_{\mathrm{c}}\left(r_{ij}\right) \times f_{\mathrm{c}}\left(r_{ik}\right) \tag{4}$$

Equations (3) and (4) are called a radial structural fingerprint and angular structural fingerprint, respectively, where r_{ik} and r_{ij} are the interatomic distances between i and k, and i and j, respectively. θ_{ijk} is the angle between bonds ij and ik. $(r_{ij} + r_{ik})^{\alpha}$ is the scalar projection of vector $(r_{ij} + r_{ik})$ along the α direction. Two parameters of radial fingerprint $V_i^{1,\alpha}$, that is, η and R_{s}, are used to control the

width of the peak and shift the peak position. Two additional parameters, that is, ζ and θ_s, are used in angular fingerprint $V_i^{2,\alpha}$. Applying the θ_s parameter allows the probing of specific regions of the angular environment in a similar manner to that accomplished by R_s in the radial part. Finally, the ML force fields were compared with DFT and MD simulations in structural optimization, it is found that the proposed angular fingerprints can significantly improve the accuracy of ML force fields for both Cu and SiO_2.

To capture the transition state during the structural phase transformation, many-body term needs to be taken into consideration. For example, Zong et al. [42] adopt three different types of local environments related to structural phase transformations are fingerprinted, namely, the change in bond length (pairwise terms), shape change (three-body terms), as well as volume change (many-body terms). For two- and three- terms, which is similar to the above treatment, so we will not explain any more here. For the many-body contributions, which is similar to the embedding energy term of the MEAM potential. The formulas is

$$V_i^{Mb}(\mu,\sigma) = \ln\left(\rho_i^m(\mu,\sigma)\right) \tag{5}$$

where μ and σ are adjustable parameters, $\rho_i^m(\mu,\sigma)$ refers to the neighborhood density of a given atom i, define as

$$\rho_i^m(\mu,\sigma) = \sum_{j\neq i} \mathrm{e}^{-\frac{(r_{ij}-\mu)^2}{\sigma^2}} f_c\left(r_{ij}\right) \tag{6}$$

4.3 Sampling and Clustering

The next step in the construction workflow is to select a representative set of atomic environments for training purposes. To do so, it is necessary to identify the redundant and noncontributing data points from within the millions sampled. Random selection of training data from reference data is the most common approach, which typically results in the selected training configurations dominated by the high-populated domains while other domains are under-represented [5]. As shown in Fig. 7a, the training data randomly selected from the reference data contains essentially no configuration with large amplitude forces. To ensure the diversity of reference data, Huan et al. [12] proposed force-binning and clustering training data selection methods. In the force-binning method, the reference data was arranged into a number of force amplitude intervals and then the training data was selected from all the intervals, as shown in Fig. 7b. In the clustering approach, the reference data divide into a given number of clusters in fingerprint space and then the training data is selected from each cluster, as shown in Fig. 7c.

Another widely used method is dimensionality reduction techniques, such as principal component analysis (PCA) to project V_i^d onto a lower dimension space.

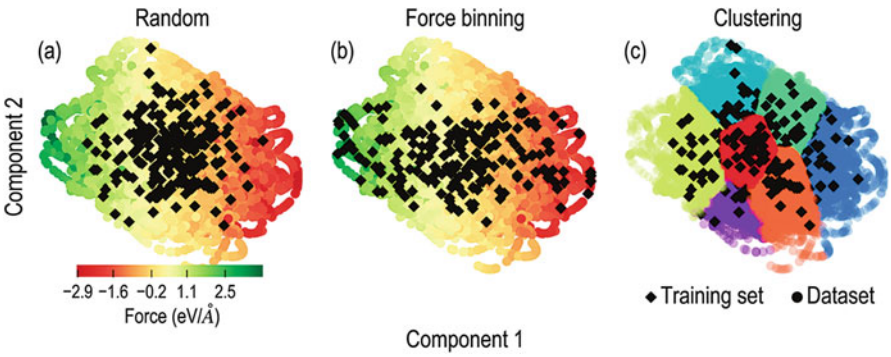

Fig. 7 An illustration of three methods for selecting a training set, including random (**a**) force amplitude sampling (**b**) and fingerprint space clustering (**c**). (Adapted with permission from ref. [12], copyright 2017 NPJ Computational Materials publishing)

In PCA the original atomic fingerprint is linearly transformed into uncorrelated and orthogonal pseudo variables. For example, Chapman et al. [43] captured more than 99% of the original fingerprint information by adopting such strategy. Moreover, some other similar dimension reduction techniques could be adopted to select representative data set, such as kernel-PCA or multidimensional scaling. Recently, using least absolute shrinkage and selection operator (LASSO) or genetic algorithm (GA) to select the important fingerprints from a large pool of candidates also have been proposed, which can have good balance between computational cost and accuracy.

4.4 Machine Learning

Once the reference data and atomic representation are in place, the final step is to carry out a learning algorithm to establish the fingerprint-force mapping. In the taxonomy of ML approaches, this is a "supervised learning" problem, because the input data (structures) are labeled (have reference energies); more specifically, it represents a regression task, because a continuous range of output values (energies) is sought. At present, various machine learning algorithms have been used in force field development, such as Kernel-based methods, linear model (LM), neural network model (NNM), and so forth. Next, we will give a brief introduction.

4.4.1 Kernel Ridge Regression

KRR is a powerful method that has widely been used in materials informatics, in which an atomic property is interpolated as a liner combination of kernel functions, as shown in Fig. 8, the latter measuring how similar a new configuration's descriptor

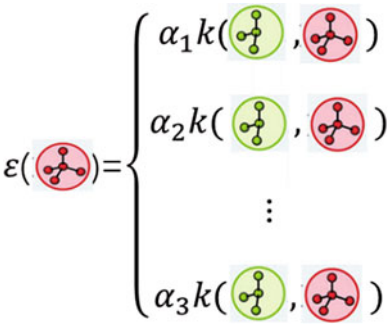

Fig. 8 Schematic of kernel methods to interpolate atomic properties by comparing an environment (red) with the reference database (green)

(red) is to those of the reference data (green). The property is typically a local energy, or a force acting on an atom, and the kernels can be understood as similarity measures (on a scale from zero to one) between the new environment and those contained in the database, both of which are represented by the descriptor. The regression coefficients that weigh each kernel basis function are computed during the fitting using simple linear algebra. Kernel ridge regression (KRR) and Gaussian Process Regression (GPR) are two currently employed techniques, differing only slightly in how these coefficients are computed.

Botu et al. [4] choose this method as the ML workhorse created a force field for six element bulk solids, including Al, Cu, Ti, W, Si, and C, and show that all of them can reach chemical accuracy. KRR predicts the atomic force F_i corresponding to the configuration i as

$$F_i = \sum_{j=1}^{N_t} a_j \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{\sigma}\right)^2\right] \tag{7}$$

where the sum runs ergodic N_t configurations in fingerprint space, d_{ij} is the "distance" between configurations i and j, here, refer to Euclidean norm. The "length" scale in this space is specified by σ.

4.4.2 Linear Model

LM is developed to describe the linear dependence between structural fingerprints and forces due to its simplicity and speed. For example, Li et al. [22] choose liner regression model to construct ML force field and compared with DFT calculation in both element (Cu) and binary (SiO_2) materials. They found that the force prediction error less than 0.1 eV/Å^{-1}. The LM takes the form

$$F_i^{\alpha} = w_1 V_i^{\alpha,(1)} + w_2 V_i^{\alpha,(2)} + \cdots + w_n V_i^{\alpha,(n)} \tag{8}$$

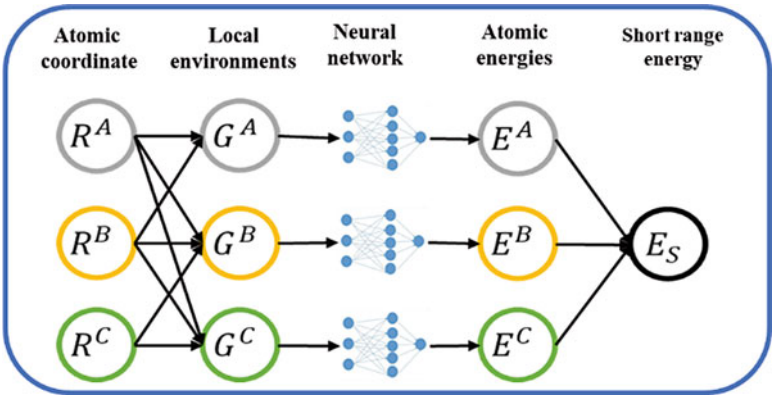

Fig. 9 High-dimensional neural network for a ternary system containing elements (**a–c**). $R^* = \left\{R_1^*, R_2^*, \cdots, R_{N_*}^*\right\}$, $G^* = \left\{G_1^*, G_2^*, \cdots, G_{N_*}^*\right\}$, $E^* = \left\{E_1^*, E_2^*, \cdots, E_{N_*}^*\right\}$, * = {A, B, C}

where w_1 is the regression coefficient, which is typically determined quickly using a standard least-squares technique. When matrix V includes the fingerprints of the reference atomic environments and F denotes the atomic forces obtained by DFT, the residual sum of squares $(\|V_w - F\|_2)^2$ is minimized in the linear regression, where $\| \bullet \|_2$ denotes the L_2 norm.

4.4.3 Neural Network Model

NNs are a set of mathematical functions that aim at resembling the functionality of neurons in brain, which was first proposed by Pro. Dr. J. Behler [3] in 2007, the structure as shown in Fig. 9, which represent a high-dimensional NN for a ternary system containing elements A, B, and C. The numbers of atoms per element are N_A, N_B, N_C, respectively. The total short-range energy E_s is the sum of all atomic energies E_i^X $(X = A, B, C)$, which are provided by individual atomic NNs. For a given element, the architecture and parameters of the atomic NNs are the same. The symmetry function vectors G_i^X provide the information about the local chemical environments of the atoms to the atomic NNs. Consequently, G_i^X depends on the Cartesian position vectors R_i^X of all the atoms within the cutoff spheres, which is represented by the black arrows. In such a method, a local atomic environment was described by generalized symmetry functions. NN potentials have been developed for many materials, such as Si, C, Cu, ZnO, TiO_2, H_2O dimers, Li_3PO_4, Cu clusters supported on Zn oxide and Au/Cu nanoparticles with water molecules. Additionally, they have been used to simulate the atom diffusion, phase transition, and search for equilibrium structures with not only MD but also the nudged elastic band (NEB) method, Monte Carlo methods and metadynamics.

5 Applications of ML Force Fields for Carbon Allotropes

5.1 ML Force Field for Graphene

Graphene has been the subject of extensive investigation since it was first isolated due to its interesting phenomena, such as the phonon-assisted diffusion of small molecules on the graphene surface, the study of thermal transport, and the incorporation of nuclear quantum effects into simulations which would benefit greatly from a highly accurate graphene model. Recently, ML-based force fields for graphene have emerged and attracted intensive attention. For example, Rowe et al. [44] constructed such an accurate interatomic potential for graphene using the GAP ML methodology. In this work, the total energy was decomposed into a sum of two-, three-, and many-body interactions, which are weighted based on their respective statistically measured contributions, the order of descriptors used in each term as follows: two atoms distance, symmetry functions, SOAP. Finally, to evaluating the accuracy of the ML model, compare the capabilities with those of empirically constructed potentials. As shown in Fig. 10, which shows the forces prediction ability of graphene GAP model and a number of other popular methods compared to the reference DFT method, black points indicate forces perpendicular to the plane of the graphene sheet (out-of-plane) while red points indicate forces oriented in the plane. The inset in the graphene GAP plot has a different scale on the y axis to show more clearly the distribution of force errors, it is clear that the predictions of the graphene GAP model align very closely with the reference DFT method. Additionally, the author also calculated the lattice parameters and in-plane thermal expansion of graphene using the developed GAP model and compared predictions of the finite temperature phonon spectra of graphene with experimental results.

In addition, Wen et al. [45] presented a hybrid potential that employs a neural network to describe short-range interactions and a theoretically motivated analytical term to model long-range dispersion for multilayer graphene. This potential can provide accurate energy and forces for both intralayer and interlayer interactions, correctly reproducing DFT results for structural, energetic, and elastic properties. Subsequently, Wen et al. [46] proposed another dropout uncertainty neural network potential for carbon and showed that it can be used to predict the stress and phonon dispersion in graphene.

5.2 ML Force Field for Diamond

As a significant member in carbon allotropes, diamond has been a hot spot in scientific research due to its fantastic mechanical and thermal properties. To better understand the atomistic behavior in diamond, an accuracy carbon potential is crucial. In 2017, Deringer et al. [47] reported a GAP model trained primarily on the

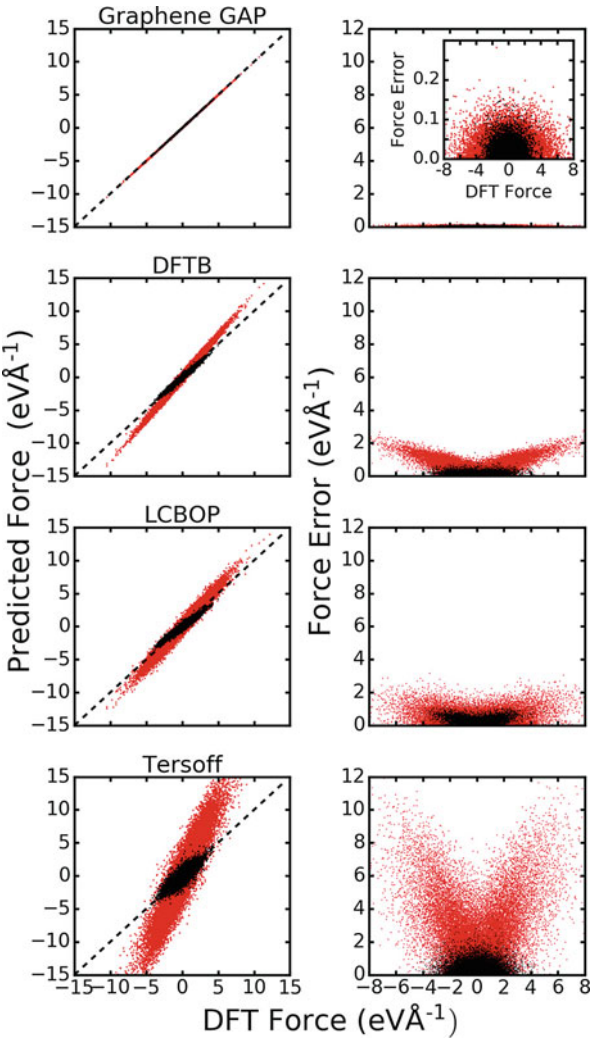

Fig. 10 The force prediction ability of graphene GAP model, DFTB, LCBOP, and Tersoff potentials compared to the reference DFT method. (Adapted with permission from ref. [44], copyright Physical Review B 2018 publishing)

amorphous and liquid phases of carbon based on DFT-local-density approximation reference data. Subsequently, Rowe and Deringer [48] proposed another improved GAP model on the basis of previous work, in which a large number of new configurations and exotic carbon allotropes are considered, such as nanotubes, cubic and hexagonal diamond and fullerene, as shown in Fig. 11a. To obtain more comprehensive database meanwhile keep the computational effort at the fitting stage tractable, they combined the farthest point sampling method with a number of mandatory configurations chosen using chemical intuition. The structural fingerprint of a configuration is quantified by SOAP descriptors. During model validation stage, they present an extensive and rigorous testing of GAP model for a wide range

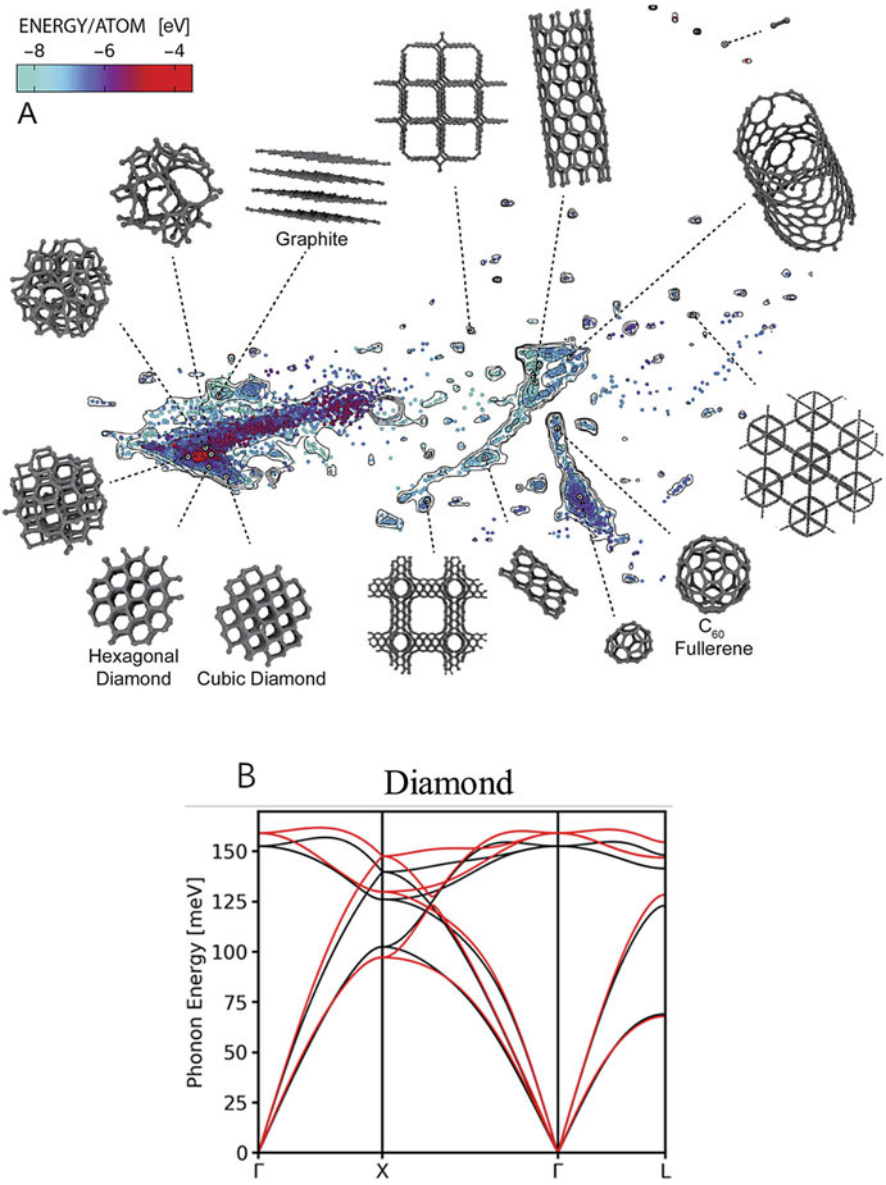

Fig. 11 (**a**) The selected configurations, as well as a representation of their position in phase space. (**b**) Phonon dispersion relation for diamond as predicted by GAP (black) with comparison to DFT reference data (red). (Adapted with permission from ref. [48], copyright 2020 Chemical Physics publishing)

of properties, as well as compare the results of GAP model to commonly used empirical potentials. As shown in Fig. 11b, the phonon dispersion of diamond is predicted successfully. In addition, the improved GAP model also correctly predicts the formation energies of diamond, graphite, fullerenes, and nanotubes, to an accuracy of a few meV, and achieves comparable accuracy for a number of crystalline and amorphous surfaces. The computed formation energies of defects are also accurate, with overall errors significantly lower than those obtained from comparable empirical models. Early, Rustam et al. [49] used an ab initio quality neural-network potential for large-scale simulations of the graphite-to-diamond transition assuming that it occurs through nucleation. The nucleation mechanism accounts for the observed phenomenology and reveals its microscopic origins. Other ML-based potentials for graphite-diamond phase study can be seen in ref. [50–52].

5.3 ML Force Field for Amorphous Carbon

The atomic structures of amorphous carbon samples depend strongly on density and are characterized by the coexistence of threefold ("sp^2") and fourfold bonded ("sp^3") carbon atoms; low- and high-density forms of amorphous carbon are loosely reminiscent of graphite and diamond, respectively. Deringer et al. [53] combined ML and DFT obtained new atomistic insight into carbonaceous energy materials. They started by modeling nanoporous carbons as used in supercapacitors. Using GAP, which has been "trained" with DFT data to fit energies and forces for amorphous and partly graphitized configurations as well as bulk graphite, they found the structural fingerprint of carbons is their atomic coordination relating to the local bonding ("sp/sp^2/sp^3"). Finally, the accuracy of GAP was tested specifically for snapshots from annealing trajectories, as shown in Fig. 12a; it achieves an energy accuracy within 2 kJ mol^{-1} of DFT data but completes the task several orders of magnitude faster. During annealing, the sp^2 count in the model systems quickly rises, as shown in Fig. 12b, which agrees well with electron energy-loss spectroscopy experiments. Comparing a calculated pair distribution function to representative experiments find that it successfully reproduces all general features, as shown in Fig. 12c. Shortly after, Deringer et al. [54] utilized this GAP fitted by a database of liquid and amorphous carbon configurations for random structure searching and readily predicted several higher- to unknown carbon allotropes. Besides, Csányi et al. [47] also introduced a similar GAP for atomistic simulations of amorphous elemental carbon and yielded accurate energetic and structural properties over a wide range of densities.

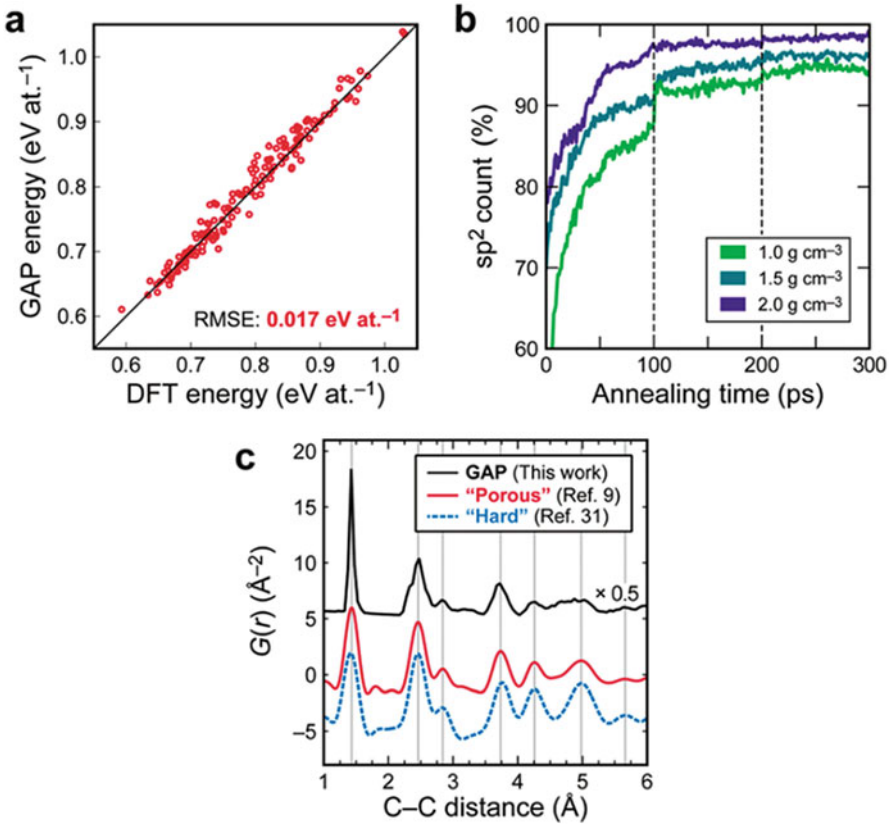

Fig. 12 (**a**) DFT versus GAP-computed energies for structures at various points of annealing trajectories. The root-mean-square error (RMSE) between these quantities is given. (**b**) Count of sp^2-bonded atoms during annealing; dashed lines indicate removal of unphysical long chains. (**c**) PDF analysis. (Adapted with permission from ref. [53], copyright 2018 Royal Society of Chemistry publishing)

6 Future Directions and Perspective

With continued increase in computing power, MD is emerging as a powerful tool for atom-level modeling as well as explore some micro mechanism without experiment. The predictive power of MD hinges strongly on the interatomic force field used to describe the atomistic interactions in the system. While the ML framework and the application cases presented above highlight the feasibility of using data-driven approaches for accurate modeling, however, there is still much room for improvement. The list below is some certain regions that would require further focused studies in the near future.

- The accuracy and speed of the ML force field depends on the choice of structural descriptors or fingerprints, so it is crucial to select the important descriptors from a large pool of candidates. For this issue, there is still controversy. Some researchers hold the view that the descriptors selected should depend on chemical or physical intuition or the basic knowledge of physic, because the essence of descriptor is physic. Some researchers proposed that blind spots exist in our intuitive judgment, for instance, the atomic force may be associated with several fingerprints combination. So ML can help us auto select descriptors.
- As materials science or chemical systems become ever increasingly complex, the configuration space for reference dataset will increase exponentially. This brings a challenge for the conventional nonlinear regression learning algorithm to handling such high dimension fitting issue, so some deep or advanced ML algorithms need to be developed.
- Most of the classical MD simulations employ predefined functional forms that can often limit the chemistry and physics that can be captured. While it appears that there can be significant improvements made by using data-driven approaches that employ extensive training data sets and advanced optimization, there will always be a ceiling limit imposed by the use of predefined functional forms. Existing force field with predefined functional form are not sufficiently flexible and cannot be transferred easily from one material class to another.
- Regardless of the application domain or area, all training data used for ML model should be carefully prepared and sufficiently diverse; for example, the reference configurations should span a wide range of energies, namely, the sampling not only includes near-equilibrium state but also consists of far-from-equilibrium configurations.
- Classical MD performs well under static or equilibrium issues while typically lacks predictive ability when it encounters dynamic and transport properties. One way to address this challenge is to include transition state configurations in the training data set. Going forward, we envisage that the temperature-dependent characteristics obtained from on-the-fly MD can also be used as part of the training program. This would allow us to directly train MD force field that can also capture dynamical and other transport properties or temperature-dependent properties of interest.
- Iterative improvement and cross-validation techniques are seldom used in the fitting of potentials. Even if with a DFT-based data set, there will always exist errors and then could be propagated to the atomistic potential model. Although higher-level theories can be introduced to generate training data and reduce errors, it is obvious that the uncertainty in prediction at various scales still needs to be quantified. Recent Strachan and his colleagues' work on quantitative methods of functional uncertainty represents an important future direction for assessing model errors. Cross-validation, sensitivity analysis, and uncertainty quantification are the key to improve the quality and prediction ability of interatomic potentials in MD.

In summary, using ML-based force field is indeed a powerful and feasible tool to accelerate atomistic simulations. Obtaining such high fidelity force prediction at a

very low cost has opened up an important way for the study of carbon materials and chemical phenomena. This can lead to revolutionary progress, enabling us to access time and length scales in carbon materials modeling that were hitherto considered to be inaccessible to MD.

References

1. Tang, Y., Zhang, D., et al. (2018). An atomistic fingerprint algorithm for learning ab initio molecular force field. *The Journal of Chemical Physics, 148*, 034101.
2. Berman, D., Sanket, A., Erdemire, A., et al. (2015). Macroscale superlubricity enabled by grapheme anoscroll formation. *Science, 14*, 126202.
3. Behler, J. (2017). First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angewandte Chemie, 56*, 12828–12840.
4. Botu, V., Batra, R., Chapman, J., et al. (2016). Machine learning force fields: construction, validation, and outlook. *The Journal of Physical Chemistry C, 121*, 511–522.
5. Botu, V., & Ramprasad, R. (2015). Adaptive machine learning framework to accelerate ab initio molecular dynamics. *International Journal of Quantum Chemistry, 115*, 1074–1083.
6. Botu, V., & Ramprasad, R. (2015). Learning scheme to predict atomic forces and accelerate materials simulations. *Physical Review B, 92*, 094306.
7. Chan, H., Narayanan, B., Cherukara, M. J., et al. (2019). Machine learning classical interatomic potentials for molecular dynamics from first-principles training data. *The Journal of Physical Chemistry C, 123*, 6941–6957.
8. Elliott, J. A. (2013). Novel approaches to multiscale modelling in materials science. *International Materials Reviews, 56*, 207–225.
9. Erdemir, A., Ramirez, G., Eryilmaz, O. L., et al. (2016). Carbon-based tribofilms from lubricating oils. *Nature, 536*, 67–71.
10. Gomez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., et al. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials, 15*, 1120–1127.
11. Hautier, G., Jain, A., & Ong, S. P. (2012). From the computer to the laboratory: materials discovery and design using first-principles calculations. *Journal of Materials Science, 47*, 7317–7340.
12. Huan, T. D., Batra, R., Chapman, J., et al. (2017). A universal strategy for the creation of machine learning-based atomistic force fields. *npj Computational Materials, 3*, 37.
13. Isayev, O., Oses, C., Toher, C., et al. (2017). Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications, 8*, 15679.
14. Jiang, Z., He, J., Deshmukh, S. A., et al. (2015). Subnanometre ligand-shell asymmetry leads to Janus-like nanoparticle membranes. *Nature Materials, 14*, 912–917.
15. Ju, S., Shiga, T., Feng, L., et al. (2017). Designing nanostructures for phonon transport via Bayesian optimization. *Physical Review X, 7*, 021024.
16. Neugebauer, J., & Hickel, T. (2013). Density functional theory in materials science. Wiley interdisciplinary reviews. *Computational Molecular Science, 3*, 438–448.
17. Smith, J. S., Nebgen, B. T., Zubatyuk, R., et al. (2019). Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications, 10*, 2903.
18. Li, Z., Kermode, J. R., & De Vita, A. (2015). Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Physical Review Letters, 114*, 096405.
19. Botu, V., Chapman, J., & Ramprasad, R. (2017). A study of adatom ripening on an Al (1 1 1) surface with machine learning force fields. *Computational Materials Science, 129*, 332–335.

20. Kruglov, I., Sergeev, O., Yanilkin, A., et al. (2017). Energy-free machine learning force field for aluminum. *Scientific Reports, 7*, 8512.
21. Suzuki, T., Tamura, R., & Miyazaki, T. (2017). Machine learning for atomic forces in a crystalline solid: transferability to various temperatures. *International Journal of Quantum Chemistry, 117*, 33–39.
22. Li, W., & Ando, Y. (2018). Comparison of different machine learning models for the prediction of forces in copper and silicon dioxide. *Physical Chemistry Chemical Physics (PCCP), 20*, 30006–30020.
23. Artrith, N., & Behler, J. (2012). High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Physical Review B, 85*, 045439.
24. Artrith, N., & Kolpak, A. M. (2015). Grand canonical molecular dynamics simulations of Cu–Au nanoalloys in thermal equilibrium using reactive ANN potentials. *Computational Materials Science, 110*, 20–28.
25. Artrith, N., & Urban, A. (2016). An implementation of artificial neural-network potentials for atomistic materials simulations: performance for TiO_2. *Computational Materials Science, 114*, 135–150.
26. Behler, J., Martonak, R., Donadio, D., et al. (2008). Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential. *Physical Review Letters, 100*, 185501.
27. Eshet, H., Khaliullin, R. Z., Kühne, T. D., et al. (2010). Ab initio quality neural-network potential for sodium. *Physical Review B, 81*, 184107.
28. Artrith, N., Morawietz, T., & Behler, J. (2011). High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Physical Review B, 83*, 153101.
29. Khaliullin, R. Z., Eshet, H., Kühne, T. D., et al. (2010). Graphite-diamond phase coexistence study employing a neural-network mapping of theab initiopotential energy surface. *Physical Review B, 81*, 100103.
30. Kondati Natarajan, S., Morawietz, T., & Behler, J. (2015). Representing the potential-energy surface of protonated water clusters by high-dimensional neural network potentials. *Physical Chemistry Chemical Physics (PCCP), 17*, 8356–8371.
31. Yao, K., Herr, J. E., Toth, D. W., et al. (2018). The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics. *Chemical Science, 9*, 2261–2269.
32. Mi, X. Y., Yu, X., Yao, K. L., et al. (2015). Enhancing the thermoelectric figure of merit by low-dimensional electrical transport in phonon-glass crystals. *Nano Letters, 15*, 5229–5234.
33. Seko, A., Togo, A., Hayashi, H., et al. (2015). Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Physical Review Letters, 115*, 205901.
34. Li, S., Yu, X., Bao, H., et al. (2018). High thermal conductivity of bulk epoxy resin by bottom-up parallel-linking and strain: A molecular dynamics study. *The Journal of Physical Chemistry C, 122*, 13140–13147.
35. Song, Q., An, M., Chen, X., et al. (2016). Adjustable thermal resistor by reversibly folding a graphene sheet. *Nanoscale, 8*, 14943–14949.
36. Yang, H., Zhang, Z., Zhang, J., et al. (2018). Machine learning and artificial neural network prediction of interfacial thermal resistance between graphene and hexagonal boron nitride. *Nanoscale, 10*, 19092–19099.
37. Ma, D., Ding, H., Wang, X., et al. (2017). The unexpected thermal conductivity from graphene disk, carbon nanocone to carbon nanotube. *International Journal of Heat and Mass Transfer, 108*, 940–944.
38. Yu, X., Li, R., Shiga, T., et al. (2019). Hybrid thermal transport characteristics of doped organic semiconductor poly(3,4-ethylenedioxythiophene): Tosylate. *The Journal of Physical Chemistry C, 123*, 26735–26741.
39. Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics, 134*, 074106.
40. Bartók, A. P., Kondor, R., & Csányi, G. (2017). Erratum: On representing chemical environments. *Physical Review B, 96*, 019902.

41. Bartók, A. P., & Csányi, G. (2015). Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry, 115*, 1051–1057.
42. Zong, H., Pilania, G., Ding, X., et al. (2018). Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning. *npj Computational Materials, 4*, 48.
43. Fan, J., Sun, Q., Zhou, W.-X., et al. (2018). Principal component analysis for big data. 1–13.
44. Rowe, P., Csányi, G., Alfè, D., et al. (2018). Development of a machine learning potential for graphene. *Physical Review B, 97*, 054303.
45. Wen, M., & Tadmor, E. B. (2019). Hybrid neural network potential for multilayer graphene. *Physical Review B, 100*, 195419.
46. Wen, M., & Tadmor, E. B. (2020). Uncertainty quantification in molecular simulations with dropout neural network potentials. *npj Computational Materials, 6*, 124.
47. Deringer, V. L., & Csányi, G. (2017). Machine learning based interatomic potential for amorphous carbon. *Physical Review B, 95*, 094203.
48. Rowe, K., Deringer, V. L., & Gasprotto, P. (2020). An accurate and transferable machine learning potential for carbon. *Chemical Pysics, 153*, 034702.
49. Khaliullin, R. Z., Eshet, H., Kuhne, T. D., et al. (2011). Nucleation mechanism for the direct graphite-to-diamond phase transition. *Nature Materials, 10*, 693–697.
50. Bartok, A. P., Payne, M. C., Kondor, R., et al. (2010). Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters, 104*, 136403.
51. Behler, J., & Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters, 98*, 146401.
52. Khaliullin, R. Z., Eshet, H., Kühne, T. D., et al. (2010). Graphite-diamond phase coexistence study employing a neural-network mapping of theab initiopotential energy surface. *Physical Review B, 81*, 100103(R).
53. Deringer, V. L., Merlet, C., Hu, Y., et al. (2018). Towards an atomistic understanding of disordered carbon electrode materials. *Chemical Communications, 54*, 5988–5991.
54. Deringer, V. L., Csanyi, G., & Proserpio, D. M. (2017). Extracting crystal chemistry from amorphous carbon structures. *Chemphyschem: A European Journal of Chemical Physics and Physical Chemistry, 18*, 873–877.