

An accurate and transferable machine learning potential for carbon

Cite as: J. Chem. Phys. **153**, 034702 (2020); <https://doi.org/10.1063/5.0005084>

Submitted: 17 February 2020 . Accepted: 22 June 2020 . Published Online: 15 July 2020

 Patrick Rowe,  Volker L. Deringer,  Piero Gasparotto,  Gábor Csányi, and  Angelos Michaelides

COLLECTIONS

Paper published as part of the special topic on [Machine Learning Meets Chemical PhysicsMACH2020](#)



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

[Machine learning for interatomic potential models](#)

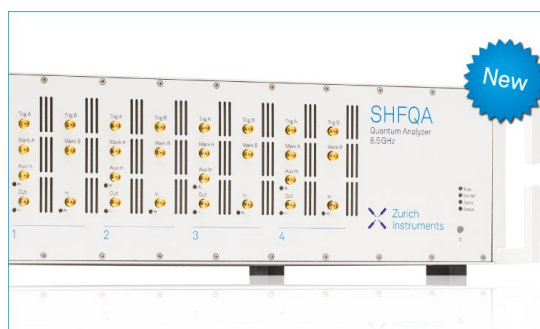
The Journal of Chemical Physics **152**, 050902 (2020); <https://doi.org/10.1063/1.5126336>

[Perspective: Machine learning potentials for atomistic simulations](#)

The Journal of Chemical Physics **145**, 170901 (2016); <https://doi.org/10.1063/1.4966192>

[Hierarchical machine learning of potential energy surfaces](#)

The Journal of Chemical Physics **152**, 204110 (2020); <https://doi.org/10.1063/5.0006498>



Your Qubits. Measured.

Meet the next generation of quantum analyzers

- Readout for up to 64 qubits
- Operation at up to 8.5 GHz, mixer-calibration-free
- Signal optimization with minimal latency

[Find out more](#)



An accurate and transferable machine learning potential for carbon

Cite as: J. Chem. Phys. 153, 034702 (2020); doi: 10.1063/5.0005084

Submitted: 17 February 2020 • Accepted: 22 June 2020 •

Published Online: 15 July 2020



Patrick Rowe,¹ Volker L. Deringer,² Piero Gasparotto,¹ Gábor Csányi,³ and Angelos Michaelides^{1,a)}

AFFILIATIONS

¹Thomas Young Centre, London Centre for Nanotechnology, and Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, United Kingdom

²Department of Chemistry, Inorganic Chemistry Laboratory, University of Oxford, Oxford OX1 3QR, United Kingdom

³Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.

a) Author to whom correspondence should be addressed: angelos.michaelides@ucl.ac.uk

ABSTRACT

We present an accurate machine learning (ML) model for atomistic simulations of carbon, constructed using the Gaussian approximation potential (GAP) methodology. The potential, named GAP-20, describes the properties of the bulk crystalline and amorphous phases, crystal surfaces, and defect structures with an accuracy approaching that of direct *ab initio* simulation, but at a significantly reduced cost. We combine structural databases for amorphous carbon and graphene, which we extend substantially by adding suitable configurations, for example, for defects in graphene and other nanostructures. The final potential is fitted to reference data computed using the optB88-vdW density functional theory (DFT) functional. Dispersion interactions, which are crucial to describe multilayer carbonaceous materials, are therefore implicitly included. We additionally account for long-range dispersion interactions using a semianalytical two-body term and show that an improved model can be obtained through an optimization of the many-body smooth overlap of atomic positions descriptor. We rigorously test the potential on lattice parameters, bond lengths, formation energies, and phonon dispersions of numerous carbon allotropes. We compare the formation energies of an extensive set of defect structures, surfaces, and surface reconstructions to DFT reference calculations. The present work demonstrates the ability to combine, in the same ML model, the previously attained flexibility required for amorphous carbon [V. L. Deringer and G. Csányi, Phys. Rev. B **95**, 094203 (2017)] with the high numerical accuracy necessary for crystalline graphene [Rowe *et al.*, Phys. Rev. B **97**, 054303 (2018)], thereby providing an interatomic potential that will be applicable to a wide range of applications concerning diverse forms of bulk and nanostructured carbon.

Published under license by AIP Publishing. <https://doi.org/10.1063/5.0005084>

I. INTRODUCTION

The same characteristics that make carbon a fascinating element for study also make it challenging to model computationally. It exhibits some of the greatest structural diversity—and associated diversity of properties—of any of the elements.^{1–7} Its allotropes range from zero to three-dimensional, have metallic, semiconducting and insulating phases, and boast mechanical properties including some of the highest tensile strengths, hardnesses, and bulk moduli measured.^{8,9} It is unsurprising, therefore, that carbon is considered to be not just an element of prime technological importance but also remains the subject of continued fundamental scientific study.^{10–16}

Current applications of elemental carbon are numerous, and they include lightweight and strong structural materials, anodes of batteries, and components in advanced optical technologies.^{9,13,16–19} The usefulness of elemental carbon has also clearly not yet been exhausted; future applications propose to make use of graphene's unique electronic properties for advanced electronics,^{11,20} carbon nanotubes' structural and optical characteristics for performance materials,^{18,21} and the thermal and optical properties of diamonds for laser optics,^{9,22,23} among others.

Atomistic simulations have played a major role in developing our understanding of carbon materials. Among the many scientific problems that have been addressed with carbon potentials, we may mention the wear process of diamond²⁴ or the compression

behavior of glassy carbon.²⁵ The first many-body (MB) interatomic potential for modeling carbon was published in 1988 by Tersoff. This potential was used to investigate the properties of carbon's crystalline and amorphous allotropes.²⁶ The reactive empirical bond order (REBO, REBO-II) potentials were built on the original Tersoff formulation including a wider range of parameters and data in the fit, as well as adding additional conjugation and torsional terms, and modifying the bond order expression for small angles.^{27,28} However, the Tersoff and REBO-II potentials only considered nearest neighbor interactions and did not account for the effects of dispersion. The adaptive intermolecular REBO (AIREBO) potential²⁹ sought to correct this by adding an additional long-range Lennard-Jones term between atoms with larger separations while making no modifications to the short-range part of the potential. The long-range carbon bond order potential (LCBOP) not only increased the range of the potential to account for longer-range interactions but also constituted a complete reparameterization of the bond order potential to improve the accuracy and transferability of the model, though long-range dispersion interactions are still omitted.³⁰

Further developments, beyond the scope of detailed discussion here, include the LCBOP-II potential, which expanded the application range of the model to include the liquid phase,^{31,32} the environment dependent interatomic potential (EDIP) for carbon,³³ which employed properties calculated from *ab initio* simulation in its parameterization, the introduction of a dynamic cut-off to bond-order potentials,³⁴ and a recent reparameterization of a carbon ReaxFF potential.³⁵ Notwithstanding the long-standing success of these potential models, there are inherent limitations to even the most advanced of them. Such issues are particularly relevant when one departs from the idealized structures (diamond, graphite, etc.), as shown in two detailed benchmark studies by de Tomas *et al.*^{36,37}

Machine learning (ML) has recently arisen as a way of addressing some of these limitations. A number of practical approaches for modeling the potential energy surface (PES) using ML have been developed in recent years, employing algorithms including artificial neural networks, Gaussian process regression, and compressed sensing.^{38–45} The demonstrated ability of ML algorithms to fit arbitrary functions with an extremely high accuracy⁴⁶ combined with recent developments in high-dimensional descriptors for atomic systems makes ML approaches for the development of interatomic potentials an increasingly popular approach.^{42,47–50} Indeed, the use of machine learning methodologies to model carbon has a significant precedent. Some of the first examples of ML potentials, using both Gaussian process regression and artificial neural networks, were fitted for graphite and diamond, and these were tested with regard to the properties of the crystalline phases and the graphite-diamond phase coexistence line.^{38,39,51} One of the first neural-network potentials was used for large-scale simulations of the diamond nucleation mechanism.⁵² We recently introduced a machine learning potential for pristine graphene constructed using the Gaussian approximation potential (GAP) framework, which achieved excellent accuracy when benchmarked against density functional theory (DFT) and experiment for a wide range of lattice and dynamical properties, including the phonon dispersion relations, thermal expansion, and Raman spectra at different temperatures.⁵³

While achieving good accuracy in a specific region of configuration space is not trivial, the problem of the transferability of a potential is much more challenging to solve from a ML perspective. In 2017, some of us reported a highly transferable GAP model trained primarily on the amorphous and liquid phases of carbon (henceforth termed GAP-17) based on DFT-local-density approximation (LDA) reference data. The focus there was somewhat complementary—to be able to describe very diverse structural environments, albeit accepting a degree of numerical error. As an example, the in-plane force errors for a pristine graphene sheet are $0.03 \text{ eV } \text{\AA}^{-1}$ with the graphene-only GAP mentioned above, as compared to $0.27 \text{ eV } \text{\AA}^{-1}$ with GAP-17. For comparison, these errors for a range of commonly used empirically fitted potentials range from $0.6 \text{ eV } \text{\AA}^{-1}$ to $3.1 \text{ eV } \text{\AA}^{-1}$ (for more details, see Ref. 53). In return, owing to the flexibility and transferability ensuing from its choice of reference database, GAP-17 enabled the study of a number of scientific problems that involve diverse structural environments, including understanding the mechanism of growth of sp^3 hybridized amorphous carbon by ion deposition,⁵⁴ extensive studies of the surface properties (and chemical reactivity) of tetrahedral amorphous carbon,^{54–56} the structure of “porous” carbonaceous materials that are relevant to applications in batteries and supercapacitors,^{57,58} and crystal-structure prediction.⁵

The model we present here, GAP-20, builds on all of the previous work applying the GAP machine learning methodology to the development of carbon potentials to achieve the accuracy required for capturing subtle differences in formation energies of nanostructures or in defect formation energies and for describing phonon dispersions to within meV accuracy—while maintaining the flexibility and transferability of GAP-17. Importantly, all data are generated using a dispersion-corrected DFT method, which properly accounts for longer-range interactions in low-dimensional carbon structures, and the fitting architecture is adapted to account for those. Our tests suggest GAP-20 to be suitable as a “general-purpose” carbon ML potential for diverse areas of study.

The detailed discussions of the construction and testing of the potential are given in Secs. II–VIII, but we take a moment to highlight the main points here. The composition of the training dataset and performance of this potential are summarized in Fig. 1. GAP-20 correctly predicts the formation energies of diamond, graphite, fullerenes, and nanotubes, to an accuracy of a few meV, and achieves comparable accuracy for a number of crystalline and amorphous surfaces. The computed formation energies of defects are also accurate, with overall errors significantly lower than those obtained from comparable empirical models. At the same time, GAP-20 can accurately predict the behavior of high temperature liquid carbon over a wide range of temperatures and densities, which will be shown below. We believe that these features make GAP-20 a useful tool for the accurate modeling of nanostructured carbons: nanotubes, graphitized carbon, and materials with varying degrees of defects and disorder.

The rest of this paper is organized as follows: We first describe our process for the construction of a training set suitable for developing such a potential. We then give details on the construction and training of the model itself, with discussion of particular aspects that required special attention or optimization. Subsequently, we present an extensive and rigorous testing of our model for a wide

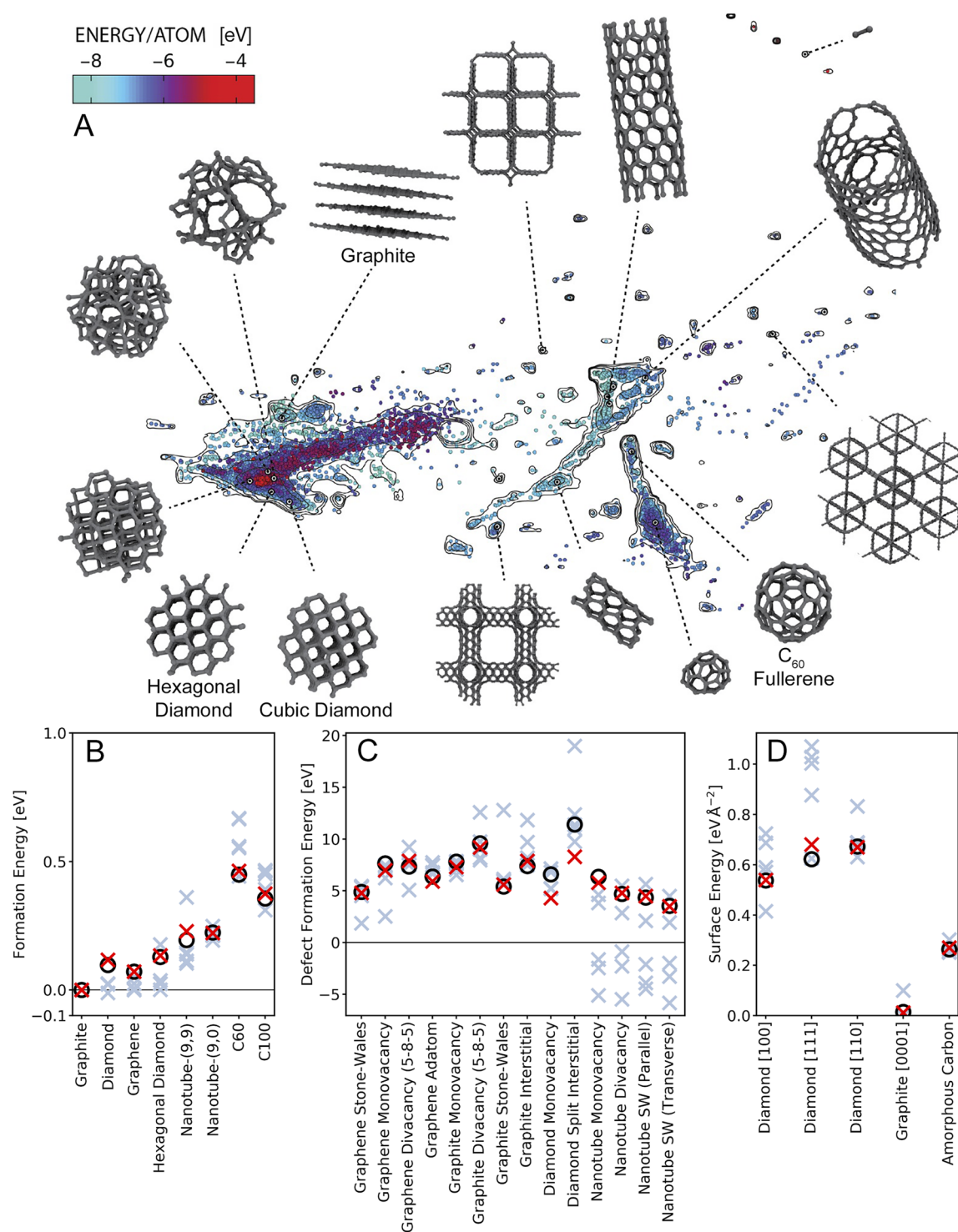


FIG. 1. Overview of some of the key structures included in the training shown through a sketch-map representation (top) as well as selected information on the performance of the potential for a variety of properties. (a) Sketch-map representation of the total dataset for carbon generated as part of this work. Select structures are identified for graphite, diamond, hexagonal diamond (lonsdaleite), amorphous carbon, and fullerenes. Points are colored according to their energy, while contours indicate the density of the database population in a particular region. (Bottom) A summary of the (b) predicted crystalline formation energies, (c) defect formation energies, and (d) surface energies, comparing the DFT (optB88-vdW) reference (black circles), GAP-20 (red crosses), and all other models (blue crosses).

range of properties. We also compare the results of our model to a selection of commonly used empirical potentials that model the interatomic interactions in carbon with differing degrees of simplification. Specifically, we choose the Tersoff, REBO-II, AIREBO, and LCBOP models. The selection of potentials considered here is by no means exhaustive and is only intended to give some basis for comparison between previous work and the model we introduce, as well as illustrating how the inclusion or exclusion of different interactions (e.g., dispersion interactions) may affect the performance of a model. A more detailed benchmarking across a wider range of potentials, complementing the existing detailed tests for amorphous and “graphitized” carbons,³⁶ may be the subject of future work.

II. GENERATION AND SELECTION OF TRAINING DATA

One of the challenges inherent in constructing a generalized potential for carbon is the enormous variety of structures that must be considered. In addition to its more commonly encountered crystalline phases, diamond and graphite, carbon may be found in forms of differing dimensionality from zero-dimensional fullerenes to one-dimensional nanotubes, two-dimensional graphene, and three-dimensional amorphous forms.⁶

Specifically, in the case of ML, one is drawn to the problem of the composition of the large database of example configurations, known as the training dataset. For a potential to be both accurate and transferable, its training dataset ought to include representative configurations from all of the thermally accessible chemical space. One might initially suggest that the problem is therefore intractable, if in order to produce a potential, which is capable of accurately modeling all of the relevant phases of carbon, we must explore the entirety of the vast 3N-dimensional chemical space. It is an empirical observation, however, that the thermally accessible and physically relevant regions of this chemical space constitute a vastly reduced subset of all of the available configurations.^{59–61} Furthermore, rather than an exploration of the 3N-dimensional space, in fitting the parameters of a ML algorithm, we are primarily concerned with an exploration of the reduced dimensionality descriptor space.^{59,62–64} In the case of atom centered descriptors such as the smooth overlap of atomic positions (SOAP), this represents the local environment around a particular atom rather than the global structure.⁶³ While the structural variability of carbon is globally almost infinite, many of these structures are constructed from similar local motifs, for example, the tetrahedral building blocks of diamond.^{65,66} Similar logic may be applied to more complex structures.

The reference configurations that comprise our structural database are drawn from a wide variety of sources. Regardless of the origin of the configuration itself (e.g., from the GAP-17 database), its properties, those being the total energy, atomic forces, and virial stresses, which comprise the actual training data, are always computed using the same level of tightly converged plane-wave DFT including dispersion corrections. We use the Vienna *ab initio* simulation program (VASP) plane-wave DFT code, and we perform spin-polarized calculations with the optB88-vdW dispersion inclusive exchange–correlation functional,^{67–70} a plane-wave cutoff of 600 eV, and a projector augmented wave pseudopotential.^{71–73} A Gaussian smearing of 0.1 eV is applied to the energy levels, and dense reciprocal space Monkhorst–Pack grids

are used.⁷⁴ In the case of the reduced dimensionality allotropes, graphene and nanotube structures, the reciprocal space sampling is only performed in the directions in which the allotrope is periodic. The properties of fullerene structures are calculated at the gamma point. For this potential, we choose the optB88-vdW functional as it has already been demonstrated to provide an excellent description of carbonaceous materials, in particular, graphitic carbon—for which its prediction of the binding energy and interlayer spacing is in good agreement with experimental values.⁷⁵

The database of configurations presented here uses as its foundation a combination of the training datasets for the two carbon potentials previously published primarily for liquid and amorphous carbon (GAP-17) and for pristine graphene, respectively.^{53,62} A large number of new configurations are considered in addition to these existing ML datasets.^{5,6,76} We endeavor to comprehensively cover all the possible crystalline phases of carbon found at moderate temperatures and pressures, including more exotic allotropes. To this end, DFT optimized structures for graphite, graphene, and cubic and hexagonal diamond are included, as well as the structures of a library of fullerenes comprised of fewer than 240 atoms and all nanotube structures with chiral indices $3 \leq n, m \leq 10$ with fewer than 240 atoms in their unit cell. The optimized structures are also included for the samara carbon allotrope database (SACADA) of exotic carbon allotropes⁷⁶ and the results of a GAP-17 driven random structure search (RSS).⁵ In addition to bulk or pristine phases, the structures of relevant low Miller-index faces of the crystalline phases are included, along with a large number of important defect structures.^{77–84}

For all of these structures, we have performed some *ab initio* and some iteratively improved GAP driven molecular dynamics simulations at a number of temperatures so as to also sample the region of phase space close to these local minima.^{53,62} The resulting database is comprised of ~17 000 configurations, each containing from 1 to 240 atoms/cell.

The choice of which structures might be important for training a potential requires for the most part chemical or physical intuition on the part of the researcher.^{42,59,85} Some of these choices may be clear, for example, the need to include configurations representing the bulk structures of diamond and graphite. Others, however, such as the inclusion or exclusion of particular defect or surface structures, will depend on the desired application of the potential (and, to some extent, on personal choice). To maximize the transferability of our model, we have produced as comprehensive a database as possible—too large to train on with current computational facilities. Rather than using the full database for sparsification, as commonly done in GAP fitting (including in the development of GAP-17), we instead allow the bulk of our training configurations to be chosen from the total dataset using a sampling method known as farthest point sampling (FPS).^{42,86} Within this set, we then carefully check the data saturation of our training with respect to the number of sparse points, which is discussed in Sec. III.

This method allows us to start with a much more comprehensive database than previously, while still keeping the computational effort at the fitting stage tractable. We wish for our training dataset to have the widest possible sampling of descriptors and forces—leaving no physically relevant configurations unsampled, while avoiding over-representation of particular regions of phase space. FPS facilitates this by allowing a selection of frames to be made based

on a measure of the global similarity (in descriptor space) between possible configurations.^{42,86} Given a set of n descriptors of type d for a number of frames, $\mathbb{Q} = \{\mathbf{q}_{i=1,\dots,n}^{d, \text{avg}}\}$, which are themselves the average of the individual descriptors of the atoms in a particular frame $\mathbf{q}_i^d$, the FPS algorithm selects configurations so that at each step, the kernel distance between previously selected configurations $\mathbb{Q}_{\text{selected}} = \{\mathbf{q}_1^{d, \text{avg}} \dots \mathbf{q}_m^{d, \text{avg}}\}$ and the new configuration $\mathbf{q}_{m+1}^{d, \text{avg}}$ is maximized. That is,

$$\mathbf{q}_{m+1}^{d, \text{avg}} = \operatorname{argmax}_{\mathbf{q}^{d, \text{avg}}} [D(\mathbb{Q}_{\text{selected}}, \mathbf{q}^{d, \text{avg}})], \quad (1)$$

where D is the kernel distance between the selected descriptors (and associated frames) $\mathbb{Q}_{\text{selected}}$ and the candidate descriptor $\mathbf{q}^{d, \text{avg}}$. In our case, we use the SOAP descriptor as a structural fingerprint of a configuration and the dot product between two SOAP descriptors as our kernel similarity measure.⁶³ As has previously been shown for molecular systems, we find that this method of selection enables the training of a potential, which demonstrates good transferability.⁵⁹ However, due to the nature of the sampling, it lacks the dense population of configurations around particular local minima which we find are important for achieving very high accuracy on particular crystalline properties. We therefore choose to augment the training dataset selected through FPS with a number of mandatory configurations chosen using chemical intuition, focused on the bulk crystalline phases and certain defect and surface structures. Specifically, we note that the optimized geometries for structures used in the validation sections of this paper are included in the training. The final database is comprised of the union of the 4000 FPS-selected points and the existing GAP-17 dataset, while further ~ 1000 configurations are manually added to target specific properties.

The selected configurations, as well as a representation of their position in phase space, are illustrated in Fig. 1. This sketch-map^{86,87}

representation of the total training dataset uses the same measure of kernel similarity as discussed above to position points in a reduced dimensionality such that points that are similar in the full high-dimensional descriptor space are closer together and those that are dissimilar are further apart.

This sketch-map representation also serves as a qualitative overview of the type of structure to which we fit our model. Structures with carbon atoms of highly varied coordination environments from sp^1 to sp^2 and sp^3 can be seen. Those allotropes that are sp^2 hybridized, such as graphene, graphite, and carbon nanotubes, are clustered together toward the right of the map. Amorphous structures can be seen as a large region in the center, with low density (sp^1 and sp^2 rich) amorphous carbon at the far right and high density sp^3 rich amorphous carbon toward the left, eventually approaching crystalline diamond at the very far left of the map. The more exotic, sometimes hypothetical structures collected from the SACADA database are often found separated from bulk crystalline or amorphous configurations. In the far top right of the map, the isolated gas phase dimer configurations are found.

III. TRAINING OF THE POTENTIAL

We choose to construct GAP-20 to represent the PES as the combination of contributions from a two body (2b), a three body (3b), and a high-dimensional many-body (MB) component. It is an empirical observation that a large proportion of the interaction in an atomistic system may be satisfactorily captured by considering 2b interactions. In particular, this is the case for the exchange repulsion experienced as interatomic distances become very small. Representing this exchange repulsion in its full high-dimensional form would be expensive from the perspective of training data generation, potential generation, and the ultimate evaluation of the potential. The

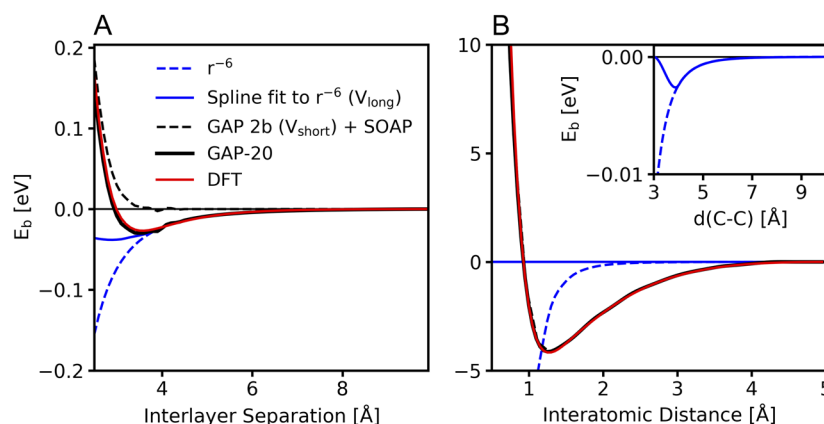


FIG. 2. Construction of the long-range 2b component of the GAP model. An analytical spline is fitted to a function, which decays as r^{-6} , designed to recover the long-range attraction between graphene layers. (a) The predicted energies for each component for the interaction of two graphene layers at different distances. The long range attraction is well characterized by the r^{-6} component of the r^{-6} potential, which is, in turn, well recovered by the analytical spline. The GAP fit using a 2b (V_{short}), 3b, and SOAP descriptor provides the appropriate repulsive potential at short distances but is too short ranged to describe the attractive tail. GAP-20 reproduces the whole curve with good accuracy across a range of distances. (b) The same decomposition for the gas phase dimer. In this case, the strong bonding interactions are dominated by the GAP 2b (V_{short}), 3b, and SOAP descriptor components. The energy of the r^{-6} component becomes large and negative for short distances. [(b), inset] A closer view how the 2b spline fit to the r^{-6} component is brought smoothly to 0 at a distance of 3 Å.

nature of bonding interactions for carbon may also be captured in an approximate way, being generally attractive between 1.2 Å and 1.6 Å, with an attractive tail at long distances. We design the 2b part of our model as a GAP fitted 2b component (V_{short}) $r < 4.0$ Å. For larger separations ($10 \geq r > 4.0$ Å), this smoothly transitions to an analytical spline potential (V_{long}), which decays as r^{-6} . This long range component is fitted to correctly reproduce (albeit without many-body contributions) the long-range attraction due to van der Waals interactions of graphitic layers. A smooth transition is achieved by first fitting the analytical form of V_{long} to the graphene bilayer interaction curve from 3.0 Å to 10.0 Å. V_{short} is then trained by first subtracting V_{long} from the total energy and fitting to the difference. The resultant 2b potential (Fig. 2) simply has the final form $V_{\text{short}} + V_{\text{long}}$.

The true subtleties of interatomic bonding are inherently many-body in character, however. We represent these higher-order contributions to the potential energy using a combination of a 3b descriptor and the aforementioned SOAP descriptor. The full details of the construction of the 3b and SOAP descriptors are given in detail elsewhere.^{53,62,63,88–91}

In short, the 3b term is a symmetrized transformation of the Cartesian coordinates of triplets of atoms, which is designed to be permutationally invariant to the swapping of the atomic indices.^{53,62} In the construction of the SOAP descriptor, the local environment around a target atom is represented by a “local neighbor density,” constructed by placing a Gaussian basis function on each neighboring atom within a certain cutoff, which we choose to be 4.5 Å. The Gaussian basis functions are scaled by a factor of $1/r^{0.5}$ to reflect the greater contribution to the material properties of atoms that are closer together.^{92–94} Other functional forms for the radial scaling exist, and the introduction of this scaling was performed independently of the optimization of the SOAP descriptor cutoff, the choice of which is motivated below. As a result, there may still be scope for further optimization of these parameter sets beyond what is performed here. The local neighbor density is expanded in a basis set of spherical harmonics, the coefficients of which form a “SOAP vector.” In our case, we use a basis set up to order $l = 4$ and $n = 12$, our motivation for which is discussed in the following paragraph. This SOAP vector constitutes a unique representation of the local environment, which satisfies the requirements of being translationally, rotationally, and permutationally invariant. The SOAP kernel, used for regression, is constructed as the scalar product of individual SOAP vectors. Such a kernel is physically interpretable, as it corresponds directly to the integral of two neighbor densities for all possible 3D rotations. Details for the specific choices for a number of associated hyper-parameters are given in the [supplementary material](#), while further details on their significance are given elsewhere.^{39,53,62,63,88–91}

We now provide the details of select convergence tests for the optimization of our GAP model. These tests consider the independent optimization of the SOAP descriptor cutoff, the number of sparse points, and the order of the radial basis set expansion. In general, we begin with a SOAP descriptor with an expansion of the neighbor density up to $l_{\text{max}} = 8$, $n_{\text{max}} = 8$, a cutoff of 4.2 Å, $\sigma_{\text{force}} = 0.01 \text{ eV Å}^{-1}$, $\sigma_{\text{energy}} = 0.001 \text{ eV}$, and $\zeta = 4$. We modify one parameter at a time in isolation while keeping the remainder fixed. We calculate the force error for the resulting models on a randomly selected independent set of test configurations, which is not included in the training.

Figure 3(a) shows the behavior of the force errors as a function of the SOAP descriptor cutoff. The force error has a minimum for a cutoff of 2.9 Å after which it begins to rise again as the increased size of the descriptor space expands beyond what can be populated with the number of available training configurations. A naïve optimization of these parameters based purely on the force errors would therefore select a cutoff radius of 2.9 Å. However, the selection of these parameters cannot be performed in isolation from the intended application of the potential but must also be motivated by the knowledge of the behavior of the material of interest. In this regard, the force (or energy) error alone may be regarded as an imperfect or incomplete target property for optimization. Specifically, we find that although the minimum in the force error is found at much shorter distances, a longer cutoff of 4.5 Å must be used in order to correctly describe graphitic structures, a feature which we

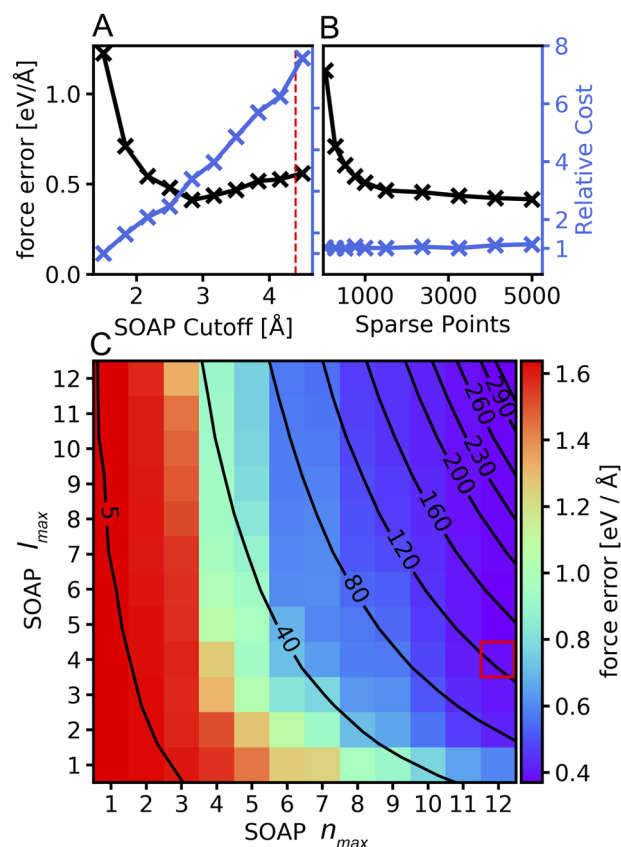


FIG. 3. Mean absolute force errors calculated for an independent test set of configurations for different SOAP descriptors. (a) Force error behavior and cost of evaluation (relative to the fastest GAP) of the resultant model as a function of the SOAP descriptor cutoff, the selected value of 4.5 Å is indicated by the red dashed line. (b) Force error convergence and relative cost as a function of the number of sparse points included in the training. (c) Dependence of the force error and model evaluation cost on the order of the SOAP neighbor density basis set expansion. Force errors are indicated by the color bar, while relative costs are shown by contour lines, and our choice $l_{\text{max}} = 4$, $n_{\text{max}} = 12$ is highlighted by the red square.

consider to be mandatory for this potential. The inter-layer spacing of graphitic structures is typically large (~ 3.3 Å), and a potential must incorporate enough of the structure of both layers to correctly model properties such as the binding curve of graphene layers or the energy difference between AA and AB stacked graphite. The effect of these short cutoffs can be seen in the unsatisfactory behavior of the Tersoff and REBO-II models when modeling the inter-layer spacing of graphite (Table I) or graphene bilayers (see Fig. 6 of the [supplementary material](#)). However, the problem of producing a single analytical metric for optimization, which satisfactorily includes properties such as the lattice parameters, defect formation energies, and phonon errors as well as the force errors themselves, is a challenging one. In this instance, the design choice of selecting an appropriate descriptor cutoff remains partly qualitative in nature.

Figure 3(b) shows the convergence of the mean absolute force errors as a function of the number of sparse points used in the training, and this may be considered as a measure of the data saturation of GAP-20. The force errors decrease rapidly up to ~ 1500 sparse points at which point they begin to level off, although we note that a further increase in the number of sparse points has a negligible impact on the cost of evaluating the model. Our choice of 9000 sparse points is therefore very tightly converged.

In Fig. 3(c), we show how the force error for our model converges as a function of the order of the basis set of radial functions used to expand the SOAP neighbor density. The relative computational cost of each basis set is indicated on the same plot by labeled contour lines. We find that the radial (n) component of the expansion typically has a greater impact on the rate of convergence than the angular (l) component. While previously in the construction of GAP models, band limits $n_{\max} = l_{\max}$ were used for the SOAP descriptor, we find that, surprisingly, an improvement in accuracy can be achieved with essentially no additional cost by making a selection for the basis set expansion, which is strongly biased to the radial ($n_{\max}$) component. Of course, the cost must also be taken into account, and the use of a larger radial component is more expensive than an identical increase in the angular component due to

the greater number of basis functions introduced. Our selection of $l_{\max} = 4$, $n_{\max} = 12$ is chosen as a compromise between accuracy and efficiency. Although a small improvement in the force errors can be achieved by selecting $n_{\max} = 12$, $l_{\max} > 4$, the resultant increase in the cost of training restricts both the size of the training dataset, which can be used, and the size and length scales to which the resultant potential can be applied.

IV. CRYSTALLINE CARBON

Among the most important material properties for any potential to predict accurately are those of the bulk crystalline phases. Table I compares the lattice parameters and bond lengths as predicted by GAP-20 to those from the reference DFT method, in addition to a number of empirical models. In Fig. 4, we also provide both the atomization energies and the formation energies of the crystalline phases relative to the thermodynamically stable state of carbon (at standard temperature and pressure), i.e., graphite. We define the atomization energy of a phase relative to the isolated gas phase carbon atom E_{at} as

$$E_f = E_{\text{bulk}} - nE_{\text{at}}, \quad (2)$$

where E_{bulk} is the energy of the bulk phase and n is the number of atoms in the bulk. Lattice parameters are calculated by independently optimizing the cell vectors for each allotrope, until the total energy is converged to less than 10^{-4} eV. GAP-20 accurately predicts the lattice parameters and bond lengths of all of the tested crystalline allotropes with an average error of 0.2% and their formation energy to within 0.5%.

Accurately modeling the graphite c lattice parameter, corresponding to the spacing between graphitic layers, proved particularly challenging for candidate GAP models, as did the formation energy. This is, in large part, due to the shallow nature of the energetic minimum characterizing the graphite inter-layer interactions and the long range of the atomic descriptor required to capture it. As discussed above, the choice of SOAP cutoff was specifically

TABLE I. Lattice parameters and bond lengths of the crystalline carbon phases. In the case of fullerenes, bond lengths are given *in lieu* of lattice parameters. Absolute values for the lattice parameters are given, with percentage errors relative to DFT in brackets. The Tersoff and REBO-II potentials have no interaction between graphite layers for any physically reasonable lattice parameters and as such these values are omitted.

	Lattice parameter(s) (Å) (% error)					
	DFT	GAP-20	Tersoff	LCBOP	REBO-II	AIREBO
Graphite (a)	2.46	2.47 (0.4)	2.53 (2.4)	2.46 (0.4)	2.46 (0.4)	2.42 (2.0)
Graphite (c)	6.65	6.71 (0.9)	...	6.36 (4.4)	...	6.72 (1.1)
Graphene	2.46	2.46 (0.0)	2.53 (2.8)	2.46 (0.0)	2.46 (0.0)	2.42 (1.6)
Diamond	3.58	3.59 (0.3)	3.57 (0.3)	3.57 (0.3)	3.57 (0.3)	3.56 (0.6)
Hexagonal diamond	2.52	2.53 (0.4)	2.52 (0.0)	2.52 (0.0)	2.52 (0.0)	2.52 (0.0)
Nanotube-(9, 9)	4.26	4.25 (0.2)	4.35 (2.1)	4.24 (0.5)	4.26 (0.0)	4.18 (1.9)
Nanotube-(9, 0)	2.41	2.39 (0.8)	2.53 (5.0)	2.47 (2.5)	2.47 (2.5)	2.43 (0.8)
C ₆₀ fullerene	1.40	1.40 (0.0)	1.46 (4.5)	1.41 (0.7)	1.42 (1.4)	1.40 (0.0)
C ₁₀₀ fullerene	1.39	1.39 (0.0)	1.39 (0.0)	1.39 (0.0)	1.39 (0.0)	1.39 (0.0)

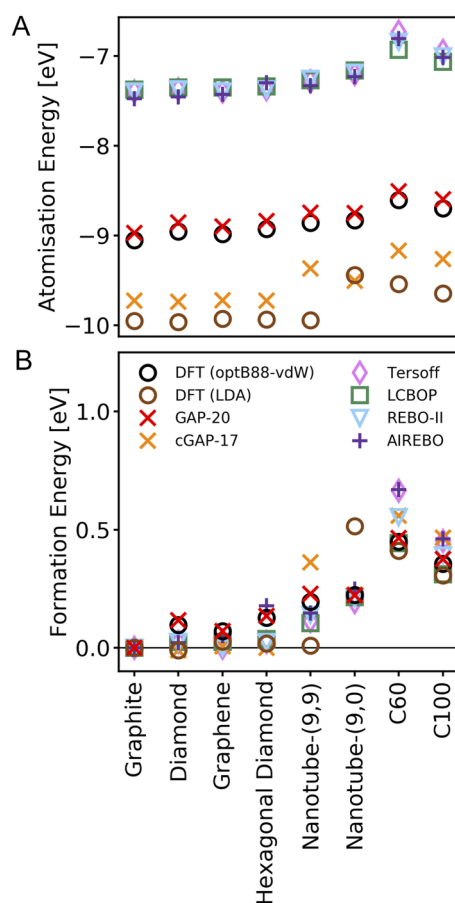


FIG. 4. Formation energies of the crystalline phases of carbon, comparing results from DFT (optB88-vdW and LDA) to those from GAP-20 and the other models tested. (a) Atomization energies using the isolated gas phase carbon atom as a reference, differences are dominated by overstabilization of the gas phase atom by empirical models. (b) Formation energies given relative to the graphite formation energy of each particular model.

informed by a desire to capture this property correctly. We consider this, in particular, to be a mandatory characteristic of a general carbon potential, which would be absent for any model with a shorter cutoff.

It is useful here to make a brief comparison to the selected empirical potentials. While we do include DFT reference data for all properties, in Secs. V–VIII, these reference values are only computed using the same level of DFT used to train GAP-20. For benchmarking GAP-20, which is our primary purpose, this is not problematic; however, we do not fully account for the potential impact of functional dependence or the errors of DFT with respect to experiment, when making comparisons to empirical models. Many of the empirical models considered are fitted to experimental data or contain values from other DFT functionals, typically LDA, which should be taken into account when comparing different model predictions to our optB88-vdW reference values. To give some indication of the functional dependence, reference values for the formation energies

in Fig. 4 are given using both the optB88-vdW and LDA functionals. We also re-iterate that the GAP-17 model was fitted to LDA data, so it would be expected to accurately reproduce DFT values at this level only.

On average, GAP-20 predicts the lattice parameters of the tested crystalline phases with an error of 0.2%, while the Tersoff, LCBOP, REBO-II, and AIREBO potentials have errors averaging 5%, 0.3%, 4%, and 1%, respectively (Table I). What the Tersoff and REBO-II potentials gain in efficiency by using short cutoffs, they lose in accuracy, notably by predicting dramatically incorrect inter-layer spacings (*c* lattice parameters) for graphite. This error is fixed by the inclusion of medium and long-range terms to account for van der Waals interactions in the LCBOP and AIREBO models, however. Despite its inaccuracy for graphene, the REBO-II potential does achieve good accuracy on the remaining lattice parameters; the additional terms included in the bond-order potential constitute a dramatic improvement over the Tersoff potential. Due to its complete reparameterization to account for the effects of long-range interactions in the bond-order part of the potential, LCBOP does outperform the other empirical potentials tested here in most cases.

In absolute terms, the atomization energies [Fig. 4(a)] from the tested empirical potentials differ significantly from those predicted by both reference DFT methods due to the very different energies of the isolated gas phase atom. In the case of GAP-17, the small offset between the LDA reference and the model prediction is the result of the isolated atom not being included in the training dataset. When using the formation energy of graphite as a reference state however [Fig. 4(b)], this offset is removed and the agreement between the empirical models and DFT improves considerably. When using both the gas phase atom and graphite as a reference, there is an excellent agreement between GAP-20 and the optB88-vdW DFT reference for all of the phases considered here. GAP-20 uniformly predicts the atomization energies of the tested allotropes to within an error of 1%, including the relatively subtle difference in energetics between normal cubic and hexagonal diamond and the energetics of nanotubes and fullerenes. The inclusion of the gas-phase atom in the training is vital to accurately predict these atomization energies. There is surprisingly little difference between the formation energies predicted by the different many-body potentials tested here, though there are a few points of note. First, due to their short cutoffs, the Tersoff and REBO-II potentials do not distinguish between graphite and graphene as the thermodynamically stable phase and as such their formation energies are predicted to be equal. Similarly, only the GAP-20, LCBOP, and AIREBO models correctly favor cubic over hexagonal diamond, although the AIREBO model overestimates the difference in energy by a factor of 5, while the other models considered do not distinguish between the two diamond phases. A more complete evaluation of the formation energies for different chiralities of nanotubes is given in the [supplementary material](#); for GAP-20, the energy errors for a significantly wider range of structure types are also given.

In addition to the static properties of the crystalline allotropes, it is an important characteristic of any potential that it accurately models the lattice dynamics of bulk crystals, i.e., their behavior at finite temperature. The phonon spectrum of a material is a direct probe of this which is experimentally measurable. In addition, a number of thermodynamically relevant properties, including the

thermal expansion coefficient and the constant volume heat capacity of a material, may be calculated directly from the phonon dispersion relation by the calculation of the free energy. It is clear therefore, why a correct prediction of the phonon dispersion relation is a highly desirable feature of an interatomic potential.

Figure 5 shows the comparison between the phonon dispersion curves calculated using the reference DFT method and those calculated using the carbon GAP model for graphene, diamond, a zig-zag (9, 9), and an armchair (9, 0) carbon nanotube. Phonon dispersion curves were computed using the finite displacement method as implemented in the Phonopy Python package.⁹⁵ Equivalent curves for the other models tested are provided in the [supplementary material](#).

We have previously reported results comparing the phonon dispersion relation for a purely graphene GAP model to those from experimental x-ray scattering data and a number of reference DFT methods.⁵³ It is useful to make a comparison between the highly targeted model previously published and the much more general GAP-20 introduced here. A particular concern might be that significantly expanding the configurational space on which we wish to train, as we have done here, would necessarily damage the quality of the predictions for graphene compared to the previous model—particularly for a property as sensitive as the phonon dispersion curves. It is demonstrated in Fig. 5(b) that this is not the case; the dispersion relation of the phonon curves for graphene from GAP-20 is comparable to that of the previously published graphene GAP model.⁵³ The energies of the phonon bands are correctly predicted across all of the high symmetry directions plotted, while the frequencies

(in particular at the high symmetry points) are found to be correct to within 4 meV, which may be compared to a value of 1 meV for the pristine graphene model.⁵³ The quality of the GAP-20 model prediction is comparable for diamond (which cannot be modeled at all with the pristine graphene model), though with marginally larger errors for the prediction of the energies of certain bands, up to 7 meV for the higher frequency modes. GAP-20 also captures the difference in vibrational behavior between armchair and zig-zag nanotubes remarkably well. There are some differences in the energy of certain splittings for some bands, but the magnitude of these errors is small, typically on the order of 2 meV–3 meV. In particular, Fig. 5 shows that the vibrational density of states for the two nanotube systems agrees well with the DFT reference.

V. SURFACES OF CARBON

From the point of view of atomistic simulation, surfaces present a major challenge, as their correct description requires a treatment of a number of competing physical interactions.^{81–84,96}

We compute the surface energy for each model by first optimizing the bulk structure for the parent crystal until the total energy is converged to 10^{-3} eV. We then cut the surface along the desired direction and compute the specific surface energy γ_s at $T = 0$ K as

$$\gamma_s = (E_n - nE_{\text{bulk}})/2A, \quad (3)$$

where E_n is the energy of the n slab layer containing two surfaces, which may be as-cut (unrelaxed) or allowed to relax, E_{bulk} is the

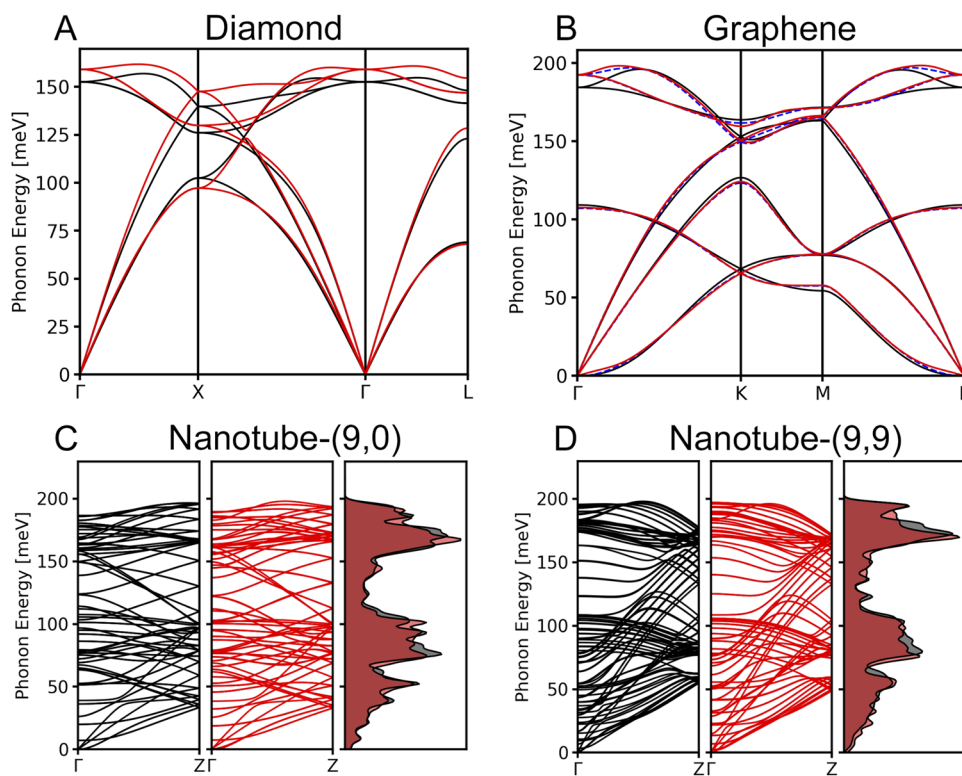


FIG. 5. (a) Phonon dispersion relation for diamond as predicted by GAP-20 (black) with comparison to DFT (optB88-vdW) reference data (red). (b) Graphene phonon dispersion relation comparing GAP-20 and DFT (optB88-vdW) reference data. The blue dashed line shows the predicted phonon dispersion curve for the graphene-only model previously published.⁵³ (c) (9, 0)-Nanotube phonon dispersion and vibrational density of states. (d) (9, 9)-Nanotube phonon dispersion relation and vibrational density of states. Equivalent comparisons for the other models tested are given in the [supplementary material](#).

energy of a single atom in the bulk structure, and A is the area of the surface structure. In the case of the amorphous surfaces, due to the extent of the surface relaxation observed, we report only the as-cut surface energies.

Graphite may be readily cleaved to expose its (0001) surface, which is remarkably stable and is by far the predominant face of graphite, while in diamond, the (100), (111), and (110) surfaces are of particular interest.⁹⁷ We also compute the as-cut surface energies for an ensemble of amorphous structures by cutting bulk amorphous systems along different directions.

The energies of several important surface cuts and their reconstructions are given in Table II. GAP-20 typically predicts the diamond surface energies correctly to within 7%, the exception being the case of the relaxed diamond (111) surface, where the error is slightly larger at 15%. The structures of the relaxed surfaces were also found to be in excellent agreement, with the average error in the positions of individual surface atoms being 10^{-3} Å. The graphite (0001) surface energy is extremely small, and it thus proved challenging to produce a model, which could correctly characterize this; however, GAP-20 predicts the unrelaxed and relaxed surface energies correctly to within an error of 3 meV Å⁻² (20%). With the inclusion of vdW interactions considered in their construction, both the LCBOP and AIREBO potentials predict the graphite (0001) surface energy rather well, with errors of 67% and 27%, respectively.

While GAP-20 achieves low errors for the surface energies of all the diamond surfaces considered, the other models generally perform well for at least one diamond surface, though none exhibit uniformly low errors. The Tersoff, REBO-II, and AIREBO models predict the energies of the diamond (110) surfaces to within 10% of the reference value. Conversely, of the empirical models, only the LCBOP potential correctly predicts the energy of the diamond (100) surface; errors for the Tersoff, REBO-II, and AIREBO potentials were 22%, 28%, and 33%, respectively. None of the empirical potentials performed well for the (111) surface of diamond. The

Tersoff and REBO-II models do not show any binding between graphitic layers for any reasonable initial geometry. This would lead to the spontaneous exfoliation of graphite layers and the eventual disintegration of graphite crystals in simulations employing these models.

VI. DEFECTIVE CARBON

A certain concentration of defects is a guarantee in any experimental material sample. Such imperfections may have a strong impact on the structural, optical, and thermal properties of a material and may be introduced into a crystal structure to induce or modify properties. The engineering of defects is of great technological importance, and consequently, their accurate modeling by an interatomic potential is highly desirable. The possibility of rehybridization, which allows carbon atoms to reconstruct with differing numbers of bonds to stabilize particular structures, allows carbon to have a wider variety of defects than most other elements.

To the best of our knowledge, there is not a set of defect formation energies for a wide range of carbon defects computed at precisely the same level of theory. Therefore, we here assemble such a reference set for which we compute defect formation energies in large supercells to avoid defect self-interaction in the computation of energies. For graphite, a $(6 \times 6 \times 2)$ supercell with 288 atoms and four graphite layers was used.^{77,98} In the case of graphene, a (10×10) supercell with 200 atoms was employed, and for diamond, a $(3 \times 3 \times 3)$ supercell with 216 atoms was used.^{53,78} Defect formation energies are calculated for the representative (9, 9) and (9, 0) index carbon nanotubes, which had 174 and 180 atoms in the supercells used, respectively.⁷⁹ For each structure, the lattice parameters and ionic positions of the pristine structures were optimized, as discussed previously. The ionic positions of the defective structures were then optimized until the energy was converged to within 10^{-5} eV while keeping the lattice parameters fixed. We compute the

TABLE II. Surface energies of low Miller index surfaces for common carbon allotropes. Reference energies are calculated using DFT, and absolute values from each model are given, with their percentage error in brackets. Note that for the amorphous surfaces, the surface energy is averaged over a large number of different surfaces. In the amorphous case, the error provided is the average of the individual point-wise errors, rather than the error between the average surface energies.

	Surface energy (eV Å ⁻²) (% error)					
	DFT	GAP-20	Tersoff	LCBOP	REBO-II	AIREBO
Diamond (100) (as cut)	0.56	0.60 (7)	0.47 (16)	0.61 (9)	0.69 (23)	0.73 (30)
Diamond (100) (relaxed)	0.54	0.56 (4)	0.42 (22)	0.59 (9)	0.69 (28)	0.72 (33)
Diamond (111) (as cut)	0.64	0.73 (14)	0.88 (38)	1.07 (67)	1.00 (56)	1.03 (61)
Diamond (111) (relaxed)	0.62	0.66 (6)	0.88 (42)	1.07 (73)	1.00 (61)	1.03 (66)
Diamond (110) (as cut)	0.68	0.70 (3)	0.70 (3)	0.89 (31)	0.74 (9)	0.74 (9)
Diamond (110) (relaxed)	0.68	0.67 (1)	0.63 (7)	0.83 (22)	0.69 (1)	0.69 (1)
Graphite (0001) (as cut)	0.015	0.013 (13)	0 (100)	0.005 (67)	0 (100)	0.011 (27)
Graphite (0001) (relaxed)	0.015	0.012 (20)	0 (100)	0.005 (67)	0 (100)	0.011 (27)
Amorphous surfaces (as cut)	0.26	0.27 (4)	0.25 (4)	0.25 (4)	0.25 (4)	0.25 (4)

formation energy E_f of a vacancy defect relative to the energy of an atom in an ideal parent structure,

$$E_f = E_d - (nE_{at} + E_{bulk}), \quad (4)$$

where E_d is the energy of the defective supercell structure, E_{bulk} is the energy of the undefective bulk structure, and E_{at} is the energy of a single atom in the bulk structure, while n is the number of carbon atoms added (positive n) or removed (negative n) to form the defect.

The simplest of defects involves the absence of one or two atoms from their regular position in the lattice, forming monovacancy and divacancy defects (Fig. 6). Monovacancy defects often result in unsaturated bonds at the defect site, while divacancy structures, particularly in sp^2 hybridized systems, can reconstruct to produce saturated configurations. In graphene, graphite, and carbon nanotubes, the 14-membered ring formed by the removal of two adjacent atoms from the structure reconstructs to form a saturated sp^2 structure with two 5-membered rings and one 8-membered ring—a more stable structure known as a 5-8-5 divacancy. In graphene, this defect may further reconstruct to remove the unfavorable 8-membered ring to form a 555-777 or 5555-6-7777 divacancy reconstruction. Monovacancy coalescence is also observed in diamond, whereupon annealing at high temperature, monovacancies migrate to form divacancy defects, with fewer unsaturated bonds per absent carbon atom.

Graphite is the only allotrope of carbon in which true interstitial defects are known wherein interstitial atoms may be found between graphite layers.⁷⁷ The most stable arrangement of this is in a “dumbbell” configuration, where the adatom displaces an atom in the graphite structure to form a symmetric arrangement of trigonally bonded carbon atoms above and below the sheet. The isolated interstitial atoms are not known either experimentally or from theory to be stable in diamond, rather a split interstitial is found, where a lattice site is shared by two carbon atoms, which are displaced along the [100] and [100] directions.⁹⁹

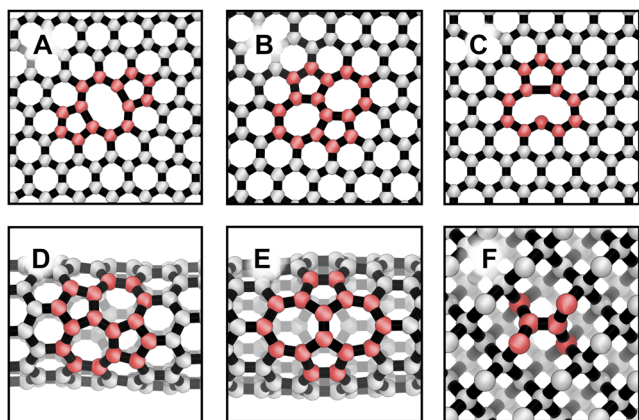


FIG. 6. Images of the selected carbon defect structures with atoms in the immediate vicinity of the defect highlighted in red. (a) Graphene divacancy defect, (b) graphene Stone-Wales defect, (c) graphene monovacancy, (d) (9, 9)-nanotube Stone-Wales defect (transverse orientation), (e) (9, 9)-nanotube Stone-Wales defect (parallel orientation), and (f) diamond split interstitial defect.

In sp^2 bonded allotropes of carbon, the rotation of a single C–C bond transforms four 6-membered rings into a cluster of two 7-membered and two 5-membered rings, forming a Stone–Wales type defect.^{100–102}

Table III compares the energies of a number of defects as computed with DFT, GAP-20, and the other models considered. In most cases, GAP-20 correctly predicts the defect formation energy to within an error of 10%. Typically, the prediction of the formation energies of Stone–Wales type defects was found to be extremely accurate, with no error (to within the precision of the values given) in either the graphite or graphene cases and only small errors for nanotubes. The errors for the formation energies of diamond defects tend to be larger, ranging from 25% to 35%, while those for defective nanotubes range from 0% to 11%. Anecdotally, we note that although relevant training data for the defects considered are represented in the training data, it proved challenging to achieve defect formation energies, which were universally accurate. In particular, this is due to the sensitivity of the formation energies to aspects such as the SOAP descriptor cutoff, specific training data included, and the number of sparse points used in the training.

Considering the empirical potentials, we find that the modifications to the Tersoff potential included in the REBO-II model dramatically improve the quality of the predicted defect formation energies; percentage errors are often decreased by an order of magnitude or more when comparing these two potentials. Surprisingly, these results show that the inclusion of the long-range Lennard-Jones term in the AIREBO model often has a negative impact on the accuracy of its predicted defect formation energies, indicating that the addition of a long-range term without reparameterization of the short-range components has adversely impacted the energetics of the model. Indeed, in the case of LCBOP, where this reparameterization of the short range bond-order potential has been performed, we find that the errors are significantly reduced and are in many cases comparable with the performance of GAP-20. The exception to this being the case of defective nanotubes, where LCBOP exhibits errors ranging from 11% to 52%. In fact, the prediction of nanotube defect formation energies proved challenging for all of the empirical models considered. In a number of cases, defect formation was found to be an energetically favorable process and was associated with a strong relaxation of the nanotube structure after defects were induced.

As well as accurately predicting the energetic cost of inducing defects in carbon structures, GAP-20 was also found to accurately predict the structures of these defects. We quantify this accuracy by calculating the structural similarity between the defect structures optimized with our GAP model and those from DFT in the form of the root mean square error (RMSE) between the two optimally overlapped structures. In all but a handful of cases, the RMSE for these defects is vanishingly small, with atoms having an error in their position of less than 10^{-2} Å, when comparing identical atoms from GAP-20 and DFT structures. In particular, the presence and height of the characteristic buckling of the Stone–Wales defect in graphene was well described, as was the structural distortion resulting from the presence of defects in both (9, 9) and (9, 0) index carbon nanotubes. Similarly, the rehybridization and reconstruction of (5-8-5), (555-777), and (5555-6-7777) graphene divacancy defects were accurately reproduced, as were the geometries of all of the diamond defects considered. Situations in which GAP-20 showed

TABLE III. Formation energies of common defects in carbon structures for GAP-20 and the other models considered, with DFT (optB88-vdW) values given as a reference. Data are given in eV, with percentage errors relative to DFT given in brackets. In each case, the value given is for the optimal geometry of the defect found with that particular model.

	Formation energy (eV) (% error)					
	DFT	GAP-20	Tersoff	LCBOP	REBO-II	AIREBO
Graphene Stone–Wales	4.9	4.8 (2)	1.9 (61)	4.5 (8)	5.3 (8)	5.4 (10)
Graphene monovacancy	7.7	7.0 (8)	2.5 (68)	6.9 (10)	7.5 (3)	7.2 (6)
Graphene divacancy (5-8-5)	7.4	7.9 (7)	5.1 (31)	7.5 (1)	7.5 (1)	9.2 (24)
Graphene divacancy (555-777)	6.6	6.9 (5)	5.2 (21)	6.6 (0)	6.8 (3)	8.7 (32)
Graphene divacancy (5555-6-7777)	6.9	7.4 (7)	7.9 (14)	7.2 (4)	7.6 (10)	9.5 (38)
Graphene adatom	6.4	5.9 (8)	6.7 (5)	6.8 (6)	7.4 (16)	7.8 (22)
Graphite monovacancy	7.8	7.3 (6)	7.1 (9)	7.8 (0)	7.9 (1)	7.6 (3)
Graphite divacancy (5-8-5)	9.6	9.2 (4)	12.6 (31)	8.2 (15)	8.0 (17)	9.7 (1)
Graphite Stone–Wales	5.4	5.6 (4)	12.8 (137)	5.7 (6)	6.0 (11)	6.0 (11)
Graphite interstitial	7.4	7.9 (7)	9.7 (31)	7.2 (3)	7.1 (4)	6.8 (8)
Diamond monovacancy	6.6	4.3 (35)	5.2 (36)	7.2 (11)	7.1 (4)	6.8 (8)
Diamond divacancy	9.1	6.6 (27)	5.1 (44)	10.6 (16)	10.7 (18)	10.1 (16)
Diamond split interstitial	11.4	8.3 (27)	12.4 (9)	9.8 (14)	11.0 (4)	11.4 (0)
Nanotube-(9, 9) monovacancy	6.4	5.8 (9)	−5.1 (180)	3.8 (41)	−1.6 (125)	−2.5 (139)
Nanotube-(9, 9) divacancy	4.7	4.8 (2)	−5.5 (217)	2.9 (38)	−0.9 (119)	−2.3 (149)
Nanotube-(9, 9) Stone–Wales (parallel)	4.4	4.5 (2)	−4.5 (202)	2.1 (52)	−2.1 (148)	−3.9 (189)
Nanotube-(9, 9) Stone–Wales (transverse)	3.5	3.5 (0)	−5.8 (261)	2.0 (44)	−3.3 (192)	−2.00 (156)
Nanotube-(9, 0) monovacancy	5.3	4.9 (8)	−0.9 (117)	4.4 (17)	4.7 (11)	3.4 (36)
Nanotube-(9, 0) divacancy	3.6	3.5 (3)	−1.0 (128)	3.0 (17)	4.1 (14)	2.8 (22)
Nanotube-(9, 0) Stone–Wales (parallel)	2.7	3.1 (15)	−1.3 (148)	3.2 (19)	3.4 (26)	3.6 (33)
Nanotube-(9, 0) Stone–Wales (transverse)	3.5	3.2 (9)	−1.1 (131)	3.1 (11)	4.2 (20)	2.6 (26)

structural inaccuracies were the nanotube monovacancy structures and the parallel Stone–Wales defect in the (9, 9) index nanotube for which the GAP model predicted a larger distortion of the bulk nanotube structure due to the presence of the defects. We also find that as with all the models considered here, GAP-20 does not correctly describe the asymmetry introduced through a Jahn–Teller distortion of the graphene monovacancy defect—instead predicting the monovacancy to have a symmetric geometry. This is perhaps unsurprising as the energy difference between the symmetric and asymmetric geometries is typically small (~ 350 meV). However, even in the cases illustrated here, the typical error in the position of any atom was found to be only 0.1 Å. This GAP-20 is capable of accurately modeling both the energetics and structural characteristics of a wide range of carbon defects, which indicates its potential usefulness in a wide range of simulations in which defective structures may be relevant, including fracture, atom bombardment, and simulations of membrane characteristics.

VII. LIQUID CARBON

As discussed previously, the requirements of a potential for the satisfactory modeling of crystalline and liquid or amorphous phases are significantly different. In the case of crystalline materials,

a highly accurate description close to a local minimum for a system is required.⁵³ Conversely, in a liquid simulation, a vastly greater number of local configurations are explored, requiring a high degree of flexibility and transferability.⁶² As in the case of GAP-17, we therefore use the liquid as a benchmark for the flexibility of our potential,⁶² scanning over a wide range of densities and (here) temperatures. The aim is to diagnose any possible issues, which might be exposed by visiting a very diverse set of configurations during the simulations. There is a strong precedent for the study of high temperature liquids, including carbon, using DFT.^{103,104} A good agreement with DFT-MD data is therefore strong evidence for the usefulness of the potential for further studies of liquid carbon, which is present only under extreme conditions, but is nonetheless vitally important, e.g., for understanding the nucleation and formation of diamond and graphite under a wide range of circumstances.^{105–108}

The radial distribution function (RDF) of a liquid represents a convenient measure of its local structuring, as does its angular distribution function (ADF). Here, we compare the results of constant volume *ab initio* molecular dynamics simulations to those of GAP-20. We perform two sets of simulations, one for a range of densities between 1.5 g cm^{−3} and 3.5 g cm^{−3} at 5000 K and the other for a range of temperatures between 5000 K and 9500 K at a fixed density of 2.5 g cm^{−3}. These simulations were performed for

216 atom systems using a chain of five Nosé–Hoover thermostats. *Ab initio* trajectories were generated using VASP, simulations were performed at the gamma point, and data were collected for 5 ps at each temperature and pressure.^{71–73} We find that there is a very good agreement between the *ab initio* data and the GAP-20 predictions for both RDF and ADF across the wide range of temperatures studied (see Fig. 7). GAP-20 correctly models the increased structuring of the liquid carbon as the temperature is reduced from 9500 K to 4500 K. At temperatures below approximately 3500 K, the GAP model predicts the liquid to form an amorphous glass that slowly graphitizes (which is entirely expected because the temperature is now below the melting line). While a full discussion on the mechanism of formation and resulting morphology of graphitized amorphous carbons generated using GAP-20 is beyond the scope of the current work, this process has previously been shown to be an excellent method of differentiation between the numerous available carbon potentials.^{36,37} Figure 8 shows the RDF and ADF computed with both GAP-20 and optB88-vdW across a wide range of

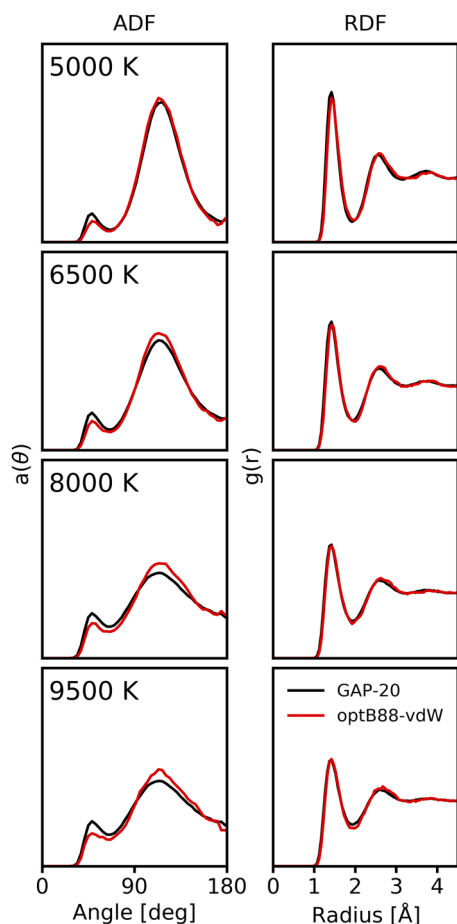


FIG. 7. Angular and radial distribution functions for liquid carbon at a fixed density of 2.5 g cm^{-3} for temperatures between 5000 K and 9500 K. GAP-20 results are shown in black, while reference DFT (optB88-vdW) data are given in red.

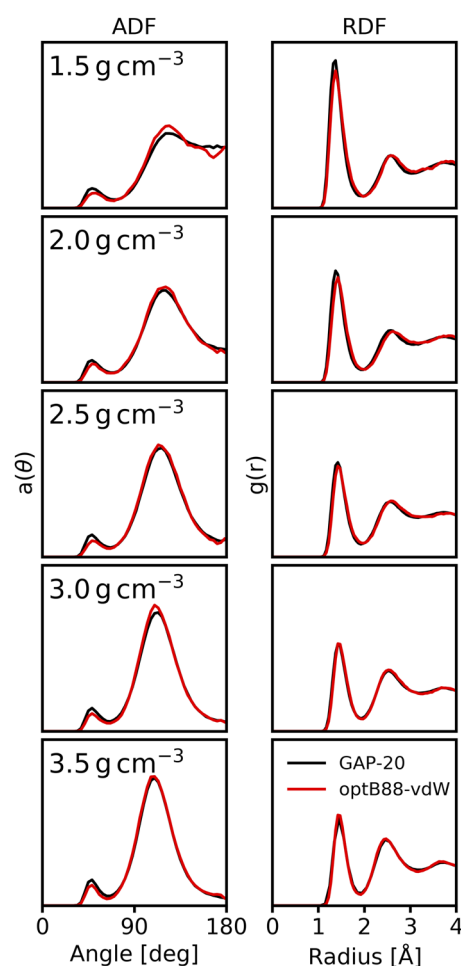


FIG. 8. Angular and radial distribution functions for liquid carbon at 5000 K for a range of densities from 1.5 g cm^{-3} to 3.5 g cm^{-3} . GAP-20 results are shown in black, while reference DFT (optB88-vdW) data are given in red.

densities from 1.5 g cm^{-3} to 3.5 g cm^{-3} at 5000 K. This test is particularly important as it represents dynamical simulations of structures from highly sp^1 and sp^2 hybridized (low density) to a predominantly sp^3 hybridized liquid at higher density. GAP-20 captures this change in the bonding characteristics of liquid carbon, in particular the increase in the sp^1 hybridized fraction of the liquid at very low densities (reflected in bond angles close to 180°), qualitatively similar to GAP-17.⁶²

That GAP-20 can model the atomistic structure of liquid carbon at a wide variety of temperatures and densities, while maintaining the ability to accurately predict properties such as the phonon relation and defect formation energies is a reflection of the flexibility of the GAP methodology. Such varied systems explore a range of characteristic energies, where important fluctuations cover many orders of magnitude; in carbon, this can be anywhere from the meV range in the case of differences between graphite defect energies and fluctuations on the order of tens of electron volts, as encountered

in the liquid. Despite these very different energy ranges, it is not unlikely that a potential may encounter all of them over the course of a single simulation (for example, during the crystallization of a solid phase directly from the liquid), and it is therefore important that they be handled correctly.

VIII. TRANSFERABILITY OF THE POTENTIAL

Ultimately, the purpose of any interatomic potential is that it may be used for the discovery of new and interesting phenomena. Consequently, in its application, it may encounter structures that were not explicitly considered in its construction, in this case meaning that it must model structures that were not included in the training database. It has therefore been a criticism of ML potentials that their poor performance in extrapolation might inhibit their use for scientific discovery. As discussed earlier, the problem of extrapolation is circumvented by the fact that we consider only the local environment around a particular atom to be important for predicting its atomic energy and the forces acting upon it. While the problem of exploring the entirety of the 3N-dimensional chemical space is indeed intractable, sufficiently sampling all of the physically relevant local environments is not.⁵⁹

We demonstrate this here by performing a diagnostic GAP driven random structure search (GAP-RSS), similar in spirit to Refs. 109 and 5, and demonstrate that the predicted energies of these structures agree well with those from DFT.^{5,60,85} We then calculate a number of high energy pathways for specific transformations not included in the training and compare these to DFT. Both of these tests serve the purpose of exploring the high energy regions of the potential energy surface, which may be explored during molecular dynamics simulations or geometry optimization and which must be well described for an ML model to be transferable. Importantly, they are both explicitly designed to include configurations that are not present in the training dataset of GAP-20.

To perform the first test, we generate a cubic unit cell with lattice parameter $a = 3 \text{ \AA}$. In this cell, we randomly place eight carbon atoms, avoiding any overlaps such that the distance between any two carbon atoms is not less than $2 \text{ \AA}$. This process is performed to generate 1000 initial randomized geometries. The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package is then used to optimize lattice vectors of the cell independently using conjugate gradient descent, while maintaining their orthogonality, until the total energy is converged to within 10^{-8} eV .¹¹⁰ Following this, the positions of the atoms in the unit cell are optimized using a FIRE algorithm,¹¹¹ until the total energy is again converged to within 10^{-8} eV . This cycle is repeated twice more before performing a final conjugate gradient optimization of the atomic positions and cell vectors until the total energy is converged to 10^{-10} eV .

To validate the results of our GAP-RSS, we recompute the energies of the structures found using the reference DFT method used to train the model. We note that for across all 1000 structures, the predicted energy agrees well with the energy predicted from DFT. It has previously been shown that correctly identifying low energy structures from a RSS is an extremely challenging task for empirical models, which often predict qualitatively incorrect behavior and fail to find physically relevant configurations due to their having many more local minima than the DFT PES.^{112,113}

Our GAP-RSS correctly identifies a range of important low-energy carbon allotropes, as well as numerous more exotic species. In particular, AB-stacked graphite was found as the lowest energy allotrope of carbon. AA- and ABC-stacked graphite allotropes are also identified in the search; their energy is correctly predicted to be higher than that of the AB stacked graphite structure. Furthermore, both diamond and lonsdaleite are correctly identified. We also identify a number of more exotic carbon allotropes, some of which are known either from experiment or theory but were not included in the training dataset, including cross-linked graphite structures, porous carbon cages, and a variety of haeckelite structures. For the vast majority of structures found during the GAP-20 driven random structure search, the predicted energies from both DFT and GAP-20 agree well (Fig. 9).

We also return at this stage to the sketch-map representation of the training dataset given in Fig. 1. In Fig. 9(b), we provide a projection of the GAP-RSS structures onto this sketch-map representation. GAP-RSS points are colored according to the GAP-20 energy error. The density of structures present in the original training dataset is indicated by black contour lines. It is clear that most structures found are clustered in the region representing the bulk amorphous and crystalline polymorphs, with very few structures representative of fullerenes or nanotubes identified. This is a reflection of the fact that only eight atoms are included in the unit cell used for the RSS. Additionally, the RSS procedure employed begins with simulation cells that are fully periodic and with no symmetry constraints imposed on the initial atomic positions. In the lower left of the sketch-map is a cluster that is structurally distinct from those present in the training data, as indicated by its large separation from other points in the sketch-map. These structures are characterized by their highly sp^1 rich character. Although a significant number of amorphous structures that are rich in sp^1 hybridized carbon atoms are included in the training data, there are indeed very few crystalline sp^1 rich structures. Despite being structurally distinct from anything included in the training dataset, the error in the GAP-20 prediction for the energy of these structures remains low. This indicates excellent performance for GAP-20 in applications where transferability to potentially novel structures is important.

We also test GAP-20 on a number of specific structural transformations. Although our GAP model is not trained explicitly on reaction barriers, it is useful to test how well the model performs for the prediction of the types of barriers, which might be encountered in the studies of the reactivity of carbon nanostructures. To this end, we compare the predictions of our GAP model to those of DFT for two approximate transformations: a rigid bond rotation in graphene and a C60 fullerene. Since these calculations are performed on rigid structures, rather than, for example, using nudged elastic band calculations, the barriers calculated here will not be true defect formation barriers. They are, however, still representations of physically reasonable points on the potential energy surface, which are not included in the training dataset and so form a useful test of the potential compared to other models.^{66,114}

Figure 10(a) shows the barrier to rigid rotation of a C–C bond within a graphene sheet, as predicted by GAP-20, and the tested empirical potentials. The performance of GAP-20 on this test is reassuring for its wider application, and it achieves excellent accuracy with respect to both the height of the local minimum in the rotation and the height of the barrier. The AIREBO and REBO potentials

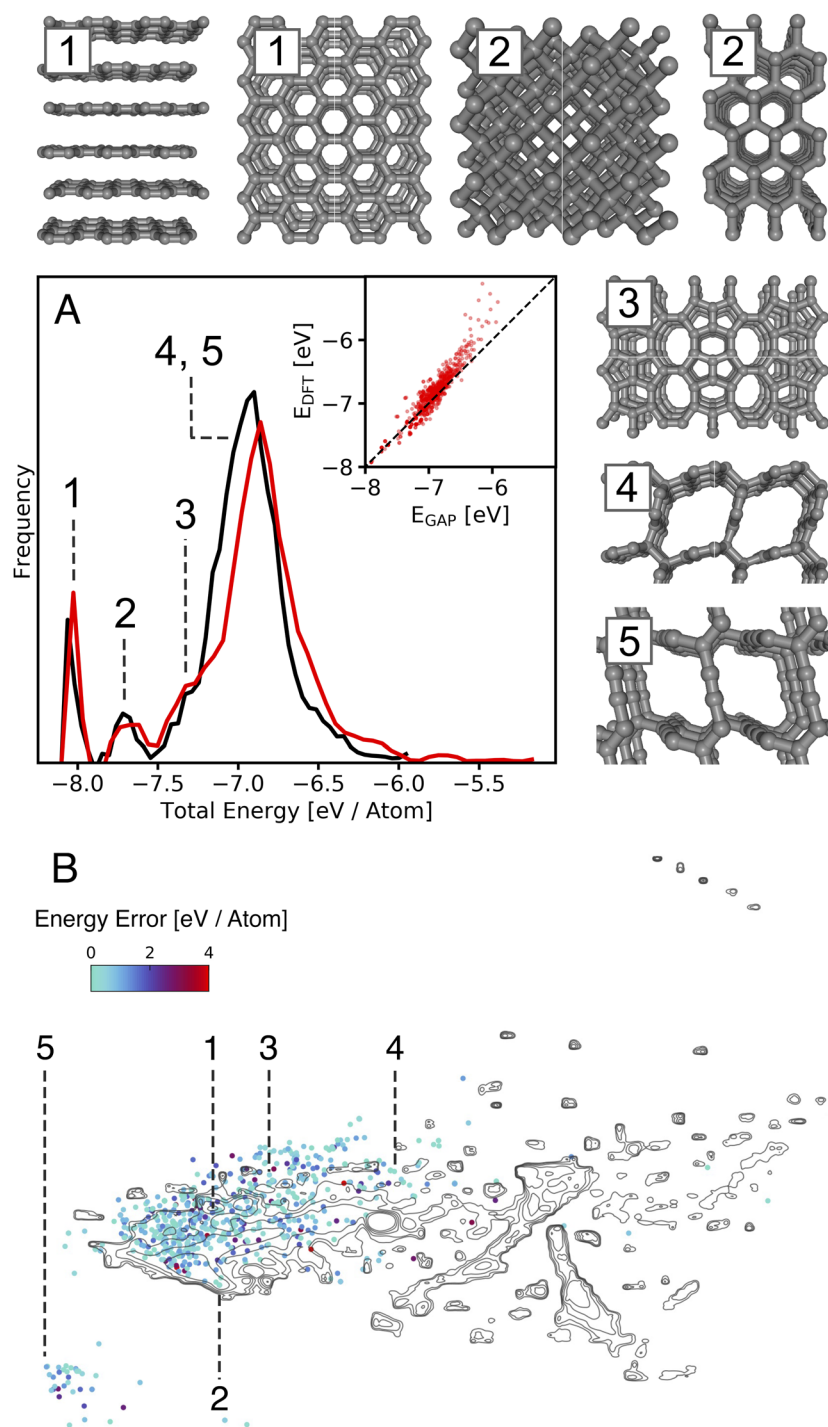


FIG. 9. (a) A comparison of the histograms of energies of structures identified by GAP-RSS, given in eV/atom, showing good agreement for the prediction of the energy of structures between GAP-20 (shown in black) and DFT (optB88-vdW) (shown in red). A number of examples of structures identified in GAP-20 driven random structure search are shown. The position of each of the example structures on the histogram is indicated by their numbering: (1) AB stacked graphite and AA stacked graphite, (2) cubic and hexagonal diamond, (3) haeckelite, (4) cross-linked graphitic structure, and (5) novel carbon structure with high proportion of sp^1 hybridized carbon atoms. (b) The structures resulting from the GAP-RSS projected into the sketch-map representation from Fig. 1. The density of the structures present in the training data is indicated by the contour lines, while the structures identified from the GAP-RSS are shown as individual points.

capture the general shape and height of the barrier but predict the jagged curves for the rotation, as compared to the smooth variation from DFT. The LCBOP and Tersoff potentials perform poorly, overestimating the energy of the rotation by more than 30 eV, and erroneously situating a maximum in the potential energy where a

minimum is found from DFT. In the case of the Tersoff potential, two additional spurious minima are located close to where the DFT maxima are located.

A similar situation is observed for the behavior of the C60 rotational barrier in Fig. 10(b). Here, it is again seen that GAP-20

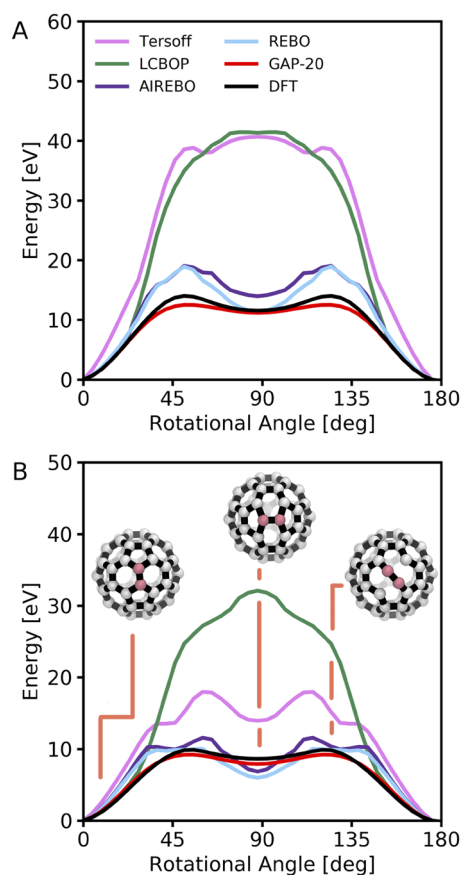


FIG. 10. Energies for rigid transformations of a C–C bond in graphene (a) and a C60 fullerene (b). Results from GAP-20, DFT (optB88-vdW), and a selection of empirical potentials are shown.

performs well, providing a good estimation of the barrier height and shape with respect to DFT. The REBO, AIREBO, and Tersoff potentials all situate spurious minima in the potential energy close to where the maxima in the reference DFT curve are located, although they do also predict a minimum at the correct rotation. LCBO again overestimates the energy of the barrier by 20 eV and situates a maximum in the potential energy surface where a minimum ought to be located.

IX. CONCLUSION

The advantages conferred by the flexibility of the Gaussian approximation potential framework are made clear by a wide variety of structures that are accurately treated by GAP-20. The variable hybridization of carbon makes it an extremely challenging element to model using empirical potentials; its structurally diverse allotropes are energetically similar, and the properties of these depend on a broad range of physical interactions from the weak van der Waals forces binding graphite to the stiff covalent bonds of diamond. We have demonstrated here a model that is equally suited to modeling not just these two bulk structures, but defects,

surfaces, and liquid carbon as well. Wherever possible, we have validated the performance of GAP-20 against the reference DFT method and shown it to perform well for a number of physical properties across the different phases. Included in these are a number of processes involving bond breaking and formation, some of which have been challenging for cheaper empirical potentials by construction. Tests for transferability, specifically by diagnostic GAP-RSS runs and the study of transformations not included in the training, suggest that GAP-20 could readily be applied to more thorough explorations of the carbon potential energy landscape, for example, in the search for larger fullerenes^{61,115} or in crystal-structure prediction by expanding on Refs. 5 and 109. Further applications may include the more detailed study of non-graphitizing or “hard” carbons,^{116–119} following on from earlier GAP-17 based studies in Refs. 57 and 58.

Despite the many potential applications of GAP-20, the model is not without its shortcomings. While it remains significantly more computationally affordable than direct *ab initio* simulation (in particular for large systems), the cost of its evaluation is much greater than that of empirical potentials (see Fig. 12 of the [supplementary material](#)), and therefore, the latter will still give access to even larger-scale systems.³⁶ We also note that “real” carbon is rarely found in isolation—hydrogenation and oxidation of carbon structures are not considered here. The expansion of the scope of the potential to treat hydrogenated or oxidized structures would complicate the process of training both by requiring a larger training dataset and by requiring the inclusion of a number of interactions not considered here. In addition to long-ranged van der Waals interactions (which are only considered approximately in the current work), the introduction of other elements introduces the associated complexities of substantial charge rearrangements: polar bonds and partial charges. The long-ranged Coulomb interactions, dipole-dipole, and higher order multipole interactions remain a challenge for ML potentials. We note that the combination of pure carbon simulations (using GAP-17) and subsequent density-functional analyses of hydrogenation and oxidation³⁶ or metal intercalation⁵⁸ has proven fruitful, and we expect that further studies of this type will be facilitated by GAP-20, particularly when low-density, dispersion-dominated nanostructures are concerned.

We believe that we have achieved an excellent compromise for our potential in that it accurately models the wide range of structures required to make it broadly applicable. We do not claim perfect accuracy for all properties; however, we accept that fitting to such a wide range of structures will necessarily impact the accuracy in some areas. Notably, a large number of structures that were generated as part of the total training dataset are excluded from the final training. Conversely, many have been included, which might be irrelevant for a researcher’s intended purpose. With this in mind, we have made freely available the total training dataset (structures, energies, forces, and virial coefficients) produced as part of this work. While we do not believe that this will typically be necessary, it is a further virtue of the GAP framework that a potential may be readily retrained to suit a particular purpose simply by modifying the composition of the training configurations used; we believe that it is beneficial to offer the opportunity for users to tune the model to target higher accuracy in a particular region of interest.

In addition to the training dataset, the potential introduced here is provided in the form of an XML file and

has been made freely available along with the GAP code (<http://www.libatoms.org>). The potential has the unique identifier GAP_2020_4_27_60_2_50_5_436 and may be used within the QUIP software package, which can be found at <https://github.com/libAtoms/QUIP>. GAP-20 may be used for simulations directly in LAMMPS, using the QUIP for LAMMPS plugin.¹¹⁰

SUPPLEMENTARY MATERIAL

See the [supplementary material](#) for full details of computed formation energies and phonon dispersion curves for all models, further information on GAP hyper-parameter selection and command line argument used, graphene bilayer separation curves, and force errors for various configurations. More information on the optimization and computational cost of GAP-20 compared to DFT is also given.

ACKNOWLEDGMENTS

A.M. was supported by the European Research Council under the European Union's Seventh Framework Programme (No. FP/2007-2013)/ERC Grant Agreement No. 616121 (HeteroIce project). V.L.D. acknowledges a Leverhulme Early Career Fellowship and support from the Isaac Newton Trust. We are grateful to the UK Materials and Molecular Modeling Hub for computational resources, which is partially funded by the EPSRC (Grant No. EP/P020194/1). We are also grateful for computational support from the UK National High-Performance Computing Service, ARCHER, for which access was obtained via the UKCP consortium and funded by the EPSRC (Grant Ref. No. EP/P022561/1). We are also grateful for the use of the UCL Grace High Performance Computing Facility (Grace@UCL) and associated support services in the completion of this work.

DATA AVAILABILITY

The data that support the findings of this study are openly available at (<http://www.libatoms.org/Home/DataRepository>), in the Cambridge Apollo Data Repository at <https://doi.org/10.17863/CAM.54529> and on request from the authors.

REFERENCES

- 1 K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric field effect in atomically thin carbon films," *Science* **306**(5696), 666–669 (2004).
- 2 H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley, "C₆₀: Buckminsterfullerene," *Nature* **318**, 162–163 (1985).
- 3 M. Martinez-Canales and C. J. Pickard, "Thermodynamically stable phases of carbon at multiterapascal pressures," *Phys. Rev. Lett.* **108**(4), 045704 (2012).
- 4 R. C. Powles, N. A. Marks, and D. W. M. Lau, "Self-assembly of sp²-bonded carbon nanostructures from amorphous precursors," *Phys. Rev. B* **79**(7), 075430 (2009).
- 5 V. L. Deringer, G. Csányi, and D. M. Proserpio, "Extracting crystal chemistry from amorphous carbon structures," *ChemPhysChem* **18**(8), 873–877 (2017).
- 6 D. Tománek, *Guide Through the Nanocarbon Jungle*, 1st ed. (Morgan & Claypool Publishers, San Rafael, 2014).
- 7 R. Mirzayev, K. Mustonen, M. R. A. Monazam, A. Mittelberger, T. J. Pennycook, C. Mangler, T. Susi, J. Kotakoski, and J. C. Meyer, "Buckyball sandwiches," *Sci. Adv.* **3**(6), e1700176 (2017).
- 8 C. Lee, X. Wei, J. W. Kysar, and J. Hone, "Measurement of the elastic properties and intrinsic strength of monolayer graphene," *Science* **321**(5887), 385–388 (2008).
- 9 M. H. Nazare and A. J. Neves, *Properties Growth and Applications of Diamond*, 1st ed. (INSPEC, London, 2001).
- 10 A. K. Geim, "Graphene: Status and prospects," *Science* **324**(5934), 1530–1534 (2009).
- 11 P. Avouris, Z. Chen, and V. Perebeinos, "Carbon-based electronics," *Nat. Nanotechnol.* **2**(10), 605–615 (2007).
- 12 Y. Yan, J. Miao, Z. Yang, F.-X. Xiao, H. B. Yang, B. Liu, and Y. Yang, "Carbon nanotube catalysts: Recent advances in synthesis, characterization and applications," *Chem. Soc. Rev.* **44**(10), 3295–3346 (2015).
- 13 W. Zhang, S. Zhu, R. Luque, S. Han, L. Hu, and G. Xu, "Recent development of carbon electrode materials and their bioanalytical and environmental applications," *Chem. Soc. Rev.* **45**(3), 715–752 (2016).
- 14 Y.-n. Zhang, Q. Niu, X. Gu, N. Yang, and G. Zhao, "Recent progress of carbon nanomaterials for electrochemical detection and removal of environmental pollutants," *Nanoscale* **11**, 11992 (2019).
- 15 Z. Yang, J. Ren, Z. Zhang, X. Chen, G. Guan, L. Qiu, Y. Zhang, and H. Peng, "Recent advancement of nanostructured carbon for energy applications," *Chem. Rev.* **115**(11), 5159–5223 (2015).
- 16 W. Yuan, Y. Zhang, L. Cheng, H. Wu, L. Zheng, and D. Zhao, "The applications of carbon nanotubes and graphene in advanced rechargeable lithium batteries," *J. Mater. Chem. A* **4**(23), 8932–8951 (2016).
- 17 A. S. Aricò, P. Bruce, B. Scrosati, J.-M. Tarascon, and W. Van Schalkwijk, "Nanostructured materials for advanced energy conversion and storage devices," *Nat. Mater.* **4**(5), 366–377 (2005).
- 18 J. N. Kholmanov, C. W. Magnuson, R. Piner, J. Y. Kim, A. E. Aliev, C. Tan, T. Y. Kim, A. A. Zakhidov, G. Sberveglieri, R. H. Baughman, and R. S. Ruoff, "Optical, electrical, and electromechanical properties of hybrid graphene/carbon nanotube films," *Adv. Mater.* **27**(19), 3053–3059 (2015).
- 19 F. Bonaccorso, Z. Sun, T. Hasan, and A. C. Ferrari, "Graphene photonics and optoelectronics," *Nat. Photonics* **4**(9), 611–622 (2010).
- 20 A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, "The electronic properties of graphene," *Rev. Mod. Phys.* **81**(1), 109–162 (2009).
- 21 C. Farrera, F. Torres Andón, and N. Feliu, "Carbon nanotubes as optical sensors in biomedicine," *ACS Nano* **11**(11), 10637–10643 (2017).
- 22 A. Courvoisier, M. J. Booth, and S. Salter, "Inscription of 3D waveguides in diamond using an ultrafast laser," *Appl. Phys. Lett.* **109**(3), 031109 (2016).
- 23 P. Latawiec, V. Venkataraman, M. J. Burek, B. J. M. Hausmann, I. Bulu, and M. Lončar, "On-chip diamond Raman laser," *Optica* **2**(11), 924 (2015).
- 24 L. Pastewka, S. Moser, P. Gumbsch, and M. Moseler, "Anisotropic mechanical amorphization drives wear in diamond," *Nat. Mater.* **10**(1), 34–38 (2011).
- 25 T. B. Shiehl, D. G. McCulloch, D. R. McKenzie, M. R. Field, B. Haberl, R. Boehler, B. A. Cook, C. De Tomas, I. Suarez-Martinez, N. A. Marks, and J. E. Bradby, "Graphitization of glassy carbon after compression at room temperature," *Phys. Rev. Lett.* **120**(21), 215701 (2018).
- 26 J. Tersoff, "Empirical interatomic potential for carbon, with applications to amorphous carbon," *Phys. Rev. Lett.* **61**(25), 2879–2882 (1988).
- 27 D. W. Brenner, "Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films," *Phys. Rev. B* **42**(15), 9458–9471 (1990).
- 28 T. C. O'Connor, J. Andzelm, and M. O. Robbins, "AIREBO-M: A reactive model for hydrocarbons at extreme pressures," *J. Chem. Phys.* **142**(2), 024903 (2015).
- 29 S. J. Stuart, A. B. Tutein, and J. A. Harrison, "A reactive potential for hydrocarbons with intermolecular interactions," *J. Chem. Phys.* **112**, 6472–6486 (2000).
- 30 J. H. Los and A. Fasolino, "Intrinsic long-range bond-order potential for carbon: Performance in Monte Carlo simulations of graphitization," *Phys. Rev. B* **68**(2), 024107 (2003).

- ³¹L. M. Ghiringhelli, J. H. Los, A. Fasolino, and E. J. Meijer, "Improved long-range reactive bond-order potential for carbon. II. Molecular simulation of liquid carbon," *Phys. Rev. B* **72**(21), 214103 (2005).
- ³²J. H. Los, L. M. Ghiringhelli, E. J. Meijer, and A. Fasolino, "Improved long-range reactive bond-order potential for carbon. I. Construction," *Phys. Rev. B* **72**, 214102 (2005).
- ³³N. A. Marks, "Generalizing the environment-dependent interaction potential for carbon," *Phys. Rev. B* **63**, 035401 (2006).
- ³⁴L. Pastewka, P. Pou, R. Pérez, P. Gumbsch, and M. Moseler, "Describing bond-breaking processes by reactive potentials: Importance of an environment-dependent interaction range," *Phys. Rev. B* **78**(16), 161402 (2008).
- ³⁵S. G. Srinivasan, A. C. T. Van Duin, and P. Ganesh, "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**(4), 571–580 (2015).
- ³⁶C. de Tomas, A. Aghajamali, J. L. Jones, D. J. Lim, M. J. López, I. Suarez-Martinez, and N. A. Marks, "Transferability in interatomic potentials for carbon," *Carbon* **155**, 624–634 (2019).
- ³⁷C. de Tomas, I. Suarez-Martinez, and N. A. Marks, "Graphitization of amorphous carbons: A comparative study of interatomic potentials," *Carbon* **109**, 681–693 (2016).
- ³⁸J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.* **98**(14), 146401 (2007).
- ³⁹A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Phys. Rev. Lett.* **104**(13), 136403 (2010).
- ⁴⁰K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," *J. Chem. Phys.* **148**, 241722 (2018).
- ⁴¹T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, "A universal strategy for the creation of machine learning-based atomistic force fields," *npj Comput. Mater.* **3**(1), 37 (2017).
- ⁴²A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," *Sci. Adv.* **3**, e1701816 (2017).
- ⁴³B. Kolb, X. Luo, X. Zhou, B. Jiang, and H. Guo, "High-dimensional atomistic neural network potentials for molecule-surface interactions: HCl scattering from Au(111)," *J. Phys. Chem. Lett.* **8**, 666–672 (2017).
- ⁴⁴P. E. Dolgirev, I. A. Kruglov, and A. R. Oganov, "Machine learning scheme for fast extraction of chemically interpretable interatomic potentials," *AIP Adv.* **6**, 085318 (2016).
- ⁴⁵A. Glielmo, Sollich, and A. De Vita, "Accurate interatomic force fields via machine learning with covariant kernels," *Phys. Rev. B* **95**(21), 214302 (2017).
- ⁴⁶V. Kůrková, "Kolmogorov's theorem and multilayer neural networks," *Neural Networks* **5**(3), 501–506 (1992).
- ⁴⁷C. M. Handley and J. Behler, "Next generation interatomic potentials for condensed systems," *Eur. Phys. J. B* **87**, 152 (2014).
- ⁴⁸J. Behler, "Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations," *Phys. Chem. Chem. Phys.* **13**(40), 17930–17955 (2011).
- ⁴⁹J. Behler, "Perspective: Machine learning potentials for atomistic simulations," *J. Chem. Phys.* **145**(17), 170901 (2016).
- ⁵⁰V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," *Adv. Mater.* **31**(46), 1902765 (2019).
- ⁵¹R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, "Graphite-diamond phase coexistence study employing a neural-network mapping of the *ab initio* potential energy surface," *Phys. Rev. B* **81**(10), 100103 (2010).
- ⁵²R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, "Nucleation mechanism for the direct graphite-to-diamond phase transition," *Nat. Mater.* **10**(9), 693–697 (2011).
- ⁵³P. Rowe, G. Csányi, D. Alfè, and A. Michaelides, "Development of a machine learning potential for graphene," *Phys. Rev. B* **97**, 054303 (2018).
- ⁵⁴M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila, and G. Csányi, "Growth mechanism and origin of high sp^3 content in tetrahedral amorphous carbon," *Phys. Rev. Lett.* **120**(16), 166101 (2018).
- ⁵⁵V. L. Deringer, M. A. Caro, R. Jana, A. Aarva, S. R. Elliott, T. Laurila, G. Csányi, and L. Pastewka, "Computational surface chemistry of tetrahedral amorphous carbon by combining machine learning and density functional theory," *Chem. Mater.* **30**(21), 7438–7445 (2018).
- ⁵⁶M. A. Caro, A. Aarva, V. L. Deringer, G. Csányi, and T. Laurila, "Reactivity of amorphous carbon surfaces: Rationalizing the role of structural motifs in functionalization using machine learning," *Chem. Mater.* **30**(21), 7446–7455 (2018).
- ⁵⁷V. L. Deringer, C. Merlet, Y. Hu, T. H. Lee, J. A. Kattirtzi, O. Pecher, G. Csányi, S. R. Elliott, and C. Grey, "Towards an atomistic understanding of disordered carbon electrode materials," *Chem. Commun.* **54**(47), 5988–5991 (2018).
- ⁵⁸J.-X. Huang, G. Csányi, J.-B. Zhao, J. Cheng, and V. L. Deringer, "First-principles study of alkali-metal intercalation in disordered carbon anode materials," *J. Mater. Chem. A* **7**(32), 19070–19080 (2019).
- ⁵⁹A. P. Bartók, J. Kermode, and N. Bernstein, "Machine learning a general-purpose interatomic potential for silicon," *Phys. Rev. X* **8**(4), 041048 (2018).
- ⁶⁰C. J. Pickard and R. J. Needs, "*Ab initio* random structure searching," *J. Phys.: Condens. Matter* **23**, 053201 (2011).
- ⁶¹D. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*, 1st ed. (Cambridge University Press, Cambridge, 2004).
- ⁶²V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon," *Phys. Rev. B* **95**(9), 094203 (2017).
- ⁶³A. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B* **87**(18), 184115 (2013).
- ⁶⁴J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *J. Chem. Phys.* **134**(7), 074106 (2011).
- ⁶⁵P. Gasparotto, R. H. Meißner, and M. Ceriotti, "Recognizing local and global structural motifs at the atomic scale," *J. Chem. Theory Comput.* **14**, 486–498 (2018).
- ⁶⁶D. J. Wales, "Closed-shell structures and the building game," *Chem. Phys. Lett.* **141**(6), 478–484 (1987).
- ⁶⁷K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, "Higher-accuracy van der Waals density functional," *Phys. Rev. B* **82**(8), 081101 (2010).
- ⁶⁸J. Klimeš, D. R. Bowler, and A. Michaelides, "van der Waals density functionals applied to solids," *Phys. Rev. B* **83**(19), 195131 (2011).
- ⁶⁹M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, "van der Waals density functional for general geometries," *Phys. Rev. Lett.* **92**(24), 246401 (2004).
- ⁷⁰J. Klimeš, D. R. Bowler, and A. Michaelides, "Chemical accuracy for the van der Waals density functional," *J. Phys.: Condens. Matter* **22**(2), 022201 (2010).
- ⁷¹G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," *Phys. Rev. B* **54**(16), 11169–11186 (1996).
- ⁷²G. Kresse and J. Furthmüller, "Efficiency of *ab initio* total energy calculations for metals and semiconductors using a plane-wave basis set," *Comput. Mater. Sci.* **6**(1), 15–50 (1996).
- ⁷³G. Kresse, "From ultrasoft pseudopotentials to the projector augmented-wave method," *Phys. Rev. B* **59**(3), 1758–1775 (1999).
- ⁷⁴H. J. Monkhorst and J. D. Pack, "Special points for Brillouin zone integrations," *Phys. Rev. B* **13**(12), 5188–5192 (1976).
- ⁷⁵G. Graziano, J. Klimeš, F. Fernandez-Alonso, and A. Michaelides, "Improved description of soft layered materials with van der Waals density functional theory," *J. Phys.: Condens. Matter* **24**(42), 424216 (2012).
- ⁷⁶R. Hoffmann, A. A. Kabanov, A. A. Golov, and D. M. Proserpio, "Homo citans and carbon allotropes: For an ethics of citation," *Angew. Chem., Int. Ed.* **55**(37), 10962–10976 (2016).
- ⁷⁷L. Li, S. Reich, and J. Robertson, "Defect energies of graphite: Density-functional calculations," *Phys. Rev. B* **72**(18), 184109 (2005).
- ⁷⁸T. Xu and L. Sun, "Structural defects in graphene," in *Defects in Advanced Electronic Materials and Novel Low Dimensional Structures* (Woodhead Publishing, 2018), Vol. 5, No. 1, pp. 137–160.

- ⁷⁹J.-C. Charlier, “Defects in carbon nanotubes,” *Acc. Chem. Res.* **35**(12), 1063–1069 (2002).
- ⁸⁰S. T. Skowron, I. V. Lebedeva, A. M. Popov, and E. Bichoutskaia, “Energetics of atomic scale structure changes in graphene,” *Chem. Soc. Rev.* **44**(44), 3143–3176 (2015).
- ⁸¹J. Ristein, “Diamond surfaces: Familiar and amazing,” *Appl. Phys. A* **82**, 377–384 (2006).
- ⁸²G. Kern and J. Hafner, “*Ab initio* molecular-dynamics studies of the graphitization of flat and stepped diamond (111) surfaces,” *Phys. Rev. B* **58**, 13167 (1998).
- ⁸³N. Ooi, A. Rairkar, and J. B. Adams, “Density functional study of graphite bulk and surface properties,” *Carbon* **44**(2), 231–242 (2006).
- ⁸⁴G. Kern and J. Hafner, “*Ab initio* calculations of the atomic and electronic structure of diamond (111) surfaces with steps,” *Phys. Rev. B* **58**(4), 2161–2169 (1998).
- ⁸⁵V. L. Deringer, C. J. Pickard, and G. Csányi, “Data-driven learning of total and local energies in elemental boron,” *Phys. Rev. Lett.* **120**(15), 156001 (2018).
- ⁸⁶S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, “Comparing molecules and solids across structural and alchemical space,” *Phys. Chem. Chem. Phys.* **18**(20), 13754 (2016).
- ⁸⁷M. Ceriotti, G. A. Tribello, and M. Parrinello, “Simplifying the representation of complex free-energy landscapes using sketch-map,” *Proc. Natl. Acad. Sci. U. S. A.* **108**(32), 13023–13028 (2011).
- ⁸⁸A. P. Bartók, M. J. Gillan, F. R. Manby, and G. Csányi, “Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water,” *Phys. Rev. B* **88**(5), 054104 (2013).
- ⁸⁹D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, “Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron,” *Phys. Rev. Mater.* **2**, 013808 (2017).
- ⁹⁰S. Fujikake, V. L. Deringer, T. H. Lee, M. Krynski, S. R. Elliott, and G. Csányi, “Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures,” *J. Chem. Phys.* **148**, 241714 (2018).
- ⁹¹W. J. Szlachta, A. P. Bartók, and G. Csányi, “Accuracy and transferability of Gaussian approximation potential models for tungsten,” *Phys. Rev. B* **90**(10), 104018 (2014).
- ⁹²M. J. Willatt, F. Musil, and M. Ceriotti, “Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements,” *Phys. Chem. Chem. Phys.* **20**(47), 29661–29668 (2018).
- ⁹³F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld, “Alchemical and structural distribution based representation for universal quantum machine learning,” *J. Chem. Phys.* **148**(24), 241717 (2018).
- ⁹⁴A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole Von Lilienfeld, “FCHL revisited: Faster and more accurate quantum machine learning,” *J. Chem. Phys.* **152**(4), 044107 (2020).
- ⁹⁵A. Togo and I. Tanaka, “First principles phonon calculations in materials science,” *Scr. Mater.* **108**, 1–5 (2015).
- ⁹⁶G. Kern and J. Hafner, “*Ab initio* calculations of the atomic and electronic structure of clean and hydrogenated diamond (110) surfaces,” *Phys. Rev. B* **56**(7), 4203–4210 (1997).
- ⁹⁷S. Thinius, M. M. Islam, and T. Bredow, “Reconstruction of low-index graphite surfaces,” *Surf. Sci.* **649**(1), 60–65 (2016).
- ⁹⁸M. I. J. Probert and M. C. Payne, “Improving the convergence of defect calculations in supercells: An *ab initio* study of the neutral silicon vacancy,” *Phys. Rev. B* **67**(7), 075204 (2003).
- ⁹⁹D. C. Hunt, D. J. Twitchen, M. E. Newton, J. M. Baker, and T. R. Anthony, “Identification of the neutral carbon (100)-split interstitial in diamond,” *Phys. Rev. B* **61**(6), 3863–3876 (2000).
- ¹⁰⁰A. J. Stone and D. J. Wales, “Theoretical studies of icosahedral C₆₀ and some related species,” *Chem. Phys. Lett.* **128**(5), 501–503 (1986).
- ¹⁰¹J. Kotakoski, J. C. Meyer, S. Kurasch, D. Santos-Cottin, U. Kaiser, and A. V. Krasheninnikov, “Stone-Wales-type transformations in carbon nanostructures driven by electron irradiation,” *Phys. Rev. B* **83**(24), 245420 (2011).
- ¹⁰²J. Ma, D. Alfè, A. Michaelides, and E. Wang, “Stone-Wales defects in graphene and other planar sp²-bonded materials,” *Phys. Rev. B* **80**(4), 033407 (2009).
- ¹⁰³A. A. Correa, S. A. Bonev, and G. Galli, “Carbon under extreme conditions: Phase boundaries and electronic properties from first-principles theory,” *Proc. Natl. Acad. Sci. U. S. A.* **103**(5), 1204–1208 (2006).
- ¹⁰⁴M. Pozzo, C. Davies, D. Gubbins, and D. Alfè, “Transport properties for liquid silicon-oxygen-iron mixtures at Earth’s core conditions,” *Phys. Rev. B* **87**(1), 014110 (2013).
- ¹⁰⁵H. M. Strong and R. E. Hanneman, “Crystallization of diamond and graphite,” *J. Chem. Phys.* **46**(9), 3668–3676 (1967).
- ¹⁰⁶A. Sorkin, J. Adler, and R. Kalish, “Nucleation of diamond from liquid carbon under extreme pressures: Atomistic simulation,” *Phys. Rev. B* **74**(6), 064115 (2006).
- ¹⁰⁷W. H. Gust, “Phase transition and shock-compression parameters to 120 GPa for 3 types of graphite,” *Phys. Rev. B* **22**(10), 4744–4756 (1980).
- ¹⁰⁸C. J. Mundy, A. Curioni, N. Goldman, I. F. Kuo, E. J. Reed, L. E. Fried, and M. Ianuzzi, “Ultrafast transformation of graphite to diamond: An *ab initio* study of graphite under shock compression,” *J. Chem. Phys.* **128**(18), 184701 (2008).
- ¹⁰⁹V. L. Deringer, D. M. Proserpio, G. Csányi, and C. J. Pickard, “Data-driven learning and prediction of inorganic crystal structures,” *Faraday Discuss.* **211**, 45–59 (2018).
- ¹¹⁰S. Plimpton, “Fast parallel algorithms for short-range molecular dynamics,” *J. Comput. Phys.* **117**, 1–19 (1995).
- ¹¹¹E. Bitzek, Koskinen, F. Gähler, M. Moseler, and Gumbusch, “Structural relaxation made simple,” *Phys. Rev. Lett.* **97**(17), 170201 (2006).
- ¹¹²C. J. Pickard, “Predicted pressure-induced s-band ferromagnetism in alkali metals,” *Phys. Rev. Lett.* **107**(8), 087201 (2011).
- ¹¹³R. J. Needs and C. J. Pickard, “Perspective: Role of structure prediction in materials discovery and design,” *APL Mater.* **4**(5), 053210 (2016).
- ¹¹⁴Y. Kumeda and D. J. Wales, “*Ab initio* study of rearrangements between C₆₀ fullerenes,” *Chem. Phys. Lett.* **374**(1–2), 125–131 (2003).
- ¹¹⁵D. J. Wales, “Global optimization of clusters, crystals, and biomolecules,” *Science* **285**(5432), 1368–1372 (1999).
- ¹¹⁶P. J. F. Harris and S. C. Tsang, “High-resolution electron microscopy studies of non-graphitizing carbons,” *Philos. Mag. A* **76**(3), 667–677 (1997).
- ¹¹⁷P. J. F. Harris, “New perspectives on the structure of graphitic carbons,” *Crit. Rev. Solid State Mater. Sci.* **30**(4), 235–253 (2005).
- ¹¹⁸P. J. F. Harris, “Structure of non-graphitising carbons,” *Int. Mater. Rev.* **42**(5), 206–218 (2014).
- ¹¹⁹X. Dou, I. Hasa, D. Saurel, C. Vaalma, L. Wu, D. Buchholz, D. Bresser, S. Komaba, and S. Passerini, “Hard carbons for sodium-ion batteries: Structure, analysis, sustainability, and electrochemistry,” *Mater. Today* **23**, 87–104 (2019).