Caro, Miguel A.

Machine learning based modeling of disordered elemental semiconductors: understanding the atomic structure of a-Si and a-C

Published in:
Semiconductor Science and Technology

DOI:
10.1088/1361-6641/acba3d

Published: 01/04/2023

Document Version
Publisher's PDF, also known as Version of record

Machine learning based modeling of disordered elemental semiconductors: understanding the atomic structure of a-Si and a-C

OPEN ACCESS
IOP Publishing
Semiconductor Science and Technology
Semicond. Sci. Technol. **38** (2023) 043001 (25pp)
https://doi.org/10.1088/1361-6641/acba3d

Topical Review

Machine learning based modeling of disordered elemental semiconductors: understanding the atomic structure of a-Si and a-C

Miguel A Caro

Department of Chemistry and Materials Science, Aalto University, 02150 Espoo, Finland

Department of Electrical Engineering and Automation, Aalto University, 02150 Espoo, Finland

E-mail: mcaroba@gmail.com

Received 13 September 2022, revised 21 January 2023
Accepted for publication 8 February 2023
Published 6 March 2023

Abstract

Disordered elemental semiconductors, most notably a-C and a-Si, are ubiquitous in a myriad of different applications. These exploit their unique mechanical and electronic properties. In the past couple of decades, density functional theory (DFT) and other quantum mechanics-based computational simulation techniques have been successful at delivering a detailed understanding of the atomic and electronic structure of crystalline semiconductors. Unfortunately, the complex structure of disordered semiconductors sets the time and length scales required for DFT simulation of these materials out of reach. In recent years, machine learning (ML) approaches to atomistic modeling have been developed that provide an accurate approximation of the DFT potential energy surface for a small fraction of the computational time. These ML approaches have now reached maturity and are starting to deliver the first conclusive insights into some of the missing details surrounding the intricate atomic structure of disordered semiconductors. In this Topical Review we give a brief introduction to ML atomistic modeling and its application to amorphous semiconductors. We then take a look at how ML simulations have been used to improve our current understanding of the atomic structure of a-C and a-Si.

Keywords: disordered carbon, disordered silicon, atomistic simulation, machine learning potentials, molecular dynamics

(Some figures may appear in colour only in the online journal)

1. Introduction

Since the inception of the first experimental semiconductor diodes in the early 1900s the presence of semiconductors in daily appliances as well as high-tech equipment has grown exponentially. Today, virtually all equipment incorporating electrical circuits or electronic components, including computers and mobile phones, have parts made of silicon. Commercially successful light-emitting diodes (LEDs) and laser diodes also use semiconductors, most often III–V compounds. Many of the familiar applications of semiconductors use their crystalline forms, and the degree of crystallinity often dictates the quality of the device. Indeed, in LEDs even tiny amounts of crystallographic defects can severely deteriorate device performance [1, 2].

On the other hand, amorphous semiconductors, notably a-C and a-Si, can have useful properties of their own. Whether they offer actual performance improvements over crystalline forms for specific applications, or a significantly cheaper and more scalable fabrication process gives them a practical advantage, these materials are widely used for applications where their electronic, chemical, mechanical and optical properties are exploited. Hydrogenated a-Si (a-Si:H) is used to fabricate low-cost solar cells [3]. More generally, a-Si and its derivatives find uses in applications where a cost-effective alternative to crystalline Si (c-Si) is desirable, or where less stringent growth conditions (e.g. lower deposition temperature) are required [4]. This includes such applications as thin-film transistors [5], liquid-crystal displays [6] and medical x-ray imaging [7]. a-C is even more versatile than a-Si since its properties can be more or less continuously tuned between those of graphitic carbon (g-C) and diamondlike carbon (DLC) [8]. Current uses of a-C and a-C thin films include biocompatible and bioimplantable devices (such as hip replacement implants) [9], electrochemical sensors for *in vivo* analysis [10] and hard coatings for tribological applications [11]. Furthermore, modified a-C such as oxygen-rich a-C (a-COx) [12], nitrogen-doped a-C (a-C:N) [13], different carbon hybrid materials [10], nanocarbons modified under extreme conditions [14–16] and the wider family of disordered carbons are starting or expected to make their way to emerging applications in energy storage [17]. More generally speaking, carbon-based materials are envisioned to be key in the transition to renewable raw materials utilization and the bioeconomy [18].

Unsurprisingly, the diversity and complexity of the atomic structure of a-C and a-Si pose serious challenges for experimental characterization. For crystalline materials, common structural characterization methods, like x-ray diffraction, rely on the *periodic* structure of crystals, and are thus less useful to characterize amorphous materials. Instead, the structure of a-C and a-Si (and other disordered materials) can be characterized using experimental techniques such as x-ray photoelectron or absorption spectroscopy (XPS and XAS, respectively), Raman spectroscopy and neutron scattering. A very complete summary of experimental structure characterization techniques for a-C has been given by Robertson [8] (these techniques are also relevant for the characterization of a-Si). In our strive to understand the atomic structure of disordered materials, computational atomistic modeling techniques arise as an obvious choice: by being able to model the interatomic energies and forces between atoms, and update or optimize their positions accordingly, we can effectively 'look' at the atomic structure. To access the length and time scales involved in modeling amorphous materials accurately, ML interatomic potentials (MLPs) have emerged in recent years as game changers in the field [19].

In this Topical Review we will first discuss general considerations pertaining to atomistic modeling of amorphous semiconductors. We will then give a brief introduction to MLPs that should be accessible to those with basic understanding of atomistic simulations, either coming from a (modest) density-functional theory (DFT) or classical molecular dynamics (MDs) background. We will then show how MLPs have enabled a new degree of realism in modeling a-Si and a-C, arguably the two most important elemental amorphous semiconductors. We will end with a brief discussion of the state of the field and an outlook for the future.

2. a-C and a-Si atomistic simulation

The main fundamental difference between a crystalline and an amorphous semiconductor is the lack of long-range atomic order in the latter. The other differences (electronic and thermal conductivities, electronic and optical band gap, mechanical properties, etc) ultimately stem from the differences in the atomic structure. In a-Si, local atomic structures are usually 4-fold tetrahedral motifs due to sp^3 chemical bond hybridization. Lower (3-fold) and higher (5-fold) coordinations in a-Si are typically considered coordination defects [20]. Thus, the structural complexity in a-Si is compounded by the interplay between the local arrangement of nearby stable 4-fold motifs and the existence of coordination defects in the amorphous network. In the case of a-C the situation is significantly more complex since stable chemical motifs in elemental carbon can be due to sp (2-fold), sp^2 (3-fold, 'graphite-like') and sp^3 (4-fold, 'diamond-like') hybridizations. The atomic structure of a-C is consequently diverse, making a-C effectively a *range* of materials, rather than just a material, typically characterized to a first approximation by the relative amount of sp^2 and sp^3 carbon. The sp^2-rich forms of a-C are low in mass density (down to 2 g cm^{-3} and less [8]), whereas the sp^3-rich forms have high mass density and are often referred to as 'diamond-like' or 'tetrahedral' a-C (DLC and ta-C, respectively) [8]. To complicate things, a-C and a-Si can exist with different degrees of hydrogenation, where some of the C–C and Si–Si bonds are replaced by C–H and Si–H bonds. These materials are usually referred to as a-C:H and a-Si:H, respectively, and their properties, especially mass density, may differ from those of the hydrogen-free forms. We note here in passing that pure a-C and a-Si do not exist in practice, and some level of impurities, mostly H and O, are always present in experimental samples [8, 21].

The standard for predicting the structure of materials at the atomic scale is DFT. DFT is a quantum mechanical method,

providing an approximation to the Schrödinger equation. Its popularity stems from the computational efficiency of the Kohn–Sham (KS) formulation of DFT [22–24], which resides at a 'sweet spot' of accuracy vs CPU cost. DFT is routinely used to study crystals and to carry out crystal structure prediction [25, 26], benefiting from the fact that crystals can be represented with small primitive unit cells, often comprising just a handful of atoms. Unfortunately, even DFT can become prohibitively expensive to model amorphous materials, which lack short-range order. In practice, DFT has been used to study amorphous materials in a limited way by employing the 'supercell' approach. A supercell is made of tens or, at most, a few hundreds of atoms in periodic boundary conditions [27–29]. Thus, effectively, amorphous compounds are modeled as crystals with very large unit cells.

The accuracy of the supercell approach to model real amorphous materials improves with system size, but not only. To provide a realistic view of an amorphous structure it is necessary to collect statistics via configurational sampling, since each individual supercell will, in general, look different from another. More critically, while a 'single-point' DFT calculation (i.e. a calculation where the atomic positions are not updated) for a given structure may be affordable even for relatively large system sizes of a couple or few thousands of atoms, placing the atoms in configurations that resemble real amorphous structures is far from trivial [28].

Computational structure-generation protocols for amorphous materials come in two flavors. On the one hand, there are direct protocols trying to mimic the experimental growth process as closely as possible. This is for instance the case for simulated deposition, where the attachment of atoms onto a growing film is simulated one atom at a time [30–33].

On the other hand, there exist indirect protocols that rely on initially randomizing the atomic positions and subsequently updating these positions. The positions can be updated by either 'relaxing' the structure (e.g. using gradient-descent optimization along the direction of the forces) [29, 34] or by carrying out MD with a rapidly decreasing temperature profile ('quench' simulations) [28, 35–37]. Sampling protocols designed for free-energy sampling at given thermodynamic conditions [38] are often not a good choice to generate amorphous structures, since (a) amorphous materials are usually *metastable* and (b) these free-energy sampling methods can be prohibitively expensive because they rely on many individual evaluations of the potential energy surface (PES). A relatively recent comparison between generation methods for a-C modeling, but lacking some of the latest developments with ML interatomic potentials [32, 33], has been given in [10].

Arguably, the most popular protocol to generate atomistic amorphous structures is the MD-based 'melt-quench' protocol [28], which resembles how glass is made in reality [39]. In a melt-quench simulation a material is first heated to high temperature T until it melts. This liquid sample is then kept at high temperature to ensure a disordered but relatively low-energy distribution of atoms (e.g. much lower in energy than random) to provide a reasonable starting point. Then, the liquid is quickly quenched down to a temperature well below

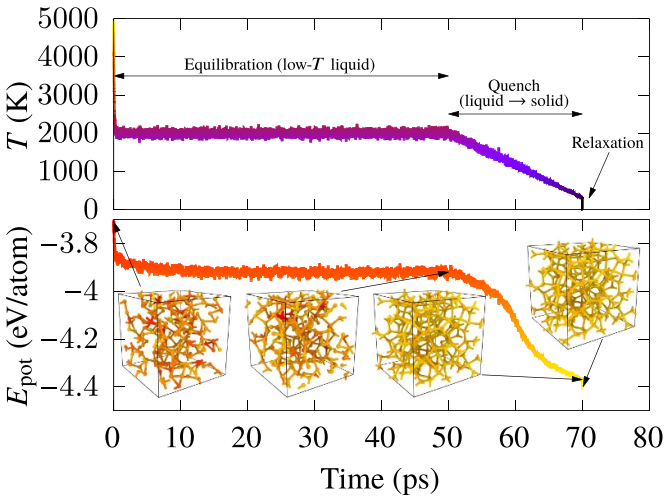

Figure 1. Melt-quench simulations of a-Si formation carried out with a GAP [45] potential refitted [46] from the database developed by Bartók *et al* [47]. The simulations were carried out with the TurboGAP code [48, 49].

the solidification temperature. Since the process is so fast, the different atomic motifs are trapped into local minima, giving an amorphous structure as a result. How fast the system is cooled down (the quench rate) will determine the quality of the structure. A very fast quench will lead to numerous defects, for example under (3-fold) and over (5-fold) coordinated atoms in a-Si [40]. A very slow quench will (theoretically) lead to formation of the thermodynamically stable allotrope of the material, for example diamond-structure silicon. Besides a temperature profile, imposed in MD through the use of a thermostat [41], one may also couple the simulation to a barostat [42], to control the pressure P. This enables exploration of phases and phase transformations within widely varying thermodynamic conditions, including some extreme conditions not accessible experimentally [43].

Additional steps can be added before or after the quench, typically some sort of annealing step. For instance, a carbon sample can be held at around 3500 K for a while before quenching to favor graphitization [35, 37, 44]. Or an a-Si sample can be annealed at a temperature below solidification (but still significantly higher than room temperature) to heal defects [40].

The melt-quench process leading to generation of a computational atomic structure is exemplified for a-Si in figure 1. Initially, the sample, containing 216 atoms, has been heated to a very high temperature of 5000 K to properly randomize the atomic positions. The temperature is rapidly brought down to 2000 K, slightly above the melting temperature of silicon, and kept there for some time (50 ps in our example). This is the equilibration stage, where we aim to homogeneously distribute the available kinetic energy among all the degrees of freedom and find local structures which are low in energy (for the given values of T and P). After equilibration, we quench the system down to 300 K using a linear temperature profile. The evolution of the potential energy does not follow this

linear trend. Instead, there is a slow initial decrease in potential energy because the temperature is too high to create stable motifs. This is followed by an accelerated decrease in energy where these stable atomic motifs, tetrahedra in a-Si, are created at a fast pace. The final stage of the quench corresponds to slow further decrease in potential energy, because of either of two reasons: (a) all the Si atoms are already part of local tetrahedra or (b) there is not enough kinetic energy to overcome local potential energy barriers and the atoms are trapped into their local metastable structures. The actual situation is a combination of both factors. Recall that, according to the virial theorem, as we linearly decrease the kinetic energy there should be a corresponding linear decrease in potential energy, *assuming that the details of the potential energy surface do not change*. Therefore, the non-linear profile observed in our example is indeed associated to the phase transition from the disordered liquid to the amorphous solid.

After the MD quench, we further relax the structure using a static relaxation of the atomic positions, following a gradient-descent minimization of the potential energy. In figure 1 we have additionally color-coded the Si atoms according to their local energy, which can be extracted from a simulation with MLPs, as we detail in section 3. The curious reader is encouraged to visit the `turbogap.fi` website for a series of tutorials on how to run this type of simulation for a-Si and a-C.

Melt-quench simulations are popular because they provide a good compromise between CPU cost and the quality of the generated structures. However, they do not (typically) reproduce the experimental growth/formation protocol of the real material, and that can have a non-negligible effect on the resulting atomic structure, as is for instance the case for a-C [10]. Unfortunately, direct simulation protocols, such as deposition in a-C [31–33], are orders of magnitude more expensive than indirect methods.

Traditionally, direct simulation has been limited to empirical interatomic potentials [31, 50–52]. These are very efficient computationally, relying on simple mathematical functions that depend on the interatomic distances and angles and are parametrized by fitting to experimental or first-principles data. Popular examples of these potentials, which can often be used for both Si and C (and even SiC) by adjusting the model's parametrization, are Tersoff [53], Stillinger-Weber [54, 55], EDIP [56] (and its carbon version C-EDIP [57]), REBO [58, 59] and ReaxFF [60–63]. However, these empirical potentials lack the accuracy of DFT, and thus provide a representation of the PES of very inconsistent quality [33, 35, 44, 47, 64]. While low-lying harmonic regions of the PES, i.e. the atomic configurations about equilibrium, can be reproduced with reasonable accuracy, chemical reactions are described very poorly. Therefore, we find ourselves at an impasse: on the one hand, the breaking and formation of chemical bonds, critical to understand the growth of amorphous materials, are not correctly described with affordable empirical potentials. On the other hand, DFT can describe chemical reactions accurately but is computationally unaffordable.

Fortunately, new atomistic simulation techniques based on ML have emerged in recent years [45, 64, 65] that bridge this huge gap in atomistic modeling of amorphous materials [36, 47]. These MLPs rely on non-parametric fits to a reference PES, typically computed at the DFT level of theory [66, 67]. While still significantly more expensive than empirical force fields, MLPs offer accuracy close to that of DFT for a tiny fraction of the CPU cost. MLPs have had, in just the last few years, a huge impact on atomistic modeling of amorphous and disordered materials, granting us atomistic insight into problems that were completely out of reach less than a decade ago.

3. ML interatomic potentials

In this section we explain the whole MLP workflow, graphically summarized in figure 2. We start with a brief general introduction to different popular ways to represent the PES, with an emphasis on DFT. We will then give an overview of the two main methodologies for learning and interpolating the DFT-PES based on (a) artificial neural networks (ANNs) and (b) the related kernel-ridge regression and Gaussian process regression (GPR) methods. To take a pedagogical approach, the introduction of these methodologies will be preceded by a general introduction to relevant ML concepts (databases and descriptors/features). We will compare ML potentials to DFT, on the one hand, and classical force fields, on the other, to get an idea of what is possible now in materials modeling, thanks to the introduction of MLPs, that was not possible just a few years ago. For more comprehensive information, the reader is referred to a recent book which nicely summarizes the current state of the field [65] including a chapter on GPR [68] and another on ANNs [69], and to several overview papers [19, 66, 67, 70–72].

3.1. Database construction (structure selection)

The creation of a new MLP starts with the generation of training data (figure 2(a), step 1). Many considerations need to go into carefully crafting a suitable database for the problem at hand. There are two main classes of MLPs depending on their scope: general- and single-purpose MLPs. A single-purpose MLP is created with a very specific application in mind. In this case, the MLP will be expected to perform with excellent (or even exquisite) accuracy for the problem of interest, but there is no guarantee that it will perform even reasonably for any other application. Recall that the MLP does not 'know' about physics, chemistry, or the Schrödinger equation; it only knows about data. Therefore, an MLP will only be able to chart the portion of configuration space corresponding to the data that it was fed. Indeed, single-purpose MLPs (and poorly designed general-purpose MLPs) will tend to 'blow up' (MD jargon for when a force field becomes catastrophically unstable) when tested on a problem for which they were not trained. A good example of a single-purpose MLP would be one trained to reproduce the phonon dispersion curves of a crystalline material, for instance to be used in thermal transport or thermal expansion calculations of

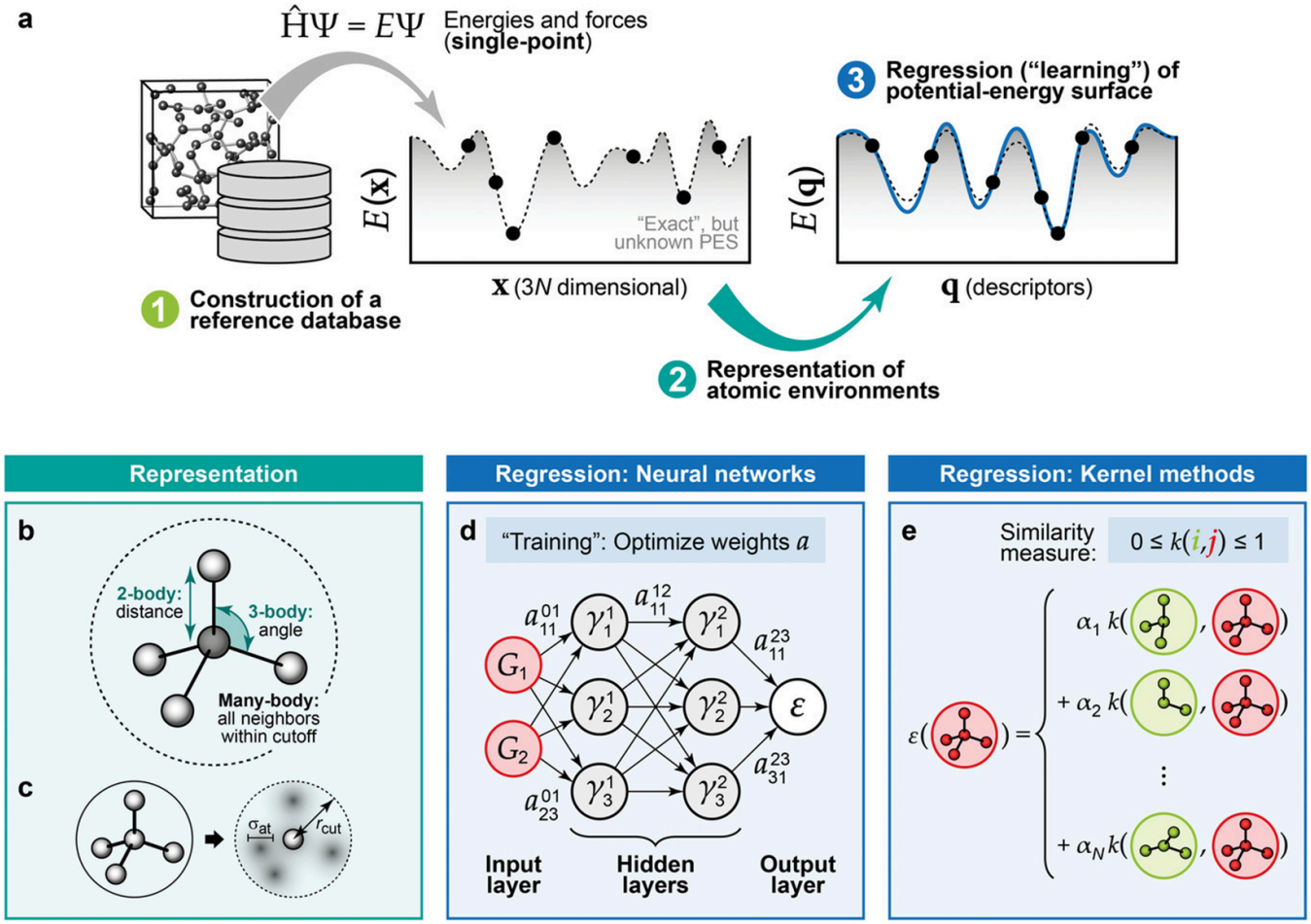

Figure 2. (a) Workflow of MLP training: a database of atomic structures and observables (energies, forces) is constructed, from which an ML algorithm is used to learn the PES as a function of atomic descriptors. (b), (c) Different kinds of descriptors commonly used to represent the atomic environments. (d) Schematics of neural network prediction. (e) Schematics of kernel-based prediction. [19] John Wiley & Sons. © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

c-Si and diamond/graphite. A database suitable to fit a good phonon MLP would typically incorporate many DFT calculations of structures distorted from the equilibrium ones, by adding either homogeneously spaced or random strain transformations to the unit cell in addition to rattling the atoms about their equilibrium positions. Furthermore, the unit cells should span from the primitive unit cell up to larger unit cells which would allow to capture interactions between distant atoms. An example of a single-purpose MLP is the graphene GAP of Rowe *et al* [73].

A general purpose MLP, on the other hand, is expected to perform reasonably accurately in as many regions of configuration space as possible, and be resistant to blowing up. A good general-purpose MLP is often very difficult to achieve because it may require prior knowledge about which these regions are, and its training is consequently difficult to automate. For instance, if one wants to fit an MLP to study the atomic structure of a-C surfaces, which can be prohibitively expensive to generate with DFT using a melt-quench simulation (cf figure 1), how are sample surfaces sourced for the (single-point) DFT calculations that will serve as reference for the MLP? In these cases, *iterative training* [36, 74] can help in improving the accuracy of the MLP in regions of interest in configuration space and also to get rid of pathological behavior. In our a-C surface example, iterative training would consist of generating surface structures with an interim (low-quality) version of the MLP via melt-quench simulations. Single-point DFT calculations are then performed on the final structures and this data added to the training set. A new interim version of the MLP is trained and the whole procedure is repeated until the MLP errors (compared to the DFT calculations) are below an acceptable threshold. Besides regular sampling of known crystal phases, iterative training can also be combined with less directed exploration of configuration space, such as random structure search [26, 75].

Finally, we can combine the features of a general-purpose MLP with those of a single-purpose MLP, to improve the accuracy of the general-purpose MLP for specific applications, as has been done for phonons in Si [76] or fullerenes in C [43].

A way to visualize these structural databases is via so-called structure maps [77, 78]. In these, the similarities between different entries in a database, i.e. between different atomic structures, can be plotted on a two-dimensional map using low-dimensional embedding techniques [77, 78].

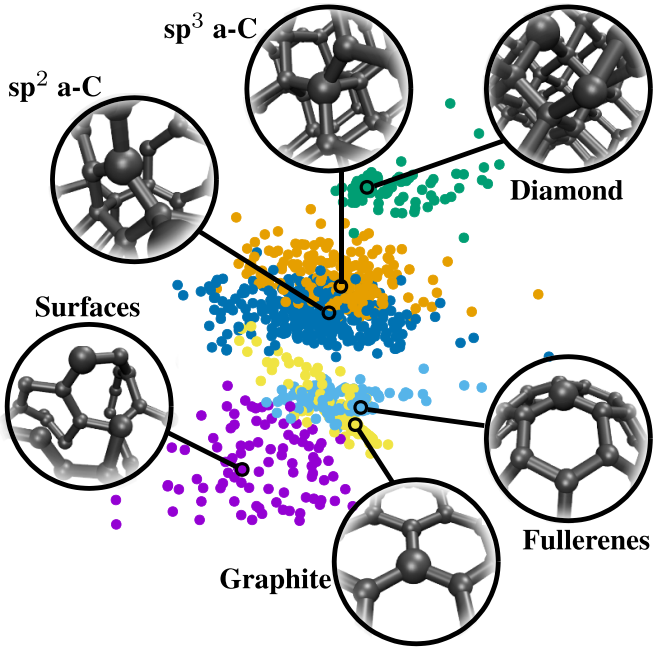

Figure 3. Low-dimensional embedding of high-dimensional data, used in this case to visualize the atomic structural diversity in a database constructed to fit an MLP for carbon. The closer two points are on the graph, the more closely the corresponding atomic environments resemble each other. Reprinted figure with permission from [43], Copyright (2021) by the American Physical Society.

An example for the fullerene-augmented C MLP mentioned earlier [43] is shown in figure 3. In this graph, each data point represents an atom-centered environment and similar structures are *clustered* together. There is a transition from diamond structures to amorphous sp^3, then to amorphous sp^2 and finally to different graphitic structures, including fullerenes. These sketchmaps are a useful tool to glimpse at the composition of an entire database and understand the relationships between the different structures.

3.2. Reference representations of the potential energy surface

A central objective of computational atomistic modeling is to gain access to an accurate representation of the Born–Oppenheimer (BO) PES of a given ensemble of interacting atoms. The BO approximation relies on decoupling the electronic and nuclear degrees of freedom. That is, the BO-PES gives the total cohesive energy of a set of interacting atoms as the *electronic* ground state (GS) for fixed nuclear positions [24]. This approximation is valid in many situations, in particular in condensed-matter physics, because of the mass difference between electrons and nuclei. The electronic degrees of freedom evolve within much shorter time scales than the nuclear degrees of freedom, and the atomic trajectories can be propagated in time treating the nuclei as classical particles, following Newton's second law. The most popular approximation used today to calculate the BO-PES is DFT [22–24]. The fundamental tenet of DFT is that the total (cohesive) energy of a system of interacting electrons in a external potential (given by the positively charged nuclei) is given by a universal functional $E[n]$ of the electron density $n(\mathbf{r})$, where the density that minimizes the functional is the GS density and the energy of the GS is given by the energy functional evaluated at the GS density [22]:

$$n_{\mathrm{GS}}(\mathbf{r}) = \underset{n(\mathbf{r})}{\operatorname{argmin}} E[n(\mathbf{r})], \qquad E_{\mathrm{GS}} = E[n_{\mathrm{GS}}(\mathbf{r})]. \tag{1}$$

The practical means for solving equation (1) are provided by the KS ansatz, which states that the density can be expressed as a combination of single-particle contributions, one per electron (or electron pair, depending on whether or not spin is explicitly modeled):

$$n(\mathbf{r}) = \sum_{i=1}^{N_{\mathrm{e}}} |\psi_i(\mathbf{r})|^2, \tag{2}$$

where $\psi_i(\mathbf{r})$ is the KS orbital of the ith electron in the system and N_{e} is the number of electrons. This approximation allows us to avoid explicitly working with the many-body wave function of the system. In the KS formulation of DFT, the many-body effects are collected into the exchange-correlation (XC) density functional $E_{\mathrm{xc}}[n(\mathbf{r})]$. The quality of the used approximation for $E_{\mathrm{xc}}[n(\mathbf{r})]$ will determine how close to the actual GS density and energy we can get.

The KS single-particle ansatz, coupled with the variational principle $\delta E[n]/\delta\psi_i^* = 0$ leads to the eigenvalue-like KS equation:

$$\epsilon_i\psi_i(\mathbf{r}) = H_{\mathrm{KS}}\psi_i(\mathbf{r}), \tag{3}$$

where the KS Hamiltonian H_{KS} contains the single-particle kinetic energy operator, the electrostatic potential and the XC potential. A deeper account of DFT is beyond the scope of this review, and the reader is referred to the excellent (nowadays almost standard) book by Martin [24] for more detailed information.

The emergence of KS DFT, together with many different approximations to the XC functional [79, 80] and efficient iterative algorithms for solving the KS equation implemented in parallel computer codes [81] have enabled quantum mechanical calculations of the properties of matter at affordable computational cost. In addition to this, the community has been very active at tackling the different shortcomings of KS DFT, such as the self-interaction error or the lack of dispersion interactions, e.g. by developing 'hybrid' XC functionals [82, 83], and van der Waals 'corrections' [84–86], respectively. While DFT is still too expensive to perform long- and large-scale simulations of atomic systems, the tradeoff between accuracy and CPU cost afforded by DFT makes it the most popular electronic structure method and, indeed, the most popular method to generate training data for MLPs.

The purpose of an interatomic potential, also referred to as a force field, is to provide a computationally affordable approximation of the BO-PES. When training MLPs we often

assume that DFT provides a 'good enough' version of the BO-PES. While we have just discussed that DFT has its own shortcomings, it is also important to recognize the limitations of the BO approximation itself. Notable breakdowns of the BO approximation occur whenever protium atoms are present (i.e. the common hydrogen isotope with one proton in the nucleus) [87] or when high-energy collisions take place (e.g. during radiation-damage events in materials) [88]. Extended MD formalisms are required when simulating these kinds of systems, for instance time-dependent DFT or Ehrenfest dynamics, where electronic and nuclear degrees of freedom are propagated simultaneously [89, 90]. In addition, electronic excitations and charge-transfer processes [91] cannot be captured out of the box by MLPs trained from DFT data. Despite these limitations, which are at the hot spot of current work by the community, MLPs have enabled incredible successes in computational materials modeling in recent years, in particular for a-Si and a-C. We review the basics of MLPs for materials and molecular modeling in the next section.

3.3. MLP architecture

The rationale for replacing a DFT calculation (or, more generally, an expensive *ab initio* calculation) by an ML prediction is that, in atomistic systems, the local atomic motifs are often repetitious. Therefore, put in simple terms, if we compute energies and forces for a series of reference structures with DFT and store those values in a database, we should in principle be able to infer a DFT-quality prediction for a new atomic structure as long as said structure is similar enough to the database entries. The interpolation should be computationally inexpensive, compared to a DFT calculation, for the procedure to be useful. The simplest example is a diatomic molecule, where a series of DFT calculations are carried out for different interatomic separations and a force field is trained from that data to predict the energy vs distance curve at arbitrary separations. An 'old-fashioned' empirical force field would often tackle this problem by fitting the data to a fixed functional form, e.g. a least-squares fit to a second-order polynomial. In that sense, MLPs can be regarded as a glorified version of an empirical force field, where the main difference is the fact that the fit is now carried out in arbitrarily many dimensions and without the user providing an explicit mathematical function. This is referred to in ML jargon as a non-parametric fit. Although the distinction between MLPs and empirical force fields may seem small in this context, in practice the flexibility of ML algorithms to fit high-dimensional data means that much more complex PESs can be learned, and to much higher accuracy.

Above, two fundamental assumptions are made whose goodness will to a large extent determine the success of MLPs. One is the assumption of locality of the PES. That is, we can construct the entire system as a collection of local fragments, each of which has an associated local energy. Physically, the local energy $\bar{\epsilon}_i$ (where the bar indicates prediction) is not a well-defined property of the system; instead, a DFT calculation will return a total energy E for a given ensemble of N_{at} interacting atoms. An MLP will build this total energy from

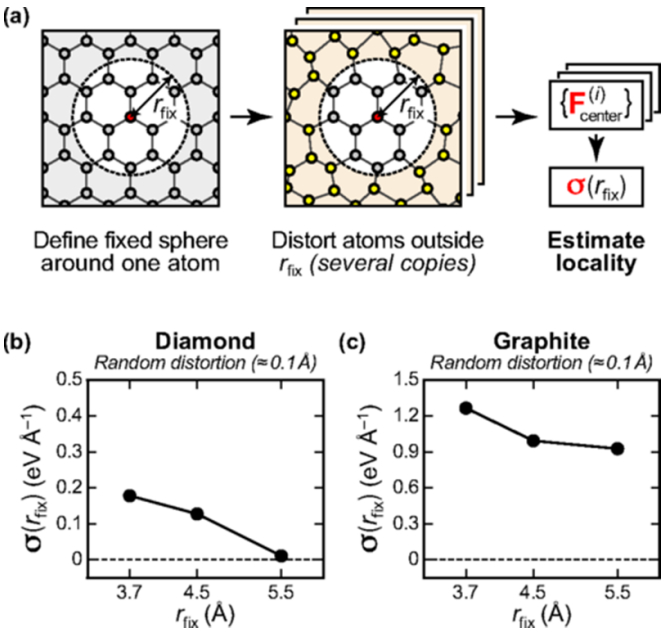

Figure 4. Locality tests in carbon-based systems. (a) Conceptualization of the locality test. (b) Convergence of the residual force acting on the central atom for diamond and graphite as a function of the cutoff radius. Reprinted figure with permission from [36], Copyright (2017) by the American Physical Society.

the sum of all the individual contributions which, in simplified terms, can be considered a sum over atom-wise contributions:

$$\bar{E} = \sum_{i=1}^{N_{\mathrm{at}}} \bar{\epsilon}_i. \tag{4}$$

An intuitive way to test the locality of the PES for a given material is to monitor the evolution of the force acting on an atom as other atoms beyond a certain cutoff distance are disturbed, as a function of said cutoff. This was done in [36] for crystalline and amorphous C. The procedure is illustrated in figure 4(a) and the results for diamond and graphite in figure 4(b) are reprinted from that reference. For diamond (as well as high-density a-C, not shown here but reported in [36]) the approximation of locality is extremely good and the errors are negligible for cutoffs around 5 Å and beyond. For graphite (and low-density a-C) the approximation is less good and convergence with the cutoff is very slow. Mathematically, this approximation implies that we can express a local (atomic) energy prediction as a function of a finite environment of the atom:

$$\bar{\epsilon}_i = f(S_i(r_{\mathrm{cut}}))\,, \tag{5}$$

where $S_i(r_{\mathrm{cut}})$ represents all the relative atomic positions within a sphere of radius r_{cut} centered on atom i. In technical terms this means that $\bar{\epsilon}_i$ has compact support.

The other main assumption is about the smoothness of the PES. That is, a small change in the positions of the nuclei should lead to a small change in the total energy of the system. In mathematical terms, the PES should be continuous

and continuously differentiable. In a data science context, smoothness is referred to as regularity.

Besides the training (DFT) data and the two central approximations for the PES, locality and smoothness, which we have already discussed, an MLP requires also two basic ingredients. The first one is the atomic structure representation, which is carried out using atomic descriptors. While in principle the Cartesian coordinates of the nuclei contain all the necessary information, in practice they are not useful because they do not fulfill the correct symmetries. Specifically, valid atomic descriptors must fulfill translational, rotational and permutational invariance. The simplest descriptor is an interatomic distance. More sophisticated descriptors, which contain increasingly more information about the environment of an atom, can be constructed with a body-order expansion [92]. An interatomic distance is a two-body (2b) descriptor, with a single degree of freedom. A 3b descriptor has three degrees of freedom and perfectly characterizes a system made of three atoms, having subtracted the translation and rotation of the center of mass, which do not affect energy and forces. Any further body-order increase adds three more degrees of freedom, and the complexity of the model (and the cost of computing descriptors) explodes with relatively low body orders. For many practical purposes in materials modeling there is no need to go beyond 3b terms [93]. However, there is another type of atomic descriptors that allow to encode the *entire* atomic environment, called many-body (mb) descriptors (cf figure 2(c)). Arguably, the most important examples are the smooth overlap of atomic positions (SOAP) [94] and atom-centered symmetry functions (ACSFs) [95]. It can be shown that these mb descriptors are formally equivalent to one another and, as constructed from 2b sums within a finite cutoff sphere, are also equivalent to an ensemble of 3b terms [96]. Two advantages over 3b descriptors are that one mb descriptor can be used instead of very many 3b ones (since the number of 3b descriptors within a cutoff sphere explodes as a function of its radius), and that mb descriptors with different numbers of atoms can be compared to one another (directly relevant in kernel regression methods, cf figure 2(e)). The topic of atomic representations is very rich and has been recently summarized in a comprehensive review paper [97].

The second basic ingredient is the ML algorithm. The first method to interpolate high-dimensional PES with close to DFT accuracy was introduced in 2007 by Behler and Parrinello [64] based on ANNs and applied precisely to model Si. The second method, based on kernel regression, was introduced by Bartók *et al* in 2010 [45] and used to model C, Si and Ge. Clearly, group-IV semiconductors have been strongly linked to the use of MLPs since their very inception, and as such it is unsurprising that the first applications of MLPs to solving outstanding problems in materials modeling have also focused on C and Si. Naturally, the methodology has advanced significantly since those two seminal papers and more recent reviews by the authors do a better job at introducing the concepts and practicalities to the beginner [66, 67, 70]. Many other methods and implementations have appeared since then. A comprehensive account of those is beyond the scope of this work and so we mention again the recent book summarizing the state of the field [65]. Below we give a brief overview of these methods, and refer the reader to the cited literature for further detail.

3.3.1. Artificial neural network potential (NNP). NNPs [64] use ANNs to interpolate the PES. An ANN consists of a series of 'layers': input, hidden and output layers. There is one input and one output layer, and one or more hidden layers. The input layer contains a vector of features (an ACSF in the case of NNPs) and the output layer returns an observable, which can be a scalar or a vector (e.g. the total energy in NNPs). Each hidden layer consists of a number of nodes, and the input data is propagated forward through the different layers by performing a series of linear and non-linear operations which depend on the connection and the node in question, respectively. This propagation procedure is illustrated in figure 2(d), where the arrows represent the connections and the circles represent the nodes. We start out with a vector of real-valued symmetry functions $\mathbf{G}$ of a certain dimension, which depends on the number of species and the quality of the representation [69, 95]. Each of these functions G_i is propagated to each of the nodes in the first hidden layer multiplied by a series of weights $a_{ij}^{0,1}$ (where 0, 1 indicates we are connecting layers 0 and 1):

$$\beta_j^1 = \sum_{i=1}^{N_0} G_i a_{ij}^{0,1} + b_j^1, \tag{6}$$

$$\gamma_j^1 = f(\beta_j^1), \tag{7}$$

where N_0 is the number of nodes in layer 0, i.e. the number of ACSFs (or, equivalently, the dimension of $\mathbf{G}$) in this case. b_j^1 is the bias of node j in layer 1 which, together with the sum in equation (6), define the function β_j^1, which is linear in the input variable G_i. This quantity, β_j^1, is used as argument to evaluate a non-linear *activation function* f. The result of this evaluation, γ_j^1, is then passed on to the next layer $n = 2$ in the same way as above:

$$\beta_j^n = \sum_{i=1}^{N_{n-1}} \gamma_i^{n-1} a_{ij}^{n-1,n} + b_j^n, \tag{8}$$

$$\gamma_j^n = f(\beta_j^n), \tag{9}$$

where we note that N can in general vary from layer to layer. We have substituted G_i by γ_i^{n-1} for generality, because G_i is the notation used for the input layer specifically in the case of ACSF for NNPs. This procedure is repeated until we reach the output layer, which in our case returns a local atomic energy. The forces can be evaluated analytically from the dependence of the symmetry functions on the atomic positions.

Training an NNP consists in the optimization of the weights $\{a_{ij}^{n-1,n}\}$ and biases $\{b_j^n\}$, and is done using *backpropagation*, for a given training set of atomic structures, to minimize the error in the corresponding observables (total energies,

forces, stresses, etc). We will not go into the details of ANN algorithms which, for most practical purposes in atomistic materials modeling, can be considered a black box.

3.3.2. Gaussian approximation potential (GAP). GAPs are based on kernel regression [45] and are arguably more interpretable than ANNs. In GAPs, the local atomic energy $\bar{\epsilon}_i$ for atom i is expressed as a linear combination of kernel functions k:

$$\bar{\epsilon}_i = \delta^2 \sum_{t=1}^{N_{\text{train}}} \alpha_t k(\mathbf{q}_i, \mathbf{q}_t) + e_0, \tag{10}$$

where δ is an energy scale, t runs over training configurations, α_t are the fitting coefficients, $\mathbf{q}_i$ and $\mathbf{q}_t$ are the atomic descriptors (often a SOAP mb descriptor) of a test and train configuration, respectively, and e_0 is a per-atom energy offset, usually taken as the reference energy of an isolated atom of a given species. The kernel can be understood as a measure of similarity between two atomic environments, as illustrated in figure 2(e), and is bounded between 0 (nothing alike) and 1 (identical up to symmetry operations). Thus, intuitively, the more a training configuration resembles the test configuration for which we want to make a prediction, the more the fitting coefficient associated with that training configuration contributes to the prediction. This is why we stated earlier that GAPs are arguably more interpretable than NNPs.

Having cast the interpolation problem as a linear problem, training a GAP simply consists in a least-squares-based inversion of equation (10):

$$\boldsymbol{\alpha} = \frac{1}{\delta^2} \mathbf{K}^{-1} (\boldsymbol{\epsilon} - \mathbf{e}_0), \tag{11}$$

where now the test index in equation (10) also runs through training configurations, and we do not use the predicted atomic energy $\bar{\epsilon}$ but the observed one ε. We note that in practice one cannot train a GAP model (or an NNP, for that matter) using local atomic energies, which are not generally available before training the GAP. Instead, the local energy in equation (10) is replaced by the *sum* over local energies leading to a total energy observable. For instance, when using training data from a DFT calculation for a supercell, we use $E_I = \sum_i \bar{\epsilon}_i$. In addition to this total energy consideration, one usually needs to use regularization and sparsification to improve the stability, transferability and efficiency of a GAP, and may combine several GAPs in the same fit. These details fall outside the scope of this paper and the reader is referred to the literature for further insight [67, 70]. Likewise, the explicit definition and discussion of atomic descriptors and kernel functions is an active research topic and better covered elsewhere [49, 94, 96–98]. As for NNPs, the forces can be computed analytically through the dependence of $\mathbf{q}_i$ on the atomic positions. Forces and stresses can also be incorporated into the inversion equation, together with total energies.

3.3.3. Other MLP approaches. The field of ML-based atomistic simulation of materials is advancing fast. Since NNPs and GAPs appeared, several other MLP flavors have been developed and we expect applications in amorphous materials modeling to follow soon. MLP methods besides NNP and GAP include 'linear' models such as the moment-tensor potential [72] and the spectral neighbor analysis potential (SNAP) [99], or MLPs based on asymptotically complete atomic descriptors like the atomic cluster expansion (ACE) [100]. The first ACE-based MLP able to simulate a-C appeared very recently [101], and it is expected that these new models and improvements thereof [102, 103] will overtake NNPs and GAPs as the state-of-the-art tools for simulating disordered materials in the near future.

In the brief discussion of NNPs and GAPs above we implicitly include short-range interactions only, since ACSFs and SOAP use radial cutoffs that exclude all interactions beyond a certain radius. We have therefore left out long-range interactions which are important beyond the typical cutoffs used to fit 'regular' MLPs. These long-range interactions include van der Waals and electrostatics and must be treated on a different footing to bonding and repulsion interactions both (a) out of necessity, to avoid the explosion of computational time with the cutoff distance, and (b) out of opportunism, since these interactions can often be cast in the form of simple analytical functions whose *parameters* can be machine learned but are in effect short ranged [43, 104–107].

4. Amorphous and disordered carbon

The precise structure of a-C and how growth conditions can be tuned to modify it have been the topic of intense debate for several decades. The reason is that, unlike in a-Si, in a-C coordination environments with different number of atomic neighbors, ranging from two to four neighbors (*sp* and sp^3 orbital hybridizations, respectively) are all possible (meta)stable motifs. Especially three- (sp^2) and four-coordinated environments can coexist in a-C thin films. Varying the relative concentration of sp^2 and sp^3 bonding in a-C allows us to tune its material properties (mechanical, electrical, optical, etc) from graphite-like to diamond-like. Thus, much of the basic characterization work on a-C has dealt with the dependence of the sp^2/sp^3 ratio on such parameters as the deposition energy during physical vapor deposition growth and, in turn, the dependence of the material properties on the sp^2/sp^3 ratio. The reference entry point into the properties of a-C, albeit a bit outdated in terms of missing atomistic simulation insights that were developed in the last few years, is Robertson's monumental review paper from 2002 [8]. Two of the most important figures in that paper are reprinted here in figure 5. The top panel shows the dependence of the sp^2/sp^3 ratio on the deposition energy (the estimated kinetic energy of incident C atoms) for different experimental techniques used to grow a-C. A common trend is the increase in sp^3 content for increasing deposition energy, up to around 100 eV, after which there is a further decline. We also note that some of these a-C samples can be grown with extremely high sp^3 contents, approaching that of diamond (100% sp^3). Therefore, diamondlike or 'tetrahedral'

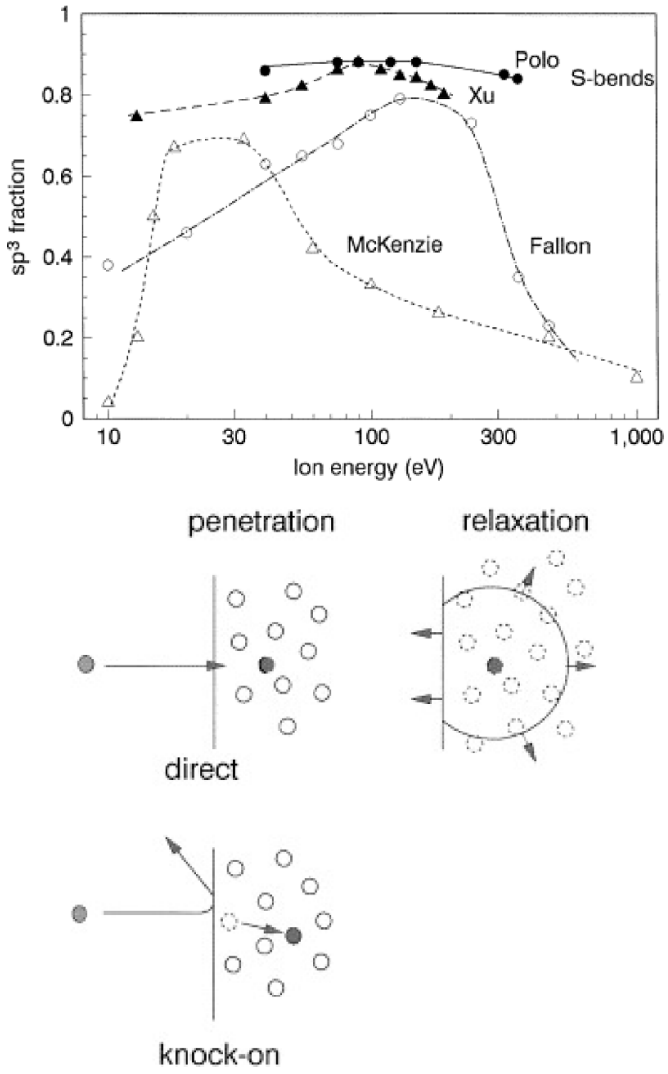

Figure 5. (Top) sp^3 fraction vs deposition energy from a series of literature works; data from Polo *et al* [108], Xu *et al* [109], Fallon *et al* [110] and McKenzie *et al* [111]. The deposition energy is the estimated kinetic energy of individual carbon atoms as they hit the substrate. (Bottom) The subplantation mechanism postulated as growth mechanism responsible for high sp^3 fractions in ta-C. Reprinted from [8], Copyright (2002), with permission from Elsevier.

a-C (DLC and ta-C, respectively) can be made experimentally, offering a route for cheap coatings with diamondlike hardness for tribological applications. The bottom panel in figure 5 shows the proposed film growth mechanism in ta-C, subplantation, which was widely regarded as the correct mechanism until recently, when MLP simulations [32] provided quantitative confirmation for earlier qualitative evidence [31] of an alternative mechanism, peening, which we will discuss later in more detail.

Thus, our journey into simulation of a-C starts with the extensive atomistic modeling efforts carried out in the pursuit of explaining how these high sp^3 contents can be achieved and, in turn, explaining the growth mechanism in a-C. The state of the art in atomistic modeling of a-C, as of 2017 (just 6 years before this review was completed), using a variety of interatomic potentials (including DFT) and modeling techniques, is summarized in figure 6. This figure was compiled right after the first MLP able to model a-C was published, also in 2017 [36]. We can see that the direct simulation route, deposition, fell short of achieving the very large sp^3 contents observed experimentally at high deposition energies. This technique had, back in the day, been restricted to fast empirical methods, such as tight binding (TB), C-EDIP and Tersoff potentials, and was (and still is) computationally unfeasible at the DFT level. Liquid quench simulations, on the other hand, were accessible to more expensive methods, like DFT, but could only predict very large sp^3 contents at unphysically high pressures, which shows as a shift towards higher mass densities on the graph. On that middle panel of figure 6 we note the first ML-based simulations of a-C generation with the MLP trained by Deringer and Csányi [36], a stepping stone in a-C modeling and key development leading to the advances that we will discuss later. Finally, pressure-corrected DFT simulations from [10, 29], based on a two-step relaxation procedure, managed to get extremely good agreement with experiment for the sp^3 content as a function of mass density but, based on an indirect simulation protocol, offered no insight whatsoever into the growth mechanism.

4.1. Explaining the growth mechanism

With the introduction in 2017 of the first MLP able to handle the structural complexity of a-C with close to DFT accuracy, but at a fraction of the computational cost, the first MLP-based simulation depositions followed soon. In 2018, Caro *et al* [32] presented MD deposition simulations of ta-C growth. In these simulations, incident atoms with varying kinetic energy (20, 60 and 100 eV) impinge on a carbon substrate, initially diamond. After each impact, the system is equilibrated back to the nominal deposition temperature (300 K) and the next deposition event takes place. After several thousands of atoms have been deposited, the size of the film is enough to collect statistics for material properties and growth mechanism. The workflow of these deposition simulations is shown in figure 7(a). Panel (b) of the figure shows a more detailed view of the impact process for a single event with a logarithmic time axis. Initially, the incident atom approaches the substrate very fast. The MLP algorithm allows us to monitor the local atomic energy, as we have discussed in section 3.3, shown in the figure offset by the average local energy in the growing film. The highly energetic initial impact is followed by equilibration of the atomic environment, where atoms settle in their new positions. GAP MLPs also allow us to monitor the predicted interpolation error, shown in the figure too. This deposition process is rather complex, with the order of 50 bond breaking/formation events taking place for each impact at around 100 eV [33]. The process is better visualized as a video animation, with several Open Access resources available from the literature, including a single impact [117], the atom-by-atom growth of a-C thin films from low to high density [118], and the resulting atomic structures in XYZ format [119] (which enable subsequent studies).

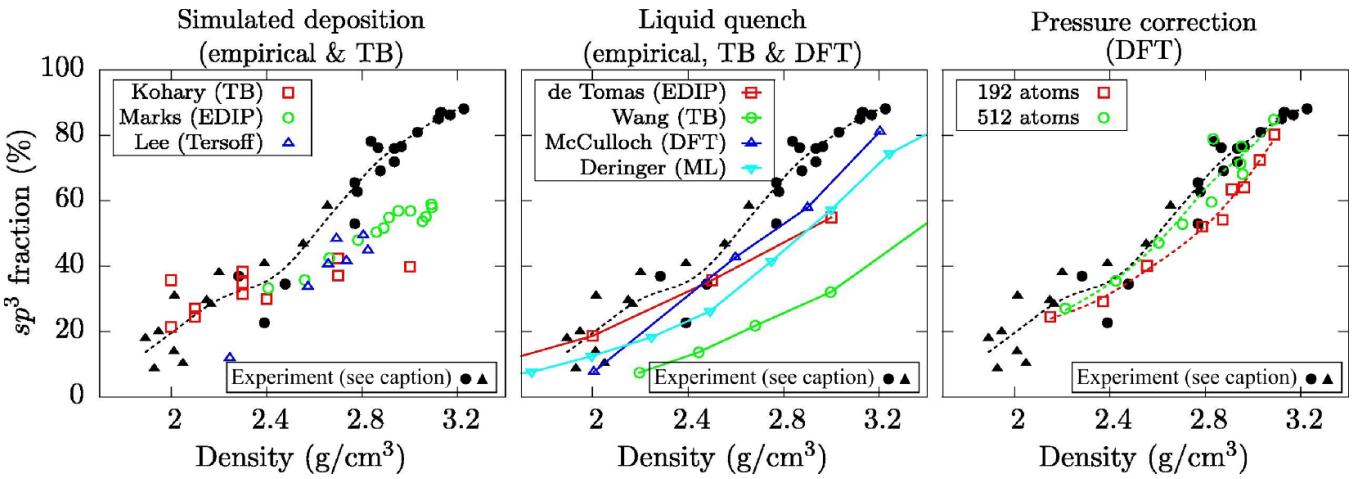

Figure 6. sp^3 fractions vs mass density for different simulation protocols, compared to experimental data (black solid dots) from Fallon *et al* [110] and (black solid triangles) Schwan *et al* [112]. Simulation data from Kohary and Kugler [113], Marks [31], Lee *et al* [114], de Tomas *et al* [35], Wang and Komvopoulos [115], McCulloch *et al* [116] and Deringer and Csányi [36]. Pressure-corrected DFT data is taken from Caro *et al* [29] for 192-atom supercells and Laurila *et al* [10] for 512-atom supercells. Reprinted from [10]. CC BY 4.0.

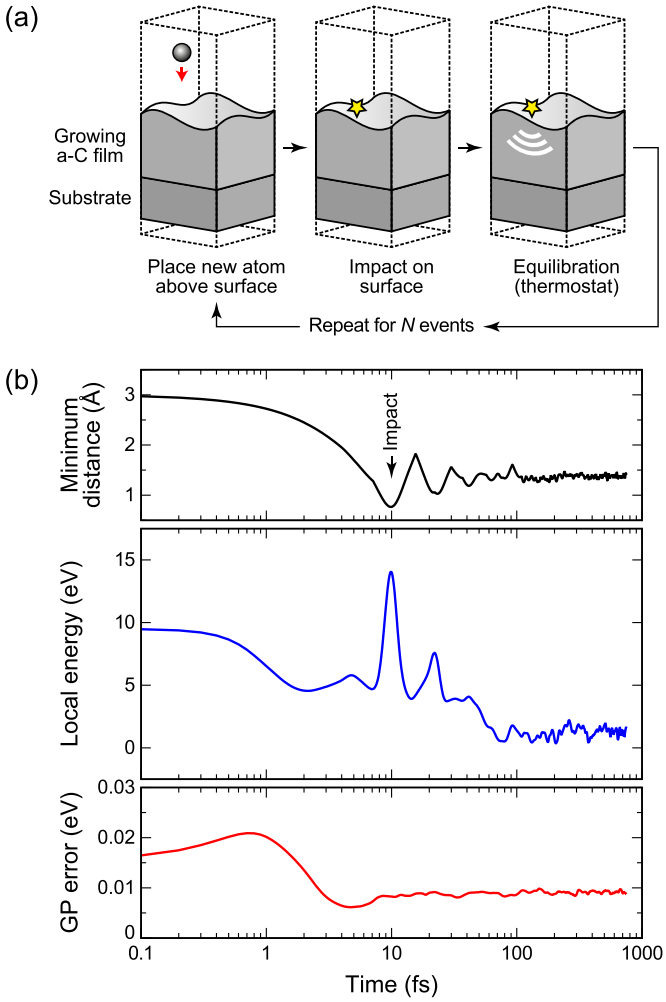

Figure 7. (a) Schematics of a deposition simulation. (b) Evolution of different observables during the course of a single impact event. Reprinted figure with permission from [33], Copyright (2020) by the American Physical Society.

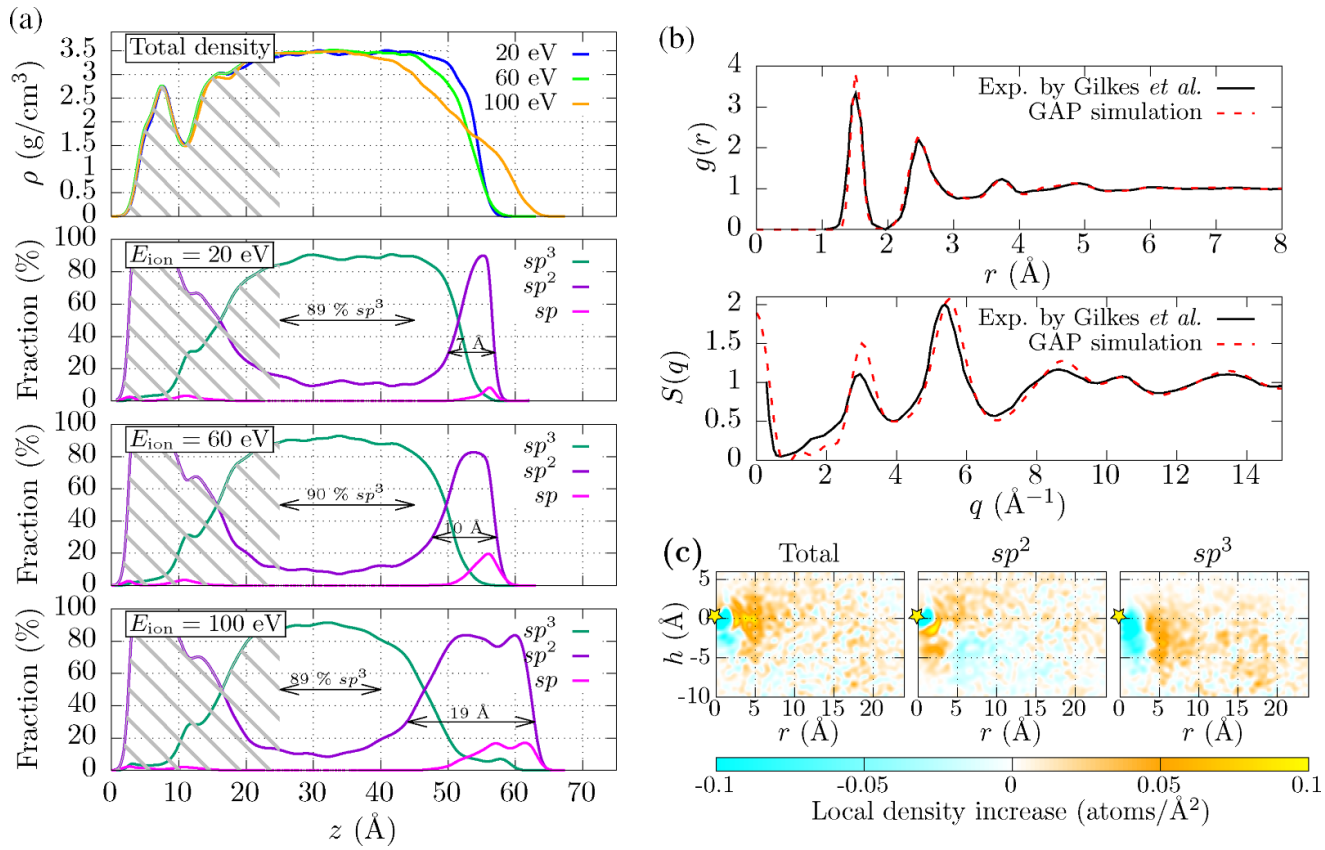

Figure 8. Results of ta-C deposition simulations: (a) evolution of mass density and coordination fractions along the film's growth direction in ta-C over three different deposition energies. The extent of the sp^2-rich surface, different for each energy, is indicated with an arrow; (b) radial distribution function and structure factor, compared to experimental results from Gilkes *et al* [120]; (c) two-dimensional pair-correlation functions indicating the regions of depletion/formation of sp^2 and sp^3 motifs as a function of distance from impact location. Reprinted figure with permission from [32], Copyright (2018) by the American Physical Society.

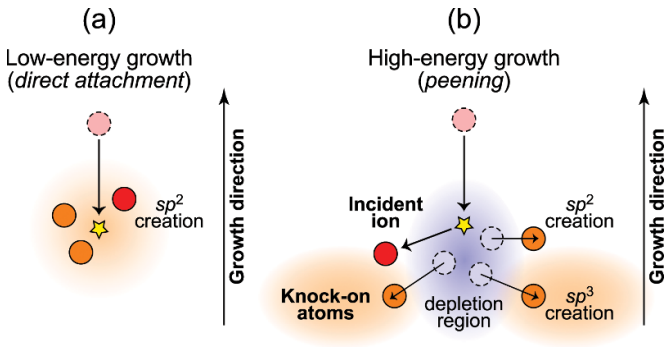

Figure 9. Observed growth mechanisms in (a) low-density (low deposition energy) and (b) high-density (high deposition energy) a-C films. Reprinted figure with permission from [33], Copyright (2020) by the American Physical Society.

Figure 8 shows the key results from these first deposition simulations [32]. Panel (a) shows the mass density and $sp/sp^2/sp^3$ fraction profiles along the growth direction for the deposition energy ranges where ta-C growth takes place. The three simulations at 20, 60 and 100 eV result in similar mass densities and coordination fractions in the bulk of the film, for the first time close to those reported experimentally for the densest ta-C films. At the same time, the data shows rather different surface morphologies, with increasing surface roughness for higher deposition energies, a result that closely follows experiment [121]. This, together with the also excellent agreement with experiment for the radial distribution function (RDF) shown in figure 8(b) gave confidence in the quality of the simulations as representative of the microscopic growth mechanism taking place experimentally. The collection of deposition statistics (up to 8000 individual events per energy) then enabled drawing a precise picture of what that growth mechanism actually looks like. In figure 8(c) we can see the mass density and coordination fraction increase/decrease before and after an impact event, averaged over all impacts, as a function of depth and lateral separation from the impact site. The color maps clearly indicate a local decrease of sp^2 and, especially, sp^3 carbons around the site of impact. In fact, these maps show that locally (around the impact site) there is an increase in the amount of sp^2 carbon, whereas the sp^3 fraction increases laterally and away from the impact site, due to pressure waves originating from the impact region. This mechanism is known as peening, and had already been proposed by Marks in 2005 on the basis of deposition simulations with the C-EDIP empirical potential [31]. Because the C-EDIP simulations lacked quantitative agreement with experiment

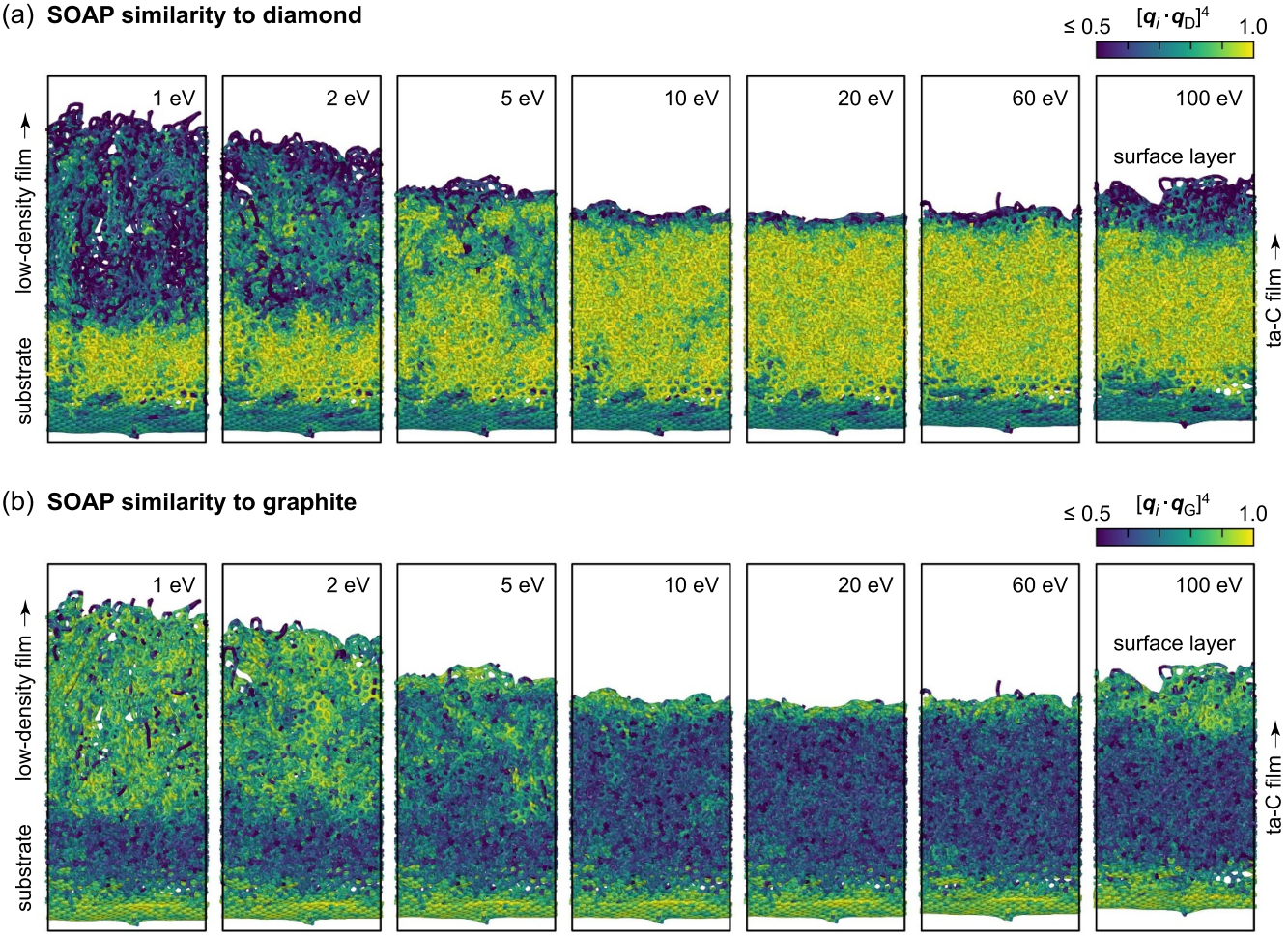

Figure 10. Evolution of a-C film nanostructure as a function of deposition energy. The top (a) and bottom (b) panels show the degree of similarity between atomic environments in the films and reference diamond and graphite, respectively. Brighter color indicates more resemblance and darker color indicates less resemblance. Reprinted figure with permission from [33], Copyright (2020) by the American Physical Society.

(cf figure 6), the peening mechanism did not gather generalized adoption. However, the GAP deposition simulations offer strong quantitative support for peening as the growth mechanism in ta-C, which is schematically illustrated in figure 9(b).

4.2. a-C structure across mass densities

Following the study of ta-C growth, these deposition simulations were extended to low-density a-C and a more comprehensive analysis of material properties and comparison with other interatomic potentials was carried out [33]. At low deposition energies (below the typical cohesive energy per atom in carbon materials, $\lessapprox$9 eV), graphitic a-C grows by direct attachment (figure 9(a)), where higher coordination increases the stability of *sp* surface motifs by creating sp^2 and sp^3 carbon. At very low deposition energies of a couple of eV the rate of sp^3 formation is small and the bulk of the a-C films is highly graphitic, with ∼80% sp^2 and similar amounts of *sp* and sp^3 motifs (∼10% each). Figure 10 shows the evolution of the structure, as well as the 'graphite likeness' and 'diamond likeness', of the simulated thin films as a function of deposition energy. The similarity to graphite and diamond is computed by calculating the SOAP kernels between each individual atomic environment and either pristine graphite or diamond [33, 36]. Since the kernel can be understood as a similarity measure, bounded between 0 and 1, as we have previously discussed, it can also be used for this kind of quantitative comparison in a very straightforward way. At and above ∼10 eV the structure is largely homogeneous and dominated by sp^3 motifs in the bulk of the film. At 5 eV sp^2 and sp^3 motifs are approximately equally frequent, and the material shows a 'patched' structure. At low deposition energies the material is graphitic in nature, as we have already discussed, and made of tubular (nanotube-like) structures and highly defective graphitic sheets.

4.3. a-C surface structure

The surface structure in a-C is always graphitic like (figures 8 and 10), but the extent of this sp^2-rich surface layer is strongly dependent on the deposition energy. The MLP deposition

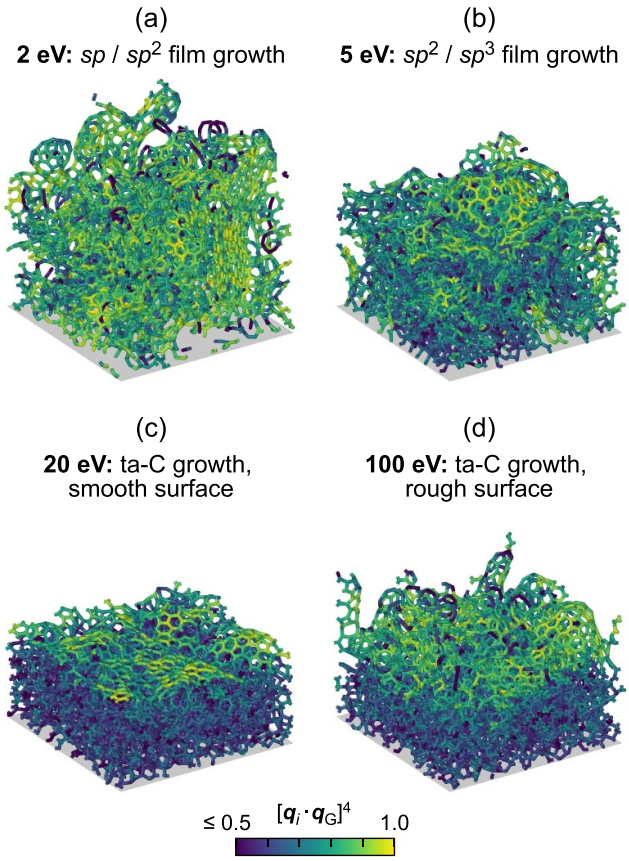

Figure 11. Detail of the surface nanostructure in a-C films grown at different deposition energies, indicating the atomic motifs similarity to graphite. Reprinted figure with permission from [33], Copyright (2020) by the American Physical Society.

simulations show a smoothest surface at around 20 eV and roughest at 100 eV (and, presumably, higher energies, which have not been studied) [32]. For very low deposition energies there is no well-defined surface region, except for a higher relative abundance of *sp* motifs within the top 10 Å or so, and the film is graphitic and highly disordered throughout. A closeup on the structure of a-C surfaces for different mass densities, again color coding according to graphite likeness, is given in figure 11.

4.4. Doped a-C

As-deposited a-C is rarely completely free of impurities. Residual elements present in the reactor chamber and sample setup lead to unintentional doping of a-C films with a wide range of chemical species, the most significant of which are H, O, N and Si (see, e.g. time-of-flight elastic recoil detection analysis [122] results of the elemental makeup of a-C:N [13]). Besides unintentional doping, it is possible to incorporate impurities in order to achieve a desired effect. The mechanical, electronic and (electro)chemical properties of a-C can be modified via H, O and N incorporation [8, 12, 13]. Intentionally doped a-C is usually denoted by a-C:X, where X stands for the dopant. The most common doped form of a-C is hydrogenated a-C, or a-C:H, where typical H contents are of the order of 30–50 at.-% [8]. This material can be made for instance by depositing C in a H/methane plasma, or by depositing acetylene or methane molecules directly, depending on the desired C/H ratio [8]. The properties of a-C:H differ from those of a-C in that H atoms will saturate many bonds, potentially leading to high sp^3 contents but relatively low mass densities, because a H atom is about 12 times as light as a C atom. The mechanical properties of a-C:H, e.g. its elastic moduli, will be inferior to those of ta-C with similar sp^3 content [8].

Computational studies on a-C:H are comparatively rare, and direct deposition simulations of a-C:H with MLPs are not yet available in the literature (although our group is currently exploring this possibility). Indirect simulations of a-C:H formation and H adsorption energetics mixing MLPs and DFT or other electronic-structure methods have appeared in recent years. Deringer *et al* [123] did grand-canonical Monte Carlo (GCMC) simulations of H adsorption using density-functional TB from a wide set of preexisting a-C surface models. These results showed that a-C:H materials with very high sp^3 fractions (∼75%) can be obtained over a relatively large range of H concentrations (ranging from 25% to 40%). The results of these GCMC trajectories and snapshots of the resulting films are given in figure 12. Caro *et al* [124] focused on the individual adsorption processes and how those depend on the geometry of the preexisting pure carbon sites, finding two separate regions of adsorption stability for *sp* and sp^2 sites each. Almost-linear and almost-planar *sp* and sp^2 motifs, respectively, are more stable and therefore less

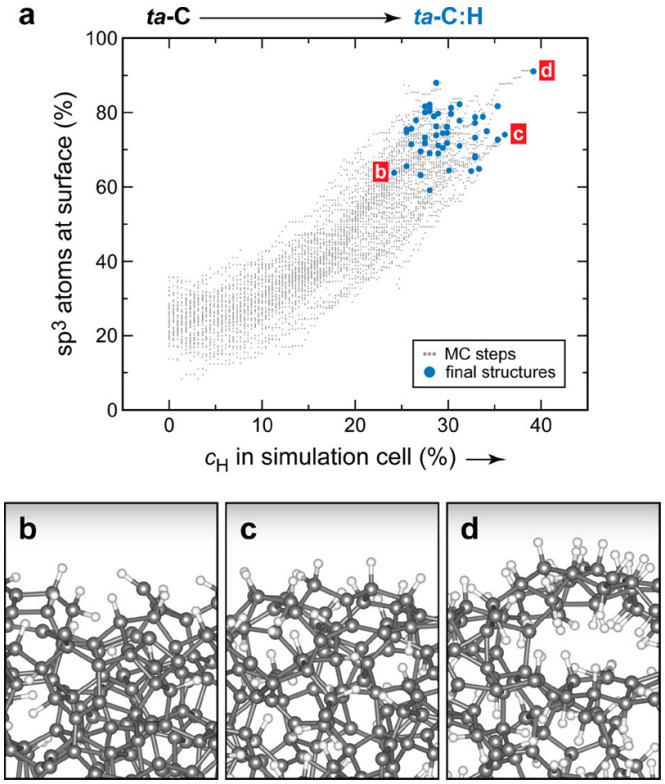

Figure 12. (a) sp^3 fraction of hydrogenated a-C as a GCMC simulation progresses where H is incorporated. The blue dots mark the end points of the simulation. (b)–(d) Ball-and-stick representation of the final a-C:H structures. Reprinted with permission from [123]. Copyright (2018) American Chemical Society.

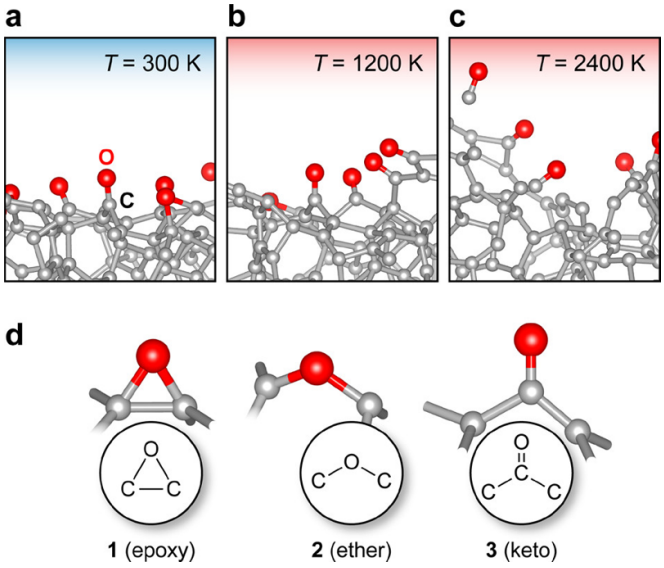

Figure 13. (a)–(c) Ball-and-stick representation of a-C:O DFT-MD simulations carried out at different temperatures. (d) Most representative O-containing motifs present in the resulting a-C:O structures. Reprinted with permission from [123]. Copyright (2018) American Chemical Society.

reactive towards H adsorption than more bendy motifs. Caro *et al* [124] also introduced a series of ML models reminiscent of MLPs to accurately predict adsorption energies. These models are based on SOAP descriptors and were augmented with electronic structure information, directly incorporating the local density of states (DOS) into the structural descriptor, which allowed to significantly increase the model accuracy.

Growth of a-C:O and a-C:N usually takes place by introducing O_2 and N_2 into the deposition chamber [12, 13]. As with a-C:H, the amounts of dopant that can be incorporated is significantly higher than in traditional doped semiconductors, with maximum values of at least 30 at.-% [12, 125] and 10 at.-% [13] for O and N incorporation, respectively, reported in the literature. By adjusting the partial gas pressure the concentration of dopants can be adjusted. Again [123, 124], appear to be the only published work where MLPs were used to study O incorporation into a-C, although this was done less comprehensively than for H. In particular, [123] did DFT-based MD simulations of a-C surface oxidation starting from a MLP-generated surface model. The resulting structures are shown in figure 13. This work established the predominant oxygen-containing motif in a-C:O surfaces being keto-like groups. ML models have also been recently used to understand the structure of H- and O-containing disordered carbon materials through atomistic simulation of XPS [125]. XPS and other spectroscopies are expected to provide increasingly stronger links between experiment and simulation as ML techniques for atomistic structural characterization continue to evolve.

We are not aware of atomistic studies of a-C:N based on MLPs, although our group is currently developing an MLP able to handle the CN system over a wide range of structures, including a-C:N. We are also developing MLPs for the CH and CO systems, which will hopefully shed light onto the structure and properties of a-C:H and a-C:O. Our more ambitious objective in the longer term is to combine these into a CHO(N) MLP able to accurately describe a wide variety of carbon-based materials and molecules under different thermodynamic conditions. This objective will likely be achieved, either by us or by others, within the next couple of years.

4.5. *Nanoporous carbon*

Nanoporous (NP) carbon is related to low-density (highly graphitic) a-C. They share the overall graphitic nature of their chemical bonds and the lack of long-range order, but NP carbons are organized in less defective graphitic layers with very low *sp* or sp^3 content. The usefulness of NP carbons resides in their porous structure and how it can be exploited in particular for ion intercalation in energy-storage solutions [17], such as Li-ion or Na-ion batteries and supercapacitors. GAP-derived structural models have already been used to understand intercalation mechanisms and diffusion in

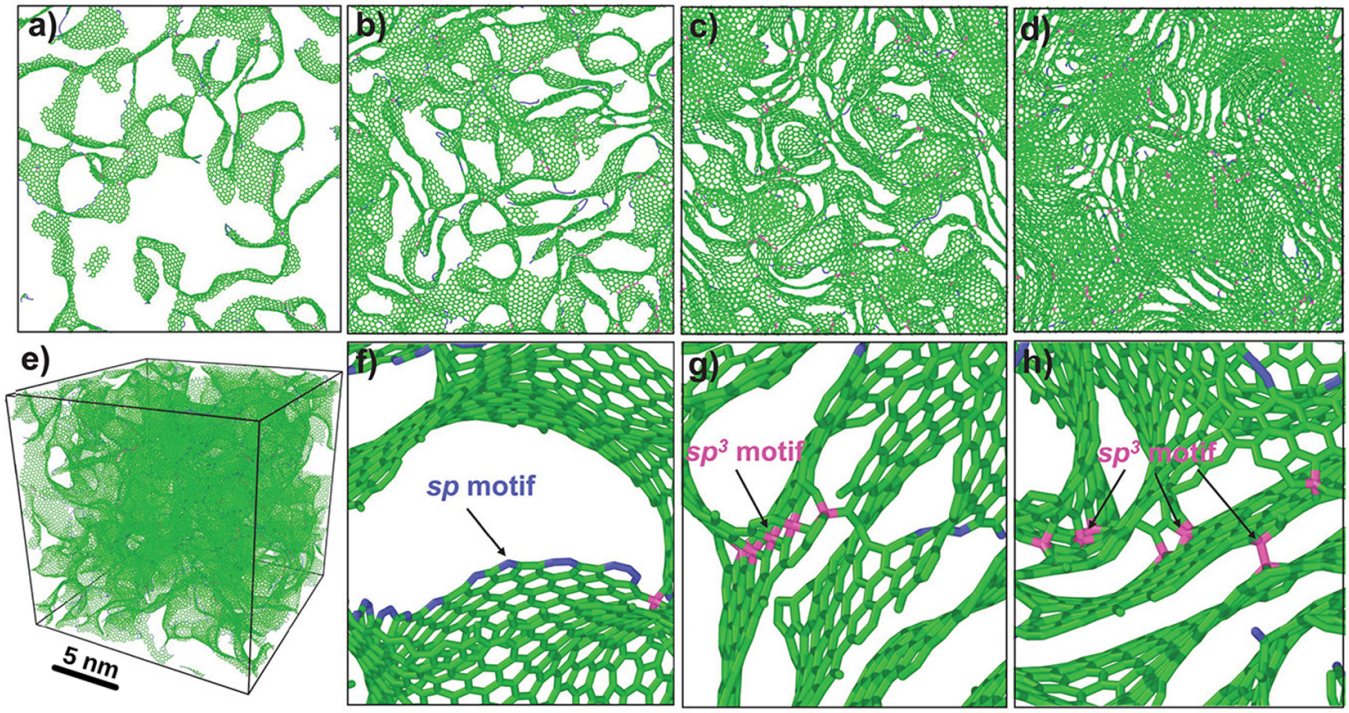

Figure 14. (a)–(d) Nanoporous carbon structures of different densities, where the density increases to the right and the average pore size decreases correspondingly. (e) 3D model of the low-density nanoporous carbon structure from (a), where the pore morphology can be more easily appreciated. (f)–(h) *sp* motifs are found in graphitic sheet termination (edges) and sp^3 motifs are found interlinking stacked graphitic layers, conferring three-dimensional rigidity to these nanoporous carbon networks. Reprinted from [37]. CC BY 4.0.

NP carbon materials [126, 127]. g-Cs of different densities can be generated computationally following a 'graphitization' protocol. This is a special kind of melt-quench simulation where there is a long annealing step at the graphitization temperature [35] which, for GAP MLPs, is around 3500 K [44]. MLPs have been used to study the intercalation of Li and other alkali-metal ions in g-Cs using small-scale structural models [128, 129]. More recently, GAP simulations by Wang *et al* [37] have produced high-quality large-scale (>130 000 atoms) structural models of NP carbon throughout a wide range of mass densities, some of which are shown in figure 14. In these materials, the relative abundance of 5- and 7-ring defects (the stable ring motif in graphite is a 6-ring) determine the curvature of the graphitic planes and thus the pore morphology. There are slightly more 5-rings than 7-rings in these materials, and nanopore sizes and morphologies seem to be rather homogeneous for a given mass density, according to the results of this study. The mechanical properties of the materials were found to evolve smoothly with density. We note that these NP structural models are already one order of magnitude bigger than those also obtained with MLPs just a couple of years prior, highlighting the rapid pace of development in the field.

4.6. Disordered carbon under extreme conditions

Atomistic simulation is a particularly attractive approach to study matter under conditions that make direct experimentation complicated or even impossible. This is the case for high-temperature and high-pressure conditions under which some allotropes are stable (notably, diamond is stable at very high pressures). The landscape of carbon allotropes is particularly rich given the flexibility of carbon covalent bonding. Traditional search strategies for new crystals can be accelerated by using MLPs which are able to navigate the PES with close to *ab initio* accuracy but orders of magnitude faster. New carbon allotropes have been found following this approach [130]. MLPs also enable us to go one step further thanks to the increased computational efficiency and chart phase transformations for disordered materials explicitly (i.e. beyond the small unit cells used for crystal structure search). An example of this is the large-scale study of the phase diagram of C_{60} carried out by Muhli *et al* [43]. In this work, phase transformations from a C_{60} precursor at high temperatures and pressures were simulated with an MLP, successfully leading to the prediction of a transformation to amorphous diamond (a-D) from the collapsed precursor, later observed experimentally at similar thermodynamic conditions [15]. The detailed phase diagram, shown in figure 15, required structural models with thousands of atoms to correctly describe the configurational disorder. Furthermore, this work exemplifies the power of MLP simulation, where accuracy can be maintained within a unified methodological framework across a wide range of thermodynamic conditions, from low pressures and temperatures where weakly bonded (e.g. van der Waals) interactions dominate to the extreme conditions at which a material phase collapses into another. This highlights the potential of MLPs to chart unknown phases of

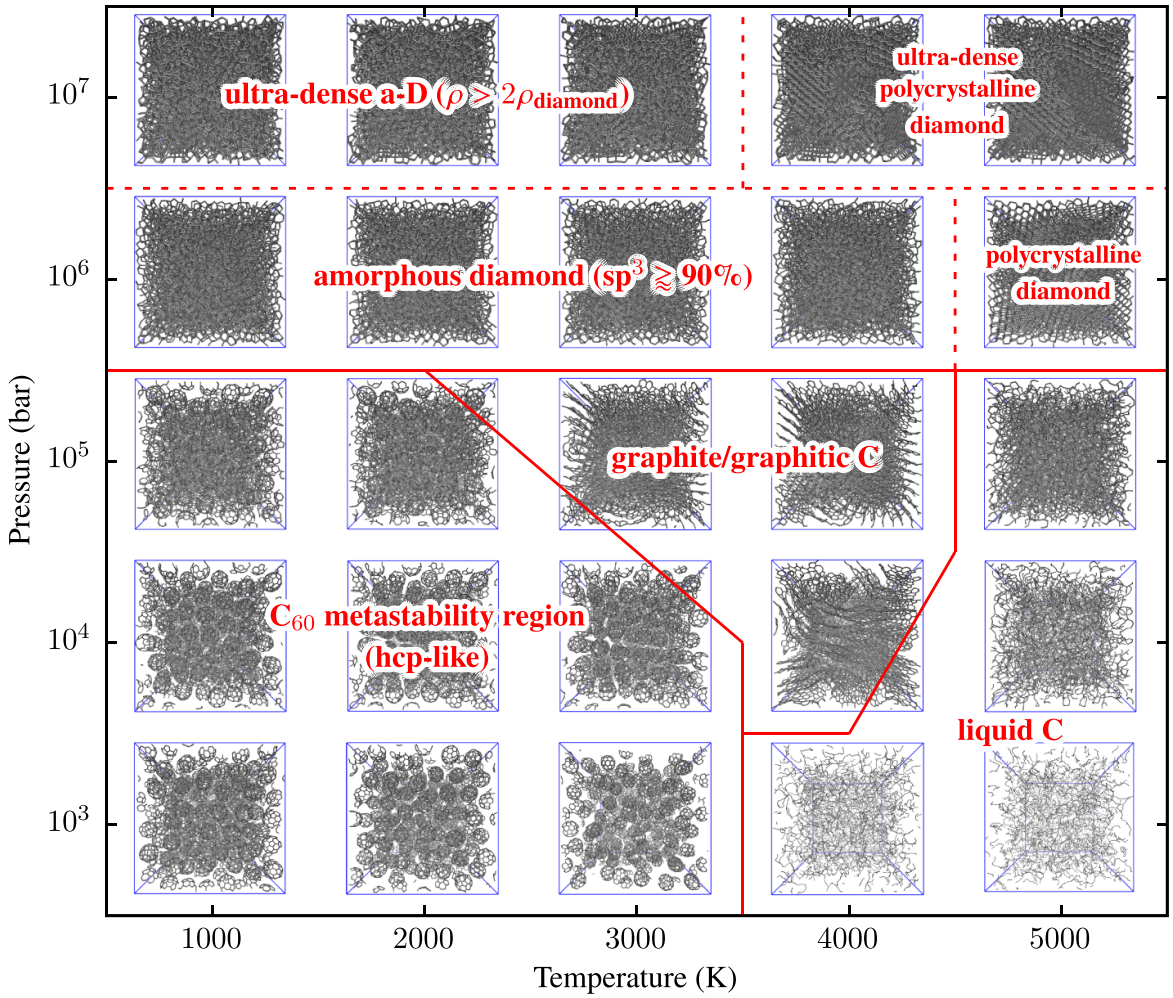

Figure 15. High-pressure/high-temperature phase diagram of C_{60}, where the molecular precursor leads to nucleation of high-density ta-C ('amorphous diamond') as the molecular C_{60} collapse at high pressure. Reprinted figure with permission from [43], Copyright (2021) by the American Physical Society.

materials taking different precursors as starting point and applying a range of physical transformations on them, which is of particular importance for the discovery of new carbon materials.

5. Amorphous silicon

Just as they have opened up new avenues in carbon simulation, MLPs have also enabled computational atomistic studies of silicon that were out of reach just a few years ago. Silicon is the archetypical semiconductor and, for this reason, the two seminal papers on atomistic ML for materials modeling used Si as a proof-of-concept material [45, 64]. Indeed, Behler and Parrinello even looked at the MLP-predicted structure of liquid Si and compared it, favorably, to DFT results [64]. Possibly, this was the first-ever MLP simulation of a 'disordered' material. Much has happened in MLP modeling of a-Si in the few years since those seminal papers appeared, which is summarized in this section.

5.1. General-purpose Si MLPs

The first prerequisite on the way to accurate atomistic simulation of an amorphous material is the availability of a general-purpose potential. The first MLP of this type for silicon was introduced in 2018 (the year following the introduction of the first general-purpose carbon MLP) by Bartók *et al* [47]. The authors lucidly proposed a materials-property benchmark as a more relevant accuracy test for their potential than the prevalent train/test splits. These comprehensive tests are summarized in figure 16, showing how this GAP MLP outperforms a wide selection of other potentials available at the time of the comparison. The tested properties include elastic moduli, surface energies, point defects, and planar defects. Further tests not shown in the figure included bulk crystal properties, liquid and a-Si RDFs, phase diagram, phonons, and more, including a simulation of crack propagation. These tests give confidence in the quality and broad applicability of the MLP to diverse problems, an important requirement for modeling the complex structure of a-Si with high accuracy, especially to obtain the

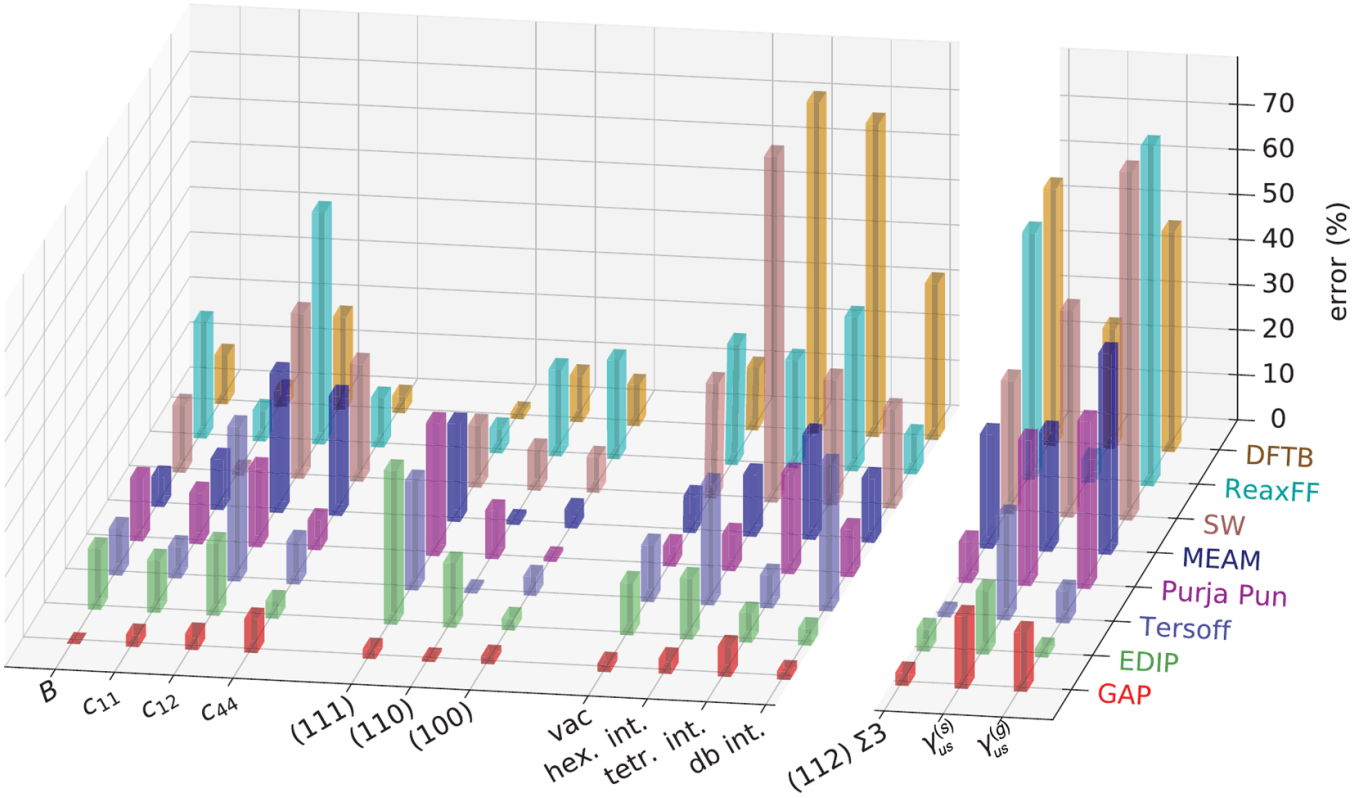

Figure 16. Comparison between the silicon GAP18 and other force fields for predicting a number of properties; from left to right: elastic moduli, surface energies, point defect formation energies, and planar defect formation energies. Reprinted from [47]. CC BY 4.0.

correct concentration of coordination defects, both under- (3-fold) and over-coordinated (5-fold) motifs. This Si GAP was extensively validated specifically for a-Si simulation in [40]. The database of structures generated by Bartók *et al* has been used to train new versions of the MLP [46, 131], and their MLP has enabled important subsequent work elucidating the atomic structure of a-Si, summarized below.

5.2. The atomic structure of a-Si

Compared to a-C, the structure of a-Si may seem relatively simple since every motif which is not made up of a 4-fold coordinated atom is a coordination defect. However, the defect concentration can have a massive impact on the optoelectronic properties of this material and, therefore, a force field's success at modeling a-Si resides in being able to capture these subtleties. In particular, the predicted relative formation energies for over (5-fold) and under (3-fold) coordinated defects under varying local strain field will determine the quality of the computational structural models that can be generated. In figure 16 we see how the GAP MLP from [47], let us call it GAP18 for short, outperforms all other available force fields in simultaneously predicting the correct elasticity and defect energetics in silicon. This basis provides the confidence required to trust the a-Si structures derived from melt-quench simulations. This confidence is reinforced by comparing the structure factor of simulated a-Si with experimental results. Figure 17 shows results from [40, 131] using GAP18 [47] and a NN MLP ('neuroevolution potential', NEP [136]) trained from the GAP18 database, respectively. Both works derive similar levels of coordination defects, 0.5% 3-fold and 1% 5-fold defects from [40], and 0.45% 3-fold and 1.45% 5-fold defects from [131]. These MLP simulations show very good agreement with experimental structure factors over a wide range of wavelengths, free of the artifacts encountered by other simulation methods, as shown in figures 17 (left-bottom) and 18 (for the RDF, which has information equivalent to that contained in the structure factor).

The medium-range order, often quantified in terms of ring counts [137], shows prevalence of the stable 6-membered ring motif (slightly less than one per atom, on average) followed by relatively large amounts of defect ring motifs: 7-membered ($\lesssim$0.6 per atom), 5-membered ($\lesssim$0.4 per atom) and 8-membered ($\lesssim$0.1 per atom), followed by almost negligible amounts of larger and smaller rings [40, 131]. Comparing to carbon, a-C has significantly broader ring distributions at low density but similar ones at high density [29], whereas NP carbon has significantly narrower distributions centered on 6-membered rings and skewed towards 5-membered rings, instead of towards 7-membered rings [37].

The most sophisticated study, to date, on the structure and structural transitions in disordered Si have been recently presented in the hallmark study by Deringer *et al* [138]. The authors used large-scale structural models with up to

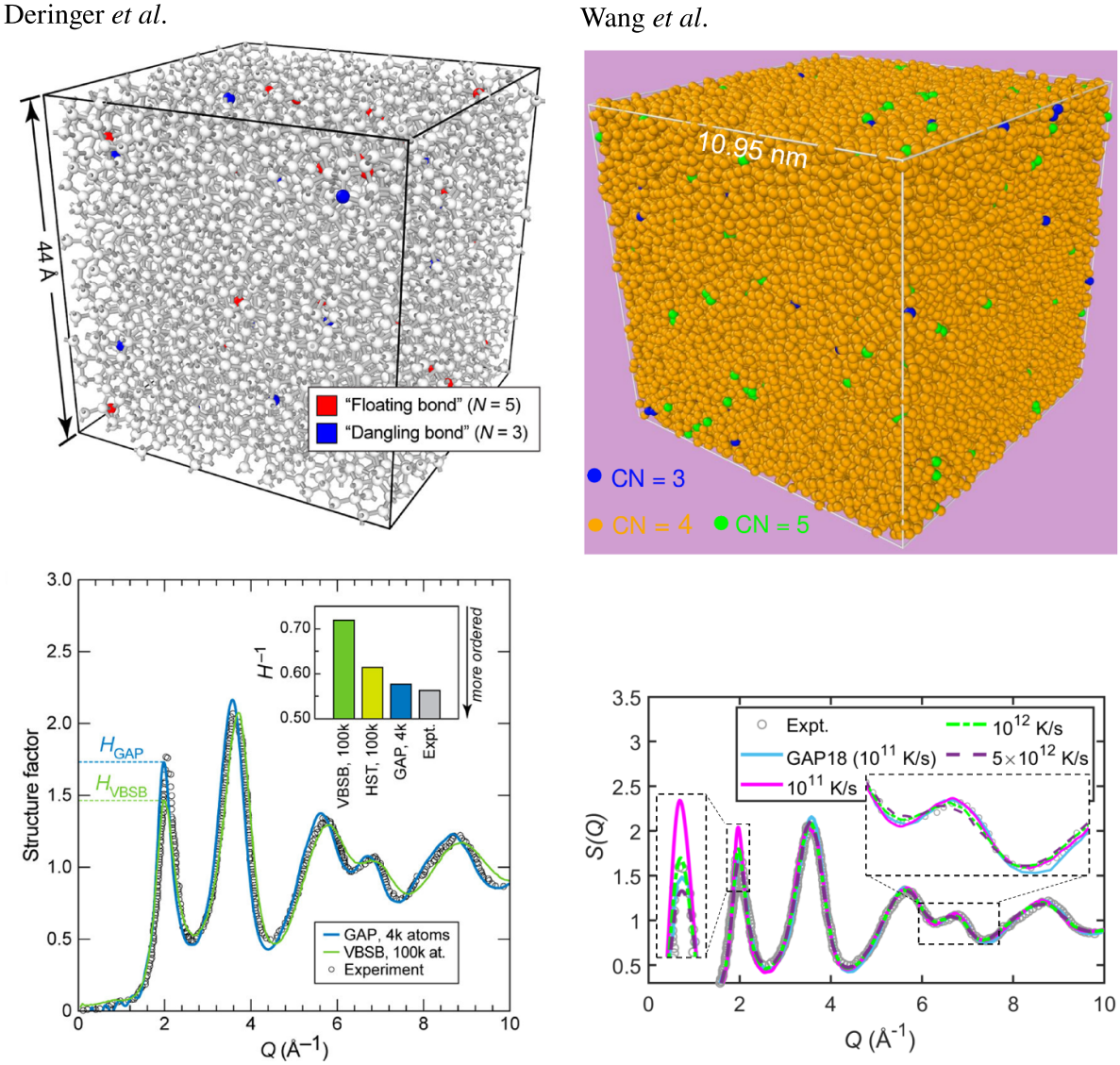

Figure 17. Computational a-Si structural models and their corresponding structure factor from Deringer *et al* [40] (left panels) and Wang *et al* [131] (right panels). Deringer's structure factor is compared to that derived from preexisting structural models from the literature, VBSB [132] and HTS [133], and to experiment. Wang shows the effect of the annealing rate on the structure factor and a direct comparison to Deringer's data (labeled GAP18). In both cases the experimental data is from Laaziri *et al* [134, 135]. Left: Reproduced from [40]. CC BY 4.0. Right: Reprinted figure with permission from [131], Copyright (2022) by the American Physical Society.

100k atoms and long-time-scale MD simulations to reproduce the liquid-to-amorphous phase transition at high temperature (around 1180 K). This transition is characterized by a rapid decrease in the number of over-coordinated ($N_{\text{neighbors}} > 4$) structural motifs and rapid increase in the resemblance of *local* a-Si atomic motifs to those in crystalline Si, even in the absence of mid- or long-range order. An even more impressive part of this work is the characterization of a highly non-trivial high-pressure phase transition between the amorphous semiconducting phase and the (poly)crystalline metallic phase via intermediate nucleation of crystallites embedded within the amorphous matrix (figure 19). The metallic nature of the high-pressure crystalline phase was established with a previously developed ML model for the electronic DOSs [139]. Indeed, integration of ML interatomic potentials with other ML-based approaches that feed on similar or the same descriptors, for example charge partitioning for model parametrization [43] or core-level energies for x-ray spectroscopy [125], is opening the door for ever more sophisticated simulation of the properties of disordered materials.

5.3. *Properties of a-Si*

While insight into the structure of a-Si is interesting on its own, for device applications we are also interested in emerging material properties, such as electronic band gap or thermal

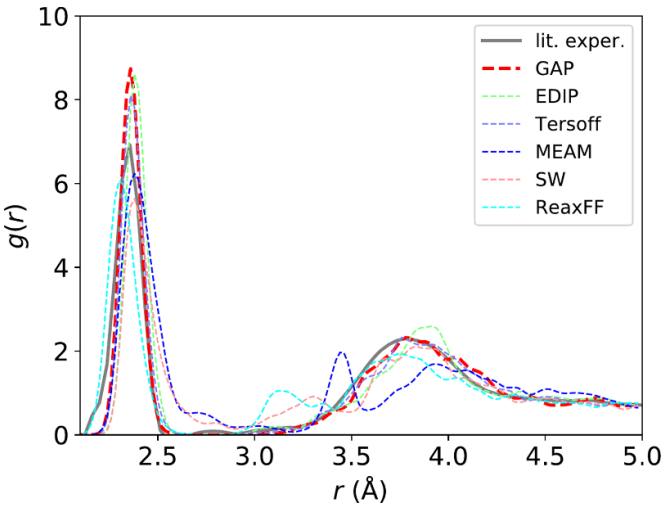

Figure 18. GAP results for the RDF of a-Si from Bartók *et al* [47], and comparison to the RDF obtained from experiment [134, 135] and using different popular interatomic force fields for silicon. Reproduced from [47]. CC BY 4.0.

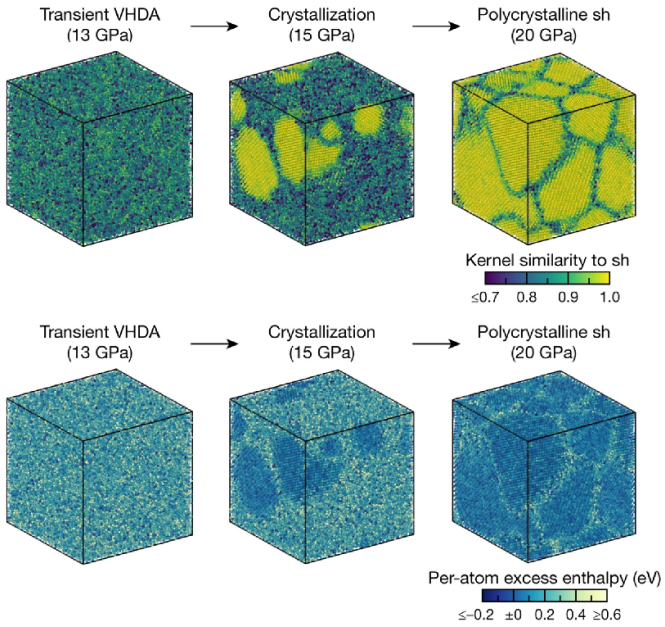

Figure 19. High-pressure phase transition in silicon from a very-high-density amorphous (VHDA) phase to a (metallic) polycrystalline phase. Reproduced from [138], with permission from Springer Nature.

transport mechanism. To predict and understand these properties the key lies in finding the link between them and the underlying atomistic structure of the material. The number of MLP-driven studies of these properties in a-Si still lags behind the more developed literature on its atomistic structure, for the simple reason that the existence of reliable structural models precedes the calculation of properties, which must necessarily rely of the availability of those models. We expect to see rapid development in characterization and prediction of the properties of a-Si within large-scale atomistic simulation in the next

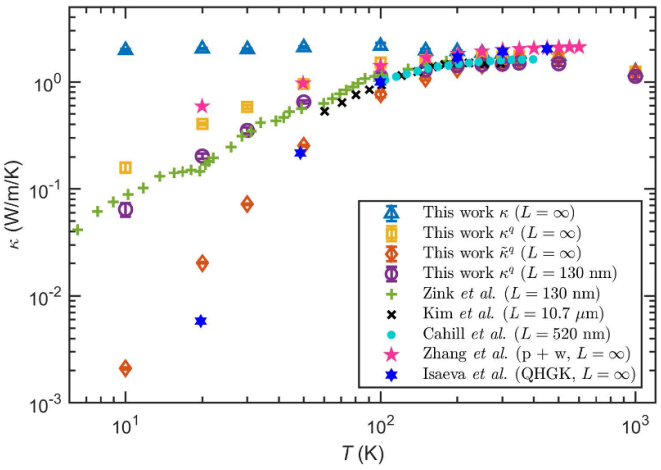

Figure 20. Evolution with temperature of the thermal conductivity of a-Si; comparison between simulation and experiment. 'This work' here refers to [131]. Experimental references are to Zink *et al* [141], Kim *et al* [142], Cahill *et al* [143], Zhang *et al* [144] and Isaeva *et al* [145]. Reprinted figure with permission from [131], Copyright (2022) by the American Physical Society.

few years. We mention here a recent example on thermal conductivity in a-Si. Wang *et al* [131] used an ANN-based NEP MLP [136], implemented in the GPUMD code [140], to perform highly efficient homogeneous non-equilibrium MD simulations of thermal transport in a-Si. This study has been able to closely reproduce the experimental evolution of the thermal conductivity of a-Si as a function of temperature and finite size (figure 20). This opens the door, in the near future, to simulating thermal properties of materials yet to be synthesized, with a high level of confidence in the accuracy of the computational results, in turn providing the basis for accelerated materials discovery and property-based materials design.

5.4. Hydrogenated a-Si

While a-Si is an interesting material from a fundamental scientific perspective, as the prototypical amorphous semiconductor, hydrogenated a-Si (a-Si:H) is arguably more important from a technological point of view. a-Si:H is commonly used in solar cells, where the intentional doping with H heals the coordination defects that are present in undoped a-Si, improving material properties towards photovoltaic applications. Unfortunately, the introduction of additional chemical species makes the development of accurate MLPs more challenging, because of the larger configuration spaced spanned. At the same time, the CPU cost of an MLP calculation with multiple species is more expensive ('curse of dimensionality') than for single species, with descriptor construction typically scaling between exponentially (worst-case scenario) and linearly (best-case scenario) with the number of species. For this reason there are comparatively few studies on a-Si:H or Si alloys compared to pure Si. On the other hand, thanks to recent developments in MLP technology, such as descriptor compression [103], and the improved collective expertise gained by the community on how to generate good databases to train

Table 1. List of some of the publicly available MLPs to simulate disordered C and Si, together with practical information. 'Code' refers to the computer software which can be used to run a simulation with the corresponding MLP. GP stands for 'general purpose', i.e., it refers to an MLP which is not tailored exclusively to simulate a disordered phase but can be reliably used for that purpose. Whenever more than one version of the MLP exists, we give the reference to the latest one (i.e., on the table v2 refers to version 2 of a given MLP).

Material	Year	MLP flavor	References	Code(s)
a-C	2017	GAP	[36]	QUIP, LAMMPS
C (GP)	2020	SNAP	[148]	LAMMPS
a-C	2021 (v2)	GAP	[37, 149]	QUIP, LAMMPS, TurboGAP
C (GP)	2021	GAP	[43, 150]	QUIP,[a] LAMMPS,[a] TurboGAP
C (GP)	2022 (v2)	GAP	[151, 152]	QUIP, LAMMPS
a-C	2022	NEP	[140, 153]	GPUMD
C (GP)	2022	ACE	[101]	LAMMPS
Si (GP)	2018	GAP	[47, 154]	QUIP, LAMMPS
Si (GP)	2021	GAP	[46]	QUIP, LAMMPS, TurboGAP
Si (GP)	2021	NEP	[136, 153]	GPUMD
a-Si:H	2022	GAP	[146]	QUIP, LAMMPS

[a] Full van der Waals corrections for this MLP are only available with TurboGAP.

transferable MLPs, these multispecies force fields are starting to emerge. Here we mention in particular the recent effort by Unruh *et al* to develop an a-Si:H GAP [146]. The new Si-H GAP shows quantitative agreement with DFT and a significant improvement upon previously available classical (non-ML) force fields. This new MLP enabled the authors to model nanopores in a-Si:H. In the near future, either this force field or an extension combining its training database with existing and new ones may enable device-size simulation of c-Si/a-Si:H heterojunctions towards mitigating degradation mechanisms [147], of major technological importance for high-efficiency solar cells.

6. Available MLPs

Table 1 provides a non-comprehensive list of available MLPs able to simulate a-C and a-Si, together with their ML 'flavor' and which code(s) they can be used with. These potentials are mostly of the GAP flavor, since the GAP community has been the most active at simulating a-C and a-Si among the different MLP developers and users base.

7. Summary and outlook

In this Topical Review we have introduced MLPs as powerful tools for the simulation of disordered materials at the atomic scale, making it possible to accurately study atomistic systems within sizes and time scales that were out of reach just a few years ago. We have discussed how these new tools have been used to shed light on important questions for understanding the structure of a-C and a-Si. MLPs have been used to elucidate the growth mechanism in DLC and the high-pressure phase transformation from a-Si to a high-coordination metallic Si phase. MLPs have been used to study phase transitions at extreme thermodynamic conditions in carbon materials, and to understand the structure of NP carbon, a material of increasing importance in battery research. These tools are being extended to more complicated systems, in particular H- and O-doped a-C and H-doped a-Si, enabling a further degree of realism in simulating these structurally complex materials. Furthermore, MLPs are being coupled to other ML approaches, including electronic-structure and spectroscopic signature prediction, improving the prospects for direct comparison and better integration between experiment and simulation. All in all, the future of atomistic modeling of disordered materials, also beyond a-C and a-Si, looks bright in the wake of MLPs. We should expect important breakthroughs in materials research in the years to come, brought about by these new powerful computational tools.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

The author is grateful to Prof. Volker L Deringer from the University of Oxford for useful comments on this manuscript, and to the Academy of Finland for personal financial support, under Research Fellow Grant #330488.

ORCID iD

Miguel A Caro https://orcid.org/0000-0001-9304-4261

References

[1] Krames M R, Shchekin O B, Mueller-Mach R, Mueller G O, Zhou L, Harbers G and Craford M G 2007 Status and future of high-power light-emitting diodes for solid-state lighting *J. Disp. Technol.* **3** 160
[2] Humphreys C J 2008 Solid-state lighting *MRS Bull.* **33** 459
[3] Stuckelberger M, Biron R, Wyrsch N, Haug F-J and Ballif C 2017 Progress in solar cells from hydrogenated amorphous silicon *Renew. Sust. Energ. Rev.* **76** 1497
[4] Rech B and Wagner H 1999 Potential of amorphous silicon for solar cells *Appl. Phys.* A **69** 155
[5] Schröder B 1991 Thin film technology based on hydrogenated amorphous silicon *Mat. Sci. Eng.* A **139** 319

[6] Chen C-W, Chang T-C, Liu P-T, Lu H-Y, Wang K-C, Huang C-S, Ling C-C and Tseng T-Y 2005 High-performance hydrogenated amorphous-Si TFT for AMLCD and AMOLED applications *IEEE Electron Device Lett.* **26** 731

[7] Karim K S, Yin S, Nathan A and Rowlands J A 2004 High-dynamic-range pixel architectures for diagnostic medical imaging *Proc. SPIE* **5368** 657

[8] Robertson J 2002 Diamond-like amorphous carbon *Mat. Sci. Eng.* R **37** 129

[9] Tiainen V-M 2001 Amorphous carbon as a bio-mechanical coating–mechanical properties and biological applications *Diam. Relat. Mater.* **10** 153

[10] Laurila T, Sainio S and Caro M A 2017 Hybrid carbon based nanomaterials for electrochemical detection of biomolecules *Prog. Mater. Sci.* **88** 499

[11] Donnet C and Erdemir A (eds) 2007 *Tribology of Diamond-Like Carbon Films: Fundamentals and Applications* (Berlin: Springer)

[12] Santini C A, Sebastian A, Marchiori C, Jonnalagadda V P, Dellmann L, Koelmans W W, Rossell M D, Rossel C P and Eleftheriou E 2015 Oxygenated amorphous carbon for resistive memory applications *Nat. Commun.* **6** 1

[13] Etula J, Wester N, Liljeström T, Sainio S, Palomäki T, Arstila K, Sajavaara T, Koskinen J, Caro M A and Laurila T 2021 What determines the electrochemical properties of nitrogenated amorphous carbon thin films? *Chem. Mater.* **33** 6813

[14] Shiell T B *et al* 2018 Graphitization of glassy carbon after compression at room temperature *Phys. Rev. Lett.* **120** 215701

[15] Shang Y *et al* 2021 Ultrahard bulk amorphous carbon from collapsed fullerene *Nature* **599** 599

[16] Sundqvist B 2021 Carbon under pressure *Phys. Rep.* **909** 1

[17] Wang G, Yu M and Feng X 2021 Carbon materials for ion-intercalation involved rechargeable battery technologies *Chem. Soc. Rev.* **50** 2388

[18] Arasto A, Asikainen A and Kaukovirta A 2021 Finnish bioeconomy on the global product market in 2035 [white paper] (VTT Technical Research Centre of Finland)

[19] Deringer V L, Caro M A and Csányi G 2019 Machine learning interatomic potentials as emerging tools for materials science *Adv. Mater.* **31** 1902765

[20] Pantelides S T 1986 Defects in amorphous silicon: a new perspective *Phys. Rev. Lett.* **57** 2979

[21] Carlson D E 1977 The effects of impurities and temperature on amorphous silicon solar cells *1977 Int. Electron Devices Meeting* p 214

[22] Hohenberg P and Kohn W 1964 Inhomogeneous electron gas *Phys. Rev.* **136** B864

[23] Kohn W and Sham L J 1965 Self-consistent equations including exchange and correlation effects *Phys. Rev.* **140** A1133

[24] Martin R M 2004 *Electronic Structure* (Cambridge: Cambridge University Press)

[25] Oganov A R and Glass C W 2006 Crystal structure prediction using *ab initio* evolutionary techniques: principles and applications *J. Chem. Phys.* **124** 244704

[26] Pickard C J and Needs R J 2011 *Ab initio* random structure searching *J. Phys.: Condens. Matter* **23** 053201

[27] Marks N A, McKenzie D R, Pailthorpe B A, Bernasconi M and Parrinello M 1996 *Ab initio* simulations of tetrahedral amorphous carbon *Phys. Rev.* B **54** 9703

[28] Marks N A, Cooper N C, McKenzie D R, McCulloch D G, Bath P and Russo S P 2002 Comparison of density-functional, tight-binding and empirical methods for the simulation of amorphous carbon *Phys. Rev.* B **65** 075411

[29] Caro M A, Zoubkoff R, Lopez-Acevedo O and Laurila T 2014 Atomic and electronic structure of tetrahedral amorphous carbon surfaces from density functional theory: properties and simulation strategies *Carbon* **77** 1168

[30] Kaukonen H-P and Nieminen R M 1992 Molecular-dynamics simulation of the growth of diamondlike films by energetic carbon-atom beams *Phys. Rev. Lett.* **68** 620

[31] Marks N A 2005 Thin film deposition of tetrahedral amorphous carbon: a molecular dynamics study *Diam. Relat. Mater.* **14** 1223

[32] Caro M A, Deringer V L, Koskinen J, Laurila T and Csányi G 2018 Growth mechanism and origin of high sp^3 content in tetrahedral amorphous carbon *Phys. Rev. Lett.* **120** 166101

[33] Caro M A, Csányi G, Laurila T and Deringer V L 2020 Machine learning driven simulated deposition of carbon films: from low-density to diamondlike amorphous carbon *Phys. Rev.* B **102** 174201

[34] Galli G, Martin R M, Car R and Parrinello M 1989 Structural and electronic properties of amorphous carbon *Phys. Rev. Lett.* **62** 555

[35] de Tomas C, Suarez-Martinez I and Marks N A 2016 Graphitization of amorphous carbons: a comparative study of interatomic potentials *Carbon* **109** 681

[36] Deringer V L and Csányi G 2017 Machine learning based interatomic potential for amorphous carbon *Phys. Rev.* B **95** 094203

[37] Wang Y, Fan Z, Qian P, Ala-Nissila T and Caro M A 2022 Structure and pore size distribution in nanoporous carbon *Chem. Mater.* **34** 617

[38] Christ C D, Mark A E and van Gunsteren W F 2010 Basic ingredients of free energy calculations: a review *J. Comput. Chem.* **31** 1569

[39] Shelby J E 2020 *Introduction to Glass Science and Technology* (Cambridge: Royal Society of Chemistry)

[40] Deringer V L, Bernstein N, Bartók A P, Cliffe M J, Kerber R N, Marbella L E, Grey C P, Elliott S R and Csányi G 2018 Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics *J. Phys. Chem. Lett.* **9** 2879

[41] Berendsen H J C, Postma J P M, van Gunsteren W F, DiNola A R H J and Haak J R 1984 Molecular dynamics with coupling to an external bath *J. Chem. Phys.* **81** 3684

[42] Martyna G J, Tobias D J and Klein M L 1994 Constant pressure molecular dynamics algorithms *J. Chem. Phys.* **101** 4177

[43] Muhli H, Chen X, Bartók A P, Hernández-León P, Csányi G, Ala-Nissila T and Caro M A 2021 Machine learning force fields based on local parametrization of dispersion interactions: application to the phase diagram of C_{60} *Phys. Rev.* B **104** 054106

[44] de Tomas C, Aghajamali A, Jones J L, Lim D J, López M J, Suarez-Martinez I and Marks N A 2019 Transferability in interatomic potentials for carbon *Carbon* **155** 624

[45] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403

[46] Caro M A 2021 GAP interatomic potential for silicon *Zenodo* (https://doi.org/10.5281/zenodo.5734463)

[47] Bartók A P, Kermode J, Bernstein N and Csányi G 2018 Machine learning a general-purpose interatomic potential for silicon *Phys. Rev.* X **8** 041048

[48] Caro M A *et al* TurboGAP: Data-driven atomistic simulations (available at: http://turbogap.fi) (Accessed 17 February 2023)

[49] Caro M A 2019 Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials *Phys. Rev.* B **100** 024112

[50] Biswas R, Grest G S and Soukoulis C M 1988 Molecular-dynamics simulation of cluster and atom deposition on silicon (111) *Phys. Rev.* B **38** 8154

[51] Luedtke W D and Landman U 1989 Molecular-dynamics studies of the growth modes and structure of amorphous silicon films via atom deposition *Phys. Rev.* B **40** 11733

[52] Ramalingam S, Sriraman S, Aydil E S and Maroudas D 2001 Evolution of structure, morphology and reactivity of hydrogenated amorphous silicon film surfaces grown by molecular-dynamics simulation *Appl. Phys. Lett.* **78** 2685

[53] Tersoff J 1988 Empirical interatomic potential for carbon, with applications to amorphous carbon *Phys. Rev. Lett.* **61** 2879

[54] Stillinger F H and Weber T A 1985 Computer simulation of local order in condensed phases of silicon *Phys. Rev.* B **31** 5262

[55] Vink R L C, Barkema G T, Van der Weg W F and Mousseau N 2001 Fitting the Stillinger-Weber potential to amorphous silicon *J. Non-Cryst. Solids* **282** 248

[56] Bazant M Z, Kaxiras E and Justo J F 1997 Environment-dependent interatomic potential for bulk silicon *Phys. Rev.* B **56** 8542

[57] Marks N A 2000 Generalizing the environment-dependent interaction potential for carbon *Phys. Rev.* B **63** 035401

[58] Brenner D W 1990 Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films *Phys. Rev.* B **42** 9458

[59] Brenner D W, Shenderova O A, Harrison J A, Stuart S J, Ni B and Sinnott S B 2002 A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons *J. Phys.: Condens. Matter* **14** 783

[60] Senftle T P *et al* 2016 The ReaxFF reactive force-field: development, applications and future directions *npj Comput. Mater.* **2** 15011

[61] Van Duin A C T, Strachan A, Stewman S, Zhang Q, Xu X and Goddard W A 2003 $\text{ReaxFF}_{\text{SiO}}$ reactive force field for silicon and silicon oxide systems *J. Phys. Chem.* A **107** 3803

[62] Srinivasan S G, van Duin A C T and Ganesh P 2015 Development of a ReaxFF potential for carbon condensed phases and its application to the thermal fragmentation of a large fullerene *J. Phys. Chem.* A **119** 571

[63] Smith R, Jolley K, Latham C, Heggie M, Van Duin A, van Duin D and Wu H 2017 A ReaxFF carbon potential for radiation damage studies *Nucl. Instrum. Methods Phys. Res.* B **393** 49

[64] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401

[65] Schütt K T, Chmiela S, von Lilienfeld O A, Tkatchenko A, Tsuda K and Müller K-R 2020 *Machine Learning Meets Quantum Physics* 1st edn (Berlin: Springer)

[66] Behler J 2017 First principles neural network potentials for reactive simulations of large molecular and condensed systems *Angew. Chem., Int. Ed.* **56** 12828

[67] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 Gaussian process regression for materials and molecules *Chem. Rev.* **121** 10073

[68] Csányi G, Willatt M J and Ceriotti M 2020 Machine-learning of atomic-scale properties based on physical principles *Machine Learning Meets Quantum Physics* (Cham: Springer) p 99

[69] Hellström M and Behler J 2020 High-dimensional neural network potentials for atomistic simulations *Machine Learning Meets Quantum Physics* (Cham: Springer) p 253

[70] Bartók A P and Csányi G 2015 Gaussian approximation potentials: a brief tutorial introduction *Int. J. Quantum Chem.* **115** 1051

[71] Von Lilienfeld O A 2018 Quantum machine learning in chemical compound space *Angew. Chem., Int. Ed.* **57** 4164

[72] Shapeev A V 2016 Moment tensor potentials: a class of systematically improvable interatomic potentials *Multiscale Model. Simul.* **14** 1153

[73] Rowe P, Csányi G, Alfè D and Michaelides A 2018 Development of a machine learning potential for graphene *Phys. Rev.* B **97** 054303

[74] Bernstein N, Csányi G and Deringer V L 2019 De novo exploration and self-guided learning of potential-energy surfaces *npj Comput. Mater.* **5** 1

[75] Pickard C J 2022 Ephemeral data derived potentials for random structure search *Phys. Rev. B* **106** 014102

[76] George J, Hautier G, Bartók A P, Csányi G and Deringer V L 2020 Combining phonon accuracy with high transferability in Gaussian approximation potential models *J. Chem. Phys.* **153** 044104

[77] De S, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754

[78] Cheng B *et al* 2020 Mapping materials and molecules *Acc. Chem. Res.* **53** 1981

[79] Perdew J P and Zunger A 1981 Self-interaction correction to density-functional approximations for many-electron systems *Phys. Rev.* B **23** 5048

[80] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865

[81] Kresse G and Furthmüller J 1996 Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set *Phys. Rev.* B **54** 11169

[82] Seidl A, Görling A, Vogl P, Majewski J A and Levy M 1996 Generalized Kohn-Sham schemes and the band-gap problem *Phys. Rev.* B **53** 3764

[83] Heyd J, Scuseria G E and Ernzerhof M 2003 Hybrid functionals based on a screened Coulomb potential *J. Chem. Phys.* **118** 8207

[84] Grimme S 2006 Semiempirical GGA-type density functional constructed with a long-range dispersion correction *J. Comput. Chem.* **27** 1787

[85] Tkatchenko A and Scheffler M 2009 Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data *Phys. Rev. Lett.* **102** 073005

[86] Tkatchenko A, DiStasio R A Jr, Car R and Scheffler M 2012 Accurate and efficient method for many-body van der Waals interactions *Phys. Rev. Lett.* **108** 236402

[87] Habershon S, Manolopoulos D E, Markland T E and Miller T F III 2013 Ring-polymer molecular dynamics: quantum effects in chemical dynamics from classical trajectories in an extended phase space *Annu. Rev. Phys. Chem.* **64** 387

[88] Darkins R and Duffy D M 2018 Modelling radiation effects in solids with two-temperature molecular dynamics *Comput. Mater. Sci.* **147** 145

[89] Casida M E and Wesołowski T A 2004 Generalization of the Kohn-Sham equations with constrained electron density formalism and its time-dependent response theory formulation *Int. J. Quantum Chem.* **96** 577

[90] Li X, Tully J C, Schlegel H B and Frisch M J 2005 *Ab initio* Ehrenfest dynamics *J. Chem. Phys.* **123** 084106

[91] Vlček A Jr and Záliš S 2007 Modeling of charge-transfer transitions and excited states in d^6 transition metal complexes by DFT techniques *Coord. Chem. Rev.* **251** 258

[92] Allen A E A, Dusson G, Ortner C and Csányi G 2021 Atomic permutationally invariant polynomials for fitting molecular force fields *Mach. Learn.: Sci. Technol.* **2** 025017

[93] Christensen A S, Bratholm L A, Faber F A and von Lilienfeld A O 2020 FCHL revisited: faster and more accurate quantum machine learning *J. Chem. Phys.* **152** 044107

[94] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115

[95] Behler J 2011 Atom-centered symmetry functions for constructing high-dimensional neural network potentials *J. Chem. Phys.* **134** 074106

[96] Willatt M J, Musil F and Ceriotti M 2019 Atom-density representations for machine learning *J. Chem. Phys.* **150** 154110

[97] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 Physics-inspired structural representations for molecules and materials *Chem. Rev.* **121** 9759

[98] Himanen L, Jäger M O J, Morooka E V, Canova F F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 DScribe: library of descriptors for machine learning in materials science *Comput. Phys. Commun.* **247** 106949

[99] Thompson A P, Swiler L P, Trott C R, Foiles S M and Tucker G J 2015 Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials *J. Comput. Phys.* **285** 316

[100] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials *Phys. Rev.* B **99** 014104

[101] Qamar M, Mrovec M, Lysogorskiy Y, Bochkarev A and Drautz R 2022 Atomic cluster expansion for quantum-accurate large-scale simulations of carbon (arXiv:2210.09161)

[102] Batatia I, Batzner S, Kovács D P, Musaelian A, Simm G N C, Drautz R, Ortner C, Kozinsky B and Csányi G 2022 The design space of E(3)-equivariant atom-centered interatomic potentials (arXiv:2205.06643)

[103] Darby J P, Kermode J R and Csányi G 2022 Compressing local atomic neighbourhood descriptors *npj Comput. Mater.* **8** 1

[104] Bereau T, DiStasio R A Jr, Tkatchenko A and Von Lilienfeld O A 2018 Non-covalent interactions across organic and biological subsets of chemical space: physics-based potentials parametrized from machine learning *J. Chem. Phys.* **148** 241706

[105] Veit M, Wilkins D M, Yang Y, DiStasio R A Jr and Ceriotti M 2020 Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles *J. Chem. Phys.* **153** 024113

[106] Xie X, Persson K A and Small D W 2020 Incorporating electronic information into machine learning potential energy surfaces via approaching the ground-state electronic energy as a function of atom-based electronic populations *J. Chem. Theory Comput.* **16** 4256

[107] Staacke C G, Wengert S, Kunkel C, Csányi G, Reuter K and Margraf J T 2022 Kernel charge equilibration: efficient and accurate prediction of molecular dipole moments with a machine-learning enhanced electron density model *Mach. Learn.: Sci. Technol.* **3** 015032

[108] Polo M C, Andújar J L, Hart A, Robertson J and Milne W I 2000 Preparation of tetrahedral amorphous carbon films by filtered cathodic vacuum arc deposition *Diam. Relat. Mater.* **9** 663

[109] Xu S, Flynn D, Tay B K, Prawer S, Nugent K W, Silva S R P, Lifshitz Y and Milne W I 1997 Mechanical properties and Raman spectra of tetrahedral amorphous carbon films with high sp^3 fraction deposited using a filtered cathodic arc *Phil. Mag.* B **76** 351

[110] Fallon P J, Veerasamy V S, Davis C A, Robertson J, Amaratunga G A J, Milne W I and Koskinen J 1993 Properties of filtered-ion-beam-deposited diamondlike carbon as a function of ion energy *Phys. Rev.* B **48** 4777

[111] McKenzie D R, Muller D and Pailthorpe B A 1991 Compressive-stress-induced formation of thin-film tetrahedral amorphous carbon *Phys. Rev. Lett.* **67** 773

[112] Schwan J, Ulrich S, Roth H, Ehrhardt H, Silva S R P, Robertson J, Samlenski R and Brenn R 1996 Tetrahedral amorphous carbon films prepared by magnetron sputtering and dc ion plating *J. Appl. Phys.* **79** 1416

[113] Kohary K and Kugler S 2001 Growth of amorphous carbon: low-energy molecular dynamics simulation of atomic bombardment *Phys. Rev.* B **63** 193404

[114] Lee S-H, Lee C-S, Lee S-C, Lee K-H and Lee K-R 2004 Structural properties of amorphous carbon films by molecular dynamics simulation *Surf. Coat. Technol.* **177** 812

[115] Wang N and Komvopoulos K 2014 The effect of deposition energy of energetic atoms on the growth and structure of ultrathin amorphous carbon films studied by molecular dynamics simulations *J. Phys. D: Appl. Phys.* **47** 245303

[116] McCulloch D G, McKenzie D R and Goringe C M 2000 *Ab initio* simulations of the structure of amorphous carbon *Phys. Rev.* B **61** 2349

[117] Caro M A 2020 Thermal spike during simulated deposition of tetrahedral amorphous carbon films *Zenodo* (https://doi.org/10.5281/zenodo.4030350)

[118] Caro M A 2017 Deposition of amorphous carbon at different energies modeled with GAP *Zenodo* (https://doi.org/10.5281/zenodo.1133425)

[119] Caro M A 2020 Amorphous carbon films generated through simulated deposition with GAP from 1 eV to 100 eV *Zenodo* (https://doi.org/10.5281/zenodo.3778153)

[120] Gilkes K W R, Gaskell P H and Robertson J 1995 Comparison of neutron-scattering data for tetrahedral amorphous carbon with structural models *Phys. Rev.* B **51** 12303

[121] Davis C A, Amaratunga G A J and Knowles K M 1998 Growth mechanism and cross-sectional structure of tetrahedral amorphous carbon thin films *Phys. Rev. Lett.* **80** 3280

[122] Mizohata K 2012 Progress in elastic recoil detection analysis *PhD Thesis* Helsinki University

[123] Deringer V L, Caro M A, Jana R, Aarva A, Elliott S R, Laurila T, Csányi G and Pastewka L 2018 Computational surface chemistry of tetrahedral amorphous carbon by combining machine learning and DFT *Chem. Mater.* **30** 7438

[124] Caro M A, Aarva A, Deringer V L, Csányi G and Laurila T 2018 Reactivity of amorphous carbon surfaces: rationalizing the role of structural motifs in functionalization using machine learning *Chem. Mater.* **30** 7446

[125] Golze D, Hirvensalo M, Hernández-León P, Aarva A, Etula J, Susi T, Rinke P, Laurila T and Caro M A 2022 Accurate computational prediction of core-electron binding energies in carbon-based materials: a machine-learning model combining DFT and *GW Chem. Mater.* **34** 6240

[126] Lahrar E H, Belhboub A, Simon P and Merlet C 2019 Ionic liquids under confinement: from systematic variations of the ion and pore sizes toward an understanding of the structure and dynamics in complex porous carbons *ACS Appl. Mater. Interfaces* **12** 1789

[127] Lahrar E H, Simon P and Merlet C 2021 Carbon–carbon supercapacitors: beyond the average pore size or how electrolyte confinement and inaccessible pores affect the capacitance *J. Chem. Phys.* **155** 184703

[128] Fujikake S, Deringer V L, Lee T H, Krynski M, Elliott S R and Csányi G 2018 Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures *J. Chem. Phys.* **148** 241714

[129] Huang J-X, Csányi G, Zhao J-B, Cheng J and Deringer V L 2019 First-principles study of alkali-metal intercalation in disordered carbon anode materials *J. Mater. Chem.* A **7** 19070

[130] Deringer V L, Csányi G and Proserpio D M 2017 Extracting crystal chemistry from amorphous carbon structures *Chem. Phys. Chem.* **18** 873

[131] Wang Y, Fan Z, Qian P, Caro M A and Ala-Nissila T 2023 Quantum-corrected thickness-dependent thermal conductivity in amorphous silicon predicted by machine-learning molecular dynamics simulations *Phys. Rev. B* **107** 054303

[132] Vink R L C, Barkema G T, Stijnman M A and Bisseling R H 2001 Device-size atomistic models of amorphous silicon *Phys. Rev.* B **64** 245214

[133] Hejna M, Steinhardt P J and Torquato S 2013 Nearly hyperuniform network models of amorphous silicon *Phys. Rev.* B **87** 245204

[134] Laaziri K, Kycia S, Roorda S, Chicoine M, Robertson J L, Wang J and Moss S C 1999 High resolution radial distribution function of pure amorphous silicon *Phys. Rev. Lett.* **82** 3460

[135] Laaziri K, Kycia S, Roorda S, Chicoine M, Robertson J L, Wang J and Moss S C 1999 High-energy x-ray diffraction study of pure amorphous silicon *Phys. Rev.* B **60** 13520

[136] Fan Z, Zeng Z, Zhang C, Wang Y, Song K, Dong H, Chen Y and Ala-Nissila T 2021 Neuroevolution machine learning potentials: combining high accuracy and low cost in atomistic simulations and application to heat transport *Phys. Rev.* B **104** 104309

[137] Franzblau D S 1991 Computation of ring statistics for network models of solids *Phys. Rev.* B **44** 4925

[138] Deringer V L, Bernstein N, Csányi G, Ben Mahmoud C, Ceriotti M, Wilson M, Drabold D A and Elliott S R 2021 Origins of structural and electronic transitions in disordered silicon *Nature* **589** 59

[139] Ben Mahmoud C, Anelli A, Csányi G and Ceriotti M 2020 Learning the electronic density of states in condensed matter *Phys. Rev.* B **102** 235130

[140] Fan Z *et al* 2022 GPUMD: a package for constructing accurate machine-learned potentials and performing highly efficient atomistic simulations *J. Chem. Phys.* **157** 114801

[141] Zink B L, Pietri R and Hellman F 2006 Thermal conductivity and specific heat of thin-film amorphous silicon *Phys. Rev. Lett.* **96** 055902

[142] Kim T, Moon J and Minnich A J 2021 Origin of micrometer-scale propagation lengths of heat-carrying acoustic excitations in amorphous silicon *Phys. Rev. Mater.* **5** 065602

[143] Cahill D G, Katiyar M and Abelson J R 1994 Thermal conductivity of a-Si:H thin films *Phys. Rev.* B **50** 6077

[144] Zhang Z, Guo Y, Bescond M, Chen J, Nomura M and Volz S 2022 How coherence is governing diffusion heat transfer in amorphous solids *npj Comput. Mater.* **8** 1

[145] Isaeva L, Barbalinardo G, Donadio D and Baroni S 2019 Modeling heat transport in crystals and glasses from a unified lattice-dynamical approach *Nat. Commun.* **10** 1

[146] Unruh D, Meidanshahi R V, Goodnick S M, Csányi G and Zimányi G T 2022 Gaussian approximation potential for amorphous Si:H *Phys. Rev. Mater.* **6** 065603

[147] Jordan D C *et al* 2017 Silicon heterojunction system field performance *IEEE J. Photovolt.* **8** 177

[148] Willman J T, Williams A S, Nguyen-Cong K, Thompson A P, Wood M A, Belonoshko A B and Oleynik I I 2020 Quantum accurate SNAP carbon potential for MD shock simulations *AIP Conf. Proc.* **2272** 070055

[149] Caro M A 2020 GAP interatomic potential for amorphous carbon *Zenodo* (https://doi.org/10.5281/zenodo.4000211)

[150] Muhli H and Caro M A 2021 GAP interatomic potential for C_{60} *Zenodo* (https://doi.org/10.5281/zenodo.4616343)

[151] Rowe P, Deringer V L, Gasparotto P, Csányi G and Michaelides A 2020 An accurate and transferable machine learning potential for carbon *J. Chem. Phys.* **153** 034702

[152] Csányi G 2022 *Research Data Supporting "An Accurate and Transferable Machine Learning Potential for Carbon"* (University of Cambridge) (https://doi.org/10.17863/CAM.82086)

[153] Fan Z NEP-data (available at: https://gitlab.com/brucefan1983/nep-data) (Accessed 12 September 2022)

[154] Csányi G 2021 *Research Data: Machine Learning a General-Purpose Interatomic Potential for Silicon* (University of Cambridge) (https://doi.org/10.17863/CAM.65004)