Accuracy and Transferability in Machine Learned Potentials for Carbon

Patrick Rowe

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Physics and Astronomy

University College London

June 21, 2021

I, Patrick Rowe, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

In this thesis, we discuss the approach taken to construct an accurate machine learning (ML) model for atomistic simulations of carbon, constructed using the Gaussian approximation potential (GAP) methodology. We begin by discussing the process for constructing a potential for a single phase, graphene. We then extend this to produce a general-purpose potential, named GAP-20, which describes the properties of the bulk crystalline and amorphous phases, crystal surfaces, and defect structures with a high degree of accuracy. 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 with the high accuracy necessary for crystalline graphene which we introduce in this thesis, thereby providing an interatomic potential that will be applicable to a wide range of applications concerning diverse forms of bulk and nanostructured carbon.

Impact Statement

The impact of the research presented in this thesis can be split into two parts. Firstly, there are the direct impacts and potential applications of the interatomic potentials produced as a part of this research. Secondly, there are the broader impacts of the scientific discoveries and developments, particularly in the application of machine learning to the development of interatomic potentials, which have resulted from the work presented here.

In terms of the direct impacts of the potentials generated, the models presented in this work have been made publicly and freely available for other researchers to download and use. Potentials for carbon which have been published similarly, including the Tersoff, REBO, AIREBO and LCBOP potentials, have gone on to facilitate research in a large number of publications. One of the primary aims of this research is that the GAP-20 carbon model goes on to be used similarly widely. Computational research into carbon physics and chemistry touches the development of batteries, fuel cells and structural and composite materials among a myriad of other applications – an interatomic potential which is employed to study these fields would have wide ranging scientific and societal impact.

While the direct connection between the research presented here and the future study of carbonaceous systems is a more obvious one, the broader impact of the methodological advances presented here for the construction of machine learning potentials is also worth noting. The asymmetric selection of smooth overlap of atomic positions (SOAP) descriptors, inclusion of long-range corrections to account for van der Waals attractions and other methodological contributions will have broader significance, as these are transferable to the application of machine learning to a wide range of other atomistic systems.

Acknowledgements

First among those deserving thanks is my supervisor, Prof. Angelos Michaelides, who has provided tireless support, guidance, advice and knowledge without which this thesis would not have been possible. The friendly and supportive environment which he has fostered within the ICE group has been invaluable throughout the four years over which this work was conducted, but none more so than in the midst of the COVID-19 pandemic, during which time this thesis was largely written.

I would also like to particularly thank Prof. Gábor Csányi, for his advice and guidance on constructing Gaussian approximation potentials. I have worked with numerous other scientific collaborators and coauthors during my PhD, all of whom deserve thanks. Prof. Alesandro Troisi, from whom I learned so much of the fundamentals of computational physics and chemistry. Prof. Dario Alfè for co-supervising me during my time at UCL. Prof. Volker Deringer and Dr. Piero Gasparotto for their help with GAP-20. Each and every member of the ICE group, in particular (and in no particular order), Fabian Thiemann, Christoph Schran, Michael Davies, Martin Fitzner, Wei Fang, Andrea Zen and Christopher Penschke. My other collaborators and coauthors, Prof. Rahul Nair, Dr. George Knee, Dr. Rocco Peter Fornari and Prof. Animesh Datta.

I am forever grateful to Edward Lamb and Josh Michel, for their enthusiasm, friendship and support. Finally, I would like to thank my family, without whom I would not be where I am today and Elise, for being the person closest to me, even when we are on opposite sides of the Earth.

List of Publications

The following publications have been published as a result of work conducted wholly or partially during the course of this PhD project (2016-2020).

Publications Included in this Thesis.

- P. Rowe, V. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, "*An accurate and transferable machine learning potential for carbon*", J. Chem. Phys., **153**, 034702, (2020)

- P. Rowe, G. Csányi, D. Alfè, A. Michaelides, "*Development of a machine learning potential for graphene*", Phys. Rev. B., **97**, 5, 054303, (2018)

Publications as a result of work conducted during this PhD but not included in the present thesis.

- F. L. Thiemann, P. Rowe, E. A. Müller, A. Michaelides, "*Machine Learning Potential for Hexagonal Boron Nitride Applied to Thermally and Mechanically Induced Rippling*", J. Chem. Phys. C., **124**, 22278, (2020)

- K. Huang, P. Rowe, C. Chi, V. Sreepal, T. Bohn, K.-G. Zhou, Y. Su, E. Prestat, P. Balakrishna Pillai, C. T. Cherian, A. Michaelides, R. R. Nair, "*Cation-controlled wetting properties of vermiculite membranes and its promise for fouling resistant oil–water separation*", Nat. Comm., **11**, 1097, (2020)

- G. Knee, P. Rowe, L. D. Smith, A. Troisi, A. Datta, "*Structure-dynamics relation in physically-plausible multi-chromophore systems*", J. Phys. Chem. Lett., **10**, 2328, (2017)

Contents

List of Figures

List of Tables

Chapter 1

Introductory Material

1.1 The Element Carbon

To many, the word 'carbon' may elicit a sense of the mundane, perhaps evoking mental images of graphite pencils or amorphous lumps of coal. At its most exciting, one might think of diamonds - to some beautiful, certainly a symbol of wealth, but surely not the subject of modern research? After all, carbon is one of the few elements known of since antiquity; surely science must have uncovered all of the secrets that this one element has to offer by now?

This same sentiment was echoed in 1984, by one Prof. Richard Smalley. When approached by a more junior researcher, asking to use his equipment for the study of a potential new allotrope for carbon, Smalley wrote 'to the modern chemist a continuing study of pure carbon would seem to offer little hope for excitement' [1]. Yet it seems that Smalley could not have been more wrong. One year later, that researcher, Harold Kroto, together with Smalley would go on to publish their identification of the C_{60} molecule, a cage-like arrangement of tesselated pentagonal and hexagonal carbon rings [2]. The molecule was famously dubbed Buckminsterfullerene for its similarity to architect Buckminster Fuller's geodesic domes. In 1996, they would be awarded the Nobel Prize for Chemistry for their work, the first in what would be a series of discoveries by numerous researchers of novel carbon phases [3]. The identification of Buckminsterfullerenes would trigger an explosion of research into their physical and chemical properties [4]. For example, their ready participation

in redox chemistry has led to their incorporation in dye-sensitised solar cells and battery applications. One synthesis which is particularly of note is that of endohedral fullerene structures. Atoms, or even small molecules may be 'trapped' within the fullerene cage, isolating them in a relatively inert chemical environment. The first of these, La@C_{60} was synthesised as early as 1985. Rather than being merely a chemical curiosity, this and similar species have since been used as biocompatible contrast enhancing agents for medical imaging, in particular nuclear magnetic resonance imaging (MRI).

The discovery of carbon nanotubes in 1991 spurred a similar surge in research [5, 6]. Where Buckminsterfullerenes may be described as spherical, carbon nanotubes are elongated along one axis to form tubes, which may be up to several micrometers in length. Depending on how the tube is rolled, the radius of a carbon nanotube may be as small as a few Angstroms, or as large as several nanometres, and may have insulating, semiconducting or metallic electronic character. Similarly to fullerenes, nanotubes are hollow within, this void may be filled with numerous species (depending on the radius of the nanotube), including smaller nanotubes or fullerenes. This has led to their application as ideal one dimensional reaction vessels, highly selective pores for water filtration and as probes for elecrochemical microscopy. In addition to their uses in research focused on fundamental chemistry and physics, the low density, combined with the high tensile strength of carbon nanotubes, has led to their incorporation into a number of structural composite materials [7, 8].

In 2010, the Nobel Prize for Physics was awarded to Andre Geim and Constantin Novoselov for the isolation of yet another hitherto unknown allotrope of carbon: graphene [9].[1] If one needs any more convincing that carbon is recognised as a unique and remarkable material in the scientific community, the awarding of two Nobel prizes for the discovery of two allotropes of a single element is an honour not (yet) bestowed on any other member of the periodic table. The significance of

[1] Regarding the mundanity of graphite pencils, it is worth noting that the first 'synthesis' of graphene involved rubbing a pencil onto Scotch tape to deposit thin layers of graphite, and then repeatedly separating the graphite layers using the adhesive tape until only one remained

the discovery of graphene was that it was not only the first time this particular form of carbon had been isolated, but also the first example of a stable two-dimensional crystal (graphene being only one atom thick), which until then had been dismissed as 'impossible'. Thus, this discovery was not just of importance for carbon science, but for materials science and fundamental physics alike. Aside from graphene's unusual topology making it the thinnest material ever made, it was also found to be the strongest, most thermally conductive and with an electrical conductivity comparable to that of copper [10]. It is the fundamental building block of all sp^2 hybridised carbon allotropes; graphene may be rolled to form nanotubes or fullerenes, or stacked to form graphite [11]. These similarities are not merely topological, but also extend to the physical properties of the materials; graphene, graphite and carbon nanotubes share many electronic and vibrational properties for this reason [12, 13, 14]. As a result of its unique mechanical, electronic and structural properties, graphene has been the subject of extensive investigation since it was first isolated [15, 16, 11]. These, combined with its characteristic 2D nature, have resulted in graphene becoming the 'poster child' for materials design in nano-electronic, mechanical and optical research [17, 18].[2]

So what of the more mundane manifestations of carbon, soot, graphite and diamond which were mentioned earlier? Each of these is unique in its own right. Soot, if studied on an atomic scale, is comprised in small part of fullerene-like, and nanotube-like structures; often these are layered to form multi-walled carbon nanotubes (MWCNT) or so-called 'carbon onions' (the fullerene equivalent of a MWCNT) [4]. Kroto himself wrote 'The fact that this new third form of carbon ... has been under our noses since time immemorial is almost unbelievable.' The layered graphitic structures found in pencil leads contain voids between them. These voids may be filled with ions, in particular Li^+, to create the electrodes of batteries. In an age of green energy, electric cars and mobile phones, the cultural, scientific and environmental importance of lithium-ion batteries cannot be overstated. Diamond meanwhile is unique among the allotropes discussed so far, in that it is com-

[2]However, it must be noted that an oft-touted joke about graphene is that it is 'capable of absolutely everything, except finding its way out of a laboratory'

prised of purely sp^3 hybridised carbon atoms. This gives diamond a much more conventional three-dimensional crystal structure than the other, predominantly sp^2 hybridised, members of the carbon allotrope family, in which atoms are arranged with a locally tetrahedral geometry. This geometry, combined with the strength and rigidity of the C−C bond, makes diamond the hardest material known to science. In turn, this has made diamond responsible for facilitating an entire field of high-pressure research, as it is used in the cores of diamond anvil cells, a role which no other element is better suited for.

Given the technological and scientific importance of carbon, it is no surprise that it has also attracted the interest of computational chemists and physicists. In fact, atomistic simulations (using both empirical models and approaches employing quantum mechanics, which will be discussed in more detail later) have played a major role in developing our understanding of carbon materials. Among myriad other scientific problems that have been addressed with carbon potentials, the wear process of diamond [19] or the compression behaviour of glassy carbon are both of note [20]. However, the same characteristics which 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 [10, 2, 21, 22, 23, 24, 25]. 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 [16, 26]. To capture all of these structures and their associated properties with the desired accuracy is challenging for traditional empirical models. As a result, modern empirical potentials often provide disparate results, with conflicting predictions made for fundamental properties such as the coefficient of thermal expansion (CTE) of graphene and graphite, even the sign of which is not reliably predicted [27, 28, 29, 30]. Such issues are particularly relevant when one departs from the idealised structures (diamond, graphite, *etc.*), as shown in two detailed benchmark studies by de Tomas *et al.* [31, 32].

There are a great number of interesting phenomena associated with carbon,

such as the phonon assisted diffusion of small molecules and graphene flakes on the graphene surface [33, 34], the study of thermal transport [35, 36, 37] and the incorporation of nuclear quantum effects into simulations which can only be studied by improving upon the accuracy of existing carbon models [38, 39]. In general, when the accuracy of the available empirical models is insufficient for a given investigation, the researcher will turn to *ab initio* molecular dynamics (AIMD). Within the AIMD framework, the forces required for evolving the dynamics of a system are obtained directly from a reference electronic structure method, typically density functional theory (DFT). The trade-off for this improvement in accuracy is the increased cost of AIMD. An empirical model, approximating the interactions between a system of atoms as a set of 'balls and springs' may be routinely evaluated for many hundreds of thousands, even millions of atoms, for nanoseconds at a time. AIMD simulations, however, must find a solution (if approximately) to the Schrödinger equation at each step, and as such are typically limited to modelling a few hundred atoms for timescales on the order of hundreds of picoseconds. This presents a hurdle to modelling carbon, as many of the systems of interest are microstructured, with hundreds or thousands of atoms required to model their structure. Machine learning (ML) approaches, which will be discussed in detail later, have recently emerged as a promising approach for bridging the gap between empirical and *ab initio* approaches. By fitting an ML model to data generated using a particular *ab initio* reference method, it is possible to perform simulations with an accuracy approaching that of AIMD, but with a cost which is many orders of magnitude lower.

1.2 Empirical Models for Carbon

Empirical and bond-order potentials have long provided an indispensable tool in facilitating molecular dynamics (MD) studies of carbonaceous materials. The first many-body potential for carbon was published in 1988 by Tersoff, which was parameterised to reproduce the experimentally determined cohesive energies of various carbon allotropes, as well as the lattice parameter and bulk modulus of diamond [40]. This potential gained rapid acceptance as research into amorphous and

other allotropes of carbon (nanotubes and fullerenes) grew [40, 41]. the Reactive Empirical Bond-Order potential (REBO) is the result of the modification and reparameterisation of the Tersoff potential,w with a fit to a broad range of molecular atomisation energies, bond lengths and reaction barriers [42]. REBO made possible the treatment of hydrocarbons and significantly improved the description of the pure carbon allotropes. While the REBO potential represented a substantial improvement over the Tersoff potential, neither of these accounted for the effects of dispersion interactions and were inherently short ranged in nature. The Adaptive Intermolecular Reactive Empirical Bond Order Potential (AIREBO) [43] potential aimed to correct this, by explicitly incorporating long-range interactions into the functional form through the use of switching functions, thereby maintaining effectively the same short-range potential as its predecessor, REBO. Parameters for the non-bonded interactions of the AIREBO potential were chosen to reproduce the experimentally determined properties of graphite. The description of the bonding behaviour of this potential was further improved upon in AIREBO-Morse (AIREBO-M) by the incorporation of a Morse pair potential (replacing the Lennard-Jones term in the original) to improve the description of anharmonicity in the bonding terms [43, 44]. A fully reparametrized bond-order potential was produced by Los and Fasolino in the form of the Long Range Corrected Bond-Order Potential (LCBOP), wherein the short range potential was fitted to a dataset comprising both experimental values and DFT results computed using the local density approximation (LDA) functional [45]. The LCBOP potential was further updated (to produce LCBOPII) in 2005 to include terms to improve the description of bond dissociation in the solid and liquid phases [46, 47]. The environment dependant interatomic potential (EDIP) for carbon, introduced in 2006 is notable in particular as it borrowed the core of its functional form an improved a REBO-style potential for silicon, and employed properties calculated from *ab initio* simulation in its parameterisation [48]. Another notable advancement in the construction of empirical models for carbon is the introduction of a dynamic cut-off to bond-order potentials [49].

In addition to these developments in traditionally constructed forcefields, a

number of different approaches have emerged which show promise as computational tools. The ReaxFF class of potentials do not represent an iterative improvement upon any of the previously discussed empirical carbon potentials, instead adopting an approach centered around the description of bond dissociation and reactivity [50]. The potential constructs the bond order from the interatomic distance, from which is derived the bond energy. Also included in the functional form are terms to account for van der Waals, Coulombic, and over- and under-coordination energies, the terms of which are fitted to quantities such as atomic charges, bond, angle and torsional energies and heats of formation [50, 51]. Density Functional Tight Binding (DFTB) represents yet another approach, it is not an interatomic potential in the traditional sense, rather an electronic method which operates on a tightly constrained set of parameterised wavefunctions. DFTB is based on a second order expansion of the DFT total energy into a distance dependent electronic Hamiltonian and two-body repulsive classical term. The diagonal elements of the Hamiltonian matrix correspond to the atomic (s, p, d) eigenenergies, while the distance dependent off-diagonal elements of the Hamiltonian - the bond energies - are parameterised to DFT and evaluated by interpolation [52, 53]. DFTB is advantageous as a methodology over DFT due to its much lower computational cost. Although both methods scale as $\mathcal{O}(N^3)$, DFTB is often two to three orders of magnitude cheaper than DFT due to DFTB's much lower prefactor [54].

1.3 Machine Learning in Materials and Molecular Simulation

Machine learning (ML) is a term which broadly applies to a set of algorithms that are designed to improve when provided with new data. In general, a mathematical model is constructed, the parameters, weights or bases of which are refined as new data are presented to the model - a process termed 'training'. Typically, ML is most usefully applied in cases where a human programmer would find it challenging or impossible to explicitly encode rules to achieve an end goal. In the vast majority of cases, this end goal is predictive in nature. The oft given ex-

amples from popular science involve image recognition, email spam filtering and advertising delivery.[3] ML algorithms have also been used extensively in the past in the physical sciences, for ^{1}H NMR spectral identification, the elucidation of quantitative structure-activity and structure-property relationships, protein folding and process control [55, 56, 57, 58]. However, it is only more recently that these algorithms have begun to find applications in theoretical chemistry and materials science. In some sense, the adoption of ML approaches in the computational chemistry and materials science has been rather slow; scientists have expressed scepticism at the largely 'black box' nature of ML algorithms, as juxtaposed to the more rigorous hierarchical series of approximations used elsewhere (e.g. in the 'ladder' of wavefunction methods used in quantum chemistry). However, as the field of ML in general has matured and further successful applications within the physical sciences have arisen, the field has snowballed in popularity. In large part this may be attributed to a concerted effort within the community, to adapt and specialise approaches which have long been popular in computer science, to work within the existing toolkit of atomistic simulation, rather than in place of it [59, 60, 61, 62]. It may also be attributed to the recognition that simply relying on the scaling of larger computers, even as the exascale era approaches, cannot hope to keep up with the ambitions of computational researchers (in particular as the limits of Moore's law are approached). By reducing the scaling of complex problems[4] to something approaching linearity, ML promises to make previously intractable problems tractable, not just more convenient [63, 64]. Aside from the construction of potentials for atomistic simulations, applications of machine learning algorithms have included structure prediction, [23, 65] property prediction (including atomisation energies, band gaps and nuclear chemical shifts) [66, 67, 68, 69, 70] and the development of DFT exchange-correlation functionals [71, 72, 73, 74] to name but a few. Here we

[3] Popular science also frequently confuses machine learning with artificial intelligence (AI), leading to questions from friends and family over whether such algorithms are likely to 'take over the world'

[4] As much as DFT is a beautiful theory, which derives real and fundamental insight about the nature of matter at the atomic scale, the argument could be made, that its impact on modern physics stems just as much from the computational effort saved by treating an electronic density rather than individual electronic wavefunctions as from the fundamental physical insight itself.

will briefly review the application of ML to problems in computational chemistry and physics, with a particular focus on approaches to reconstructing the potential energy surface (PES). Throughout this section, we will pay particular attention to applications involving carbon. Because carbon arises so frequently as a test-case for the development of new methodologies, we will discuss it here, *in situ* and in context, rather than devoting a section to the discussion of the application of ML methodologies to carbon specifically. The literature surrounding this topic has exploded in recent years and so only a brief overview of the methods will be provided here, the interested reader is directed to a number of excellent reviews on the topic for a more thorough treatment, e.g. Refs [75, 76, 77, 78, 79, 80, 81].

ML based approaches to the generation of intermolecular potentials are by their very nature parametrised exclusively to *ab initio* data - but the differences between an ML and a bond-order or empirical potential extend far beyond this. Indeed, although the application of non-parametric machine learning algorithms to construct interatomic potentials is relatively new, the value of using *ab inito* data to parameterise potentials (in place of experimental data) is not. Such an approach was first applied in the 1990's in the form of 'force matching' potentials; empirically derived functional forms were parameterised not to reproduce experimentally measured properties, but to minimise the model error on a set of reference forces and energies available from an *ab initio* method [82]. This force matching approach was successfully applied to produce accurate Glue and Embedded Atom Method (EAM) potentials for systems such as Al and Mg [82, 83]. Force matching may be considered a hybrid approach between contemporary non-parametric machine learning approaches and traditonal empirical methods of parameterisation; the underlying functional forms used in force matching are still fixed and therefore limited in their flexibility, but the approach to their parameterisation through minimising a loss function is strongly reminiscent of machine learning approaches. Similarly, while potentials such as LCBOP may optimise the parameters of (for example) a Morse style functional form based on a fit to *ab inito* data, such an approach will always be fundamentally limited by the assumption that the two-body part of such an

interaction is describable by a specific closed mathematical form. This assumption - while physically motivated - does not arise from a first principles consideration of the shape of the PES, but from empirical observations and will therefore incorporate a physical bias. This limits the quality of the resulting potential, but does serve to improve the stability of the model in many cases. ML approaches, however, make no such assumptions about the functional form into which the PES may be decomposed - beyond that it must be a regular function of the atomic coordinates (continuously differentiable) and that interactions become infinitesimal as interatomic distances become very large. Machine learning methodologies have been shown to be capable of the reproduction of arbitrary functions with arbitrarily high accuracy [84]. Since the Born-Oppenheimer approximation states that the Hamiltonian, and therefore the ground-state PES is entirely defined by the atomic positions, charges and total charge of the system, there is a physical motivation for assuming that a suitably flexible ML model will be able to accurately reproduce the results of direct electronic structure calculation.

1.3.1 Neural Network Potentials

The first neural network potentials were trained on global energies, using Cartesian coordinate systems [85, 55, 86, 87]. As a result, they were limited in the number of degrees of freedom which they could treat, and lacked the properties of rotational, translational and permutational invariance (the importance of which is discussed in detail in section 3.1). An implication of this was that such approaches were only suitable for the treatment of systems with the same number of atoms as the reference *ab initio* method [86]. Nevertheless, much was achieved with these approaches, and they demonstrated the potential applications of MLPs in principle; a full review of these earlier approaches can be found in ref [77]. The first 'high dimensional' MLP, which decomposed the total energy into a sum of atomic contributions came in the form of neural network potentials (NNPs) from Behler and Parrinello in 2007 [61]. They employed a set of 'symmetry functions' to provide a mathematical description of the local environment surrounding each atom. In the first instance, the Behler-Parrinello NNPs were applied to bulk crystalline and liquid silicon. Their model

accurately predicted the ordering of the crystalline phases of silicon, as well as the radial distribution function of the liquid and the energies encountered during a metadynamics simulation [88, 61]. The Behler-Parrinello approach to constructing NNPs would go on to be applied to numerous systems, including the diamond-graphite phase transition for carbon [89, 90], the vibrational properties of water at surfaces [91], the crystallisation of GeTe phase-change compounds [92] and metallic sodium [93]. Other notable applications include a metadynamics study of the phase diagram of silicon, in which the transitions between numerous bulk silicon phases were modelled [94]. Despite the number of systems to which this approach has been applied, these potentials were often highly specialised; requiring careful tailoring to one specific set of calculations [91, 89, 90]. A major development was the application of NNPs to amorphous systems at a range of temperatures [92], which indicated the potential for more general MLPs. One hurdle to the wider adoption of Behler-Parrinello potentials was the lack of access to both the code used for training and evaluation and to the models themselves. This meant that while the results obtained with these MLPs was undoubtedly of great scientific interest, their usefulness as a tool to the wider scientific community was somewhat hindered [95].

The successes of Behler-Parrinello NNPs led to the publication of a menagerie of other applications of artificial neural networks aimed at fitting the PES. In addition to direct third party implementations of the original Behler-Parinello approach, such as n2p2 [96], Amp [97], ANI [98] and PiNN [99] a number of further developments have been realised. Deep tensor neural networks (DTNNs) [100] have been successfully applied to the fitting of PESs for small molecules. Rather than using a fixed representation of the chemical environment for the input layer of the neural network (see section 3.1 for a discussion of how this is approached in the present work) DTNNs instead use vectors of atomic numbers and two-body (2b) distances as their input. The many-body interactions are then 'inferred' by repeatedly refining a high-dimensional convolutional neural network interpretation of these distances. This was later extended to allow for dynamical simulations of crystalline and periodic materials [101]. A similar approach was adopted in the DeepMD approach

[102], in which the input layer of the neural network was instead fed a three-body representation of the local environment. The subtle difference here to the previous examples is that in addition to learning the relationship between the descriptors and the forces, additional layers in DeepMD are responsible for learning the representation of the environment.[5] A further notable example of deep learning is PhysNet [103], one application of which has been the folding of a small protein fragment by extrapolating from training data comprised of smaller peptide oligomers. Recently, committee NNPs were introduced [104], in which a number of Behler-Parrinello NNPs are trained in parallel, by taking the average of the the predicted energies and forces from this committee, a significantly more accurate model can be produced from very few training configurations. By measuring the disagreement between the individual NNPs, new training data can be selected in the regions of phase space where the model is least accurate (a process known as active learning which will be discussed later), the model may also be 'biased' by this error to avoid straying into regions of phase space where predictions will be inaccurate. Aside from the differing approach to the input layer, it bears noting that the size of the networks; the number of hidden layers and nodes within a layer, is significantly larger in the later iterations of NNP methods than in the original Behler-Parrinello approaches. The increased size would in principle offer greater flexibility to the model, but also requires more sophisticated approaches to their optimisation. What remains unclear in comparing the DeepMD [102], DTNNs [100], PhysNet [103] and SchNET [101] to their predecessor, the Behler-Parinello potentials [61], is to what degree the differences in their approaches to the descriptors/input layers along with the very different sizes of the networks contributes to differences in the quality of the final models, as no direct comparison has been performed between them.

1.3.2 Gaussian Approximation Potentials

Although neural networks were by a significant margin the first ML methods applied to fitting the PES of molecules and materials, they are by no means the only

[5] More colloquially, 'deep learning' is sometimes also used to simply refer to neural networks with many hidden layers.

example of success. Gaussian approximation potentials (GAPs), the application of Gaussian process regression to the problem of PES fitting, were first introduced in 2010 [60]. Although formally speaking, neural networks form a subset of Gaussian processes [105], in practical terms the approaches vary significantly in how they are trained and evaluated; the fitting and evaluation of GAPs will be discussed in section 5.3. GAP potentials were first applied to modelling the vibrational and thermal properties of crystalline diamond and graphite [60]. Later, GAP models were also constructed for the amorphous and liquid phase of carbon, and for crystalline graphene [106, 107]. A number of applications of the amorphous carbon GAP (GAP-17 in this work) would become manifest, including crystal structure prediction [23], studying the mechanism of the growth of tetrahedral amorphous carbon [108], modelling the vibrational and elastic properties of amorphous carbon [109, 110] and as a structure generator for DFT studies of carbon surfaces [111]. The widespread application of GAP-17 can be attributed in large part to the transferability of the model. Until this point, MLPs had suffered from an inability to model any system other than that which they were specifically designed for; a lack of transferability. This transferability arose not necessarily from a difference in the methodology (GAPs versus NNPs, or SOAP descriptors vs symmetry functions), but more from a difference in the approach used to generate the training data. By generating many structures with amorphous character, one explores the phase space of local environments with a much lower degree of redundancy and correlation than by focusing on specific crystalline structures. This transferability was exemplified in two publications benchmarking the performance of a number of carbon potentials for the graphitisation of amorphous carbon [112, 32]. An extension of this development was applied to elemental boron, which has an extremely complicated PES with many nearly degenerate minima, where instead of generating amorphous structures from quenched liquid simulations, GAP and *ab initio* random structure searching (RSS) was instead used [113, 114]. The end goal of a general purpose interatomic potential is not just transferability, however. Numerous empirical models are highly transferable by construction, the advantage that ML affords is combining both trans-

ferability and accuracy into a single model. The first example of a potential which was both transferable (able to model the bulk crystalline structures, liquid, surfaces, defects etc. of an element) while also maintaining the accuracy required to distinguish between different surface reconstructions, predict defect formations energies and model the liquid, was for silicon [115]. The transferable silicon GAP model represented a milestone for MLPs; previously, MLPs had either achieved high accuracy within a narrow region of phase space, or more rarely transferability to many different structures, but not both. The achievement is also notable for the fact that silicon has a complex phase diagram, with well-studied (and therefore with a high bar set for accuracy) surface reconstructions and defects.[6]

1.3.3 Learn-on-the-Fly and Other Approaches

A counterpoint to the argument that an ideal MLP must be able to visit any point in the phase space of a material and predict the energy of that point with a high degree of accuracy, is the fact that many studies *in practise* only visit a very small region of the possible space of configurations. Why invest the significant time and effort required to produce a comprehensive dataset of *ab initio* data, when most potential applications will not make use of it? Similarly, as there will be a high degree of correlation between frames in an MD trajectory, it is wasteful to repeatedly perform expensive *ab initio* calculations on very similar geometries. These points are recognised by learn-on-the-fly approaches, in which a model is continuously refined over the course of a simulation by augmenting and refitting a potential during the course of an MD simulation [116]. For the first few timesteps, no ML model exists and so direct AIMD must be performed, however, at each step, the energies and forces are kept (rather than being discarded as in a traditional AIMD simulation) and used as a training dataset for a model. Subsequently, at each timestep (or at regular intervals of several timesteps) the confidence of the model at predicting

[6]The situation is comparable, or perhaps more challenging, for carbon. While silicon exhibits predominantly tetrahedral bonding, carbon frequently exhibits linear, trigonal planar and tetrahedral coordination geometries. This leads to both a complex phase diagram of thermodynamically stable phases alongside a menagerie of metastable nanostructures (fullerenes, nanotubes etc). This, combined with the fact that many carbon structures are characterised by weak van der Waals interactions, makes carbon a particularly interesting material to construct MLPs for.

the energy of the configuration is evaluated. If the confidence of the model is high (i.e. a similar configuration has been observed previously and the reference data previously calculated) then the model is used to evaluate the energies and forces. However, if the confidence is low, rather than using an erroneous model prediction, a new *ab initio* calculation will be performed to evolve the dynamics - the energies and forces of this will be saved and the model refined; if this configuration is observed again, it will not require a new *ab initio* calculation. This approach was first applied to study the diffusion of defects and the propagation of a brittle crack in silicon, though in the first application the parameters of the empirical Stillinger-Weber potential for silicon were augmented, rather than constructing a high-dimensional MLP [116].[7] This approach was extended significantly in 2015 to make use of high-dimensional Gaussian process regression for the underlying MLP and applied to bulk silicon [117]. Remarkably, it was demonstrated that over the course of a simulation oscillating between 300 and 800 K, the model had effectively eliminated the need to directly perform *ab initio* calculation after just a few tens of picoseconds. The learn-on-the-fly approach has also been used successfully for crystalline Li, employing moment tensor potentials (introduced below) as the underlying ML framework [118]. A particular benefit of learn-on-the-fly approaches such as these is the potential for their ready incorporation into QM/MM embedding schemes, further improving on the efficiency of the approach [119].

Although the literature surrounding the construction of MLPs is dominated by approaches employing nonlinear regression approaches, namely neural networks and Gaussian process regression, these are not the only approaches which have been explored. Some novel approaches to have emerged in recent years employ linear regression, in contrast with the nonlinear nature of GAPs and NN potentials. These include moment tensor potentials (MTPs) [120], atomic cluster expansion (ACE) [121] and permutationally invariant polynomials (PIPs) [122]. Although the specific approach varies between the methodologies, in general, the total energy is expressed as a linear combination of terms corresponding to a body-order expansion. A basis

[7] This still very much fits within the remit of 'machine learning' as the parameters of the Stillinger-Weber potential are continuously improved with new data without manual input

set of symmetric, permutationally and rotationally invariant polynomial functions is used, in which a linear fit is performed. A particularly attractive feature of these is that by including polynomials of varying order, one can readily include 2b, three body (3b), or higher order interactions in the model as explicit low-dimensional functions. This is in contrast to nonlinear approaches, which typically focus on trying to produce a complete, high-dimensional fit of the PES - occasionally with additional terms corresponding to explicit two and three-body contributions. This low dimensionality allows the fitting of the polynomial functions using a reduced amount of training data; this in turn results in a significant improvement in the transferability of these linear models when compared to their nonlinear alternatives (potentially at the cost of reduced flexibility). Furthermore, they have recently been shown in a comparative study to achieve accuracies comparable to those of GAP potentials, all while incurring a much reduced cost [79]. MTPs in particular have been successfully applied to a number of systems, including crystalline Li [118], crystal structure prediction [123], vacancy diffusion in metals [124] and for the study of chemical reactivity including nuclear quantum effects [125].

Message passing networks, although in some sense a manifestation of NNPs, have some key features which differentiate them from the other NNPs discussed in this chapter. While the internal geometry of the NNPs inspired by Behler-Parrinello potentials is to a degree arbitrary (they are certainly carefully designed, but an optimal node structure cannot be chosen *a priori*), the structure of a message passing network is a graph representation designed to replicate the connectivity of atoms in a molecule or crystal [126, 127]. For small molecules, this connectivity may involve connecting all atoms to all others within the molecule, while for crystalline systems it is possible to construct a graph in such a way that it respects the periodicity of the crystal structure [128, 129]. A particular advantage that message passing networks have over other neural network potentials is the interpretability of the network, which makes them not just useful for materials discovery, but also for post-rationalisation of material properties and the elucidation of structure-property relationships. One particular example of note, is their application to perovskite

materials by Grossman and Xie [128], where the site energies of metal atoms representing a significant portion of the periodic table in perovskite structures were accurately predicted from graphs trained on existing DFT data from the materials project.

The final approach which we will discuss here is rather different to all of the approaches previously introduced, in that rather than searching for a high-dimensional, black box representation of the PES, interpretability of the final model is a core aim. Symbolic regression is an implementation of genetic programming algorithms which aims to bring the application of machine learning to the development of MLPs full circle, by finding (in a generative, data driven fashion), closed mathematical forms akin to those of traditional empirical models. Initial applications of this focused on 'rediscovering' preexisting empirical model functional forms, including the Lennard-Jones and Stillinger-Weber potentials [130, 131]. Recently, the Potential Optimization by Evolutionary Techniques (POET) algorithm has been applied to copper, where two accurate potentials, with errors on validation data of 3.5 meV/atom and 2.7 meV/atom respectively, were found by training on just 75 DFT reference configurations [132]. A notable advantage of these approaches is of course that their computational cost will be less than that of any other ML approach, but more subtly, they may offer the opportunity to learn something fundamental about the physical interactions characterising a system. Where in the past, empirical models have been constructed using chemical and physical intuition, perhaps in reversal, the functional forms uncovered using symbolic regression might inform our understanding of physics and chemistry.

Chapter 2

Computational Methods

In this chapter, we will introduce the core methodological aspects used in this thesis. Because the application of ML presented in this work aims to bring together the favourable aspects of a diverse range of methodologies, there are a number of large fields to introduce, which will be presented in just enough detail so that the content of the thesis can be understood without requiring reference to external text. Firstly, in section 2.1, we will introduce molecular dynamics (MD). Understanding the fundamentals of MD is core to understanding the significance of the other methodologies we introduce. Our electronic structure method, density functional theory (DFT), which we introduce briefly in section 2.1 and in more detail in section 2.2 allows us to perform accurate molecular dynamics by computing atomic forces from first principles, but is problematically expensive. Finally, to circumvent the cost associated with using DFT directly, we discuss the machine learning approaches applied in this thesis in sections 5.3, 3.1 and 3.1.2. Here we show how Gaussian process regression (section 5.3) can be used alongside developments in the quantitative description of atomic environments (section 3.1) to perform accurate regression or fitting of the DFT potential energy surface, to perform accurate and efficient MD simulations. In section 3.1.2, we briefly discuss farthest point sampling, which allows us to select reference configurations upon which to train the models discussed in section 5.3, with a minimal bias, further improving the efficiency of the models.

2.1 Molecular Dynamics

2.1.1 Integrating the Laws of Motion

Important to an understanding the work presented in this thesis is molecular dynamics. Newton's laws of motion show (if implicitly) that given a system of interacting bodies, if their locations, momenta and the forces acting between them are known, it is possible to predict the positions and momenta of those bodies at some time in the future. Newton first applied these laws of motion to understanding the motions of the planets, but we can apply them today just as effectively to understand the motion of a system of atoms.

In the very simplest case, we could consider an atom, moving with some velocity v_0 in the absence of any external interaction. If we observe it in a position r_0 at some initial time t_0, we know from Newton's first law that at some later time t_1, the atom will have moved by an amount proportional to its velocity, [133]

$$r(t_1) = r(t_0) + v_0(t_1 - t_0) = r(t_1) + v_0\Delta t. \tag{2.1}$$

In the absence of any external force, it will continue moving at this velocity for eternity - computationally straightforward but not hugely interesting from a chemical point of view. If we now consider that some force might act upon this atom, Newton's second law defines this as,

$$F = m\frac{dv}{dt} = ma \tag{2.2}$$

Where m is the mass of the atom, and a is the acceleration (the change in velocity over time) of the atom under the influence of the force. This equation defines the force in terms of what input is required to change the velocity of an object, in molecular dynamics, we rather know the force acting on the object, and wish to calculate how its velocity is changing,

$$v(t_1) = v(t_0) + \Delta t a(t_0). \tag{2.3}$$

Or in terms of the force acting on the atom,

$$v(t_1) = v(t_0) - \Delta t \frac{F(t_0)}{m}. \tag{2.4}$$

In the simplest case, these equations are all that is required [133]. By repeatedly evaluating equations 2.1 and 2.4 for times $t_1, t_2, t_3 \ldots t_N$ to move the atom of interest by some small amount, each time calculating the change in the force acting on it and updating the velocity of the atom accordingly, one can simulate essentially any classical process imaginable.

A rather more elegant integration of Newton's equations of motion tends to be used in practise. Many such approaches exist, but we will discuss here only one - the Verlet algorithm. By performing a Taylor expansion of these equations of motion around time t, we may arrive at the following,

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \frac{d\mathbf{r}}{dt} + \frac{1}{2}\frac{d^2\mathbf{r}}{dr^2} + \frac{1}{6}\frac{d^3\mathbf{r}}{dr^3}\ldots \tag{2.5}$$

Where we have made two changes to our notation, the first is that $\mathbf{r}$ is now a vector, to reflect the fact that in practise, we are concerned with systems of many particles moving in more than just one dimension. Secondly, we have substituted $t_1 - t_0$ for δt, the timestep. The timestep is a centrally important quantity in a practical molecular dynamics simulation, in general, the smaller the timestep is, the closer the computed dynamics will approximate the exact dynamics of the system (ignoring for the time being fundamental inaccuracies resulting from the forcefield or finite size effects, for example), however, if it is too small, the computational cost of the simulation will increase dramatically, limiting the timescales which can be studied. We may rewrite equation 2.5 in terms of the more familiar quantities encountered in equations 2.1 and 2.4, as

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta(t)\mathbf{v}(t) + \frac{\delta t}{2}\mathbf{a}(t) + \frac{\delta t}{6}\frac{d\mathbf{a}}{dt}. \tag{2.6}$$

Written as is, equation 2.6 is not much more than a restatement of what we have

discussed previously, however, the Verlet algorithm eliminated the terms of odd order, by summing together equation 2.6 with the equivalent terms for $t - \delta t$. That is,

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2 \mathbf{a}(t), \tag{2.7}$$

or, as before, in terms of the force,

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2 \frac{\mathbf{F}(t)}{m}. \tag{2.8}$$

It is clear from equation 2.8, that at any point, only positions of the current and previous timestep, along with the forces are required to evolve the dynamics of the system. Although it may at first appear as though we have lost our knowledge of the velocities of the particles (which are indeed useful for the computation of many statistical properties), however these can be computed from the finite difference between the positions as,

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\delta t) - \mathbf{r}(t-\delta t)}{2\delta t}. \tag{2.9}$$

The cancellation of the odd-order terms results in an equation which is more accurate than simply integrating the Taylor expansion. The equation for the position (2.8) has an associated error which is order $\mathscr{O}(\delta t^4)$, compared to the error of $\mathscr{O}(\delta t^3)$ for the Taylor expansions, while that of the velocities in equation 2.9 is of order $\mathscr{O}(\delta t^2)$. A number of refinements and improvements of the Verlet algorithm exist, as well as entirely different approaches to the integration of Newton's laws of motion, for a more thorough detailing of some of these, the interested reader is directed to Allen and Tildesley [134].

2.1.2 Interatomic Potentials and Forces

The forces required for evolving a systems of interacting atoms or molecules are available from the interatomic potential energy as the negative of the derivative of the potential,

$$\mathbf{F} = -\frac{dV}{d\mathbf{r}}. \tag{2.10}$$

As such, when discussing the use of forces for integrating the equations of motion (e.g. equation 2.8), it is convention to interchangeably refer to the interatomic potential energy or the forces. From the perspective of empirical models, it is convenient to consider the interatomic potential energy as being comprised as the sum of terms arising from the interaction between pairs, triplets, etc.:

$$V_{tot} = \sum_i V_{(1b)}(\mathbf{r}_i) + \sum_i \sum_{j>i} V_{(2b)}(\mathbf{r}_i, \mathbf{r}_j) + \sum_i \sum_{j>i} \sum_{k>j>i} V_{(3b)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) \dots \quad (2.11)$$

Here, the term $V_{(1b)}$ represents the effect of any external field, which does not depend on the position of other atoms in the system. Examples of this might be an external electric field, or an artificial confining potential imposed on the system. The $V_{(2b)}$ represents the interaction between pairs of atoms; it depends only on the distance between neighbouring atoms and typically accounts for the largest contribution to the interatomic potential. In traditional empirical potentials, higher order terms typically refer to angular (3b) or dihedral (4b) terms, however there are many other ways to construct functional forms (or descriptors, as discussed in section 3.1 and 1.3) which depend on the positions of triplets or quartets of atoms but which are not directly interpretable as angular or dihedral terms. Possibly the most well-known interatomic potential is the Lennard-Jones potential. The Lennard-Jones potential assumes that for certain systems (notably those comprised of noble gas atoms) only the 2b interaction term is required to characterise the potential; 3b and higher order interactions are assumed to be negligible. The Lennard-Jones potential is comprised of two terms, a repulsive component, raised to the 12th power, which is designed to approximate the short range repulsive interaction between atoms and an attractive component, raised to the 6th power, which mimics the long-range attractive effects of van der Waals interactions [135, 136].

$$V_{2b}^{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6} \right] \quad (2.12)$$

Where σ is a term defining the position of the attractive minimum between atoms and ε defines the strength of the interaction. The precise values of these parameters

may be determined in many ways, however they have traditionally been determined empirically, hence the name, empirical potentials. The values of σ and ε would be systematically optimised so that a simulation of, say, liquid argon, would reproduce some experimentally determined property, for example the liquid-vapour phase equilibrium. In general, the simple functional form of the Lennard-Jones potential only works well for weakly interacting atoms or molecules, such as noble gasses or methane. Other empirical models, such as the Tersoff potential for carbon, have much more complicated functional forms, with many more than two parameters, which include three or even four-body terms in their construction [40]. Despite these added layers of complexity, the parameters are still optimised empirically, to reproduce certain 'bulk' properties of the material. It is important to note, therefore, that in both of these cases (and in the case of all empirically fitted models,) the parameterised potential is an *effective* potential, V^{eff} rather than a rigorous decomposition into 2b, 3b and higher order effects. As such, for empirical potentials, their parameterisation may depend on on the temperature, density, or target property for which they are fitted, while the true potential implied by equation 2.11 would not.

Not all forces for molecular dynamics necessarily come from empirical potentials, however. Although adopted and developed much later, it is perhaps more natural to consider computing the forces for molecular dynamics directly from *ab initio* calculations - so called *ab initio* molecular dynamics (AIMD). In principle, any electronic structure method could be used to compute the forces acting on a system of atoms - in practise, it is often the case that DFT strikes the best balance between cost and accuracy, which has led to its widespread application in many systems. We will discuss DFT in more detail in the subsequent section (section 2.2), but for now suffice to say that it allows us to readily compute the electron density for an atomic or molecular system. The Hellman-Feynman theorem states that the force acting on a nucleus is equal to the negative derivative of the total energy in that direction. For example, for the force in the x direction acting on a nucleus A, the force would be given by,

$$F_{A,x} = -\frac{\partial E}{\partial x_A}. \tag{2.13}$$

Which is readily obtainable from a knowledge of just the electron density $\rho(\mathbf{r})$, and the positions and charges of the other nuclei in the system. For example, for a system of with M atomic nuclei at points $\mathbf{R}$ in space, and nuclear charges Z the force in the x-direction on nucleus A would be given by,

$$F_{A,x} = Z_A \left(\int d\mathbf{r}\, \rho(\mathbf{r}) \frac{r_x - \mathbf{R}_{A,x}}{|\mathbf{r} - \mathbf{R}_A|^3} - \sum_{A \neq B} \frac{\mathbf{R}_{A,x} - \mathbf{R}_{B,x}}{|\mathbf{R}_A - \mathbf{R}_B|^3} \right). \tag{2.14}$$

Again, the details of DFT, including the significance of $\rho(\mathbf{r})$ and how it is computed will be discussed in section 2.2. Put colloquially for the time-being, suffice to say that the Hellman-Feynman theorem allows us to compute the forces acting on a system of atoms from first principles, knowing just the positions of the nuclei and probable positions of the electrons around them.

Clearly, obtaining the forces for evolving a dynamical system from first principles is hugely desirable from an accuracy point-of-view. No approximations must be made about into what functional form the interatomic potential might decompose; in fact we can compute the potential with no approximations other than those which are present in our chosen electronic structure theory.[1] From an application point of view of course, the dramatically increased cost of electronic structure methods, compared to empirical potentials, prohibits their application to many problems of interest. As only very small numbers of atoms can be routinely simulated, a highly reductivist approach must typically be employed - idealised surfaces, single solvated molecules and perfect crystalline structures for example. The reduced cost of empirical models, meanwhile, facilitates the study of systems with far more degrees of complexity - while the interatomic forces themselves become less accurate, the physical model we are able to construct of a system becomes far closer to reality. The folding of whole proteins in solution, the growth of crystals with uneven and defective facets and many other complex phenomena may all be modelled

[1] For DFT, one might argue that these approximations are numerous, ranging from the usually sound (Born-Oppenheimer approximation) to the potentially highly problematic (the choice of exchange-correlation functional). However, these are the approximations of DFT, in *principle* one could compute forces from an electron density obtained using configuration interaction, in which case they would be essentially exact.

with empirical potentials, while such a thing would be impossible with AIMD. It bears mentioning that even if we doubled, or even tripled the world's computational power tomorrow, this gulf in applicability would not close appreciably. While the cost of empirical potentials scales more-or-less linearly with the number of atoms, the cost of a DFT calculation scales with the cube of the number of electrons; simply obtaining greater computational resources is not necessarily the right solution. Among the many promising approaches for modelling larger and more complicated systems with *ab initio* accuracy (linear scaling DFT is one particularly promising area of development [137]) the construction of machine learning potentials is the one that this thesis is concerned with. We will discuss in detail the methods for using forces computed using *ab initio* methods to construct interatomic potentials in sections 5.3 and 3.1.

2.1.3 Ensembles and Thermostats

The equations set out in the previous section establish a regime for modelling systems in which the total energy and total linear momentum are conserved - in thermodynamic terminology this is called the microcanonical, or NVE ensemble. In the NVE ensemble, the thermodynamic variables N, the number of atoms, V, the volume and E, the total energy are conserved. Other thermodynamic quantities, such as the pressure or temperature, are not fixed and can therefore fluctuate considerably. In designing computational simulations which approximate experimental systems, this is often not desirable. Systems tend to be inexorably connected to their environment, unless deliberate steps are taken to isolate them; materials are often in thermal equilibrium with their environment and pressures tend to vary slowly based on macroscopic changes rather than rapid microscopic fluctuations. For these cases, we turn to other ensembles. In the canonical, or NVT ensemble, the temperature is instead kept constant, while in the isothermal-isobaric, or NPT ensemble, the temperature and pressure are both kept constant while the volume and total energy vary.

When we wish to sample these thermodynamic ensembles over the course of a molecular dynamics simulation, we must employ thermostats for controlling the

temperature or barostats for the pressure. In this thesis, we are primarily concerned with sampling the NVT ensemble, so barostats will not be discussed further. In the simplest of cases, one might approach the problem of controlling the temperature of an ensemble by simply rescaling the velocities $\mathbf{v}$ of the atoms in a simulation to match the desired temperature T,

$$\mathbf{v}_{rescaled} = \mathbf{v}_0 \left(\frac{T_0}{T_{ave}} \right)^{\frac{1}{2}}. \tag{2.15}$$

Where T_0 is the temperature desired for the simulation and T_{ave} is the average temperature measured in the simulation since the velocities were last rescaled. This algorithm is termed 'velocity rescaling'. Problematically, the velocity rescaling algorithm perturbs the dynamics of the system considerably from those which would be found in a real equilibrium. As such, statistical averages of, for example, the structure of a liquid, vary significantly from those measured experimentally. Langevin dynamics is one approach which attempts to modify the velocities of the atoms in a system in a way which is more motivated by physical phenomena. The stochastic nature of the interaction of a system and its environment (with which it is in thermal equilibrium) is simulated by stochastically adding and subtracting forces from each atom at each timestep, proportionally to the temperature, along with a frictional or 'damping' constant. The stochastic forces are chosen with zero mean and a standard deviation given by,

$$\sigma_F = 2k_B T_0 m_i b_i \delta t \tag{2.16}$$

Where b_i and m_i represent the frictional coefficient and mass of atom i in the system and k_B is the Boltzmann constant. The frictional force is similarly applied directly to the velocities of the atoms themselves,

$$\mathbf{v}_{i,rescaled} = m_i b_i \mathbf{v}_i \tag{2.17}$$

In this way, assuming that the damping coefficients b are chosen appropriately, the velocities of the atoms in a system may be modified in such a way that the dy-

namics of the system are not significantly perturbed from those which might be found experimentally. This is particularly important in cases where, for example for graphene, we wish to compute the vibrational density of states to compare to experimental results.

One potential criticism of the Langevin approach to thermostatting is that it is not truly deterministic, because there is a stochastic element in choosing the force.[2] The Nosé-Hoover thermostat is an example of a deterministic thermostat which operates in a conceptually similar way to Langevin dynamics [138]. Again, the velocities in our system are modified by a frictional coefficient b,

$$m_i \mathbf{v}_i = \mathbf{F}_i - b m_i \mathbf{v}_i. \tag{2.18}$$

However, in the Nosé-Hoover thermostat, the changing value of the frictional coefficent b is chosen determinstally from the difference between the measured and target temperatures,

$$\frac{db}{dt} = \frac{1}{Q}\left[\sum_{i=1}^{N} m_i \frac{\mathbf{v}_i^2}{2} - \frac{3N+1}{2} k_B T_0\right]. \tag{2.19}$$

In which Q defines how quickly the frictional coefficient b relaxes, and T_0 is the target temperature of the simulation. When the temperature of the system matches the target temperature, the rate of change of the frictional coefficient tends towards 0. In this thesis, we make use of both Langevin dynamics and the Nosé-Hoover thermostat, depending on the application. In either case it will be clearly stated which thermostat is employed.

2.2 Density Functional Theory

In principle, any number of electronic structure methods could be used as the reference for training a GAP model. Methods such as the random phase approximation, coupled cluster, or configuration interaction certainly provide a better agreement

[2] In practise, due to the fact that the random numbers generated in a computer are only quasi-random, Langevin dynamics often is, in practise, deterministic. However, this is as a result of our failure to apply the method in a sufficently rigorous way!

with experiment for some properties than does DFT. However, their cost is highly prohibitive. In the first instance, there are many relevant carbon structures which do not have small, well-defined unit cells. Fullerenes, nanotubes and amorphous carbon structures may all require several hundred atoms in order to provide an appropriate structural model. Computations of such sizes are simply not computationally feasible with wavefunction-based electronic structure methods. Furthermore, as is discussed in detail in chapter 4, we find that DFT, when employing the optB88-vdW functional, provides excellent agreement with a number of experimental properties for graphene. The manageable computational cost of DFT, combined with its high-accuracy for carbon, make DFT an eminently practical method for producing reference data.

The goal of any electronic structure method is to determine an approximate (for any system more complicated than the $H_2{}^+$ ion) solution to the time-independent, non-relativistic Schrödinger equation.

$$\hat{H}\Psi = E\Psi \tag{2.20}$$

Where $\hat{H}$ is the Hamiltonian operator for the system of interest, which is composed of M atomic nuclei, with N electrons. Due to the significant differences in the masses of the electrons and nuclei, we can make use of the Born-Oppenheimer approximation to consider only the electronic Hamiltonian in this equation, which is given as,

$$\hat{H}_{elec} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} = \hat{T} + \hat{V}_{Ne} + \hat{V}_{ee} \tag{2.21}$$

Where A and B count summations over atomic nuclei, and i, j count summations over electrons. $\hat{T}$ represents the electron kinetic energy, $\hat{V}_{Ne}$ represents the external field acting on the electrons due to the positively charged nuclei, and $\hat{V}_{ee}$ represents the electron-electron interaction. The total energy E_{tot} of the system is then the sum of the electronic energy and the repulsive term characterising the in-

teractions between atomic nuclei, E_{nuc}.

$$E_{nuc} = \sum_{A=1}^{M} \sum_{B>A}^{M} \frac{Z_A Z_B}{r_{AB}} \tag{2.22}$$

The challenge inherent in operating directly on the wavefunction as is suggested by the Schrödinger equation is that it is a $3N$ dimensional function, where N is the total number of electrons in the system. The key revelation underlying density functional theory lies in the first Hohenberg-Kohn Theorem, to quote the 1964 paper of Hohenberg and Kohn directly "*the external potential* $V_{ext}(\mathbf{r})$ *is (to within a constant) a unique functional of* $\rho(\mathbf{r})$; *since, in turn* $V_{ext}(\mathbf{r})$ *fixes* $\hat{H}$, *we see that the full many particle ground state is a unique functional of* $\rho(\mathbf{r})$" [139], in which the quantity $\rho(\mathbf{r})$, is defined. $\rho(\mathbf{r})$ represents the electron density, which is the integral over the spin and spatial coordinates of the wavefunctions of the electrons of the system.

$$\rho(\mathbf{r}) = N \int \dots \int |\Psi(\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_N)|^2 ds_1, d\mathbf{x}_1, \mathbf{x}_2 \dots d\mathbf{x}_N. \tag{2.23}$$

$\rho(\mathbf{r})$ is at its heart a probability density (though it is often called the electron density, and will be termed such from hereon), it describes the probability of finding any of the electrons of the system within some infinitesimal volume element. From this, we can write an analogue of equation 2.21 in terms of the electron density,

$$E_0[\rho(\mathbf{r})] = T[\rho(\mathbf{r})] + E_{ee}[\rho(\mathbf{r})] + E_{Ne}[\rho(\mathbf{r})]. \tag{2.24}$$

Where $E_0[\rho(\mathbf{r})]$ is the ground state energy given by the electron density of the system. While the first Hohenberg-Kohn theorem establishes that the electron density in principle contains all of the information required to obtain all properties of interest, it does not establish that only a single electron density uniquely gives the correct energy for a given external potential. The second Hohenberg-Kohn theorem, often termed the variational principle, states that the functional operating on the electron density to provide the energy of a system, provides this lowest energy only if the input density is the true ground-state density $\rho(\mathbf{r})_0$. Put another way, for any trial

electron density $\rho(\mathbf{r})_{trial}$, the energy obtained from the functional (equation 2.24) represents only an upper bound on the ground state energy E_0

$$E_0 \leq E[\rho(\mathbf{r})_{trial}] = T[\rho(\mathbf{r})_{trial}] + E_{ee}[\rho(\mathbf{r})_{trial}] + E_{Ne}[\rho(\mathbf{r})_{trial}]. \tag{2.25}$$

That the ground state properties of a system are uniquely defined by $\rho(\mathbf{r})$ rather than the wavefunction, reduces dramatically the computational cost of evaluations of the energy for different systems, as the quantity of interest $\rho(\mathbf{r})$, is defined in just 3-dimensions, as opposed to 3N (plus N spin) dimensions for the wavefunction. It is this decreased dimensionality which underpins the reduced computational cost of DFT, compared to wavefunction based methods.

While the above has introduced the theoretical background for DFT, there are a number of practical aspects which must also be discussed. The first of these is the Kohn-Sham orbital. Rather than evaluating equation 2.21 explicitly, Kohn and Sham chose to represent the wavefunction underlying the density in a basis of non-interacting single electron orbitals $\{\phi_i\}$, known as the Kohn-Sham orbitals. Of course, the real wavefunctions describing individual electrons *do* interact. The connection between this artificial system of orbtials and the system we would actually like to treat, is in finding an effective potential V, such that the density resulting from the Kohn-Sham orbitals equals the density of the real system of interacting electrons,

$$\rho_s(\mathbf{r}) = \sum_{i}^{N} \sum_{s} |\phi_i(\mathbf{r}, s)|^2 = \rho_0(\mathbf{r}). \tag{2.26}$$

In which $\rho_s(\mathbf{r})$ is the Kohn-Sham density, and $\rho_0(\mathbf{r})$ is the true ground state electron density and we are summing over the spatial and spin coordinates of the N electrons in our system. A further advantage of this linear combination of atomic orbitals (LCAO) approach in which a system of non-interacting electrons is assumed, is that the kinetic energy for a system with the same electron density as our system of interest, T_s (which cannot be obtained for an interacting system of electrons) is readily available. Of course, this kinetic energy is not the same as the true kinetic energy of the fully interacting system, even if the densities themselves are the same, $T_s \neq T$.

Kohn and Sham accounted for this discrepancy, along with a discrepancy present due to the fact that the electron-electron contribution to the total energy ($E_{ee}[\rho(\mathbf{r})]$) is non-trivial to evaluate by introducing what is known as the exchange-correlation (XC) energy. Without going into too much detail, the exchange correlation energy E_{XC} is the functional which contains all of the aspects of the electron-electron interaction which are too complex (due mostly to their many-body character) to be calculated explicitly. Specifically the quantum mechanical electron exchange and correlation energies, as the name would suggest, but also the many-body contribution to the kinetic energy. There are many hierarchical approximations to the XC energy, which will not be discussed in detail here. The simplest approximation is known as the local density approximation (LDA), in which the absent quantities are calculated for a homogeneous electron gas (which can be treated exactly) and it is assumed that these do not differ too much for the very inhomogeneous electron density surrounding atomic nuclei. Surprisingly, this approach works remarkably well. The generalised gradient approximation (GGA) takes into account this inhomogeneity, adding terms which account for the local derivative of the electron density. Further developments include hybrid functionals, which incorporate some portion of the 'exact' exchange energy, which is available from the wavefunction methods (Hartree Fock) which predate DFT. A further branch of particular note is the inclusion of dispersion interactions. These may be included integrally within the functional, in the form of a semi-local dispersion inclusive functional, or added as an analytical correction after a calculation using a local XC functional has been performed. From a practical point of view, we have selected just such a dispersion inclusive functional for this work named 'optB88-vdW', which provides accurate material properties for carbon with respect to both experiment and accurate wavefunction methods [140, 141]. In particular, we note that the energetics and geometry of layered materials such as graphite, which is particularly affected by dispersion interactions is well characterised by optB88-vdW [142]. In practise, in this work, we use an implementation of DFT found in the VASP code, which is capable of determining accurate numerical approximations to the ground state energy [143].

In practical DFT (and in other wavefunction based methods) the Kohn-Sham orbitals, and hence the density, are expressed by expanding them in a basis set η. The quality of a DFT calculation is directly linked to the number or completeness of the basis functions used in this expansion. Numerous basis sets exist and much work has gone into improving the accuracy and efficiency of these, here we will discuss only one, plane waves. These are simple exponential functions, which are solutions to the Schrödinger equation for a free particle,

$$\eta^{PW} = e^{i\mathbf{k}\cdot\mathbf{r}} \tag{2.27}$$

Where the wavevector $\mathbf{k}$ is related to the momentum $\mathbf{p}$ of the wave through $\mathbf{p} = \hbar\mathbf{k}$. These are a particularly attractive basis for the periodic systems in materials science, as they naturally incorporate periodic boundary conditions into their construction. Other basis sets, for example Gaussian basis sets, do not have this quality and are therefore more frequently applied to small molecules. The completeness of the plane-wave basis set is proportional to the highest wavevector and hence the frequency and energy of the wave. The total energy computed will eventually converge to a given value if plane-waves of a sufficiently high frequency are included. However, if all of the electrons in an atomic system are to be treated, this convergence is extremely slow, due to the shape of the wavefunction close to the nucleus requiring very high frequency (and hence a very large number of) basis functions. This brings us to our final important approximation for performing practical DFT calculations of the sorts of systems we are concerned with here, the pseudopotential. The pseudopotential is a fictitious potential constructed to approximate the average effects of the core electrons (those closest to the nucleus) so that they do not need to be explicitly treated in a DFT calculation. By removing these core electrons from the calculation, the maximum frequency of the plane-wave basis functions needed to converge the total energy is dramatically reduced; which is required in order for plane-waves to be a practical choice of basis set. Although it may initially appear drastic to abstract away the very electrons whose properties we are interested in computing, the physical motivation surrounding this approximation is very sound.

The character of the core electrons typically does not change dramatically due to the effects of chemical bonding - they are simply too tightly held by the nucleus of the atom. If they do not participate in bonding, or are otherwise unaffected by the external potential V_{ext} then their contribution to the total energy must remain approximately constant. Combining these aspects of fundamental theory with numerous improvements to make DFT practical while maintaining high accuracy is what has led to it being such a useful tool in the present thesis. There is essentially no other method available to us with which a dataset of the size *and* accuracy of the one presented in this work could be generated.

2.3 Gaussian Process Regression

Gaussian Approximation Potentials are the product of the application of the Gaussian process regression machine learning methodology to the problem of function interpolation of the Born-Oppenheimer PES [60, 144]. The *ab initio* PES is sampled using a database of observations of quantum mechanical (often DFT) atomic forces and total energies (additionally sometimes virial stresses) on structures representative of the desired regions of phase space to be studied. These data are used to train the GAP model which can be used to accurately interpolate energies and forces between the previously observed reference data points, the resulting prediction can be used to generate MD trajectories; much like an empirical potential. This method circumvents a problem inherent in empirical potentials wherein assumptions must be made about the functional forms into which the PES can be decomposed. No prior supposition is made, for example, that the microscopic interactions between two atoms must be representable by a harmonic, Morse or Lennard-Jones type function. This allows for a faithful and unbiased (so far as any *ab inito* method may be called unbiased) representation of the PES to be built, which may be conveniently evaluated to accurately predict the energies and forces acting on arbitrary configurations within the sampled phase space.

Traditional empirical models use parametric, closed functional forms to approximate an underlying function (the PES). As discussed, for most systems, the

PES cannot be expressed as such a linear combination of simple terms (not to be confused with the recent success in employing linear machine learning algorithms to produce interatomic potentials, for example using permutationally invariant polynomials [122], which is discussed in section 1). In such cases, non-parametric regression techniques are desirable. Gaussian process regression is one such approach, which allows for the approximation of the underlying PES without prior knowledge about the functional form into which it might decompose. As a supervised learning technique, Gaussian process regression aims to estimate the value of an unknown function $y(\mathbf{x})$, the value of which can be measured at specific locations $\mathbf{x} = \{x_i\}_{i=1}^{N}$ to give a set of known values $\mathbf{y} = \{y_i\}_{i=1}^{N}$ which comprise our training data set. Using this training data set, Gaussian process regression aims to make a prediction of the function values $\tilde{y}(\mathbf{x}')$ at arbitrary points in the training space $\mathbf{x}'$ (note that we use $\tilde{y}(\mathbf{x}')$ to indicate predicted values of the true function $y(\mathbf{x}')$). In practical terms, for a system of interacting atoms, the points $\mathbf{x}$ represent the atomic configurations, or positions in phase space. To a first approximation, for the purposes of predicting the potential energy surface of an atomic system, the measured function values $\mathbf{y}$ would be the total energies from a reference *ab initio* method. However, as will be discussed in more detail later, in principle any quantity which is a function of the atomic coordinates can be learned, including linear functional observables of the total energy such as the atomic forces.

2.3.1 Weight-Space View of Gaussian Process Regression

In the following, we will discuss the weight space view of Gaussian process regression in general terms, following the derivations given in Mackay (2005) and a recent review covering the application Gaussian process regression to learning the properties of molecules and materials [145, 146]. We first introduce this without without a discussion of any the specifics related to learning a function which depends on atomic coordinates, which will be elaborated on in subsequent sections. We begin by making the assumption that our function $y(\mathbf{x})$ can be approximated as a linear

combination of M basis functions k, placed at locations $\mathbf{x}_m$ in the input space,

$$\tilde{y}(\mathbf{x}) = \sum_{m=1}^{M} c_m k(\mathbf{x}, \mathbf{x}_m). \tag{2.28}$$

where we recall that $\tilde{y}(\mathbf{x})$ is the predicted value of the true function $y(\mathbf{x})$. Here, c_m are we weights or coefficients associated with each of the kernel functions $k(\mathbf{x}, \mathbf{x}_m)$. Note that the set of $\{\mathbf{x}_m\}_{m=0}^{M}$ data points in equation 2.28 need not be the same as the set of data points in the training data $\{\mathbf{x}_n\}_{n=0}^{N}$. The set of M data points is often in practise much smaller and is termed the sparse set (sometimes also called the representative or active set). For the time-being, we will not specify the form of the basis functions and focus instead on the task of determining the unknown weights $\mathbf{c}$, where we define the vector of weights $\mathbf{c} = (c_1, c_2 ... c_M)$. The accuracy of a candidate estimator with a set of weights $\mathbf{c}$, may be measured by a loss function l which depends on the error of the predicted values $\tilde{y}(\mathbf{x})$ relative to the measured function values at the same point $y(\mathbf{x})$:

$$l = \sum_{n=1}^{N} \frac{[y_n - \tilde{y}(\mathbf{x}_n)]^2}{\sigma_n^2} + R. \tag{2.29}$$

Here, we have introduced the variables σ_n which allow us to assign a relative importance to different data points. The first term of this equation penalises predictions $\tilde{y}(\mathbf{x})$ which are not close to the measured function values $y(\mathbf{x})$, which ensures that l is low for close fits to the data points, but will quickly lead to overfitting. To compensate for this and ensure smoothness, an additional regularising term R is added, which depends on the weights $\mathbf{c}$ and the corresponding kernel functions,

$$R = \sum_{m,m'}^{M} c_m k(\mathbf{x}_m, \mathbf{x}'_m) c_{m'}. \tag{2.30}$$

The regularising term penalises situations in which the weights $\mathbf{c}$ become very large, thereby helping to ensure the smoothness of the estimator, while the parameters $\{\sigma_n\}$ can be adjusted to ensure a close fit to individual data points. Note that different values of σ may be applied to each data point individually, allowing the esti-

mator to be biased towards fitting certain data points more closely. We can rewrite equation 2.29 in matrix notation for simplicity,

$$l = (\mathbf{y} - \mathbf{K}_{NM}\mathbf{c})^{\top}\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{K}_{NM}\mathbf{c}) + \mathbf{c}^{\top}\mathbf{K}_{MM}\mathbf{c}. \tag{2.31}$$

In this equation, $\mathbf{y}$ is a vector of measured function values $(y_1, y_2, ...y_N)$ and $\mathbf{c}$ is the vector of weights $(c_1, c_2, ...c_N)$. $\mathbf{K}_{NM}$ is the kernel matrix with elements $[K_N M]_{nm} = k(\mathbf{x}_n, \mathbf{x}_m)$, where N is the total number of data points in the training data, and M is the number of data points in the sparse set. The values of σ_n have here been collected as the diagonal elements of the matrix $\mathbf{\Sigma}$, such that $\mathbf{\Sigma}_{nn} = \sigma_n^2$. Recall that our aim is to find the set of weights $\mathbf{c}$ which minimises the loss function defined in equation 2.29. Since equation 2.31 is a linear system of equations, we can find minima of the loss function l by taking the derivative of equation 2.31 and solving for 0, doing so and rearranging for $\mathbf{c}$ gives,

$$\mathbf{c} = (\mathbf{K}_{MM} + \mathbf{K}_{MN}\mathbf{\Sigma}^{-1}\mathbf{K}_{NM})^{-1}\mathbf{K}_{MN}\mathbf{\Sigma}^{-1}\mathbf{y}. \tag{2.32}$$

Here, we have exploited the relation $\mathbf{K}_{NM}^{\top} \equiv \mathbf{K}_{MN}$ resulting from the symmetric nature of the kernel function to simplify the expression. Using the computed weights, a prediction at some new point in the input space $\mathbf{x}$ is given by,

$$\tilde{y}(\mathbf{x}) = \mathbf{c}^{\top}\mathbf{k}(\mathbf{x}, \mathbf{x}_M). \tag{2.33}$$

Where $\mathbf{k}(\mathbf{x}, \mathbf{x}_M)$ is the vector of kernel functions between the points in the input space where a prediction is to be made $\mathbf{x}$ and the set of sparse points $\{\mathbf{x}_m\}_{m=0}^{M}$. Equations 2.32 and 2.33 highlight one of the key benefits to Gaussian process regression, that in the end the problem of solving for the weights and making predictions is simply one of linear algebra. In particular, the ability to determine the weights $\mathbf{c}$ analytically avoids the need for multivariate and often highly convex optimisations encountered in determining the weights in neural networks.

2.3.2 Function-Space View of Gaussian Process Regression

In addition to the previously provided weight-space derivation of Gaussian Process Regression, we provide here the alternative function space view, which has the advantage of highlighting some of the other, probabilistic aspects of the method. As before, we begin with the assumption that our unknown function $y(\mathbf{x})$ can be sampled at $\mathbf{x} = \{\mathbf{x}_n\}_{n=0}^{N}$ points in input space, giving $\mathbf{y} = \{\mathbf{y}_n\}_{n=0}^{N}$ associated measurements of the function we wish to learn. Similarly, we assume that the function $y(\mathbf{x})$ may be approximated by a function $\tilde{y}(\mathbf{x})$ which has the form,

$$\tilde{y}(\mathbf{x}) = \sum_{h}^{H} w_h \phi_h(\mathbf{x}). \tag{2.34}$$

This is similar in form to the equation given in the weight-space derivation of Gaussian Process Regression, but in which the functions ϕ are not representing the kernel functions $k(\mathbf{x}, \mathbf{x}'$ in equation 2.28, nor can the weights $\mathbf{w} = \{w_h\}_{h=0}^{H}$ be equated to the weights $\mathbf{c}$. Rather, the functions ϕ are as-yet unspecified basis functions which serve primarily as a mathematical tool in the derivation - we will show later that the derivation follows by showing that it is not actually necessary at any stage to assign a specific functional form to ϕ. We focus instead on the coefficients of these functions $\mathbf{w}$. We may a prior upon these weights that they are sampled from a Gaussian distribution, with a variance σ_w^2 and a mean of 0. This collection of random variables, any number of which have a joint Gaussian distribution, is the definition of a Gaussian process [105]. In this general case, for weights in which the Gaussian prior is imposed it can be shown that the covariance of any two predictions $\tilde{y}(\mathbf{x})$ and $\tilde{y}(\mathbf{x}')$ is given by,

$$\langle \tilde{y}(\mathbf{x}) \tilde{y}(\mathbf{x}') \rangle = \sum_{h,h'} \phi_h(\mathbf{x}) \phi_h(\mathbf{x}') \int d\mathbf{w} P(\mathbf{w}) w_h w_{h'}. \tag{2.35}$$

Given the Gaussian prior ($P(w_h) \propto \mathrm{Norm}(0, \sigma_w^2)$) imposed on $\mathbf{w}$ we have,

$$\sum_{h,h'} \sigma_w^2 \delta_{h,h'} \phi_h(\mathbf{x}) \phi_h(\mathbf{x}') = \sigma_w^2 \sum_{h} \phi_h(\mathbf{x}) \phi_h(\mathbf{x}'). \tag{2.36}$$

We can define the kernel function k as the second term,

$$k(\mathbf{x},\mathbf{x}') \equiv \sigma_w^2 \sum_h \phi_h(\mathbf{x})\phi_h(\mathbf{x}') \tag{2.37}$$

Note here that we abstract our the specific basis functions into the 'kernel function' as before - so that the function ϕ are never explicitly required. The specific functional form of the kernel function will not be given here, but is examined for the specific cases used in this thesis in section 3.1. As in the previous derivation, it is important to acknowledge the fact that our training data set is not free from noise. We can impose a similar Gaussian prior upon the distribution on this noise, again with the distribution $\text{Norm}(0,\sigma_w^2)$, from which we can write the covariance of any two input training data points $y(\mathbf{x_n})$ and $y(\mathbf{x'_n})$ as,

$$\langle y(\mathbf{x}_n) y(\mathbf{x}_{n'}) \rangle = k(\mathbf{x}_n,\mathbf{x}_{n'}) + \sigma^2 \delta_{n,n'} \tag{2.38}$$

Which we can again rewrite in matrix notation for simplicity as, $\mathbf{K}_{NN} + \sigma^2\mathbf{I}$. The distribution of all of the training data $\mathbf{y}$ is thus,

$$P(\mathbf{y}) \propto \exp[-\frac{\mathbf{y}^\top(\mathbf{K}_{NN}+\sigma^2\mathbf{I})^{-1}\mathbf{y}}{2}]. \tag{2.39}$$

Given this probability distribution, and given our known function values $\mathbf{y}$ how might we proceed in making a prediction about a new value of the function y_{N+1} at the point $\mathbf{x}_{N+1}$? This may be expressed as a problem of conditional probability, in which case Bayes' theorem provides us with,

$$P(y_{N+1}|\mathbf{y}) = \frac{P(y_1,y_2,y_3 \ldots y_N,y_{N+1})}{P(\mathbf{y})}. \tag{2.40}$$

Which of course only provides a probability distribution of the value y_{N+1} which we would like to predict. We take our prediction to be the mean of this distribution, which is given by,

$$y_{N+1} = \mathbf{k}(\mathbf{x}_{N+1},\mathbf{x_N})(\mathbf{K}_{NN}+\sigma^2\mathbf{I})^{-1}\mathbf{y}. \tag{2.41}$$

Where as before, $\mathbf{k}(\mathbf{x}_{N+1})$ represents the vector of kernel functions between the point in the input space where we wish to make a prediction $\mathbf{x}_{N+1}$ and points in the training data $\mathbf{x}_N$ Note that equation 2.41 is equivalent to the expression for prediction given in the previous weight-space derivation. In addition to taking the mean, it is useful to note that since the prediction in this case is a Gaussian probability distribution, we can also compute the variance of the prediction at $\mathbf{x}_{N+1}$, which is

$$\text{var}(y_{N+1}) = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \sigma^2 - \mathbf{k}(\mathbf{x}_{N+1}, \mathbf{x}_N)^\top (\mathbf{K}_{NN} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}_{N+1}, \mathbf{x}_N). \tag{2.42}$$

Which provides a metric for the uncertainty of the prediction using the Gaussian process.

2.3.3 Gaussian Process Regression based on Linear Functional Observables

When training using the data which are readily available from quantum mechanical calculations such as DFT, we do not have direct access to certain atomic-scale observables, particularly the 'atomic energy', as only the total energy of a system is directly calculable. This is challenging for a number of reasons. Firstly, it is necessary for the total energy of a to be representable as a sum over individual atomic contributions, if the resulting models are to be scaleable. Secondly, we would like to train on the atomic forces, which are available from electronic structure calculations via the Hellman-Feynman theorem (equation 2.13) and virial coefficients. The forces are derivatives of the total energy with respect to local atomic displacements, while the virial coefficients are the derivative of the energy with respect to distortions of the lattice vectors. Similarly, the atomic energies themselves are not directly observable, but are only available as a linear functionals of the local energy. The question becomes one of how to estimate the value of a function when the function itself is not measurable, only its derived properties. This problem was tackled in ref [60] in which Gaussian Approximation Potentials were first introduced. The prediction is still made using equation 2.33, but expressions must be found for the covariance of the local energy with respect to the available total energy and its

derivatives.

To illustrate how this is performed for derivative observations, we first consider the derivative of equation 2.34 with respect to an arbitrary component α of the input vector $\mathbf{x}$.

$$\nabla_\alpha \tilde{y}(\mathbf{x}) = \sum_h^H w_h \nabla_\alpha \phi_h(\mathbf{x}) \tag{2.43}$$

Which might represent the derivative of the total energy with respect to the displacement of an atomic coordinate, or in the case of fitting to virial coefficients, its derivative with respect to the lattice vectors of the cell. We wish to find the covariance of two such derivative observations (i.e. the covariance between two forces rather than two energies as before), which can be expressed as,

$$\langle \nabla_\alpha \tilde{y}(\mathbf{x}) \nabla_\beta \tilde{y}(\mathbf{x}') \rangle = \sigma_w^2 \sum_h \nabla_\alpha \phi_h(\mathbf{x}) \nabla_\beta \phi_h(\mathbf{x}'). \tag{2.44}$$

Recalling from equation 2.37 that we have defined our kernel function for the general case to be $k(\mathbf{x},\mathbf{x}') = \sigma_w^2 \sum_h \phi_h(\mathbf{x})\phi_h(\mathbf{x}')$, from this it follows that the kernel for the derivative case is the double derivative of the kernel defined previously:

$$\sigma_w^2 \sum_h \nabla_\alpha \phi_h(\mathbf{x}) \nabla_\beta \phi_h(\mathbf{x}') = \frac{\partial}{\partial x_\alpha} \frac{\partial}{\partial x_\beta} k(\mathbf{x},\mathbf{x}'). \tag{2.45}$$

The expressions for the kernel functions in equation 2.45 and 2.37 facilitate regression using only derivative observations or only function measurements respectively. Following the derivation given above, one can also arrive at an expression for the covariance between a function observation and a derivative observation (i.e. the covariance between an energy and a force) as,

$$\langle \nabla_\alpha \tilde{y}(\mathbf{x}) \tilde{y}(\mathbf{x}') \rangle = \frac{\partial}{\partial x_\alpha} k(\mathbf{x},\mathbf{x}') \tag{2.46}$$

With these equations, it is possible to construct a covariance matrix for arbitrary observations of a function and its derivatives. In general for any linear operator $\hat{L}$, which might include differentiation, scaling, etc, the coefficients that must be found

in analogy to equation 2.32 are given by,

$$\mathbf{c} = (\mathbf{K}_{MM} + \hat{L}\mathbf{K}_{MN}\boldsymbol{\Sigma}^{-1}\hat{L}\mathbf{K}_{NM})^{-1}\hat{L}\mathbf{K}_{MN}\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{2.47}$$

2.4 Computational Cost and Sparsification of Gaussian Process Regression

At this point, we may highlight some of the key computational considerations in training and evaluating such a Gaussian process model. Throughout the derivation in section 2.3.1, we make use of a sparse set of data points $\{\mathbf{x}_m\}_{m=0}^{M}$, in which the number an locations of the basis functions used for fitting and evaluation need not coincide with the training data points $\{\mathbf{x}_m\}_{m=0}^{M}$.[3] This is because the limiting factor in computing the weights in equation 2.32 is the inversion of a matrix with dimension (M,M). Matrix inversion is a costly process in terms of both memory, which scales as $\mathcal{O}(M^2)$ and computational time, which scales as $\mathcal{O}(M^3)$. The choice of letting $\{\mathbf{x}_m\}_{m=0}^{M} = \{\mathbf{x}_n\}_{n=0}^{N}$ rapidly becomes computationally intractable for even moderately sized data sets (particularly when dealing with the high-dimensional descriptors as discussed in section 3.1), as such, all GAP models are sparse kernel models. The key limiting factor in the fitting is, in-practise, not the computational time but the available memory. Since the fitting need only be performed once, it is easy to rationalise a computation which may take many days to complete as a 'one-off' cost, easily compensated for by the computational effort saved by using a GAP model over direct *ab initio* simulation. However, the memory requirements of the matrix inversion in equation 2.32 rapidly approaches the maximum practical capabilities of most modern HPC facilities (1-1.5 Tb per node).

A further consideration is of course the evaluation of the resulting model - the limiting factor for which is the computation of the vector of kernel values in equation 2.33, $\mathbf{k}(\mathbf{x},\mathbf{x}_m)$. This scales linearly with the number of points in the sparse data set, M. While not computationally intractable, it should be noted that this cost is

[3] For the special case where $N = M$, these expressions are equivalent to those of Kernel Ridge Regression.

incurred at every step of an MD simulation, so a small increase in M can potentially add up to a significant increase in computational time used when considering the total number of simulations in which a model may be applied over the course of its lifetime. In practise, however, it may be the case that the evaluation of the kernel function itself is not the limiting step of the overall calculation. Rather, the computation of the high-dimensional descriptors (discussed in section 3.1 used as inputs to the kernel functions may be computationally limiting, depending on the type of descriptor used (see figure 5.3).

2.4.0.1 Selection of Representative Points

A key consideration upon which we have not yet touched is the issue of how one should select the M representative points from the total N available. It is important that the range of sparse points is as diverse as possible and thereby contains examples of many varied atomic environments. In low-dimensional cases, it is sufficient to select these points regularly on a grid in descriptor space; the dimensionality of the problem is such that a regular grid may tile the space sufficiently. In the case of the two-body descriptors (discussed in section 3.1) we do precisely this - it is sufficient to ensure that all of the interatomic distances represented in the training data are also represented as sparse points. Here, the benefits of sparsification become very clear: rather than requiring that a GAP model retain in its kernel functions corresponding to interatomic distances for all pairs of atom (within a certain cutoff) in practise only 20-30 *representative* points (each for a different distance) are necessary.

In higher dimensional cases however, such a grid-based approach performs poorly, the volume of the descriptor space rapidly becomes too large to populate in this way. One potential way to approach this problem for higher-dimensional descriptors might at first seem to be instead randomly sampling from the available points. However, a key weakness of this approach is that the selected points are heavily biased by the composition of the training data set. If a training data set contains a disproportionately large number of configurations of one kind (e.g. 30% of the configurations are graphene while the remaining 70% represent the whole of the

remaining carbon phase space) then the sparse points, if selected randomly, will reflect this, despite this being highly sub-optimal. For these situations, leverage score CUR decomposition has been shown to be a powerful tool for representative point selection. In the spirit of principle component analysis and other dimensionality reduction algorithms, CUR decomposition aims to construct a low-rank approximation of the original kernel matrix [147]. The key differentiating feature of CUR decomposition is that the low-rank approximation is constructed in terms of a small number of the actual rows and columns of the kernel matrix (rather than a linear combination of many of them). This ensures that the representative points selected during sparsification correspond to actual atomic environments, which, while not necessary for the fitting of a sparse Gaussian process, is beneficial from the point of view of interpretability of the model. In general, when given a kernel matrix K_{NM}, CUR decomposition will aim to represent it as a product of 3 matrices, C, U, and R. Here, C and R are constructed from a small number of columns and rows of K respectively. U is an additional matrix which has been constructed to ensure that that the product CUR closely approximates the original kernel matrix K_{NM}, which is the sparse kernel matrix used for fitting. Rows and columns are selected for inclusion in the matrices C and R based upon their 'statistical leverage', in order of decreasing importance. Note that the aims of this sparsification are conceptually similar to the pre-filtering of configurations performed using farthest point sampling (FPS) as discussed in more detail in 3.1.2. The key difference is that sparsification using CUR decomposition occurs during the training process; environments which are not selected as representative points at this stage will still have an impact on the overall fit of the Gaussian model. This is in contrast to FPS, which occurs prior to any fitting of any kind and serves to eliminate entire configurations from the training database if they are highly correlated with others. Configurations eliminated using FPS will not be considered in the calculation of the model weights at all - though it should be noted that configurations which are eliminated in this way are likely to be drawn from regions of configuration space which are already well represented in the training data.

Chapter 3

Gaussian Approximation Potentials

In this chapter, the specific methodological approaches required to apply Gaussian process regression to the problem of constructing a practically useful machine learning potential will be discussed. The topics discussed in this chapter often favour practical aspects of the application of the the methodology, for example the selection of model hyperparameters, descriptors and the construction of training datasets. It is the aim of this chapter to cover in a practical manner the necessary techniques and approaches which would allow one to construct and optimise a training database, identify suitable model parameters for a particular material and piece these various aspects together into a final model.

3.1 Kernel Functions and Representation of Chemical Environments

Until this point, we have considered the kernel in rather abstract terms, here we will discuss its construction in more detail with our specific application of atomic systems in mind. In order to facilitate the simulation of systems of larger sizes than those upon which *ab initio* calculations are feasible, the GAP model total energy is decomposed into a sum of local contributions, computed from kernel functions which represent the similarity between chemical environments. It is here important to briefly discuss why such a transformation is necessary, when the Cartesian coordinates used for performing the reference *ab initio* calculation form such a convenient (and unique) representation for some calculations. Recall that in training

a GAP model, we seek a measure of similarity between two structures, in order to compute the kernel function element between them. Such a comparison is challenging in a Cartesian basis, where identical structures, with associated identical physical properties, may be described by very different sets of Cartesian points. By taking any single structure and permuting the rows in which atom coordinates are listed, translating the entire frame of reference, or rotating the frame of reference, such sets of coordinates, corresponding to identical structures, can be produced. The physical characteristics (e.g. the total energy) of these systems remain unchanged however. We therefore seek a mathematical description of our system which is invariant under these transformations: permutation, rotation and translation.

Behind the sometimes florid language used in the discussion of machine learning, is a concept which is recognised in the very earliest molecular dynamics simulations performed. The simplest example of a rotationally, translationally and permutationally invariant descriptor, is the interatomic distance.

$$q_{ij}^{(2b)} = |\mathbf{r}_j - \mathbf{r}_i| \equiv r_{ij}, \tag{3.1}$$

where $\mathbf{r}_j$ indicates the position vector of atom j. This expression is clearly invariant to rotations and translations. No rotation of the frame of reference will change the distance between the points within that frame of reference, nor will any translation. Similarly a permutation in the ordering of the atoms will give the same distance,

$$r_{ij} = |\mathbf{r}_j - \mathbf{r}_i| = r_{ji} \tag{3.2}$$

The total energy of the system is simply represented by a sum over these pairwise contributions.

$$E = \sum_i \sum_{j>i} V^{(2b)}(r_{ij}) \tag{3.3}$$

We can see here that a permutation of the order of atoms simply affects a change in the order of summation, leaving the total energy unchanged - the two-body distance therefore satisfies our requirements for a descriptor.

In this work we decompose the total energy function into a sum of two body (2b), three body (3b) and many-body (MB) interactions, which are weighted (in terms of their contribution to the total energy and atomistic forces) based on their respective statistically measured contributions. The largest portion of the energy is described by pairwise interactions, then 3b, then MB contributions, each of which is represented by a distinct descriptor and associated kernel function [60, 148, 75]. 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. We also found in the development of our potentials that combined descriptors additionally facilitated greater accuracy - a higher quality potential - thereby making more efficient use of the training data as compared to single descriptor methods. Descriptors vary greatly in their complexity, the 2b term used here is simply the distance between two atoms, while the MB term takes the form of the smooth overlap of atomic positions (SOAP) descriptor, which provides an overcomplete mapping of general n-body configurations. There are many other possible descriptors in the literature, including symmetry functions, Coulomb matrices and bispectra [149, 150, 151]. The specific parameters used for the descriptors in this work are different for each potential, and are discussed in specifics in their respective chapters.

The fundamental feature defining an interatomic potential is that the total energy is the sum of individual atomic contributions. This is also true of GAP models. The local atomic energy expression for the GAP model is a linear combination over the contributions from each kernel function $\mathbf{K}^{(d)}$ associated with a descriptor d:

$$\varepsilon^{(d)}(\mathbf{q}_i^{(d)}) = \sum_{t=1}^{N_t^{(d)}} \alpha_t^{(d)} K^{(d)}(\mathbf{q}_i^{(d)}, \mathbf{q}_t^{(d)}), \tag{3.4}$$

in which the sum over t runs over the N_t basis functions. $K^{(d)}(\mathbf{q}_i^{(d)}, \mathbf{q}_t^{(d)})$ is the

covariance kernel quantifying the similarity between the descriptor of the atomic environment for which the prediction is to be made, $\mathbf{q}_i^{(d)}$, and the prior observation, $\mathbf{q}_t^{(d)}$, which has associated with it a weighting α_t obtained during the fitting process. The total energy expression for a system is then given by the sum of each of the contributions of each descriptor used in the model, weighted by a corresponding factor δ

$$\begin{aligned} E = &\delta^{(2b)} \sum_{ij} \varepsilon^{(2b)}(\mathbf{q}_{ij}^{(2b)}) \\ &+ \delta^{(3b)} \sum_{ijk} \varepsilon^{(3b)}(\mathbf{q}_{ijk}^{(3b)}) \\ &+ \delta^{(MB)} \sum_{i} \varepsilon^{(MB)}(\mathbf{q}_{i}^{(MB)}). \end{aligned} \tag{3.5}$$

The indices i, j and k run over all atoms in the system. We now introduce the mathematical form of each of the descriptors used. The two body descriptor is simply the distance between any two atomic pairs i and j, as shown in equation 3.1. The 3b term ($\mathbf{q}^{(3b)}$) used here involves a symmetrized transformation of the Cartesian coordinates, which is designed to be permutationally invariant to the swapping of atoms j and k, given by [144]

$$\mathbf{q}_{ijk}^{(3b)} = \begin{pmatrix} r_{ij} + r_{ik} \\ (r_{ij} - r_{ik})^2 \\ r_{jk} \end{pmatrix}. \tag{3.6}$$

Many body interactions are described using the smooth overlap of atomic positions (SOAP) descriptor [150, 148]. For this descriptor, we begin with the atomic neighbor density around an atom i, which is constructed by the placement of a Gaussian function on each neighbor atom j within a given cut-off r_{cut},

$$\rho_i(\mathbf{r}) = \sum_j f_{\text{cut}}(r_{ij}) \exp\left[-\frac{(\mathbf{r}_i - \mathbf{r}_{ij})^2}{2\sigma_{\text{at}}^2} \right]. \tag{3.7}$$

Here, σ_{at} determines the width of the Gaussian and f_{cut} is any function which goes smoothly to 0 at the cut off distance (we note that all descriptors in this work

use this same cut-off function). For example,

$$f_{\text{cut}}(r_{ij}) = \begin{cases} 1 & \text{if } r_{ij} \leq r_{\text{cut}} - w_{\text{cut}} \\ g_{\text{cut}}(r_{ij}) & \text{if } r_{\text{cut}} - w_{\text{cut}} < r_{ij} \leq r_{\text{cut}} \\ 0 & \text{if } r_{ij} > r_{\text{cut}} \end{cases} \tag{3.8}$$

in which w_{cut} specifies the width of the region over which the function goes to 0, and where $g_{\text{cut}}(r_{ij})$ may be any function which decreases monotonically from 1 to 0 between $r_{\text{cut}} - w_{\text{cut}}$ and r. We choose

$$g_{\text{cut}}(r_{ij}) = \frac{1}{2}\left[\cos\left(\pi \frac{r_{ij} - r_{\text{cut}} + w_{\text{cut}}}{w_{\text{cut}}}\right) + 1\right]. \tag{3.9}$$

The neighbor density is then expanded in a basis set of radial functions $g_n(r)$ and spherical harmonics $Y_{lm}(\mathbf{r})$ as

$$\rho_i(\mathbf{r}) = \sum_{nlm} c^{(i)}_{nlm} g_n(r) Y_{lm}(\mathbf{r}), \tag{3.10}$$

in which $c^{(i)}_{nlm}$ are the expansion coefficients for the atom i. The descriptor itself is formed from the power spectrum of these coefficients

$$\mathbf{q}^{\text{MB}}_i = p^{(i)}_{nn'l} = \frac{1}{\sqrt{2l+1}} \sum_m c^{(i)}_{nlm} (c^{(i)}_{n'lm})^*. \tag{3.11}$$

To obtain a power spectrum for $n < n_{\max}, l < l_{\max}$, the expansion of the atomic neighbor density into radial basis functions can employ a truncated basis set. In the local energy expression (eq. 3.4) the covariance kernel $K^{(d)}_t$ provides a quantitative measure of the similarity between two chemical environments $\mathbf{q}^{(d)}$ and $\mathbf{q}^{(d)}_t$. The functional form of the covariance kernel differs depending on the descriptor, for the 2b and 3b descriptors, we choose the squared exponential kernel,

$$K^{(d)}(\mathbf{q}^{(d)}_i, \mathbf{q}^{(d)}_t) = \exp\left[-\frac{1}{2}\sum_\xi \frac{(q^{(d)}_{\xi,i} - q^{(d)}_{\xi,t})^2}{\theta^2_\xi}\right]. \tag{3.12}$$

The index ξ runs over either the single value of the 2b descriptor, or the three

components of the 3b descriptor. For the many-body SOAP descriptor, the natural choice of covariance function is the dot product of the two power spectra $\mathbf{p}_i$ and $\mathbf{p}_t$ with elements $p^{(i)}_{nn'l}$ and $p^{(t)}_{nn'l}$, as this corresponds analytically to an integrated overlap over all possible 3D rotations of the two associated neighbor densities, that is

$$K^{(\mathrm{MB})}(\mathbf{q}_i^{(\mathrm{MB})}, \mathbf{q}_t^{(\mathrm{MB})}) = [\mathbf{p}_i \cdot \mathbf{p}_t]^{\zeta} = \left[\int d\hat{R} \left| \int dr^3 \rho_i(\mathbf{r}) \rho_t(\hat{R}\mathbf{r}) \right|^2 \right]^{\zeta}. \tag{3.13}$$

3.1.1 Database Construction

By far the most important element of any ML potential is the composition of the training database. Regardless of whether neural networks, Gaussian process regression, linear or nonlinear machine learning methods are employed, training data of some kind will be employed in the fitting. The quality of the eventual model predictions is highly dependent on the content and quality of this training data. The selection of training data for the construction of ML potentials is as much an area of ongoing research as the development of the algorithms themselves. As such, there are many potential approaches to the generation of such databases, the details of which may vary depending on the nature of the intended application of the model and the ML algorithm employed. For example, a potential which is aimed at producing an accurate description of a single crystal structure (e.g. graphene) will require a very different approach to the generation of the database than a 'general-purpose' potential (i.e. all of carbon). Two broad approaches have proven to be useful in the construction of the GAP models presented within this thesis, both of which will be discussed in this section. These two approaches may be broadly summarised as a 'bottom-up' and 'top-down' approach. The former is well suited to ensuring that small, well-defined regions of phase space (e.g. ground state crystal structures, specific defects and surfaces) are accurately described and is important in the context of a general potential for ensuring a high degree of accuracy is attained for these important structures. The latter, 'top-down' approach, aims to build the robustness

and transferability of the potential, by generating configurations at points which are both high and low in energy on the DFT PES. A combination of both of these is key to achieving both accuracy and transferability in a general model; the two approaches are described in detail below.

3.1.1.1 The Bottom-Up Approach to Training Data Generation

In the early stages of the literature surrounding the GAP methodology, potentials were predominantly fitted using a 'bottom-up' approach [60, 115, 107]. This approach has also been used in the construction of the GAP models presented in this thesis, in particular that for pristine graphene in Chapter 4. In the so-called 'bottom-up' approach, an initial set of training configurations is generated for particular ground states of the material which are known *a priori* to be scientifically relevant. In the case of carbon, these ground states might be the known bulk allotropes; graphite and diamond. The ground state structures alone do not provide sufficient data to train a GAP model however, something more is required. A suitable set of training data might be generated for these structures by performing, for example, AIMD at a fixed temperature, to generate a set of reference structures close to the energy minima but with some fluctuations in the atomic positions, forces and total energy. A GAP model trained on these configurations could be reasonably expected to reproduce the properties of the ground state crystal structures, for example the phonon dispersion curves. The data set may subsequently be diversified to include a wider variety of configurations, thereby increasing the number and type of properties which the GAP model might reproduce. These could include manually selected configurations such point or line defects, low-Miller index crystal surfaces, and other carbon allotropes. A diversified training data set might also include crystal structures at a range of lattice parameters (either selected manually or sampled from AIMD performed in the NPT ensemble), which would be expected to improve the description of properties such as elastic constants and lattice parameters. The advantage of this method is that at all times, a good understanding of the model performance in the region of phase space to which the model will be applied is maintained - one knows precisely which systems the model would be ex-

pected to accurately model. The key is that at each step in the bottom-up approach, the types of structures to be included in the training data set are guided by a prior knowledge of the behaviour of the material. In the case of well-studied systems, this can be highly beneficial. Specifically in the case of carbon - it is clear that a model must properly describe the properties of diamond and graphite in order to be useful. It is also known from the literature which point defects are relevant, one can therefore ensure that these are correctly described. In this way, the top-down approach is useful for targeting specific structures and properties which are known to be relevant from the perspective of materials science and can serve as a useful way to impose beneficial bias on a potential to ensure accuracy of these properties. However, since the model is only ever fitted to a narrow region of phase space, it will only be accurate for properties which depend entirely on the included geometries; the transferability of potentials trained entirely in this way will be poor to non-existant.

A specific example in the context of the work presented in this thesis is the case of graphene. The crystal structure of graphene, including the lattice parameters and atomic positions, was first optimised, using the same level of DFT which was later used to compute the reference data for training. From this optimised ground-state crystal structure, a short molecular dynamics simulation was performed at 300 K, in the case of the work discussed in chapter 4 of this thesis, this was performed using an available empirical model for carbon, LCBOP, however a general procedure (as later used for other crystal structures such as diamond in chapter 5) would instead employ AIMD. Configurations were drawn from this molecular dynamics simulation and the relevant quantities used for training, the energies and forces, were computed using tightly converged DFT. A GAP model fitted to this data might well appear to perform well if validated against the energies and forces from a independent validation set of configurations drawn from a trajectory at the same lattice parameter and temperatures. However, the model predictions would quickly breakdown when used to compute simple material properties, for example if performing a geometry optimisation to determine the lattice parameter. To remedy this, con-

figurations drawn from molecular dynamics simulations performed at a range of lattice parameters (between 2.460 and 2.480 Å) were included, rather than just the optimised geometry. Similarly, by gradually adding degrees of complexity, including sampling at higher temperatures, further perturbations to the lattice parameters (e.g. to induce shear stresses), a model capable of describing a number of important characteristics of graphene was obtained.

The final data set was comprised of pristine graphene, containing configurations drawn from molecular dynamics performed at temperatures between 300 and 4000 K and lattice parameters between 2.460 and 2.480 Å(as discussed in detail in chapter 4). Such a model will be highly accurate within that region of phase space, but have uncontrolled errors elsewhere - for example, the introduction of a monovacancy or Stone-Wales defect will produce nonsensical predictions. One could consider manually constructing each of the important defects by hand (and indeed in some cases this has been done), but this is laborious and time consuming - furthermore, it does little to improve the transferability of the potential. A model trained on pristine graphene, and graphene with a single Stone-Wales defect, might well reproduce the energies and other properties of each of these extrema correctly, but will provide no useful information on closely related properties not included in the training dataset; for example the activation barrier to introducing a Stone-Wales defect, the energy of a Stone-Wales defect in a carbon nanotube, or the formation energy of a vacancy defect. As the intended scope of a model increases, the complexity of the configurations which must be included if employing an entirely bottom-up approach increases dramatically, and in practise another approach is required.

3.1.1.2 The Top-Down Approach and Iterative Training

While the bottom-up approach presented in section 3.1.1.1 is in many ways the most intuitive approach to database generation, it is not terribly robust; that is to say it does not lead to the generation of particularly transferable or stable models. The reason for this is that performing small perturbations on known crystal structures is a highly inefficient approach to exploring the phase-space of an element (or set

of elements). For a GAP model to be transferable, it must be trained on a data set which contains within it as varied a set of environments as possible; the challenge is in generating environments which are both varied and physically relevant. One approach which has been shown in the literature to be highly efficient at doing precisely this is the *ab initio* random structure searching (AIRSS) approach. In the AIRSS algorithm, one first begins with a randomised (and therefore very high-energy) initial configuration of atoms. The geometry of this configuration is then optimised - typically with DFT, in order to identify various minima in the PES. Because of the highly stochastic nature of this approach, a large number of minima may be identified in a computationally efficient manner. Often a number of these minima coincide with the known ground states of a material (indeed a primary appliation of the method is the identification of *unknown* ground states), but still more will correspond to higher energy local minima. Such an approach could employ an entirely DFT driven structure search, in which case the inclusion of a wide variety of structures will produce a more transferable model than one produced using the bottom-up approach, however this is still not optimal. For a GAP (or other ML model) to be truly robust, it must be provided with information on not just which structures are stable, i.e. low in energy, but also those which are high in energy. Examples of such high-energy structures could be as simple as atomic dimers at short separation, or more complicated such as 5-coordinated carbon atoms; for a GAP model to predict a reliably high-energy for such configurations, they must be included in the training data, however they will never be identified in a DFT driven structure search (they are too unstable!).

The solution is to drive the optimisation using GAP energies and forces, rather than directly those from DFT [152]. In the initial stages of the development of a model, many of the minima in the PES which will be found by a RSS employing a GAP model will be spurious; early GAP models will wrongly predict arrangements of atoms to be minima in the PES in fact they would be very high in energy if the energies were computed using DFT. By sampling these configurations and computing the correct energies and forces with DFT, these high-energy structures can

be included in the training database. Subsequent GAP models will then correctly predict the energies of these high-energy structures in the future. We note here that iterative training is often employed using configurations from GAP driven MD trajectories as well as random structure searches, particularly if these MD trajectories are performed at high enough temperatures for significant bond reorientation to occur. The concept of iterative training is most naturally described in the context of the 'top-down' approach to database generation, but is a broadly applicable feature of almost all training data sets for GAP models.

This process of structure searching (either with an RSS-type algorithm or MD) is iterated multiple times, to slowly accrue a larger database of configurations, which are highly structurally diverse and represent not just important global minima but various important high-energy structures as well. A model trained on a data set produced in this way could be reliably expected to perform well even at high temperatures, without generating unphysical configurations, it might also be expected to make accurate predictions of the energies of unstable configurations or local maxima (e.g. reaction barriers). At the same time, since some of the minima identified in the random structure search will be highly likely coincide with ground state crystal structures (e.g. diamond, graphite) a reasonable degree of accuracy could be expected here also, though perhaps less than if configurations sampled from targeted MD simulations were included. GAP-RSS alone will rarely explore all of the scientifically important minima, however - point and line defects in otherwise crystalline materials are not commonly identified, yet they are vital for many simulation studies and material applications. In situations where these configurations are known to be relevant to the intended application, they must be included manually as described in section 3.1.1.1. In practise, the optimal approach is often a hybrid between both of these methods; random structure searching and iterative training is employed to ensure transferability and stability, while directed simulations and sampling are performed for the particular regions of phase space where higher model accuracy is a requirement.

3.1.2 Farthest Point Sampling

Whether the 'bottom-up' or 'top-down' approach is employed for the generation of training data, or some combination thereof, it will often be the case that a greater amount of training data has been generated than can reasonably be used as input configurations for the training - it is necessary to select a sub-sample of these configurations prior to beginning the fitting process. In the case of GAP-20, we allow the bulk of our training configurations to be chosen from the total dataset using a sampling method known as farthest point sampling (FPS) [148, 62]. Since the cost of not only training, but also evaluating a GAP model scales with the number of training configurations used, it is highly desirable to make an informed choice about which configurations to include and exclude from training. Most importantly, we wish to exclude the use of the many highly correlated configurations which might be explored during a typical molecular dynamics run used to generate training configurations. It is important to distinguish the selection process involved in FPS from the sparsification procedure employed during the fitting itself. The sparsification (selection of representative points) discussed in section 2.3.1 operates on individual atomic environments (in the case of SOAP, this corresponds to individual descriptors) and occurs *during* the determination of the weights, i.e. during the fitting procedure. The FPS discussed here is an approach aimed at pre-filtering entire configurations prior to any training being performed. This means that configurations eliminated from the training data via FPS have no impact on the weights of the Gaussian process whatsoever. Of course, since the FPS is aimed at eliminating very highly correlated configurations, it is highly probable that frames which are eliminated from the training dataset correspond to regions of phase space which are already represented by other, similar, configurations.

Given a set of n descriptors of type d for a number of frames, $\mathbf{Q} = \{\mathbf{q}^{\mathrm{d,avg}}_{i=1\ldots\mathrm{n}}\}$, which are themselves the average of the individual descriptors of the atoms in a particular frame $\mathbf{q}^{\mathrm{d}}_{\mathrm{i}}$, the FPS algorithm selects configurations so that at each step, the kernel distance between previously selected configurations $\mathbf{Q}_{\mathrm{selected}} =$

$\{\mathbf{q}_1^{\mathrm{d,avg}} \ldots \mathbf{q}_{\mathrm{m}}^{\mathrm{d,avg}}\}$ and the new configuration $\mathbf{q}_{\mathrm{m+1}}^{\mathrm{d,avg}}$ is maximised. That is,

$$\mathbf{q}_{\mathrm{m+1}}^{\mathrm{d,avg}} = \mathrm{argmax}_{\mathbf{q}^{\mathrm{d,avg}}}[\mathrm{D}(\mathbf{Q}_{\mathrm{selected}}, \mathbf{q}^{\mathrm{d,avg}})], \tag{3.14}$$

where D is the kernel distance between the selected descriptors (and associated frames) $\mathbf{Q}_{\mathrm{selected}}$ and the candiate descriptor $\mathbf{q}^{\mathrm{d,avg}}$. 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 [150].

3.1.3 Selection of Model Hyperparameters

An important practical aspect of training a GAP model which is both stable and transferable enough to perform large-scale molecular dynamics while still achieving the desired target accuracy is the selection of the model hyperparameters. While an oft-discussed approach in the broader machine learning literature to the selection of hyperparameters for Gaussian process regression is their selection via a log-marginal likelihood optimisation, this is in practice not possible with GAP models due to the prohibitive cost which would be associated with the process. It is also an empirical observation that for the specific case of interatomic potential generation, such approaches to hyperparameter optimisation are not strictly necessary. Generally, GAP model hyperparameters can be effectively chosen from considerations of the physical characteristics of the system of interest and the distribution and quality of the training data available. They key hyperparameters which must be considered can be separated into two broad categories. There are those associated with the Gaussian process regression itself and those which are more related to the specific choice of descriptor, which we will consider individually in detail below. Considering the first, the key hyperparameters which are inherent in any sparse Gaussian process regression are the values for σ used to regularise the kernel and the number of representative points (sparse points). We also include here the values for δ used to weight each descriptor's contribution to the total energy. The cutoff radius is another key hyperparameter which is common to any interatomic potential (ML or otherwise). In addition to the question of which descriptors to use, a number of

descriptor-specific hyperparameters are also worthy of consideration. Here we will focus primarily on those related to the SOAP descriptor, which is used and referenced frequently throughout this work, specifically, these are the basis set expansion used to represent the neighbour density (l_{max} and n_{max}) and the power to which the SOAP kernel is raised, ζ.

3.1.3.1 Representative Points

The first and most obvious hyperparameter which is intrinsic to any application of sparse Gaussian process regression is the choice of the number of representative points used in the fitting process. In principle, the accuracy of a GAP model would be expected to improve systematically with an inclusion of a greater number of representative points. However, as noted, the cost of both model fitting and evaluation also increases as the number of representative points is increased. In this regard, the number of representative points may be considered as a convergence parameter; the optimal selection for the number of representative points will almost always be the greatest number which is computationally tractable. In all of the work presented in this thesis, it is this approach which has been used to select the number of representative points, the largest number which is computationally feasible to train on with the available computational resources is selected. A point of note is of course the question of how the representative points should be selected - a GAP model trained on the same number of representative points selected using different methods would not be expected to perform equivalently. A discussion of the specific methodologies used to select each individual representative points is given in section 2.4.0.1.

3.1.3.2 Descriptor Weights

We show in equation 3.5 that the potentials presented in this thesis are constructed as the sum of energetic contributions from three separate descriptors; a two-body, three-body and SOAP descriptor. The proportion of the total energy which comes from each descriptor is tuned by the hyperparameter δ, of which there is one for each descriptor. Here, again, our choice of the values of these hyperparameters is physically motivated. It is a common feature of other interatomic potentials and an empirical observation of many physical systems that much of the interaction energy

can be captured from simple two-body interactions. Three-body and higher order n-body interactions typically have a decreasing contribution when expressing the total energy of a system as a body-order expansion. We use this motivation to select initial values for δ for the GAP potentials given here, the largest δ is given to the two-body interaction, reflecting the fact that a large portion of the total energy may be expressed in terms of this low-dimensional descriptor. The absolute value selected for δ is not terribly important, the key factor is the ratio of the values of δ for the different descriptors. A common procedure for selecting these hyperparameters might be to initially set the δ_{2b} to a value of 1.0 and train a GAP model using just this descriptor. The performance of the model is then evaluated on an independent test set of configurations, and the error in the energy prediction measured. One then sets the δ for the next highest-order term to be equal to the square of the energy error and a new model is trained using just the first term (typically this will correspond to a 2b + 3b potential). This process is repeated to determine an initial value of δ_{MB} for the SOAP (or other higher order) descriptor.

While this process is useful for determining approximate initial values for δ, further modification is sometimes necessary. For example, if a GAP model frequently produces spurious results in the form of unphysical minima, a suitable solution might be to reduce the weight of the SOAP descriptor, to place a greater significance on the lower-dimensional descriptors. This is usually successful due to the fact that spurious minima often occur for regions of phase space not suitably sampled in a given descriptor space, resulting in large fluctuations in the predicted energy. Lower-dimensional descriptors are populated much more effectively with fewer data and are thus less likely to have such 'holes' in their training. Conversely, if an important physical observable is inaccurately predicted by a GAP model and additional training data does not improve the quality of the prediction, it may be advisable to increase the δ for the higher-dimensional descriptors. An example of this might be surface reconstructions or defect formation energies; the properties of specialised configurations such as these will not be accurately reproduced without a sufficiently flexible (high-dimensional or non-linear) model.

3.1.3.3 Kernel Regularisation

The regularisation of the kernel is another key hyperparameter (or set of hyperparameters) which must be chosen. Recall that regularity broadly refers to the smoothness of the potential, a more regular model will result in the energies and forces being smoother functions of the descriptor space. One of the key points in approaching the regularisation of GAP models is by taking the view introduced in section 2.3.2 of the regulariser as being proportional to the noise present in the dataset. Broadly speaking, a larger regulariser would correspond to a larger assumed variance present in the input data. Although in principle, a single value of σ may be used to regularise the kernel in equations 2.37 and 2.32, it is in practise valuable to specialise these values to account for the type and source of the data considered; recall that each individual derivative observation and atomic energy can have associated with it a separate value for σ.

In identifying potential sources of noise in the training data and fit, it is first worth noting that the computed values for the total energies and forces from the DFT calculations themselves will have errors associated with them. Ignoring for the time being any other potential sources of error related to interatomic potentials themselves (e.g. locality) we can consider the potential sources of noise present in the DFT data itself. As a variational method, the energy computed from a given DFT calculation will be converged to within a particular energy threshold. This threshold (the DFT convergence criterion) represents a lower bound on the accuracy which is potentially achievable by a GAP model trained on that data, typically $10^{-4} - 10^{-6}$ eV. Furthermore, since the forces are computed from the local derivative of this energy, they can be expected to have a larger error than the energy itself, it would thus be reasonable to attribute to this a larger value for σ. Similarly, the virial coefficients are given as derivatives of the total energy with respect to a particular lattice distortion and will thus also have a different associated error. A practical example of this is used in chapter 4, where different values for σ are chosen for the energies, forces and virial coefficients respectively (a similar technique is also employed in chapter 5). Although in practise the selection process required some

trial and error, eventually values of $\sigma_E = 0.001$ eV, $\sigma_F = 0.01$ eV and $\sigma_V = 0.05$ eV were chosen for the energy, force and virial stresses regularisation respectively.

However this is not the only source of noise in the training data which must be taken into consideration. A further important factor is that not all configurations present in a training data set are necessarily from comparable physical conditions (in fact it is frequently the case that they are not). Consider, for example, the case of a potential such as GAP-20, which is trained on (amongst other things) crystalline graphene and high-temperature liquid carbon. Both the requirements of the accuracy of the potential and the reasonable accuracy which could be expected from a GAP model for each of these regions of phase space will be very different. The desired accuracy for crystalline regions of phase space is much higher than those of the high-temperature liquid or amorphous phases. For example, in order to reliably reproduce the phonon dispersion curves of graphene, a candidate model might be expected to have errors in the forces on the order of 10 meV Å^{-2}. Such a degree of accuracy is neither necessary nor feasible to obtain for a high-temperature liquid phase, where the fluctuations in the energies and forces measured are orders of magnitude greater than those necessary for modelling low-temperature crystal structures. Furthermore, in systems with a greater conformational variety, such as liquids, the magnitude of the errors resulting from the finite cutoff of the interatomic potential will also become greater (an excellent in-depth analysis of this locality effect for carbon is given in Ref. [106]). This physical difference in the sources of training data can be reflected in the regularisation of those data points during fitting. For example, in the work presented in chapter 5, rather than the value of σ_E = 0.001 eV used for the crystalline phases, a value of $\sigma_E = 0.05$ eV was selected for liquid carbon. Similar groupings of configurations and their associated regularisers are given in table B.2. This selection of regularisation parameters allows the GAP model to be regular and transferable in regions where high accuracy is unnecessary, while maintaining a high degree of accuracy for those regions of phase space where it is required.

In general, an approximate approach to selecting values of σ for a well defined

crystal structure would be to say that the target accuracy for σ_E should be roughly equal to the square of the values selected for σ_F and σ_V - reflecting the fact that the error in the energy scales as the square of the atomic displacement, while the error in the force scales linearly with atomic displacement. For well-converged DFT calculations with a plane-wave code, it is an empirical observation that values for σ_F of approximately 0.01 eV Å^{-2} are reasonable, with associated values for σ_E of 0.001 eV. When considering to amorphous and liquid structures, it is usually advisable (again, from empirical observation) to increase these values by roughly one order of magnitude to 0.01 eV and 0.1 eV Å^{-2} for the energies and forces respectively; this approximately accounts for both the lower requirements for accuracy and the additional sources of error introduced.

3.1.3.4 Descriptor Cutoff Radius

The radial cutoff distance is a parameter which is common to almost all interatomic potentials, ML based or otherwise. It is no surprise, then, that it is also one of the most important hyperparameters to consider in parameterising the descriptors for a GAP model. The GAP methodology describes the total energy as a sum over a number of local contributions, whether these are represented as bond distances, angles or SOAP vectors, the radial cutoff determines what is considered in these local sums and therefore which interactions contribute to the total energy. By construction, the magnitude of the interaction between atoms outside of the cutoff radius will be 0. That locality is imposed on such a fundamental level in a GAP model has implications for the sorts of physical interactions which can be expected to be modelled. A key point of note is that long-ranged interactions, particularly charge-charge interactions and van der Waals dispersion interactions cannot easily be modelled for atoms at large separation within the base GAP framework.

As with the hyperparameters discussed previously, the selection of the cutoff radius must take into account the physical length scales relevant for the properties of interest of the system, rather than utilising the cross-validation or Bayesian optimisation protocols found elsewhere. While this may at first seem trivial, it has implications for the construction of GAP potentials which are at first unintuitive

from a purely machine learning perspective. In particular in the case of carbon, as is discussed in section 5.3, to correctly capture important physical effects, a value for the the cutoff radius is chosen which is sub-optimal from a purely energetic point of view - but which is absolutely necessary for capturing the fundamental structure of graphite. This is discussed in more detail in the relevant section of the thesis (section 5.3) but can be summarised as follows. In figure 5.3 the minimum force error on a validation set of structures is obtained for a SOAP descriptor cutoff radius of 2.9 Å. This distance is significantly smaller than the distance between individual layers in bulk graphite. To select a cutoff radius of 2.9 Åwould mean that by construction, graphitic layers would be non-interacting in the bulk crystal; the implication being that graphite would be unstable with respect to its dissociation into individual graphene layers. Clearly, although in absolute terms of energy this error is small (the interaction strength between graphitic layers is very weak, on the order of 50 meV/atom), the implications for simulations employing a GAP model trained in this way would be significant.

It might initially be tempting to try and solve both the problems of long-ranged interactions by increasing the cutoff radius of the descriptors to very large values. This is problematic for two reasons, both of which are exemplified in figure 5.3. Firstly, note that the force error of the GAP model shown here does not decrease monotonically with increasing descriptor cutoff, despite the inclusion of a greater number of neighbour atoms. Instead, as mentioned, a minimum at 2.9 Åis observed, with a gradual increase in the force error seen for greater cutoffs. This error arises from the fact that the volume of the descriptor space of a SOAP descriptor (and indeed any descriptor) increases significantly as the cutoff is increased. This greater descriptor volume requires a greater number of data points to populate it completely (i.e. a greater number of training configurations *and* representative points) which may not always be computationally feasible. There is also the fact that the cost of the GAP model increases significantly as the radius of the SOAP descriptor increases; a doubling of the SOAP descriptor cutoff roughly corresponds to a doubling of the cost of model evaluation. These factors combined play a key role in the

choice of descriptor cutoff employed in both chapters 4 and 5 and are significant motivators for the decision to employ a semi-analytical long-range term to account for vdW interactions in chapter 5

3.1.3.5 SOAP-Specific Hyperparameters

Although three types of descriptors, two-body, three-body and SOAP, feature heavily in the work presented in this thesis, only the SOAP among these has additional hyperparameters (beyond the cutoff radius) which are worthy of discussion. In the construction of the SOAP descriptor, the neighbour density is expanded in a basis set of spherical coordinates (equation 3.10). This basis set is truncated for some value n_{max} for the radial and l_{max} for the angular component. As this basis set is made more complete, both the accuracy and cost of the potential increase; a choice of larger l_{max} and n_{max} produces a larger descriptor vector and a more accurate GAP model. Panel C in figure 5.3 shows the behaviour of both the cost and accuracy of a potential trained on a moderately sized carbon database as both l_{max} and n_{max} are increased. A key point of note is that, although a choice of $l_{max} = n_{max}$ has frequently been used in the literature, this is not in fact that optimal choice (at least for carbon). Rather, it is beneficial to have a larger radial component of the radial basis set expansion with a smaller angular component ($n_{max} > l_{max}$). This asymmetry has a potentially very significant impact on the performance of a model (as can be seen in panel C of fig 5.3, and should be considered when selecting the hyperparameters of the SOAP descriptor.

Another hyperparameter worthy of brief discussion is the power to which the dot product of two SOAP vectors is raised when computing the kernel function, ζ (equation 3.13). A value of $\zeta = 1$ corresponds to a linear kernel, which results in a model which can be written as a sum over triplets of atoms (a three-body model). The reasoning behind this is complex and is discussed in more detail elsewhere [153]. Raising the the SOAP dot product kernel to a power of 2 ($\zeta = 2$) results in a dependence on four neighbours (five including the central atom). As ζ is increased further, the body-order of the interactions considered by the SOAP descriptor is given by $2\zeta + 1$. Although there is no additional cost associated with increasing the

value of ζ used, to date, no systematic study has been performed examining how GAP model performance varies for different values. As an empirical observation, most successful models have employed values of either $\zeta = 2$ or $\zeta = 4$, particularly in the case of the carbon potentials discussed in chapters 4 and 5, a value of $\zeta = 4$ is chosen.

3.1.4 Testing and Validation

In practise, a significant portion of the time and effort which must be devoted to the construction of a GAP potential is spent on testing and validating the model. In fact, the bulk of the results chapters of this thesis (chapters 4 and 5) is devoted to precisely this process. As such, to avoid redundancy we will not provide here a detailed discussion of precisely what properties should be computed to validate a GAP model, but rather give a general overview of the important points to consider which may be lacking elsewhere.

When attempting to validate a model, one should aim to test on as widely varied a set of properties as is possible (within the intended scope of application of the potential). Typically, this process should go beyond the obvious step of validating that a GAP potential is able to correctly predict the forces on an independent test set of configurations, which is a necessary but not sufficient test of a potential. A suitable strategy often involved first identifying material properties which will be relevant to the intended application of the model. For example, if surface interactions or interfaces are to be studied, then a suitable model might be expected to reproduce the surface energies of the low Miller-index faces of the crystal structures, along with any reconstructions associated with them. Other suitable physical properties to compute might be geometrical properties of any species adsorbed on the surface. For example, the distribution of the density or angular orientation of the species as a function of the distance from the surface. Care must always be paid to ensure that suitable reference values for the computed properties can be found - in the case of dynamical and structural properties of liquids or amorphous systems, this can often involve performing AIMD simulations to produce reference values, which can be prohibitively expensive. Further consideration must be paid to the fact

that the quality of the fit to a particular reference electronic structure method is not necessarily the end-goal of a particular model, rather it is the ability of the model to predict experimental quantities is desired. It may be tempting to consider the reference electronic structure method to be the 'ground truth' against which accuracy should be measured, but in practise there may often be a discrepancy between the physical quantities computed using electronic structure theory calculations and experiment. As such, even an idealised 'perfect' GAP model, which achieves a perfect fit to the reference data, would still have an error when compared to experimentally derived quantities. It is often, therefore, desirable to instead validate a potential against experimental quantities where applicable, as is done in chapter 4 of this thesis. A particular example of this is the case of the phonon dispersion curves, for which a comparison to experimentally determined x-ray diffraction data is given. In this case, the larger portion of the error is associated with the discrepancy between DFT and experiment, rather than the quality of the GAP fit to the reference data. Such an approach necessitates that the system of interest has been suitably studied by experimental groups, however, which does somewhat limit the approach of GAP model validation using experimentally available quantities to systems which are already well-known and studied. Graphene is one example of a system where this works well, but it is often the case that an interested user will wish to develop a GAP model because a material is new, interesting and therefore poorly studied experimentally.

From a practical perspective, during the development of a GAP potential, it will often be necessary to evaluate the performance of numerous candidate models repeatedly. In some cases this may have to be performed many hundreds of times, in which case a manual approach would be at best a highly tedious process. A framework has been developed by other members of the community, which is available at `https://github.com/libAtoms/testing-framework` to aid with this process. The testing framework provides a convenient way to evaluate many GAP model, or indeed any model callable from the Atomic Simulation Environment (ASE) in an automated process. It also facilitates the development of 'tests' as modular scripts

which can be executed independently from one another. A set of basic tools, which employ functionality from ASE, are already present for performing common computational tasks such as geometry optimisations, phonon calculations and the computation of elastic properties. There numerous additional examples for materials such as phosophorous, silicon and boron present, from which inspiration may be drawn. The framework in written in Python and therefore has all of the flexibility and integration with other available libraries and packages which one might require.

3.1.5 Using the GAP Code for Training and Evaluation

Aside from the more formal, general or methodological points discussed above, we take here a moment to discuss in practical terms the GAP code which is used for training and fitting. The code used for model training and evaluation in this thesis is available as a software package written in FORTRAN, called the Gaussian approximation potential (GAP) code, which is available online under a non-commercial license at the following link: `http://www.libatoms.org/gap/gap_download.html` GAP itself is not a standalone piece of software but a plugin, in order to use it, it must be compiled together with the open source software package named QUantum mechanics for Interatomic Potentials (QUIP). QUIP is a collection of tools written in FORTRAN, aimed at performing molecular dynamics simulations and serves as a useful interface between codes used to compute energies and forces (e.g. GAP, DFTB) and the end user. Similarly, QUIP is freely available for download: `https://github.com/libAtoms/QUIP`. The functions comprising QUIP (and hence a GAP potential) may be called in Python (through quippy), FORTRAN or compiled as a library to act as an interface with other common simulation packages, for example the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) code [154]. This is particularly useful as it allows simulations employing GAP potentials to make use of the mature, well-developed and highly maintained environment of simulation techniques available within LAMMPS. As before, LAMMPS is freely available to donwload at the following link, `https://www.lammps.org/download.html`. Detailed instructions on how to compile GAP and QUIP together will not be provided here, as that is out of the scope of

this thesis and may become outdated, up-to-date set of instructions can be found online, however: `https://libatoms.github.io/GAP/installation.html`.

Chapter 4

A Machine Learning Potential for Graphene

In this chapter we use the GAP methododology [60] to generate an accurate ML interatomic potential for graphene, with the aim of examining the limit of the attainable accuracy of the GAP methodology for a single carbon phase. A secondary aim is to establish how this accuracy compares with those of empirically constructed many-body potentials. We evaluate the quality of the prediction of atomic forces of our GAP model and a number of empirical potentials versus a reference DFT method. We also compare predictions of the finite temperature phonon spectra of graphene with experimental results, where we find excellent agreement. We further compare the predictions of our GAP potential to those from *ab initio* molecular dynamics (AIMD) simulations of the thermal expansion of graphene - a property which has historically been very challenging for interatomic potentials to predict[155, 30, 156, 157]. We show thereby that for the case of graphene, machine learning potentials have the capability to act as a substitute for direct *ab initio* calculation, at a much reduced cost and only marginally compromised accuracy. This capability will be particularly valuable in instances where accurate descriptions of dynamics are mandated, such as the description of the diffusion of small molecules on the graphene surface [34] and the treatment of nuclear quantum effects via path integral molecular dynamics [38, 39].

The remainder of this chapter will be structured as follows; in section 5.3 we

provide an outline of how the GAP model is constructed, section 4.1 outlines how the *ab initio* configurations and training data were generated. Sections 4.2 to 4.4 are concerned with the evaluation and benchmarking of the potential, considering first the force accuracy, followed by the phonon spectra and thermally induced Raman band dispersion, lattice parameters and thermal expansion. We give our conclusions in section 4.5.

4.1 Generation of Training Data

Our training data are generated from tightly converged plane-wave DFT calculations performed on configurations sampled from various molecular dynamics trajectories. While the atomic configurations herein are generated using a variety of methods (MD with existing potentials and various iterations of our GAP model) the values for atomic forces, virial stresses and energies which comprise the training dataset have all been calculated using precisely the same level of DFT. For these calculations, we use the VASP plane-wave DFT code [158, 159, 143], with the optB88-vdW dispersion inclusive functional [160, 141] with a projector augmented wave potential [161], a plane wave basis cutoff of 650 eV and Gaussian smearing of 0.05 eV [162, 140]. We use a dense reciprocal space Monkhorst-Pack grid [163] with a maximum spacing of 0.012 Å^{-1}. In order to ensure a low degree of noise on the calculated forces, the energy convergence criterion for the SCF iterations was set to 10^{-8} eV. We choose the optB88-vdW functional as it has already been shown to provide an excellent description of graphitic carbon [142]. We further evaluate the sensitivity of our predictions to this choice, by comparing against other common exchange-correlation functionals, which is discussed briefly in section 4.2 and the details of which are given in the Appendix B. We have generated our training data so as to have a dense sampling of a specific region of phase space, with the aim of exploring the optimum accuracy possible for a particular allotrope, this approach is distinct from that used in the generation of the previously published amorphous carbon potential wherein the training set was chosen to maximise the transferability of the potential [106].

The first set of training data was generated from three MD simulations of a free-standing graphene sheet comprised of 200 atoms with lattice parameter a = 2.465 Å. Simulations were performed in the NVT ensemble at temperatures of 1000, 2000 and 3000 K. Trajectories were generated using the LAMMPS [154] open source molecular dynamics program, interactions were modelled using the LCBOP many-body potential for carbon and a Nosé-Hoover thermostat was used to maintain a constant temperature over the simulation. A total of 100 configurations were sampled from each of the three 2 ns trajectories at 20 ps intervals, the total energies and forces of these atomic configurations were then calculated using VASP as outlined above.

An initial GAP model was generated using the *ab initio* quantities computed on the 300 configurations. A further set of molecular dynamics trajectories were generated as above, but with interactions now computed using the preliminary GAP model. Simulations were performed between 300 and 3000 K at a fixed lattice parameter of a = 2.465 Å, a sample of *ab initio* energies, forces and virial stresses from these configurations was added to the training set to produce a second GAP model. A number of iterations of improvement were performed using this approach, the final dataset was comprised of 1083 configurations of 200 atoms at temperatures between 300 and 4000 K and lattice parameters between 2.460 and 2.480 Å.

A random sample of 5% of these configurations was withheld as a validation set to benchmark the quality of the GAP fitting procedure. The parameters used for the fitting of the GAP model are shown in Table. 4.1. Additionally, we choose the expected error (analogous to the target closeness of the fit of the GAP model to the training data) in energies to be $\sigma_{\mathrm{E}} = 10^{-3}\,\mathrm{eV}$, for forces we choose $\sigma_{\mathrm{f}} = 5 \times 10^{-4}\,\mathrm{eV}$, and for virial stresses $\sigma_{\mathrm{v}} = 5 \times 10^{-3}$. For full reproducibility, we provide here the complete training command line provided to the gap_fit program used for fitting the graphene GAP model discussed here:

```
at_file=Graphene_GAP_v2.2.1.xyz gap={distance_2b cutoff=4.2
n_sparse=50 covariance_type=ard_se delta=10.0 theta_uniform=1.0
sparse_method=uniform : angle_3b cutoff=4.2 n_sparse=200
```

	2b	3b	SOAP
δ(eV)	10	3.7	0.07
r_{cut}(Å)	4.0	4.0	4.0
w_{cut}(Å)	1.0	1.0	1.0
Sparse Method	uniform	uniform	CUR
N_t	50	200	650

Table 4.1: Additional parameters used for the training of the GAP model. δ indicates the relative weighting of the different descriptors, r_{cut} indicates the cutoff width of the descriptor and w_{cut} indicates the characteristic width over which the descriptor magnitude goes to 0. 2b, 3b and MB indicate the two body, three body and many body descriptors used in the construction of the potential. N_t indicates the number of sparse points chosen for each descriptor during training, while the sparse method denotes the method by which sparse points were chosen. More information can be found in the GAP code documentation at [http://www.libatoms.org].

```
covariance_type=ard_se delta=3.663 theta_uniform=1.0 :
soap l_max=8 n_max=8 atom_sigma=0.4 zeta=2 cutoff=4.2
cutoff_transition_width=1.0 n_sparse=650 delta=0.1
covariance_type=dot_product sparse_method=cur_points add_species=F}
default_sigma={0.005 0.0005 0.005 0} sparse_jitter=1.0e-12
hessian_parameter_name=dummy virial_parameter_name=virial
energy_parameter_name=energy force_parameter_name=force
gp_file=Graphene_GAP_v2.2.2.xml
```

4.2 Force Prediction

The first natural metric for the quality of a potential - in particular one of a machine learning origin - is the quality of the forces it predicts relative to an appropriate reference. We choose a random sample of 1.5×10^4 atomistic reference points from our data and compare the forces as predicted by our model to those from DFT. Additionally, we compare the forces predicted by a number of other popular methods for atomistic modelling; DFT with common exchange correlation functionals, density functional tight binding (DFTB), a number of empirical many body potentials; Tersoff, REBO, AIREBO, AIREBO-Morse and LCBOP, a ReaxFF potential parameterized for condensed carbon and the recently published GAP model for the

Potential	Force Error (In-plane) [eV Å^{-1}]	Force Error (Out-of-plane) [eV Å^{-1}]	Lattice parameter (0 K) [Å]	Time (Relative)
Graphene GAP	0.028	0.019	2.467 (+0.003)	340
Amorphous GAP	0.270	0.258	2.430 (-0.03)	380
Tersoff	3.122	0.542	2.530 (+0.08)	1
REBO	0.722	0.187	2.460 (-0.004)	1.2
AIREBO	0.548	0.414	2.419 (-0.05)	1.9
AIREBO-Morse	0.720	0.568	2.459 (-0.005)	2.9
LCBOP	0.595	0.306	2.459 (-0.005)	2.3
ReaxFF	1.226	0.311	2.462 (-0.002)	23
DFTB	0.693	0.162	2.470 (+0.006)	950
DFT (optB88-vdW)			2.464	2×10^7 (AIMD)
Exp.[142] Graphite, 300 K			2.462	

Table 4.2: Root mean squared force errors, lattice parameters predicted and relative costs of empirical many-body and GAP models. The details for other common DFT functionals tested are available in the SM.

amorphous phase, all of which have been used in their originally published forms [51, 40, 41, 42, 45, 47, 43, 106]. Force errors for the graphene GAP, LCBOP, Tersoff and DFTB methods are shown in Fig. 4.1, where we have separated the data into forces in the 'in-plane' directions and those in the 'out-of-plane' direction. Root mean squared errors (RMSE) are given for all methods in Table 4.2, plots of force correlations and errors for all methods can be found in the SM. We calculate the cost of each of the methods over 10^4 identical MD steps for 200 atoms, which we normalise for the number of cores on which the simulation was run.

Fig. 4.1 shows that the predictions of the graphene GAP model align very closely with those of the reference DFT method. Forces are obtained with an RMSE of 0.028 eV Å^{-1} in the in-plane direction, and 0.019 eV Å^{-1} in the out of plane direction. The errors obtained from the DFTB and LCBOP methods are much larger, RMS errors in forces are 0.69 and 0.55 eV Å^{-1} respectively and maximum errors of 2 eV Å^{-1} are observed in the worst cases. Errors are largest for the Tersoff potential, for which the RMSE is measured as 3.1 eV Å^{-1} with a maximum in excess of 11 eV Å^{-1}. Despite the AIREBO-Morse potential being a more recent iteration of

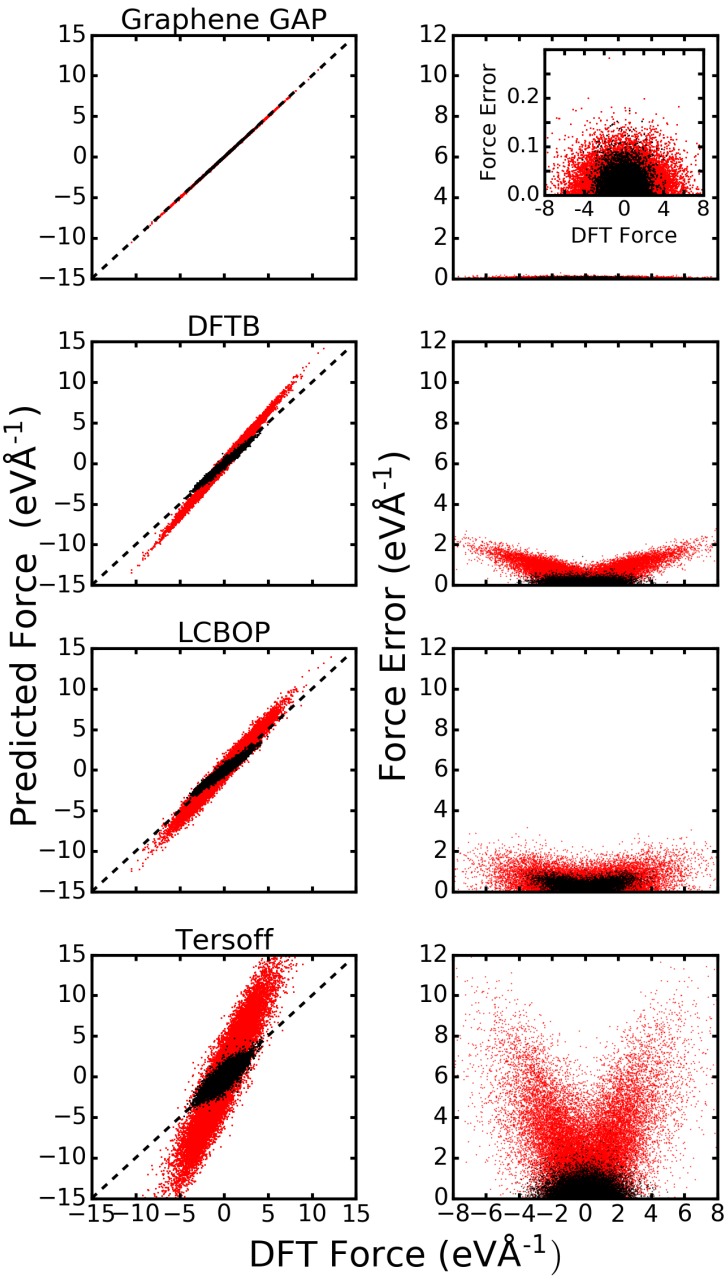

Figure 4.1: Force correlations (left) and associated force errors (right) on an independent reference dataset of configurations for the graphene GAP model, DFTB, LCBOP and Tersoff potentials as compared to the reference DFT method, the plots for all methods considered can be found in Appendix A. Black points indicate forces perpendicular to the plane of the graphene sheet (out-of-plane) while red points indicate forces oriented in the plane. The inset in the graphene GAP plot has a different scale on the y-axis to show more clearly the distribution of force errors, which are smallest for large forces with a Gaussian distribution.

the AIREBO potential (including a Morse potential to model bonding interactions) we find that the modifications are actually a detriment to the quality of the predicted forces, despite the increased cost (Table 4.2).

It is important to briefly consider how these conclusions may be affected by the choice of reference method; there are many instances in the literature of disagreement between various exchange correlation functionals and it is important to evaluate the importance of this in the context of graphene, the details of which we give in Appendix B. We find that there is a minimal dependence of the measured forces on the choice of exchange correlation functional for this system, on average 0.026 eV Å^{-1} in the in-plane and 0.018 eV Å^{-1} in the out of plane direction - indicating that the relative ranking of the benchmarked methods would be the same irrespective of the chosen reference method. Furthermore, the expected performance of the graphene GAP model would also be insensitive to this choice. This is supported by the similarity in the phonon spectra calculated with each of the functionals, which are also available in Appendix B.

4.3 Lattice Parameters and In-Plane Thermal Expansion

The lattice parameter is a fundamental property for any atomistic model of a material to predict. Many intrinsic properties of materials such as graphene are affected by the lattice constant, while the degree and type interaction between two distinct materials can vary dramatically based on the degree of lattice matching between their two structures [164]. In addition to the ground state lattice parameter, the thermal expansion of graphene is also of interest as it provides insight into the relative strengths of the in-plane and out of plane forces, the anharmonicity of the bonding interactions and the coupling between harmonic and anharmonic vibrational modes.

The nature of the thermal expansion of graphene is, however, a topic wherein many conflicting computational reports may be found [155, 30, 156, 157]. The experimental coefficient of thermal expansion of freestanding graphene is generally accepted to be negative at moderate temperatures - low lying bending phonon modes

cause graphene to 'crumple' and thus shrink in the in-plane direction [155, 30]. Graphene has been found from Raman spectroscopy and micromechanical measurements to have a negative in-plane coefficent of thermal expansion at temperatures between 30 and 500 K [156, 157]. However, graphene must typically be investigated experimentally while adsorbed on a substrate material, the strain induced from this significantly affects both its 0 K lattice parameter and the thermal expansion of the material, leaving the study of freestanding graphene as a particularly attractive topic for theoreticians [29, 165]. *Ab initio* investigations broadly agree in their prediction that the CTE of graphene is negative over a moderate temperature range - but differ in their predictions at higher temperatures. Results from DFPT show non-monotonic behavior, a negative and in-plane coefficient of thermal expansion up to 2000 K, with a minimum at 300 K [166]. Green's function lattice dynamics calculations have found the sign of the CTE to change from negative to positive at temperatures above 500 K and AIMD simulations have found the CTE to be weakly negative over a large temperature range [167, 29]. Results from studies employing empirical potentials vary more substantially, the REBO potential predicts a positive CTE over a wide temperature range, the Stillinger-Weber and LBOP potentials predict the CTE to be entirely negative and the LCBOP and LCBOPII [46] potentials predict a change in the sign of the CTE around 500 K [27, 30].

We now compare to lattice parameters over a range of temperatures as predicted by *ab initio* molecular dynamics simulations of graphene sheets using the method established in Ref. [29]. In-plane lattice parameters were averaged over AIMD simulations on freestanding graphene sheets containing 200 atoms between 60 and 2500 K. Calculations were performed at the gamma point, using the optB88-vdW functional and a projector augmented wave potential with a plane wave cutoff of 400 eV, in the NPT ensemble as implemented in VASP, with the constant pressure algorithm applied only in the lateral directions (in-plane) [29, 27, 168]. Three independent simulations at each temperature were conducted and statistics were collected for between 40 and 95 ps depending on the temperature until the lattice parameter was converged to within 10^{-4}Å. We note that this approach neglects the

effect of the zero-point vibrational energy (ZPE) on the calculated lattice parameter and thermal expansion. The inclusion of this has previously been found to increase the ground state lattice parameter of graphene by 0.3% [166]. The effect of ZPE could be included via path-integral type methods, but we consider this unnecessary for the benchmarking purposes of the current study.

Lattice parameters for the empirical and GAP potentials were determined similarly. We performed NPT simulations using the Nosé-Hoover thermostat on free-standing graphene sheets containing 200 atoms. Simulations were equilibrated for 5 ns and statistics collected on three replica simulations over a further 5 ns for each potential, in each case the time averaged lattice parameters were converged to within 10^{-4}Å. The coefficient of thermal expansion of graphene is calculated as,

$$\mathrm{CTE} = \frac{1}{A_T}\frac{\partial A_T}{\partial T}. \tag{4.1}$$

Here, A denotes the area of the graphene sheet and T the temperature in Kelvin. To calculate the CTE we interpolate between calculated data points by fitting splines to the data - we take the derivatives of the fitted splines to evaluate equation 4.1. The optimized lattice parameters at 0 K for graphene for all methods are also given in Table 4.2 for comparison.

The calculated lattice parameters from ground state optimization are given in Table 4.2. The majority of the empirical potentials considered accurately predict the 0 K lattice parameter (with errors typically less than 0.2%), which is found from DFT to be 2.464 Å. The exceptions to this are the Tersoff, AIREBO and Amorphous GAP potentials. The Tersoff potential is found to overestimate the lattice parameter of graphene by 3.2%, while the AIREBO and amorphous carbon potentials underestimate by 2.0% and 1.2% respectively. DFTB would generally be expected to represent an improvement over empirical potentials, however in this instance predicts the lattice parameter of graphene with an error of +0.3%, representing an improvement over only the three worst empirical potentials. The Graphene GAP and ReaxFF potentials are both in excellent agreement with our *ab inito* results with errors of 0.1%.

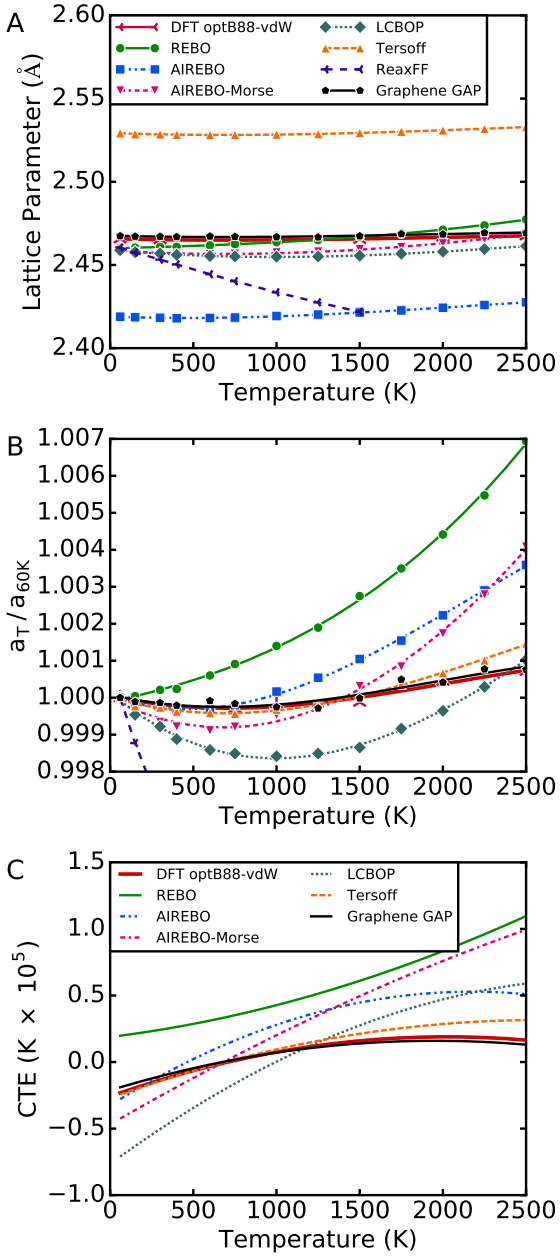

Figure 4.2: A) Thermal dependence of the lattice parameter of graphene between 60 and 2500 K, for a range of potentials as compared to the reference value calculated from *ab initio* molecular dynamics calculations. B) Thermal dependence of lattice parameter, a, normalized according to the predicted value at 60 K, emphasising the relative behavior of the different methods - a range of predictions is observed, from monotonically increasing or decreasing lattice parameters to more complex non-monotonic behavior in the case of GAP, LCBOP and AIMD calculations. C) Computed thermal expansion coefficients for graphene as a function of temperature calculated using equation 4.1, for DFT and various potentials.

Most of the potentials considered predict a much larger dependence of the in-plane lattice parameter on the temperature than is calculated from AIMD, which predicts an overall maximum change in value of 0.1% as can be seen from Figure 4.2B. Our first principles calculations predict a contraction of the graphene sheet up to approximately 1750 K, above which we observe expansion in the in-plane direction. Our graphene GAP model is in excellent agreement with the predictions of the first principles calculations both in terms of the absolute and relative lattice parameters. The relative predictions of the Tersoff potential are also found to be in good agreement with *ab initio* results at low temperatures, despite the significant overestimation of the absolute lattice parameter. The AIREBO and AIREBO-Morse potentials significantly overestimate the in-plane expansion of graphene at moderate temperatures, while the REBO potential predicts an in-plane lattice parameter which increases over the entire observed temperature range. The predictions of the LCBOP potential are in line with those of previous studies, it predicts a strongly negative thermal expansion with a minimum close to 1000 K [30]. The ReaxFF potential considered here is observed to predict a very strong, negative thermal expansion coefficient and predicts the fragmentation of the graphene sheet at temperatures above 1500 K, well below the experimentally determined melting point. Between temperatures of 60 and 1500 K, ReaxFF predicts a strong contraction of the in-plane lattice parameter as a result of large out-of-plane displacements. Figure 4.2C shows the values for the CTE of graphene as calculated with each of the potentials and with *ab initio* calculations. The LCBOP, AIREBO and AIREBO-Morse potentials predict CTEs which are strongly temperature dependent, switching from negative to positive at temperatures between 500 and 1000 K. The REBO potential similarly predicts a strong temperature dependence, however in this case the CTE is predicted to be positive over the entire measured range. In contrast, the GAP, Tersoff and AIMD simulations predict a much weaker temperature dependence of the CTE, with a change in sign close to 1000 K. The Tersoff potential predicts a continued increase of the in-plane CTE throughout the measured temperature range, while the GAP and AIMD calculations predict a slowdown in the increase and a

plateau above 1500 K. Overall it is clear that, the lattice expansion of graphene represents a challenging property to evaluate with molecular dynamics, the GAP model introduced here quantitatively reproduces the results of the reference calculations.

4.4 Prediction of Phonon Spectra

A correct description of the lattice dynamics of a material is a fundamental requirement for any atomistic model. This experimentally measurable property of a material is obtained computationally directly from the derivative of the forces acting upon the atoms. There is thus a natural and close link between the quality of the phonon spectrum and the quality of the predicted forces with respect to experiment. This makes the prediction of the phonon spectrum an excellent independent metric of the overall quality of a potential. Furthermore, a number of thermodynamic properties of materials, for example the heat capacity, may be obtained directly from dispersion relations via calculation of the free energy. We note here that two definitions of dispersion are used in this text, when referring to dispersion in the context of phonon dispersion curves, we refer to the rate of change of the energies of the various modes as a function of reciprocal space, rather than the effect of van der Waals interactions.

We use two methods to calculate the phonon spectrum of graphene. To calculate the 0 K phonon spectrum, we use the finite displacement method as implemented in PHON [169]. In order to predict the anharmonic phonon spectrum at finite temperature, we evaluate the elastic constants and thus the phonon spectrum directly from the forces and displacements sampled from MD trajectories [170, 171]. This is performed using the ”fix phonon” method implemented by Kong Et. Al,. in LAMMPS. Rather than making finite displacements, displacements are observed naturally over the course of a molecular dynamics simulation. The positions of the atoms are Fourier transformed at regular intervals during the simulation and averages of the atomic positions and associated forces are performed, from this, the dynamical matrix is also computed at regular intervals before finally being averaged at the end of the simulation. Full technical details on the method are given

in references [170] and [171]. As our reference, we compare our results to those determined from the fifth nearest neighbor force constant fit to data measured experimentally using x-ray diffraction (XRD) on graphite [12, 14]. The phonon spectrum of graphene is comprised of six branches; ZA, TA, LA, ZO, TO and LO. At the Γ point, the LO and TO phonon branches take on the symmetry label E_{2g}, the ZO branch is labelled B_{2g} and the lowest energy LA, TA and ZA branches together as A_{2u} and E_{1u}. [1]

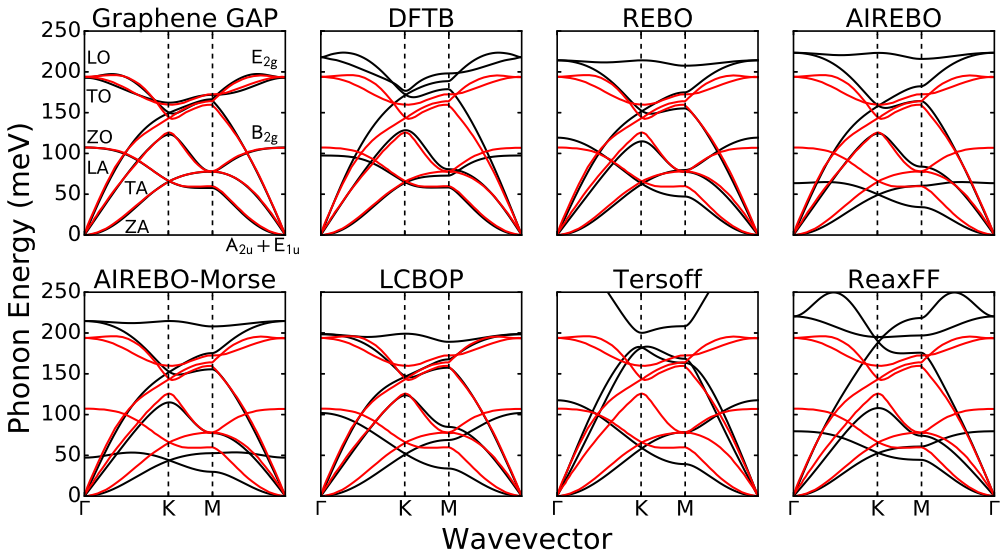

Figure 4.3: Comparison of model predictions using the finite displacement method [169] to phonon dispersion from XRD [170, 171]. Black lines represent the calculated phonon spectrum, red is the reference XRD. The GAP model accurately reproduces the experimentally determined phonon spectrum over all of the high symmetry directions considered. Labels for branches are shown on the Graphene GAP plot (left) along with symmetry labels at the Γ point (right). Note that the highest energy LO branch is not shown for the Tersoff potential in this figure - this branch crosses the Γ point at approximately 350 meV.

Figure 4.3 shows the phonon spectra predicted using each of the potentials compared to the reference XRD data. The graphene GAP model achieves excellent agreement with experiment; it correctly predicts the phonon frequencies at almost all of the high symmetry points with sub-meV accuracy. The dispersion behavior of

[1] The label 'Z' denotes an out-of-plane vibration, 'L' a longitudinal, in-plane vibration and 'T' a transverse shear mode. Each of these modes may be either acoustic or optical in nature, indicating the phase of the displacements of adjacent nuclei relative to one another. Acoustic phonons represent in-phase vibrational modes, while an optical phonon represents an out-of-phase normal mode of vibration, wherein any two atoms are seen to move against each other.

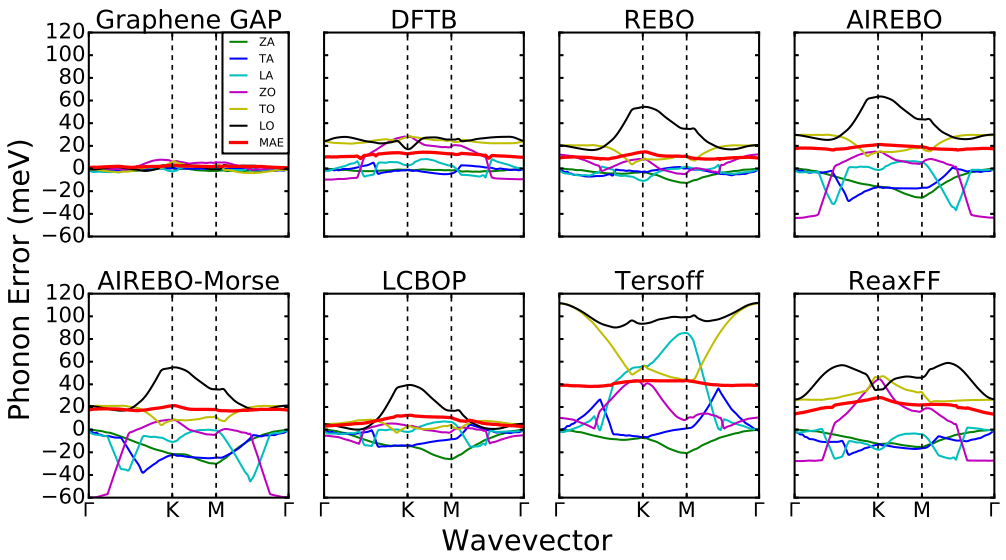

Figure 4.4: Absolute errors in prediction of phonon band frequencies along the high symmetry directions in the graphene Brillouin zone, separated by phonon branch type. The thick red line denotes the mean absolute error (MAE) summed across all bands. Notable similarities in the error predicting the character of the LO branch can be seen across the LCBOP, REBO and AIREBO(-Morse) potentials (black line).

each of the bands is also accurately predicted across all of the sampled regions of the Brillouin zone. The LCBOP and REBO potentials perform comparably to one another, qualitatively correctly predicting the shape and dispersion character of most of the phonon branches. What can be seen in more detail from Figure 4.4 is that LCBOP achieves a greater accuracy than REBO close to the Γ point, but amasses more significant errors overall, on the order of 20 meV, towards the *K* and *M* high symmetry points. Conversely, the error in the prediction of the phonon frequencies made by the REBO potential is a much flatter function of k-space with an overall mean absolute error (MAE) of 10 meV. However, both potentials exhibit significant errors in the prediction of the highest energy longitudinal optical (LO) branch, with peak errors of 40 meV and 60 meV for LCBOP and REBO respectively. As would be expected, both the AIREBO and AIREBO-Morse potentials perform comparably, with notable underestimations of the transverse optical (ZO) phonon modes at the Γ point. The MAE of each potential is again a relatively flat function of k-space, at 20 meV in both cases. The dispersive character and B_{2g} Γ point frequency

predicted by DFTB are in good agreement with the experimental results, the most most notable error being the overestimation of the E_{2g} symmetry frequency at the Γ point, which is overestimated by 20 meV. We find that the ReaxFF potential provides a reasonably good estimate of dispersion of the low frequency phonon modes, however fails for the highest energy LO and TO branches. This is the case in particular away from the Γ point, for which peak errors in the LO branch are found to be in excess of 60 meV. The Tersoff potential, finally, is shown to fail in predicting the energies and dispersion behaviors of all but the two lowest energy branches of the phonon spectrum. Band errors are as large as 110 meV for the E_{2g} symmetry (LO and TO) bands at the Γ point, with a MAE across the sampled region of k-space of 40 meV. Although a modified version of the Tersoff potential has been constructed which was optimized to reproduce the lowest energy phonon dispersion modes of graphene, we find that the stability of this potential is not satisfactory due to the reparametrization, and have therefore not included it here [172, 27]. We note that an error common to all of the empirical potentials is a failure to describe the dispersive behavior of the high energy LO branch of the phonon spectrum - which the graphene GAP model predicts with negligible error.

In addition to a consideration of the phonon spectrum at a single temperature, we can compare the behavior of particular phonon modes as a function of temperature to experimental observations from Raman spectroscopy. The G band of the graphene phonon spectrum may be unambiguously assigned to the frequency of the E_{2g} symmetry phonon mode at the Γ point. We may therefore make a direct comparison between the experimentally measured thermal softening of this mode and the softening predicted by each of the potential models. The correct description of the thermal character of this band is of great importance for the technological application of graphene - the degree of population of the E_{2g} band has implications for the ballistic energy transport which makes graphene so attractive as an electronic material [155, 173]. One aspect of this characterization is the correct prediction of the energy of this mode at the Γ point, the comparison for which shown in Figure 4.5 where the phonon spectra for graphene from 60 to 2500 K are given.

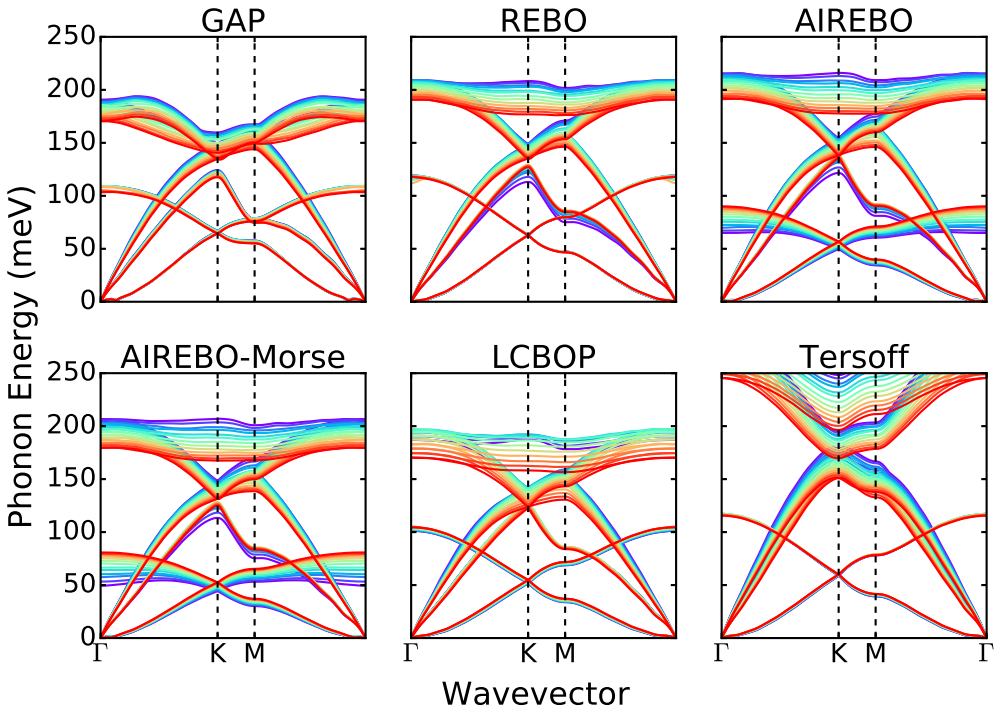

Figure 4.5: Finite temperature phonon calculations for graphene simulations between 60 and 2500 K derived directly from molecular dynamics simulations. Strong thermally induced dispersion is seen for the highest energy E_{2g} symmetry phonon modes across all potentials, corresponding to the observed thermally induced dispersion of the Raman G band of graphene. Varying predictions are made for the transverse optical (ZO) branch's dependence on temperature - the AIREBO(-Morse) potentials predict this to have a strong thermal dispersive character. Blue corresponds to simulations at 60 K, through to 2500 K for red in a linear scale.

For each temperature we use the lattice parameter determined for each potential for the given temperature as calculated using the same procedure for determining the lattice parameter described above. Simulations were run for each lattice parameter and each potential in the NVT ensemble using Langevin dynamics. Configurations were first equilibrated for 2 ns until the temperature had equilibrated and statistics were collected over 30 ns trajectories at each temperature, in each case the phonon frequencies of the degenerate LO/TO (E_{2g}) branches at the Γ point were converged to within 1 meV.

We observe that all potentials predict a large degree of thermally induced dispersion in the highest energy LO/TO branches (Figure 4.5). The AIREBO and

AIREBO-Morse potentials both predict a strong dependence of the transverse optical (ZO) branch on temperature, which is not observed for the other methods considered. We compare quantitatively the results of our calculations to those obtained from the variable temperature Raman scattering measurements [165]. The thermally induced dispersion of the Raman G band was measured between 150-900 K for graphene sheets adsorbed on a SiN substrate. The effect of the substrate on the position and thermal dispersion of the G band is two-fold, a constant offset induced by the mismatched lattice parameter and interlayer interactions between the substrate and the graphene and an effect due to the thermally induced strain from the different thermal expansions of the two materials. To account for the first effect, we simply report the change in G band frequency rather than the absolute value. The effect of the differing lattice expansion of the materials may be accounted for by calculating the induced strain and correcting the data using the known biaxial strain coefficient of the graphene G band.[165, 156]

$$\Delta\omega_G^s(T) = \beta \int_{T_0}^{T} [\mathrm{CTE_{sub}}(T) - \mathrm{CTE_{gr}}(T)]dT \tag{4.2}$$

Where $\mathrm{CTE_{sub}}$ and $\mathrm{CTE_{gr}}$ represent the CTEs of the substrate (SiN) and graphene respectively, and β is the known biaxial strain coefficient of graphene ($\beta = -70 \pm 3\,\mathrm{cm}^{-1}/\%$) [174, 175]. We use values for the CTE graphene as determined by our earlier *ab initio* calculations.

Figure 4.6 shows the thermally induced dispersion of the $\mathrm{E_{2g}}$ symmetry phonon modes at the Γ point. Our graphene GAP model is seen to be in good agreement with the experimentally observed effects as are the predictions of both the AIREBO and REBO potentials. The AIREBO-Morse potential slightly overestimates the degree of dispersion while the Tersoff potential predicts a significantly enhanced effect. Surprisingly, despite the good predictions of the shape of the phonon dispersion curves by the LCBOP potential using the finite displacement method, we find here a strong qualitative disagreement with the experimental results.

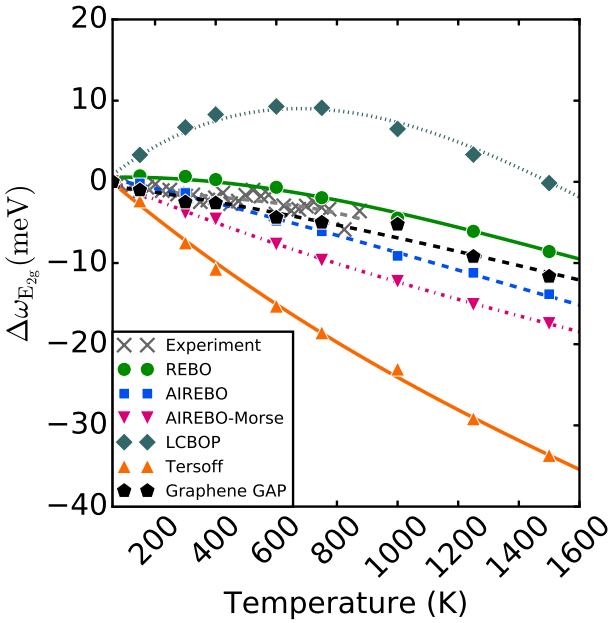

Figure 4.6: Change in Γ point frequencies for graphene E_{2g} symmetry vibrational mode in the region of 150-1400 K. Compared with results from variable temperature Raman spectroscopy, which have been corrected for the strain induced by the adsorption of the graphene sheet onto the SiN substrate.

4.5 Conclusions and Discussion

We have used the Gaussian Approximation Potential method to construct a machine learning potential for graphene, which we have trained using energies, forces and virial stresses calculated using high quality vdW inclusive DFT calculations. We have benchmarked the quality of this potential alongside a number of other commonly used potentials against both *ab initio* and experimental references. We find that the graphene GAP model predicts quantitatively the lattice parameter, coefficient of thermal expansion and phonon properties of graphene. Among the other potentials considered, many of them provide reasonable predictions of one property, but none is successful in predicting the whole range of properties considered. We find the REBO potential to be the best empirical model, providing a good overall description of the lattice dynamics of graphene, including accurately describing the effect of temperature on these. However, despite accurately predicting the 0 K lattice parameter, the REBO potential's predicted dependence of the in-plane lattice

parameter is in qualitative disagreement with the results of *ab inito* calculations. In fact, we find that none of the empirical many-body potentials accurately predicts both the 0 K lattice parameter of graphene and the lattice expansion at finite temperature.

The GAP method is computationally more demanding than the empirical many-body potentials considered here, but approximately four orders of magnitude cheaper than direct *ab initio* molecular dynamics, for 200 atoms. Even taking into consideration the computational cost of the generation of the training database, this represents a significant reduction in computational cost with only a marginal compromise on accuracy. Since the scaling of the cost of the GAP model with system size is the same as that of a force-field MD simulation, compared with the $O(N_{\text{electron}}^3)$ scaling of DFT, this reduction in cost would be more effective for larger system sizes. The purpose of the GAP framework is to provide an accuracy close to that of AIMD at a much reduced cost, rather than offering a universally applicable alternative to empirical potentials. Such a potential would be best put to use in cases where a highly accurate description of dynamics is mandated. One such example may be the description of adsorbate diffusion on or confined by graphene sheets, a process which is in some cases strongly enhanced by a coupling between adsorbed molecules and particular graphene phonon modes [176, 25]. In this instance, the accurate finite temperature description of the phonon modes provided by the GAP model would be highly desirable. The GAP model would also be ideally suited to modelling thermal transport in graphene nanoelectronic devices, such as transistors. Such systems require highly accurate modelling of heat dissipation, but involve systems of sizes which are beyond the reach of routine *ab initio* calculations [35, 36, 37]. In many cases, such as for exotic or newly discovered materials, computational investigations may be hampered by the absence of a well parametrised empirical potential. The GAP framework provides a systematic pathway for the development of specialized potentials in these cases.

Despite the promising behaviour of the GAP model considered here, it is important to note that the transferability of the various models may also be an impor-

tant property. While the GAP model presented here is exemplary in its treatment of free-standing graphene, it is (by construction) not transferable to other phases of carbon i.e. diamond, which the other empirical potentials are capable of. That many machine learning models fail to extrapolate into foreign regions of chemical space is a well documented one, and great care and attention must be paid to generate a machine learning potential which is capable of treating a wide range of phases of a material [106, 23]. In the following chapter, this aspect of the GAP model will be addressed. We will discuss how it is possible to combine the accuracy of the graphene GAP model presented in this chapter, with the transferability of the GAP-17 amorphous carbon model [106] to produce a single GAP model for carbon, which is both accurate for the crystalline phases while being transferable enough to extend into new regions of chemical space.

Chapter 5

An Accurate and Transferable Machine Learning Potential for Carbon

5.1 Introduction

In the previous chapter we introduced a machine learning potential for pristine graphene constructed using the Gaussian approximation potential (GAP) framework, which achieved excellent accuracy when benchmarked against 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 [107]. 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, Deringer *et al.* reported a highly transferable GAP model trained primarily on the amorphous and liquid phases of carbon (termed GAP-17), based on DFT-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 eV Å^{-1} with the graphene-only GAP mentioned above, as compared to 0.27 eV Å^{-1} with GAP-17. For comparison, these errors for a range of commonly used empirically fitted potentials range from

0.6 to 3.1 eV $Å^{-1}$ (more details are in Ref. [107]). 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 which involve diverse structural environments, including understanding the mechanism of growth of sp^3 hybridised amorphous carbon by ion deposition [108], extensive studies of the surface properties (and chemical reactivity) of tetrahedral amorphous carbon [111, 108, 177], the structure of "porous" carbonaceous materials which are relevant to applications in batteries and supercapacitors [178, 179], and crystal-structure prediction [23].

The model we present in this chapter, 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.

While detailed discussions of the construction and testing of the potential will be given in subsequent sections, we take a moment to highlight the main points here. The composition of the training data set and performance of this potential is summarised in figure 5.1. 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 behaviour 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 modelling of nanostructured carbons; nanotubes, graphitised carbon and materials

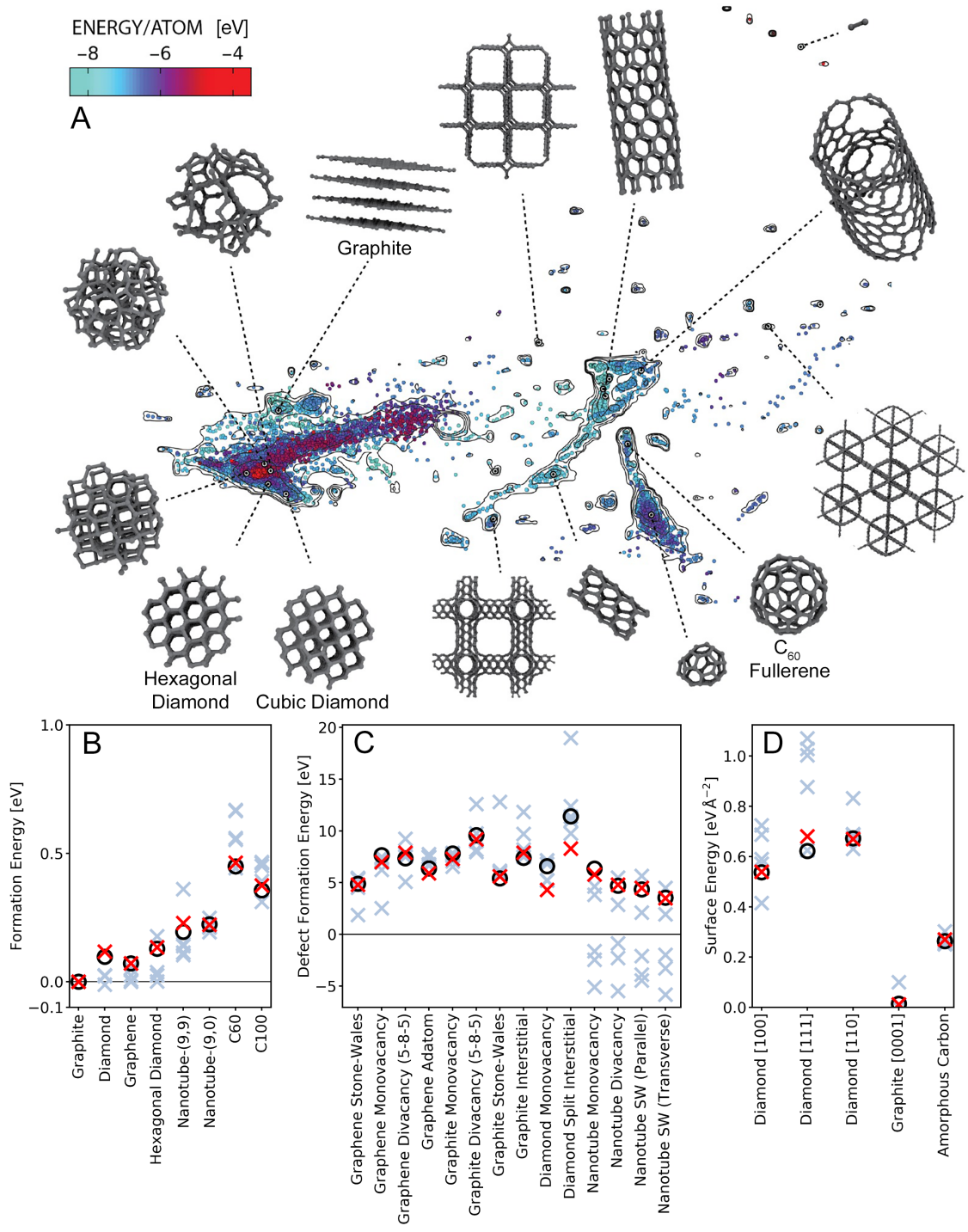

Figure 5.1: Overview of some of the key structures included in the training shown through a sketch-map representation (top) as well as selected information on the performance of the potential for a variety of properties. (a) Sketch-map representation of the total data set for carbon generated as part of this work. Select structures are identified for graphite, diamond, hexagonal diamond (Lonsdaleite), amorphous carbon and fullerenes. Points are coloured according to their energy, while contours indicate the density of the database population in a particular region. Bottom, is a summary of (b) the 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).

with varying degrees of defects and disorder.

The rest of the chapter will be organised 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 which required special attention or optimisation. Subsequently, we present an extensive and rigorous testing of our model, for a wide range of properties. We also compare the results of our model to a selection of commonly used empirical potentials which 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 "graphitised" carbons [31], may be the subject of future work.

5.2 Generation and Selection of Training Data

One of the challenges inherent in constructing a generalised potential for carbon is the enormous variety of structures which 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 [24].

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 data set. For a potential to be both accurate and transferable, its training data set 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 modelling 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 [115, 114, 180]. Further, 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 [106, 115, 150, 149]. In the case of atom centred descriptors such as the smooth overlap of atomic positions (SOAP), this represents the local environment around a particular atom rather than the global structure [150]. 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 [181, 182]. Similar logic may be applied to more complex structures.

The reference configurations which 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 VASP plane-wave DFT code, we perform calculations with the optB88-vdW dispersion inclusive exchange-correlation functional [183, 140, 160, 141], a plane-wave cut-off of 600 eV and a projector augmented wave pseudopotential [143, 159, 161]. A Gaussian smearing of 0.1 eV is applied to the energy levels and dense reciprocal space Monkhorst-Pack grids are used [163]. 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 [142].

The database of configurations presented here uses as its foundation a combi-

nation of the training data sets for the two carbon potentials previously published primarily for liquid and amorphous carbon (GAP-17), and for pristine graphene, respectively [106, 107]. A large number of new configurations are considered in addition to these existing ML data sets [184, 23, 24]. We endeavour to comprehensively cover all the possible crystalline phases of carbon found at moderate temperatures and pressures, including more exotic allotropes. To that end, DFT optimised structures for graphite, graphene, 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 ≤ 10 with fewer than 240 atoms in their unit cell. Optimised structures are also included for the SAmara Carbon Allotrope DAtabase (SACADA) database of exotic carbon allotropes [184] and the results of a GAP-17 driven random structure search [23]. 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 [185, 186, 187, 188, 189, 190, 191, 192].

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 [107, 106]. The resulting database is comprised of ca. 17000 configurations each containing from 1 to 240 atoms per 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 [113, 62, 115]. 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 maximise 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), which is detailed in section 3.1.2 [148, 62]. Within the set of configurations chosen from FPS, we then carefully check the data saturation of our training with respect to the number of sparse points, which is discussed in section 3. 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 [148, 62]. 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 [115]. 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 optimised geometries for structures used in the validation sections of this chapter 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 a further ca. 1000 configurations are manually added to target specific properties.

The selected configurations, as well as a representation of their position in phase space, can be seen illustrated in figure (Fig. 1). This sketch-map [193, 148] 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 which are similar in the full high-dimensional descriptor space are closer together, and those which are dissimilar are further apart.

This sketch-map representation also serves as a qualitative overview of the type

of structures 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 which are sp^2 hybridised, such as graphene, graphite and carbon nanotubes are clustered together towards the right of the map. Amorphous structures can be seen as a large region in the centre, with low density (sp^1 and sp^2 rich) amorphous carbon at the far right, high density sp^3 rich amorphous carbon towards 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, isolated gas phase dimer configurations are found.

5.3 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 costly from the perspective of both training data generation, potential generation and the ultimate evaluation of the potential. The nature of bonding interactions for carbon may also be captured in an approximate way, being generally attractive between 1.2 - 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 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. 5.2) simply has the final form $V_{short} + V_{long}$

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 is given in detail elsewhere [150, 194, 195, 106, 107, 196, 197].

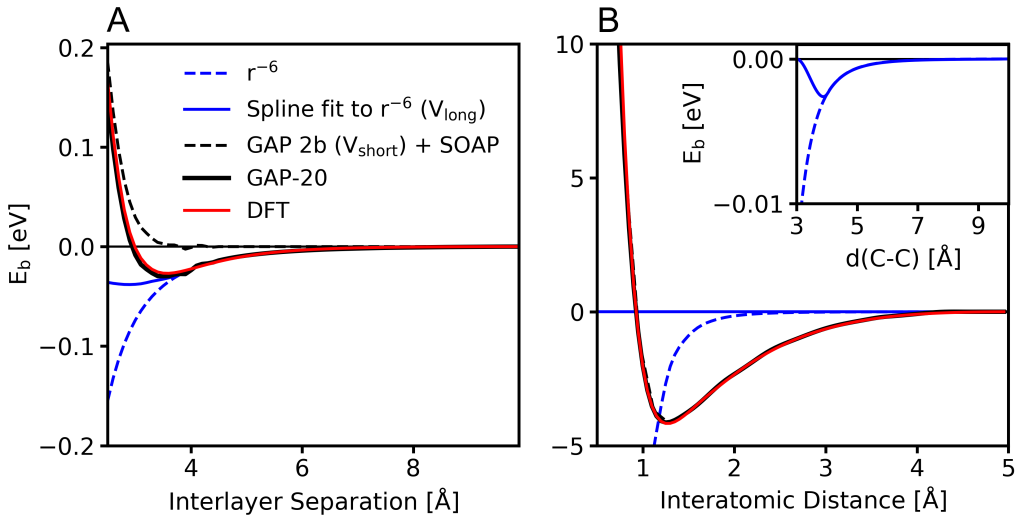

Figure 5.2: Construction of the long-range 2b component of the GAP model. An analytical spline is fitted to a function which decays as r^{-6}, designed to recover the long-range attraction between graphene layers. (a) Shows the predicted energies for each component for the interaction of two graphene layers at different distances. The long range attraction is well characterised 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) Shows 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) Provides a closer view of panel B, and shows how the 2b spline fit to the r^{-6} component is brought smoothly to 0 at a distance of 3 Å.

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 the atomic indices [106, 107]. In the construction of the SOAP descriptor, the local environment around a target atom is represented by a ‘local neighbour den-

sity', constructed by placing a Gaussian basis function on each neighbouring 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 material properties of atoms which are closer together [198, 199, 200]. Other functional forms for the radial scaling exist and the introduction of this scaling was performed independently of the optimisation of the SOAP descriptor cutoff, the choice of which is motivated below. As a result, there may still be scope for further optimisation of these parameter sets beyond what is performed here. The local neighbour 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, then raised to a power ζ. Such a kernel is physically interpretable, as it corresponds directly to the integral of two neighbour densities for all possible 3D rotations. Details for the specific choices for a number of associated hyperparameters are given in the supplementary material, while further details on their significance is given elsewhere [60, 150, 194, 195, 106, 107, 196, 197].

We now provide the details of select convergence tests for the optimisation of our GAP model. These tests consider the independent optimisation of the SOAP descriptor cutoff, 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 neighbour density up to l_{max} = 8, n_{max} = 8, a cutoff of 4.2 Å, $\sigma_{\text{force}} = 0.01$ eV Å^{-1} and $\sigma_{\text{energy}} = 0.001$ eV and $\zeta = 4$. Motivations for the selection of those hyperparameters which aren't discussed here are given in section 3. 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 5.3(a) shows the behaviour of the force errors as a function of the SOAP

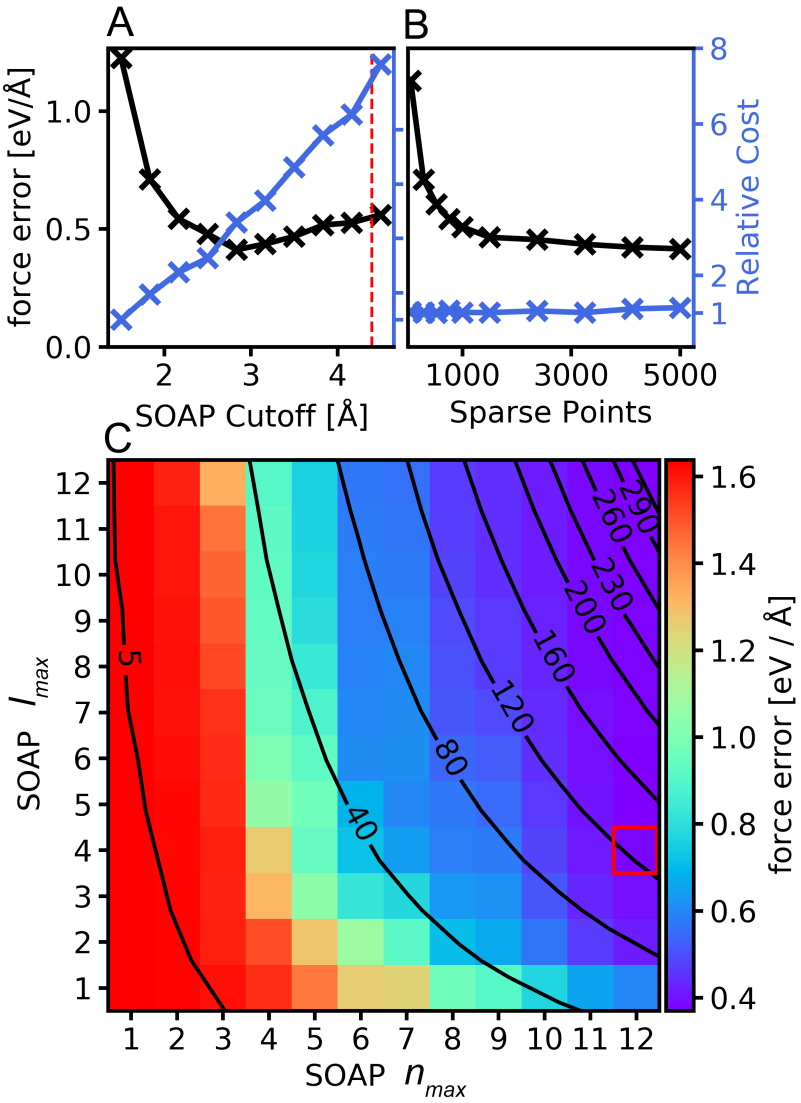

Figure 5.3: Mean absolute force errors calculated for an independent test set of configurations for different SOAP descriptors. (a) Force error behaviour and cost of evaluation (relative to the fastest GAP) of the resultant model as a function of the SOAP descriptor cutoff, the selected value of 4.5 Å is indicated by the dashed red 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 neighbour density basis set expansion. Force errors are indicated by the colour bar, while relative costs are shown by contour lines, our choice $l_{\mathrm{max}} = 4$, $n_{\mathrm{max}} = 12$ is highlighted by the red square.

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

optimisation of these parameters based purely on the force errors would therefore select a cutoff radius of 2.9 Å. However, selection of these parameters cannot be performed in isolation from the intended application of the potential, but must also be motivated by knowledge of the behaviour 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 optimisation. 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 consider to be mandatory for this potential. The inter-layer spacing of graphitic structures is typically large (approx. 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 behaviour of the Tersoff and REBO-II models when modelling the interlayer spacing of graphite (Table 5.1), or graphene bilayers (Supplementary fig. 6). However, the problem of producing a single analytical metric for optimisation, which satisfactorily includes properties such as the lattice parameters, defect formation energies or 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 5.3(b) shows the convergence of the mean absolute force errors as a function of the number of sparse points used in the training, this may be considered as a measure of the data saturation of GAP-20. The force errors decrease rapidly up to approx. 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 figure 5.3(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 neighbour density. The relative computational cost of each basis set is indicated on the same plot by labelled 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, 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 data set which can be used and the size and length scales to which the resultant potential can be applied.

To aid reproducibility, we provide here the complete training command line provided to the gap_fit program used for fitting the GAP-20 model discussed here:

```
at_file=General_Carbon_V10_2_4000_All_GAP_17_Unique_LD_Iteration_1.xyz
gap={distance_2b n_sparse=15 theta_uniform=1.0 sparse_method=uniform
covariance_type=ard_se cutoff=4.5 delta=2.0:angle_3b n_sparse=200
theta_uniform=1.0 sparse_method=uniform covariance_type=ard_se cutoff=2.5
delta=0.05:soap n_max=12 l_max=4 atom_sigma=0.5 zeta=4.0 cutoff=4.5
cutoff_transition_width=1.0 central_weight=1.0 n_sparse=9000 delta=0.2
covariance_type=dot_product sparse_method=cur_points radial_decay=-0.5}
default_sigma={0.001 0.01 0.05 0.0} energy_parameter_name=energy
force_parameter_name=force virial_parameter_name=virial do_copy_at_file=F
sparse_jitter=1.0e-8 gp_file=Carbon_GAP_20.xml core_ip_args={IP Glue}
core_param_file=r6_innercut.xml
config_type_sigma={Liquid:0.050:0.5:0.5:0.0:
Liquid_Interface:0.050:0.5:0.5:0.0: Amorphous_Bulk:0.005:0.2:0.2:0.0:
Amorphous_Surfaces:0.005:0.2:0.2:0.0: Surfaces:0.002:0.1:0.2:0.0:
Dimer:0.002:0.1:0.2:0.0: Fullerenes:0.002:0.1:0.2:0.0:
Defects:0.001:0.01:0.05:0.0:Crystalline_Bulk:0.001:0.01:0.05:0.0:
```

```
Nanotubes:0.001:0.01:0.05:0.0: Graphite:0.001:0.01:0.05:0.0:
Diamond:0.001:0.01:0.05:0.0: Graphene:0.001:0.01:0.05:0.0:
Graphite_Layer_Sep:0.001:0.01:0.05:0.0: Single_Atom:0.0001:0.001:0.05:0.0}
```

5.4 Crystalline Carbon

Among the most important material properties for any potential to predict accurately are those of the bulk crystalline phases. Table 1 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 figure 5.4, we also provide both the atomisation 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 atomisation energy of a phase relative to the isolated gas phase carbon atom E_{at} as,

$$E_f = E_{bulk} - nE_{at}, \tag{5.1}$$

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 optimising 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 modelling 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 characterising the graphite inter-layer interactions and the long range of the atomic descriptor required to capture it. As discussed above, the choice of SOAP cut-off was specifically 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 cut-off.

It is useful here to make a brief comparison to selected empirical potentials. While we do include DFT reference data for all properties, in subsequent sections 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 figure 5.4 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, AIREBO potentials have errors averaging 5%, 0.3%, 4% and 1% respectively (Table 5.1. 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 in part to its complete reparameterisation 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 atomisation energies (fig. 5.4(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

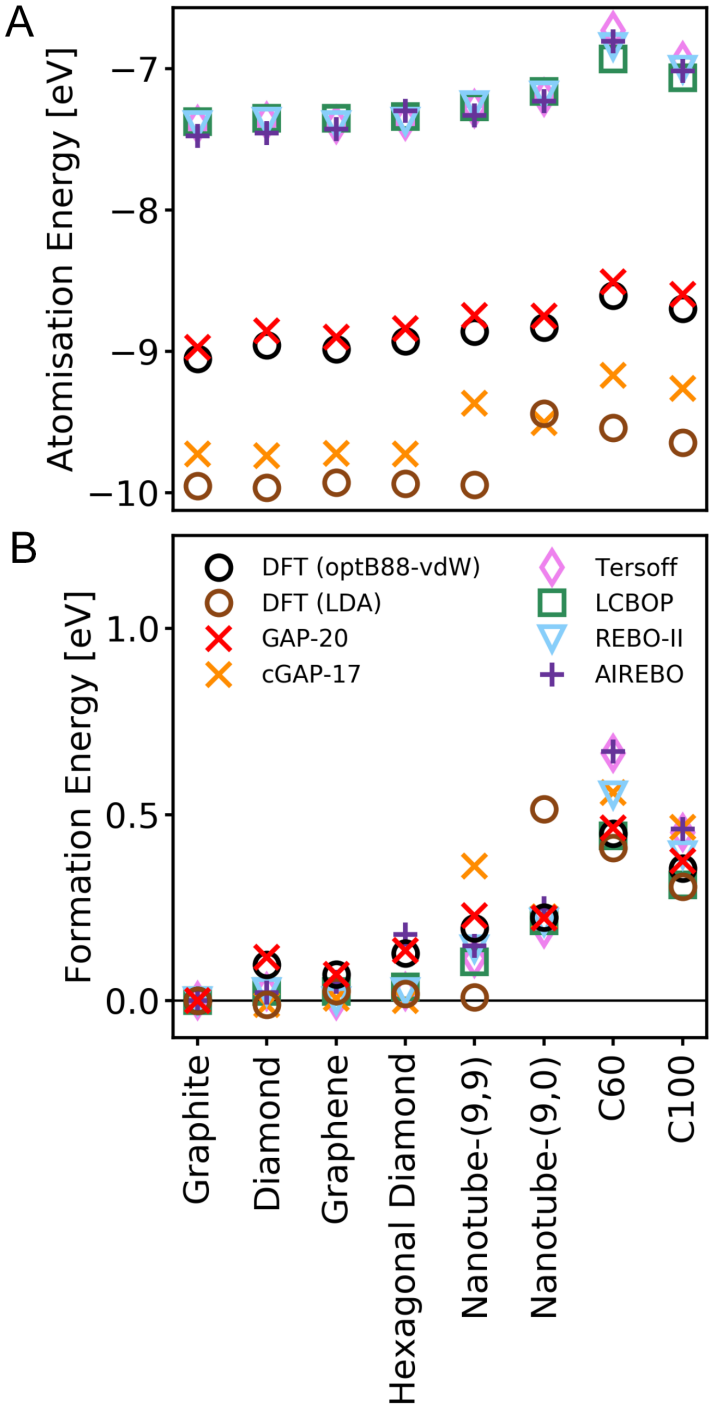

Figure 5.4: Formation energies of the crystalline phases of carbon, comparing results from DFT (optB88-vdW and LDA) to those from GAP-20 and the other models tested. (a) Atomisation energies using the isolated gas phase carbon atom as a reference, differences are dominated by overstabilisation of the gas phase atom by empirical models. (b) Formation energies given relative to the graphite formation energy of each particular model.

Table 5.1: 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)

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. 5.4(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 atomisation 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 atomisation 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. Firstly, 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 favour cubic over hexagonal di-

amond, 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 supplemental 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 model the lattice dynamics of bulk crystals, i.e. their behaviour 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 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.

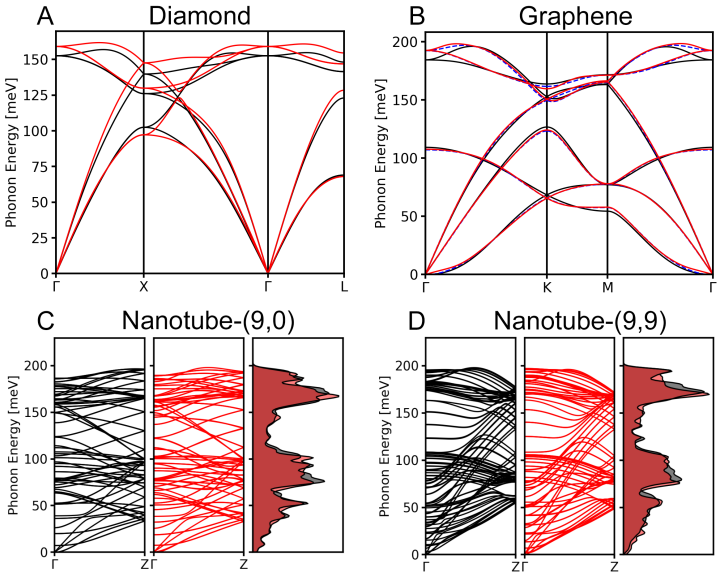

Figure 5.5: A. Phonon dispersion relation for diamond as predicted by GAP-20 (black) with comparison to DFT (optB88-vdW) reference data (red). B. Graphene phonon dispersion relation comparing GAP-20 and DFT (optB88-vdW) reference data. The dashed blue line shows the predicted phonon dispersion curve for the graphene-only model previously published [107] 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.

Figure 5.5 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 [201]. 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 [107]. 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 figure 5.5(b) that this is not the case; the dispersion relation of the phonon curves for graphene from GAP-20 are comparable to those of the previously published graphene GAP model [107]. 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 [107]. The quality of the GAP-20 model prediction is comparable for diamond (which cannot be modelled 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 behaviour 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 a 2-3 meV. In particular, it can be seen from fig. 5.5 that the vibrational density of states for the two nanotube systems agrees well with the DFT reference.

5.5 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 [189, 202, 192, 190, 191].

We compute the surface energy for each model by first optimising 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,

$$\gamma_{\mathrm{s}} = (\mathrm{E}_n - \mathrm{nE}_{\mathrm{bulk}})/2\mathrm{A}, \tag{5.2}$$

where $\mathrm{E_n}$ is the energy of the n slab layer containing two surfaces, which may be as-cut (unrelaxed) or allowed to relax and $\mathrm{E_{bulk}}$ is the 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 [203]. 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 5.2. 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 characterise this, however, GAP-20 predicts the unrelaxed and relaxed surface energies correctly to within an error of 3 meV $\mathrm{Å}^{-2}$ (20 %). With the inclusion of vdW interactions considered in their construction, the LCBOP and

Table 5.2: Surface energies of low Miller index surfaces for common carbon allotropes. Reference energies are calculated using DFT, 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)

AIREBO potentials both 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.

5.6 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 modelling by an interatomic potential is highly desirable. The possibility of rehybridisation, which allows carbon atoms to reconstruct with differing numbers of bonds to stabilise 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 [185, 204]. 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 [186, 107]. 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 [187]. For each structure, the lattice parameters and ionic positions of the pristine structures were optimised as discussed previously. The ionic positions of the defective structures were then optimised until the energy was converged to within 10^{-5} eV, while keeping the lattice parameters fixed. We compute the formation energy $\mathrm{E_f}$ of a vacancy defect relative to the energy of an atom in an ideal parent structure:

$$E_{\mathrm{f}} = E_{\mathrm{d}} - (\mathrm{n}E_{\mathrm{at}} + E_{\mathrm{bulk}}) \tag{5.3}$$

where $\mathrm{E_d}$ is the energy of the defective supercell structure, $\mathrm{E_{bulk}}$ 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. Monovacancy defects often result in unsaturated bonds at the defect site, while divacancy structures, particularly in sp^2 hybridised 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 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 unfavourable 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.

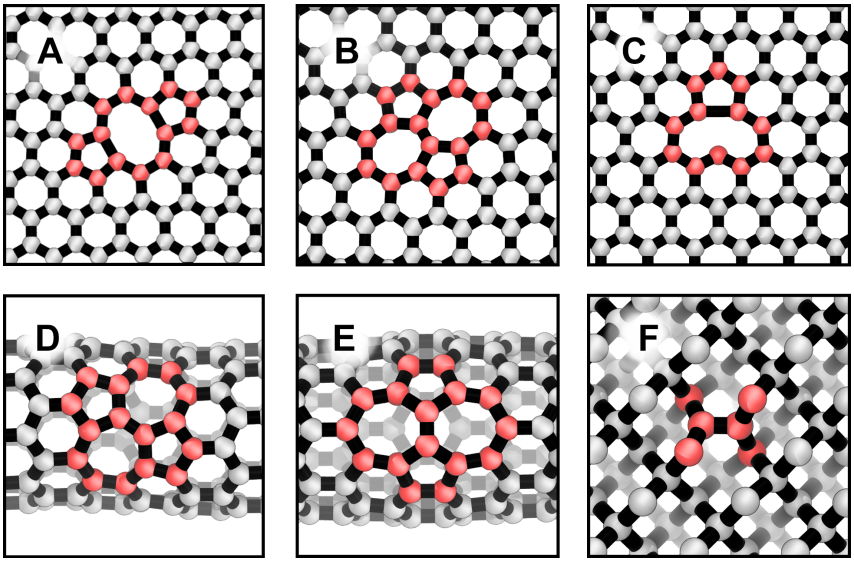

Figure 5.6: Images of selected carbon defect structures, with atoms in the immediate vicinity of the defect highlighted in red. (a) graphene divacancy defect (b) graphene Stone-Wales defect (c) graphene monovacancy (d) (9, 9)-nanotube Stone-Wales defect (transverse orientation) (e) (9, 9)-nanotube Stone-Wales defect (parallel orientation) (f) diamond split interstitial defect

Graphite is the only allotrope of carbon in which true interstitial defects are known, wherein interstitial atoms may be found between graphite layers [185]. 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

Table 5.3: Formation energies of common defects in carbon structures for GAP-20 and the other models considered, with DFT (optB88-vdW) values given as 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)

trigonally bonded carbon atoms above and below the sheet. 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 $[\bar{1}00]$ directions [205].

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 [206, 207, 208, 207].

Table 5.3 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 25-35%, while those for defective nanotubes range from 0-11%. 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 reparameterisation of the short-range components has adversely impacted the energetics of the model. Indeed, in the case of LCBOP, where this reparameterisation 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 11-52%. 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 favourable 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 car-

bon 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 optimised with our GAP model and those from DFT, in the form of the root mean squared 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 rehybridisation and reconstruction of (5-8-5), (555-777) and (5555-6-7777) graphene divacancy defects was accurately reproduced, as were the geometries of all of the diamond defects considered. Situations in which GAP-20 showed 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 (ca 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. That GAP-20 is capable of accurately modelling both the energetics and structural characteristics of a wide range of carbon defects 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.

5.7 Liquid Carbon

As discussed previously, the requirements of a potential for the satisfactory modelling 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 [107]. Conversely, in a liquid simulation, a vastly greater number of local configurations are explored, requiring a high degree of flexibility and transferability [106]. As in the case of GAP-17, we therefore use the liquid as a benchmark for the flexibility of our potential [106], 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 [209, 210]. 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 [211, 212, 213, 214].

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 $1.5 - 3.5\,\mathrm{g\,cm^{-3}}$ at 5000 K and the other for a range of temperatures between 5000 - 9500 K at a fixed density of 2.5 $\mathrm{g\,cm^{-3}}$. These simulations were performed for 216 atom systems using a chain of 5 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 [143, 159, 161]. We find that there is a very good agreement between the *ab initio* data and the GAP-20 predictions for both the RDF and ADF across the wide range of temperatures studied (see figure 5.7). GAP-20 correctly models the increased structuring of the liquid carbon as the temperature is reduced from 9500-4500 K. At temperatures below approximately 3500 K, the GAP model predicts the liquid to form an amorphous glass which slowly graphitises (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 graphitised amorphous car-

bons 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 [32, 31]. Figure 5.8 shows the RDF and ADF computed with both GAP-20 and optB88-vdW across a wide range of densities, from 1.5 to 3.5 $\mathrm{g\,cm^{-3}}$ at 5000 K. This test is particularly important as it represents dynamical simulations of structures from highly sp^1 and sp^2 hybridised (low density) through to a predominantly sp^3 hybridised liquid at higher density. GAP-20 captures this change in the bonding characteristics of liquid carbon, in particular the increase in the sp^1 hybridised fraction of the liquid at very low densities (reflected in bond angles close to 180 degrees), qualitatively similar to GAP-17 [106].

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 wildly different 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 to 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 crystallisation of a solid phase directly from the liquid) and it is therefore important that they be handled correctly.

5.8 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 which were not explicitly considered in its construction, in this case meaning that it must model structures which were not included in the training data base. 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

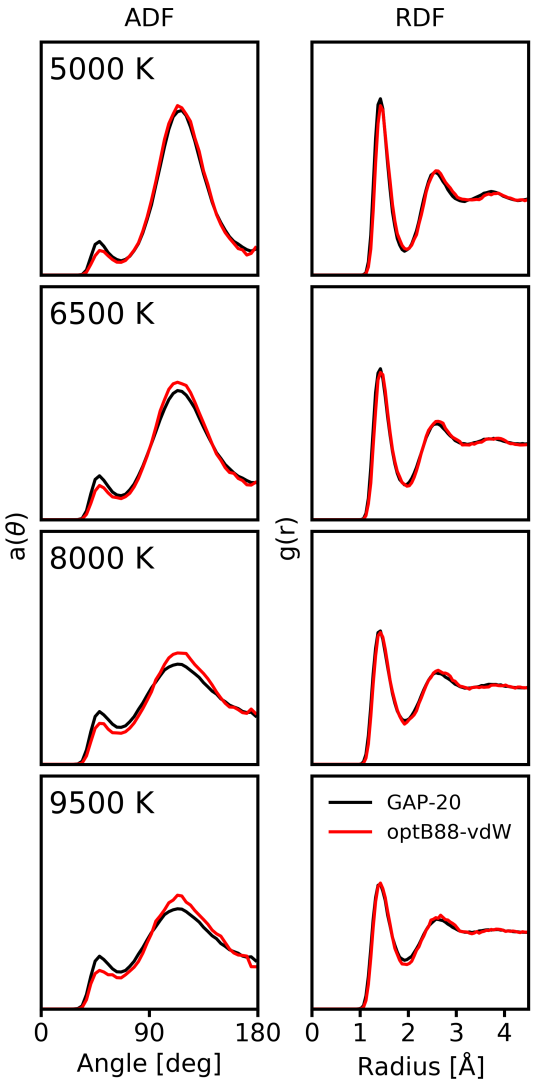

Figure 5.7: Angular and radial distribution functions for liquid carbon at a fixed density of 2.5 g cm^{-3} for temperatures between 5000 - 9500 K. GAP-20 results are shown in black, while reference DFT (optB88-vdW) data are given in red.

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 [115].

We demonstrate this here by performing a diagnostic GAP driven random structure search (GAP-RSS), similar in spirit to Refs. [215] and [23], and demonstrate that the predicted energies of these structures agree well with those from DFT [114, 113, 23]. We then calculate a number of high energy pathways for specific

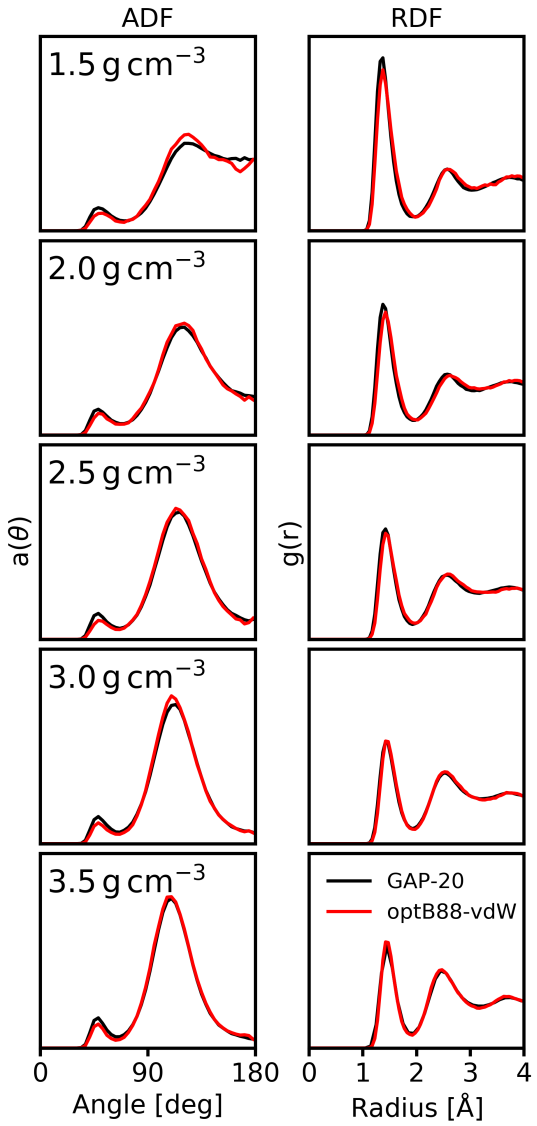

Figure 5.8: Angular and radial distribution functions for liquid carbon at 5000 K for a range of densities from 1.5 to 3.5 g cm^{-3}. GAP-20 results are shown in black, while reference DFT (optB88-vdW) data are given in red.

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 optimisation and which must be well described for an ML model to be transferable. Importantly, they are both explicitly designed to include configurations which are not present in the training data set of GAP-20.

To perform the first test, we generate a cubic unit cell with lattice parameter a

= 3 Å. In this cell, we randomly place 8 carbon atoms, avoiding any overlaps such that the distance between any two carbon atoms is not less than 2 Å. This process is performed to generate 1000 initial randomised geometries. The LAMMPS package is then used to optimise 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 [154]. Following this, the positions of the atoms in the unit cell are optimised using a FIRE algorithm [216], until the total energy is again converged to within 10^{-8} eV. This cycle is repeated twice more before performing a final conjugate gradient optimisation 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 behaviour and fail to find physically relevant configurations due to their having many more local minima than the DFT PES [217, 218].

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 both 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 dataseset, including crosslinked graphite structures, porous carbon cages and a variety of haekelite 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 (figure 5.9).

We also return at this stage to the sketch-map representation of the training

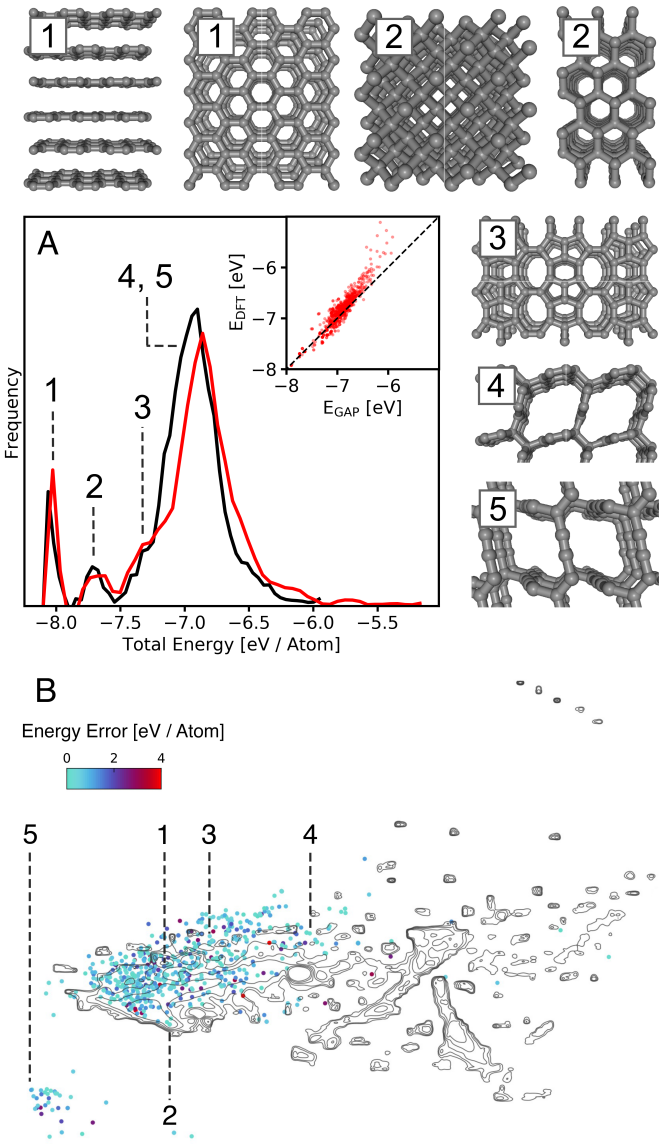

Figure 5.9: (a) A comparison of the histograms of energies of structures identified by GAP-RSS, given in eV / atom, showing good agreement for the prediction of the energy of structures between GAP-20 (shown in black) and DFT (optB88-vdW) (shown in red). A number of examples of structures identified 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, AA stacked graphite. 2) cubic and hexagonal diamond. 3) Haekelite 4) crosslinked graphitic structure. 5) Novel carbon structure with high proportion of sp^1 hybridised carbon atoms (b) The structures resulting from the GAP-RSS projected into the sketch-map representation from Fig 5.1. The density of the structures present in the training data are indicated by the contour lines, while the structures identified from the GAP-RSS are shown as individual points.

dataset given in figure 5.1. In figure 5.9(b) we provide a projection of the GAP-RSS structures onto this sketch-map representation. GAP-RSS points are coloured 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 8 atoms are included in the unit cell used for the RSS. Additionally, the RSS procedure employed begins with simulation cells which 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 which 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 characterised by their highly sp^1 rich character. Although a significant number of amorphous structures which are rich in sp^1 hybridised 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 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 [182, 219].

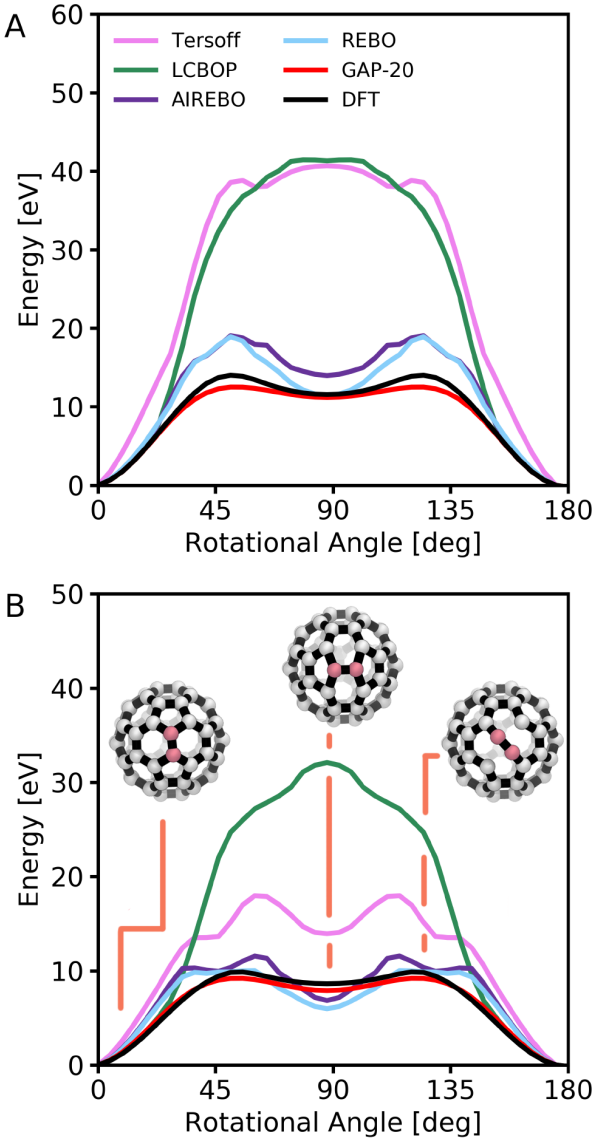

Figure 5.10: Energies for rigid transformations of a C-C bond in graphene (a) and in a C60 fullerene (b). Results from GAP-20, DFT (optB88-vdW) and a selection of empirical potentials are shown.

Figure 5.10 (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, 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 capture the general shape and height of the barrier, but predict jagged curves for the rotation, as compared to the smooth variation from DFT. The LCBOP and Ter-

soff 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 behaviour of the C60 rotational barrier in figure 5.10 (b). Here, it is again seen that GAP-20 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. LCBOP 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.

5.9 Conclusion

The advantages conferred by the flexibility of the Gaussian approximation potential framework are made clear by the wide variety of structures which are accurately treated by GAP-20. The variable hybridisation 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 an 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 which is equally suited to modelling 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 [180, 220] or in crystal-structure prediction by expanding on Refs. [23] and [215]. Further applications may include the more detailed study of non-graphitising or "hard" carbons [221, 222, 223, 224], following on from earlier GAP-17 based studies in Ref. [178] and [179].

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 (Supplementary Fig. 12), and therefore the latter will still give access to even larger-scale systems [31]. We also note that 'real' carbon is rarely found in isolation – hydrogenation and oxidation of carbon structures is not considered here. The expansion of the scope of the potential to treat hydrogenated or oxidised 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. 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 oxidation [31] or metal intercalation [179] 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 which 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 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 at `https://doi.org/10.17863/CAM.54529`, it 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 [154].

Chapter 6

General Conclusions

The application of novel machine learning techniques to the study of carbon has significant precedent in the physical sciences. From the first neural networks [61], to the first GAP models [60] and moment tensor potentials [123], carbon has often been used as a test system. What all of these prior applications have in common, is that they are in essence a showcase of the potential of a given ML algorithm. A brief look through the literature quickly shows, however, that those methodologies which are historically widely adopted by the scientific community - and therefore have the greater impact overall - are not those for which many isolated specialised applications exist, but for which a universally available implementation is made available to the community. In part, this is due to the highly specialised nature of modern research. By example: a major contributor to the impact of the research making use of DFT is down to the large number of universally available (often open source) implementations - one does not have to be an expert in DFT in order to apply it to a problem of interest. As such, a given researcher may be an expert in catalysis, excited state dynamics or any number of other fields, while leveraging DFT as a tool. It is a major aim of this thesis that by providing a broadly applicable implementation of machine learning for a particular system, which is of continued scientific interest, that others may be able to leverage the benefits of machine learning without necessarily being an expert in it.

In order to meet this aim of producing an accurate and broadly applicable model, fulfilling all of the criteria we have discussed a number of methodologi-

cal and conceptual advances had to be made. The structural diversity of carbon produces some unique challenges as previously discussed. Since, with the GAP methodology, we are limited (due to computational cost) in the number of configurations we can use for training, one must be very careful in selecting those configurations which are most important for producing a good model. As a researcher, we might be led to believe that we are a good authority for what configurations are 'important' for modelling carbon and which are not. It turns out (perhaps unsurprisingly) that this is not the case. Early on in the work presented here, our approach was to generate training structures which we believed would be important - individual defects, crystalline structures, clean surfaces, etc. The results of the application of models trained with datasets generated using this reductionist approach were often dissapointing; models tended to be inaccurate or unstable away from the sorts of well-defined, clean structures commonly encountered in DFT applications. For example, a GAP model trained on isolated Stone-Wales and monovacancy defects in graphene will perform well at those well-defined state points, but extremely poorly for a monovacancy adjacent a Stone-Wales defect. The solution to this problem was to approach the construction of a dataset from an entirely different perspective - generating as many varied and stochastic structures as possible, downloading carbon structures from online databases, melting crystal surfaces, etc. In doing so, we constructed a database of 17,000 configurations, representing almost 2.5 million local atomic environments. This is more than could ever reasonably be used for training a GAP model, so to thin these structures down to only the most important, we employed farthest point sampling; 'allowing the computer to decide' based on a structural metric which configurations were most important for the modelling. This process appears to be key to achieving the greatest measure of transferability in a ML model.

Closely connected to the above, is selecting the parameters of the descriptors (in particular the SOAP descriptors), to maximise the accuracy of the model while reducing the size of the descriptor. In this work, we showed for the first time that in the expansion of the SOAP neighbour density, the radial component of the spherical

harmonics used are far more important than the angular components. This allowed us to significantly improve the accuracy of the model, with no modifications to the fundemantal methodology required. This approach has since been used in the construction of two other GAP models using SOAP descriptors, one for hexagonal boron-nitride and a general model for phosophorus [225, 152].

A further necessary advancement was the treatment of long-range van der Waals effects in a computationally efficient way, in particular for layered structures like graphite. Many potential approaches could be conceived, including the direct use of D3 corrections as found in DFT. However, these corrections require the use of specific XC functionals and don't necessarily provide accurate results for the system of interest; in particular the use of a D3 correction would preclude the use of vdW inclusive functionals used in this work, which have been demonstrated to provide a far better agreement for layered materials with respect to experiment. Therefore, in this thesis we have developed a simple and broadly applicable approach that can be used to implement effective long-range van der Waals into any model - including older machine learning models which may not have been trained with van der Waals interactions in mind. We first compute a semianalytical spline, which matches the r^6 functional form for long-range interactions in the layered phase. We subtract the energy of this semianalyical function from the entire training dataset, and fit a GAP model as normal to the remaining energy. In this way, short-range van der Waals effects are fitted in a manybody fashion by the short-range SOAP descriptor, while long-range interactions are fitted with a semianalytical two-body potential - we are also guaranteed a smooth crossover between these two regimes because of the character of the SOAP cutoff function. Although similar semianalytical terms have been included in the repulsive component of GAP mdoels in the past, to the best of our knowledge, this is the first application in which the approach has been used to include long-range van der Waals interactions in a machine learned potential. This approach has also been subsequently applied to the construction of two other GAP potentials for layered materials, with good success [225, 152].

In terms of the potentials constructed, we have first investigated how to con-

struct a model for carbon for one particular phase, graphene. We have extensively tested this model against both experimental and theoretical values for the lattice parameter, vibrational properties and thermal expansion - demonstrating good agreement in all cases. We have then extended this model and combined it with existing datasets, to produce GAP-20, a potential for carbon which is accurate enough to model the crystalline phases of carbon, while maintaining the transferability required to study a wide range of liquid, amorphous and novel structures. This has again been rigorously tested for numerous physical properties, including lattice parameters, formation energies, phonon spectra, surface energies, liquid radial and angular distribution functions, activation barriers and random structure searches. We also note here, that the benchmark for quality and testing required when producing a widely-available model which claims broad applicability is arguably higher than in cases where a purpose-build model is constructed and used for one application 'in-house'. As such, additional stability testing has been performed to search for potential spurious behaviour, examples of this include liquid quenches from very high temperature, annealing simulations of amorphous structures and crystal surfaces, simulations of C_{60}@CNT 'pea pods', nudged elastic band calculations of defect formation and structure searches of carbon nanoclusters. This testing, combined with the quantitative testing analyses reported in the thesis, give confidence to the statement that the GAP-20 model can be applied to a wide range of carbon structures and be expected to produce accurate results.

That is not to say that GAP-20 is at all perfect. We have shown in this thesis that the prediction of diamond defect formation energies leaves something to be desired - GAP-20 often predicts the energies of these to be about 1-2 eV lower than is found from DFT. Similarly, the Jahn-Teller distortion known to be present in the graphene graphene is not reproduced. Perhaps the largest oversight of the GAP-20 model for studying 'real' carbonaceous systems is its elemental purity. Carbon is rarely, if ever, found in isolation. An extension of the GAP-20 model to include, for example, hydrogen and oxygen, would greatly extend the scope of application of the model. Even more desirably, an inclusion of the other elements commonly found

in living organisms, proteins and DNA: nitrogen, phosophorus, etc. would result in a model with the potential for significant scientific impact. Doing so would not necessarily be a straightforward 'extension' of the present model, however. Such is the nature of carbon, that the structures which it adopts in the presence of other elements bear little resemblance to its pure crystalline forms. The carbon backbone found in the double-helix of DNA could not be less similar in its structure or properties to those of a diamond. A further challenge would be that of long-range charge interactions, one which is only recently being addressed in the ML community. In pure carbon there is little chance of generating charged structures, dipoles or polar surfaces - but in the case of carbon bonded to hydrogen, oxygen or nitrogen, these are key to understanding the structure and properties of the material. Finally, we must address the choice of electronic structure method. DFT is a convenient choice, but if many thousands of CPU hours will be taken up in applying a ML model, it is ultimately economical to ensure that the highest quality reference configurations are performed. Problematically, there is no clear 'ideal' candidate for what those high quality reference configurations ought to be. For gas-phase molecules and non-periodic systems, coupled cluster calculations have been successfully applied [104, 78]. However, in the condensed phase, we require calculations which implement periodic boundary conditions naturally; such calculations are at present extremely challenging with coupled cluster. Methods like the random phase approximation (RPA) are appealing for this reason, but have the notable disadvantage of not quite reaching the 'gold standard' of accuracy achieved by coupled cluster. Recent advances have made Quantum Monte-Carlo (QMC) appear much more promising, as fully periodic calculations involving hundreds of electrons may now be performed in the condensed phase with a cost close to that of RPA [226, 227]. However, the lack of availability of atomic forces with QMC presents great challenges to its use for ML.

Overall, we believe that the work presented here achieves an excellent balance between accuracy and transferability. A number of methodological advances for the construction of ML models have also resulted from the work conducted over

the course of this PhD. Although there are many improvements which could - and hopefully will - be made, these are generally of the sort that require either entirely new projects, or methodological advances, or both. Moreover we believe that the final 'product' of the thesis, GAP-20, is a useful and practical tool for the scientific community to use; it is my hope that GAP-20 will have a life beyond the work presented in this document.

Appendix A

Additional Information on Graphene Model Testing

A.1 Force Errors and Lattice Parameters of Empirical Models

The reference DFT method (optB88-vdW) overestimates the lattice parameter by 0.002 Å (0.08%) with respect to the experimentally determined graphite lattice parameter of 2.462 Å [142]. The remaining DFT functionals considered also overestimate the lattice parameter by between 0.002 and 0.008 Å, with the exception of LDA which underestimates the lattice parameter by 0.016 Å (0.65%) with respect to the experimentally determined value. We note that the force errors associated with the quality of the fit of the GAP model to the DFT reference is in general smaller than or comparable to the force error if measured between two different exchange-correlation functionals.

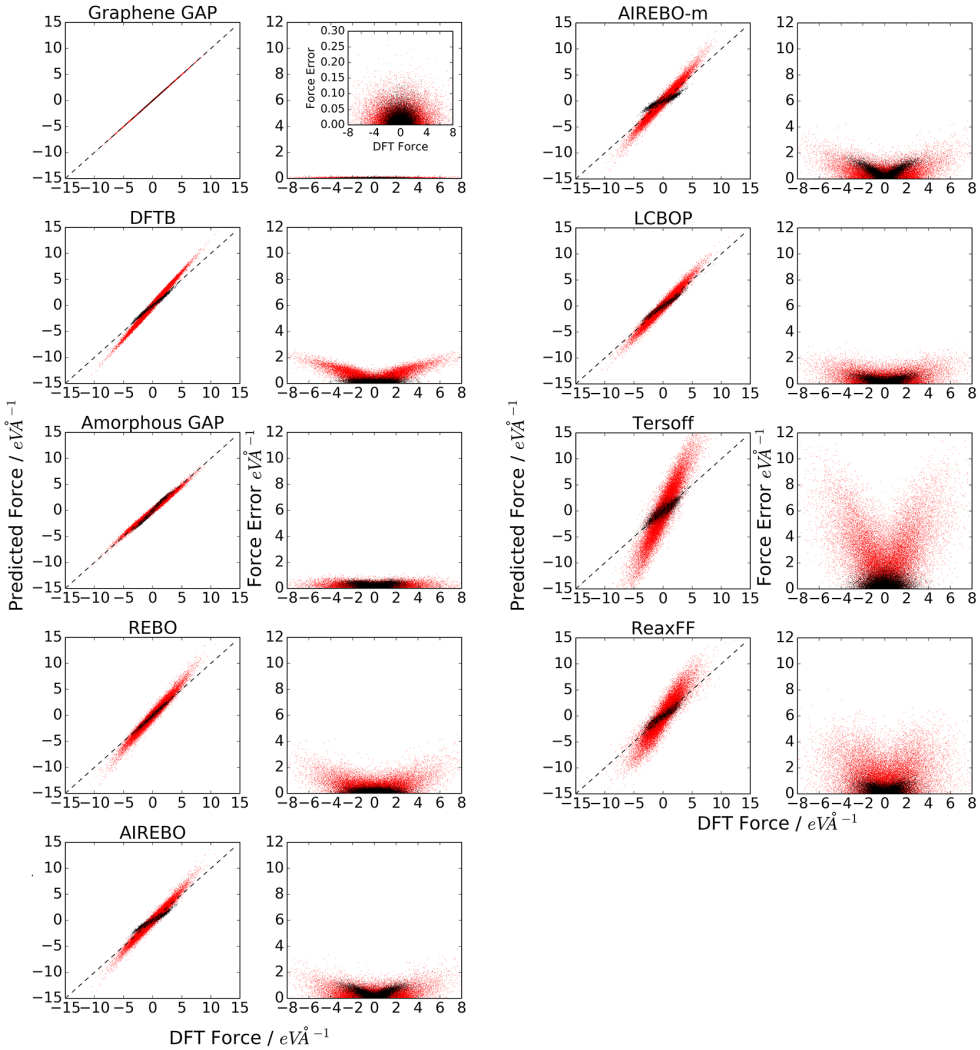

Figure A.1: Graphene force errors for empirical potentials compared to DFT
Force errors for all tested classical potentials, Amorphous and Graphene GAP potentials and DFTB, compared to optB88-vdW dft.

A.2 Functional Sensitivity of Forces; Correlation of Forces with Different Functionals

Potential	RMSE (In-plane) eV Å^{-1}	RMSE (Out-of-plane) eV Å^{-1}	Lattice parameter (0 K)	Time (Relative)
Graphene GAP	0.028	0.019	2.467 (+0.003)	344
Amorphous GAP	0.270	0.258	2.430 (-0.03)	-
Tersoff	3.122	0.542	2.530 (+0.08)	1
REBO	0.722	0.187	2.460 (-0.004)	1.2
AIREBO	0.548	0.414	2.419 (-0.05)	1.9
AIREBO-m	0.720	0.568	2.459 (-0.005)	2.9
LCBOP	0.595	0.306	2.459 (-0.005)	2.3
ReaxFF	1.226	0.311	2.462 (-0.002)	23
DFTB	0.693	0.162	2.470 (+0.006)	950
optB88-vdW	-	-	2.464	2×10^7 (AIMD)
LDA	0.015	0.058	2.446 (-0.018)	-
PBE	0.032	0.008	2.467 (+0.003)	-
optB86b-vdW	0.027	0.010	2.466 (+0.002)	-
optPBE-vdW	0.017	0.017	2.471 (+0.006)	-
PBE-D3[228]	0.032	0.008	2.467 (+0.003)	-
PBE-TS[229]	0.031	0.008	2.464 (+0.0)	-
Exp. (Graphite, 300 K)			2.462	

Table A.1: Root mean squared force errors, lattice parameters predicted and relative costs of empirical many-body and GAP models. Bracketed values represent the difference in lattice parameter associated with choosing a different exchange correlation functional rather than optB88-vdW as the reference DFT method for fitting and benchmarking. Dashed entries for timings have not been measured.

A.3 Functional Sensitivity of Graphene Phonon Dispersion Curves

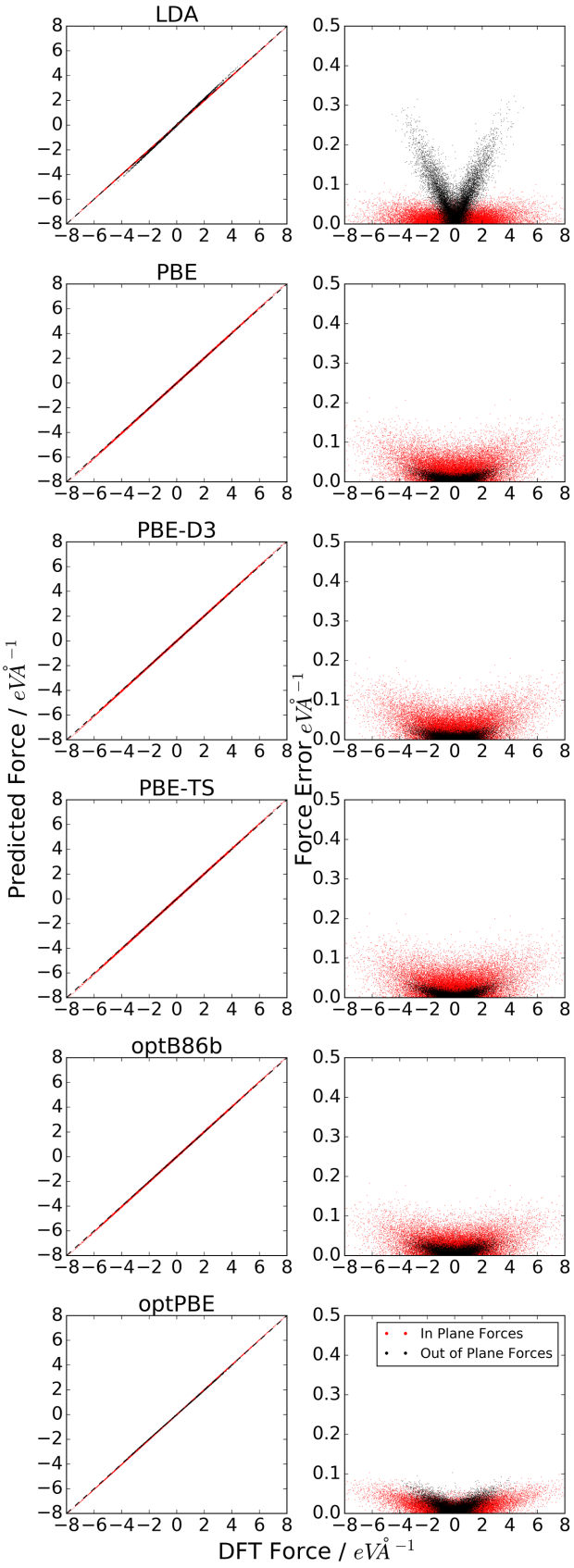

Figure A.2: Force errors for other choices of DFT functional versus the chosen optB88-vdW reference.

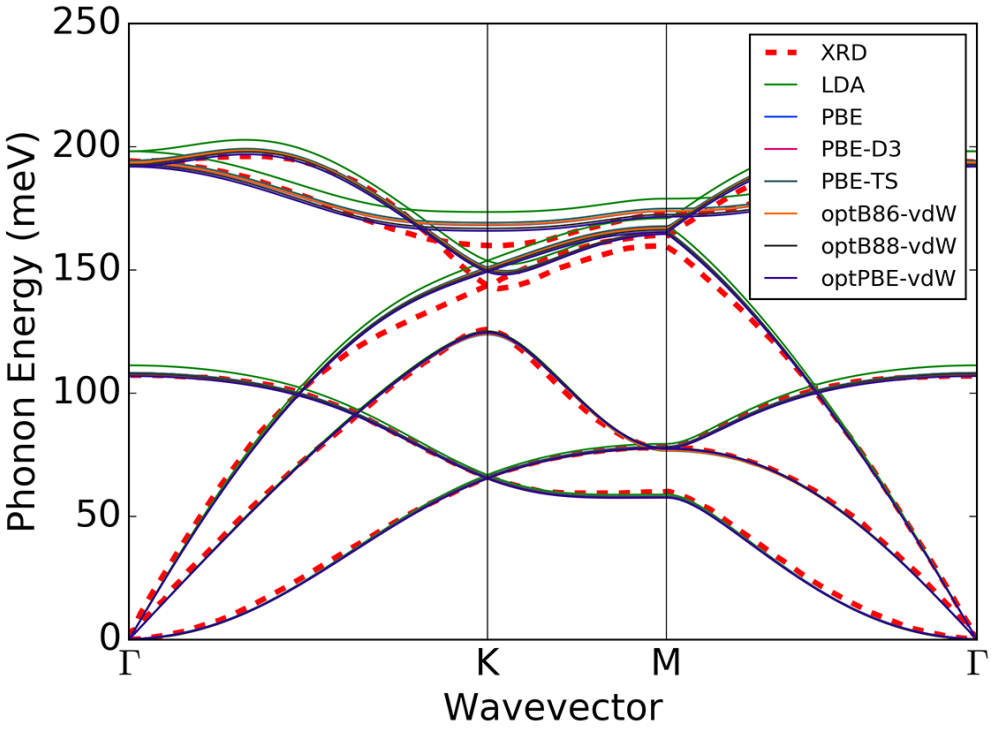

Figure A.3: Comparison of phonon spectra as calculated using the finite displacement method for a variety of DFT functionals, highlighting the robustness of these results to this choice. We note that the errors present in the graphene GAP versus experiment in the main text are the result of a close fit to the reference DFT method - the same inaccuracies close to the K high symmetry point are observed both for the DFT reference and the graphene GAP.

Appendix B

Additional Information on GAP-20 Carbon Model Testing

In this appendix, we include data on further tests performed on GAP-20, in addition to further data for comparison with the other models considered in this work. In section 1, we give details of the GAP model parameters used in this text. In section 2, we provide tabulated data for the crystalline formation energies presented in the main text, this is followed in section 3 by further information on the predicted formation energies of a range of armchair and zig-zag nanotubes. Sections 4, 5, 6 and 7 give the phonon dispersion curves for graphene, diamond, a 9,9-index nanotube and a 9,0-index nanotube, for all of the models considered. In section 8, we plot the graphene bilayer separation curves for the models tested. In Section 9, the force and energy errors for a wide range of configurations considered in the training set are provided. In section 10, we give some further details on the optimisation of the GAP model. Finally, in section 11, we provide some information on the efficiency of the model compared to direct *ab initio* simulation.

B.1 Training of the Potential

Table B.1: Hyperparameters of the GAP Model. Note that modified values for σ_{energy}, σ_{force} and σ_{virial} are used for a number of sets of configurations.

GAP 2b Descriptor	
Cutoff Å	4.5
δ	2.0
Sparse Method	uniform
Covariance	Gaussian
Sparse points	15
3b Descriptor	
Cutoff Å	2.5
δ	0.05
Sparse Method	uniform
Covariance	Gaussian
Sparse points	200
SOAP Descriptor	
Cutoff Å	4.5
Cutoff width Å	1.0
δ	0.2
Sparse method	CUR
Sparse points	9000
l_{max}	4
n_{max}	12
ζ	4
Global Parameters	
σ_{energy} [eV]	0.001
σ_{force} [eV / Å]	0.1
σ_{virial} [eV]	0.2

Table B.2: Due to the differing energy scales involved in training, specific sets of configurations are fitted with customised values for σ_{energy}, σ_{force} and σ_{virial}. The values for these are tabulated here.

Config Type	σ_{energy} [eV]	σ_{force} [eV / Å]	σ_{virial} [eV]
Graphite	0.001	0.01	0.05
Diamond	0.001	0.01	0.05
Graphene	0.001	0.01	0.05
Crystalline Bulk	0.001	0.01	0.05
Nanotubes	0.001	0.01	0.05
Fullerenes	0.002	0.1	0.2
Amorphous Bulk	0.005	0.2	0.2
Liquid	0.050	0.5	0.5
Defects	0.001	0.01	0.05
Surfaces	0.002	0.1	0.2
Amorphous Surfaces	0.005	0.2	0.2
Liquid Interface	0.050	0.5	0.5
Graphite Layer Separation	0.001	0.01	0.05
Dimer	0.002	0.1	0.2
Single Atom	0.0001	0.001	0.05

B.2 Crystalline Properties

Table B.3: Formation energies of the crystalline carbon phases computed with GAP-20 and the tested empirical models compared to those from DFT. In the first table, absolute values of the formation energies relative to the gas phase atom are given, with errors relative to the DFT value in brackets. In the second part of the table, the values for the formation energies are instead given relative to graphite.

	Formation Energy [eV] (% Error)					
	DFT	GAP-20	Tersoff	LCBOP	REBO-II	AIREBO
Graphite	-8.98	-8.97 (0.1)	-7.40 (18)	-7.38 (21)	-7.40 (18)	-7.48 (17)
Diamond	-8.85	-8.85 (0.1)	-7.37 (18)	-7.35 (18)	-7.37 (18)	-7.46 (17)
Graphene	-8.91	-8.9 (0.1)	-7.40 (18)	-7.35 (18)	-7.40 (18)	-7.43 (17)
Hexagonal Diamond	-8.85	-8.84 (0.1)	-7.37 (17)	-7.34 (18)	-7.37 (17)	-7.30 (18)
Nanotube-(9, 9)	-8.73	-8.74 (0.1)	-7.28 (17)	-7.27 (17)	-7.26 (17)	-7.33 (16)
Nanotube-(9, 0)	-8.75	-8.75 (0.0)	-7.20 (18)	-7.16 (19)	-7.18 (19)	-7.23 (18)
C60 Fullerene	-8.53	-8.51 (0.2)	-6.73 (22)	-6.93 (19)	-6.84 (20)	-6.81 (21)
C100 Fullerene	-8.52	-8.60 (0.9)	-6.94 (20)	-7.06 (19)	-7.00 (20)	-7.02 (19)
	Formation Energy Rel. Graphite [eV] (% Error)					
Graphite	0	0	0	0	0	0
Diamond	0.12	0.12 (1.2)	0.03 (67)	0.03 (67)	0.03 (67)	0.02 (78)
Graphene	0.07	0.07 (0.4)	0.00 (100)	0.03 (100)	0.00 (100)	0.05 (29)
Lonsdaleite	0.13	0.13 (1.6)	0.03 (75)	0.04 (67)	0.03 (75)	0.18 (50)
Nanotube-(9, 9)	0.25	0.22 (7.5)	0.12 (57)	0.11 (61)	0.14 (50)	0.15 (46)
Nanotube-(9, 0)	0.22	0.22 (1.8)	0.20 (9)	0.22 (0)	0.22 (0)	0.25 (14)
C60 Fullerene	0.45	0.45 (2.3)	0.67 (49)	0.45 (0)	0.56 (24)	0.67 (49)
C100 Fullerene	0.35	0.37 (5.7)	0.46 (31)	0.32 (9)	0.40 (14)	0.46 (31)

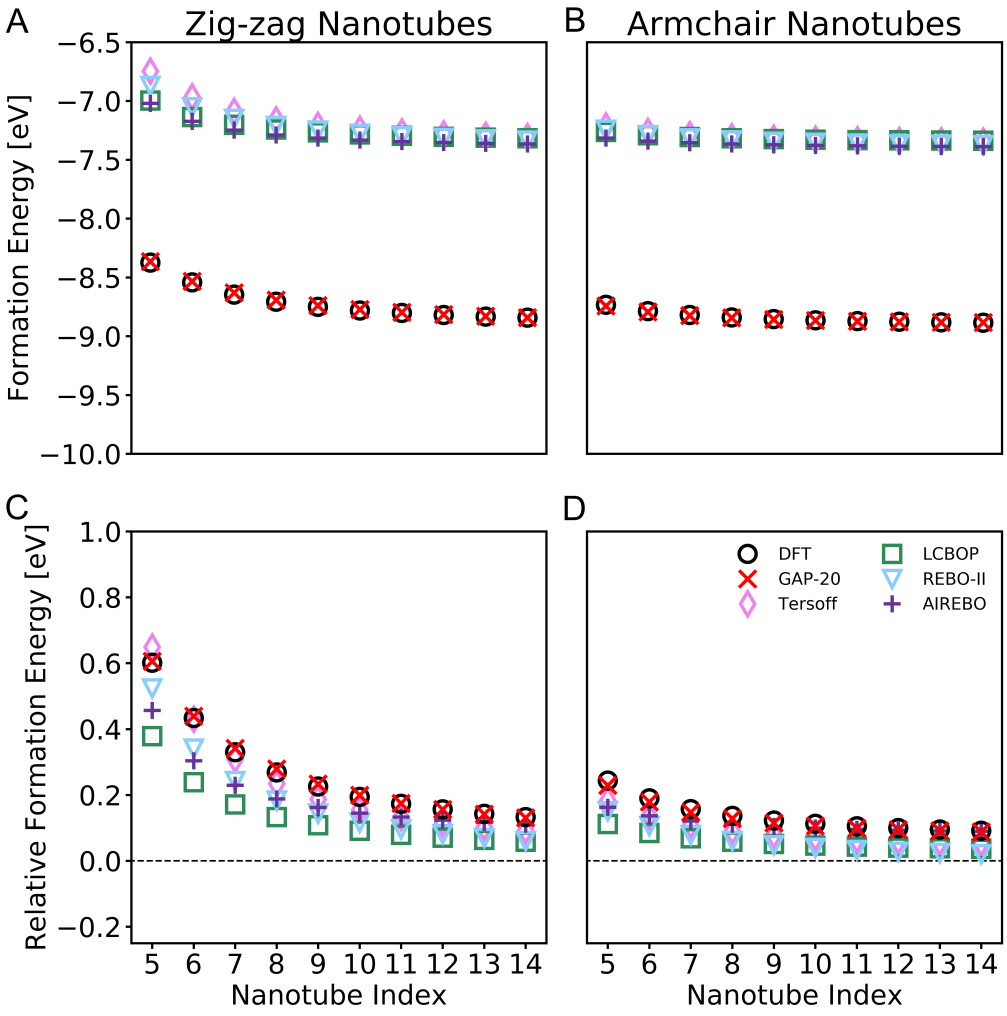

Figure B.1: Formation energies of zig-zag and armchair chirality carbon nanotubes compared to those from DFT. (A) Zig-zag nanotubes formation energies. (B) Armchair nanotubes formation energies. (C) Zig-zag nanotubes formation energies given relative to the formation energy of graphite for each model. (D) Armchair nanotubes formation energies given relative to the formation energy of graphite for each model.

B.3 Graphene Phonon Dispersion Curves

While the phonon dispersion curves for graphene as calculated with a number of empirical models have been reported previously, they are provided here in figure B.2 again for completeness - we have additionally computed the phonon dispersion relation for graphene as predicted with the existing GAP-17 potential. Many of the potentials achieve a reasonably good description of the shape of the low frequency phonon modes, however they struggle to describe the spare and energy of higher

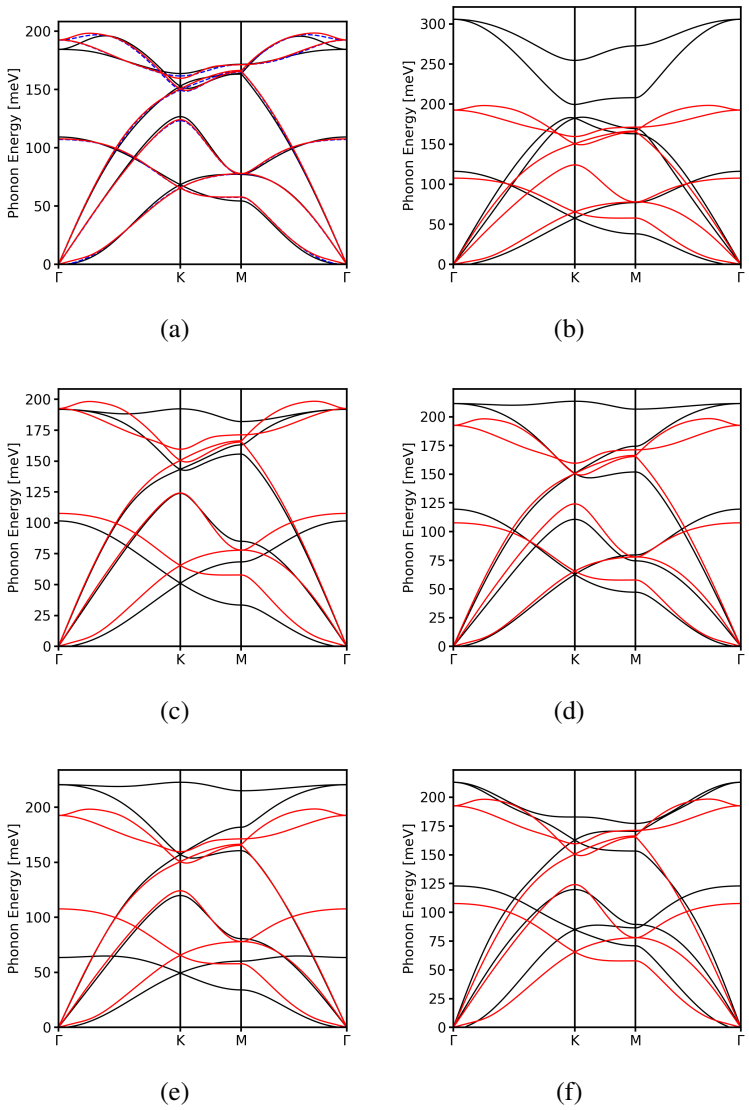

Figure B.2: Phonon dispersion curves for graphene calculated with (A) GAP-20 (B) Tersoff Potential (C) LCBOP (D) REBO (E) AIREBO (F) GAP-17. DFT (optB88-vdW) Reference data are shown in red, while model test data are shown in black. Note that the in this instance, we compare GAP-17 (trained using the local density approximation (LDA) DFT functional) to optB88-vdW reference data, so some disagreement is inevitable. A comparison of the functional dependence of these data for graphene can be found in [107].

frequency modes.

B.4 Diamond Phonon Dispersion Curves

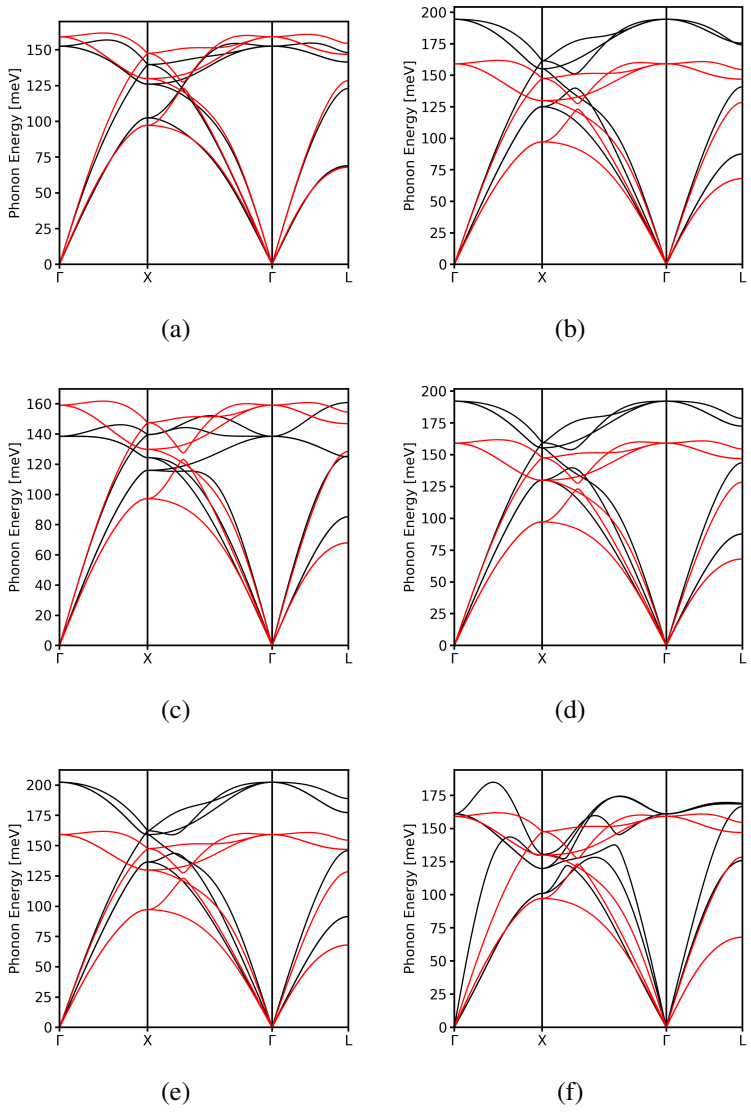

Figure B.3: Phonon dispersion curves for diamond calculated with (A) GAP-20 (B) Tersoff Potential (C) LCBOP (D) REBO (E) AIREBO (F) GAP-17. DFT Reference data are shown in red, while model test data are shown in black. Note that the in this instance, we compare GAP-17 (trained using the LDA DFT functional) to optB88-vdW reference data, so some disagreement is inevitable. A comparison of the functional dependence of the phonon spectrum of graphene can be found in [107].

In figure B.3 we provide the calculated phonon modes for diamond using our GAP-20 model, the empirical models tested and the GAP-17 potential. In most cases, the empirical models struggle to predict the shape of the phonon dispersion relations and the phonon band energies are often inaccurate.

B.5 Nanotube-(9, 9) Phonon Dispersion Curves

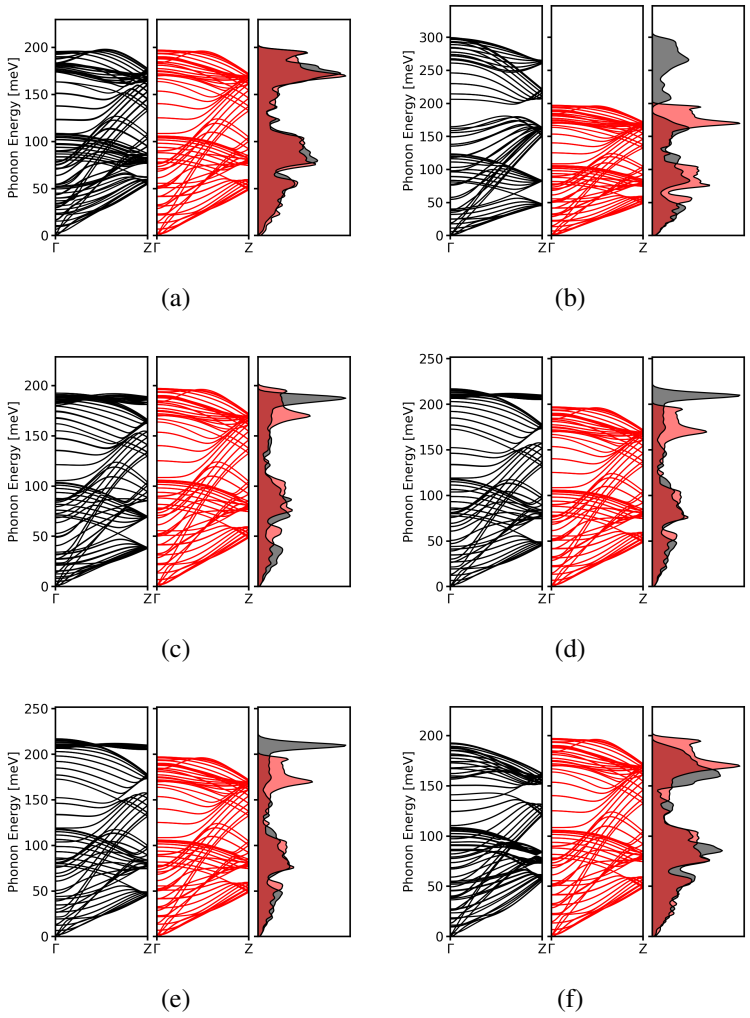

Figure B.4: Phonon dispersion curves and density of states for (9,9) index nanotube calculated with (A) GAP-20 (B) Tersoff Potential (C) LCBOP (D) REBO (E) AIREBO (F) GAP-17. DFT Reference data are shown in red, while model test data are shown in black. Note that the in this instance, we compare GAP-17 (trained using the LDA DFT functional) to optB88-vdW reference data, so some disagreement is inevitable. A comparison of the functional dependence of the phonon spectrum of graphene can be found in [107].

In figure B.4 we report the phonon dispersion relations for a (9, 9) index carbon nanotube, additionally showing the density of states. It is challenging to make a detailed comparison here regarding specific modes, due to the number and similarity of the various phonon modes. It is more convenient to refer to the density of states in order to draw broad conclusions. Similarly to the case of graphene, the Tersoff

potential vastly overestimates the energy of the highest frequency vibrational modes of carbon nanotubes as well as predicting the degeneracy of many phonon modes close to the Brillouin zone edge. The LCBOP, REBO and AIREBO potentials all predict a large spurious peak in the vibrational density of states at high energies. The GAP-17 model correctly predicts the energies of the highest energy modes, but again predicts degenerate vibrational states near the Brillouin zone edge. The GAP-20 model introduced here shows some broadening of the peaks in the vibrational density of states, but generally provides an accurate description of the shape and energies of the phonon dispersion.

B.6 Nanotube-(9, 0) Phonon Dispersion Curves

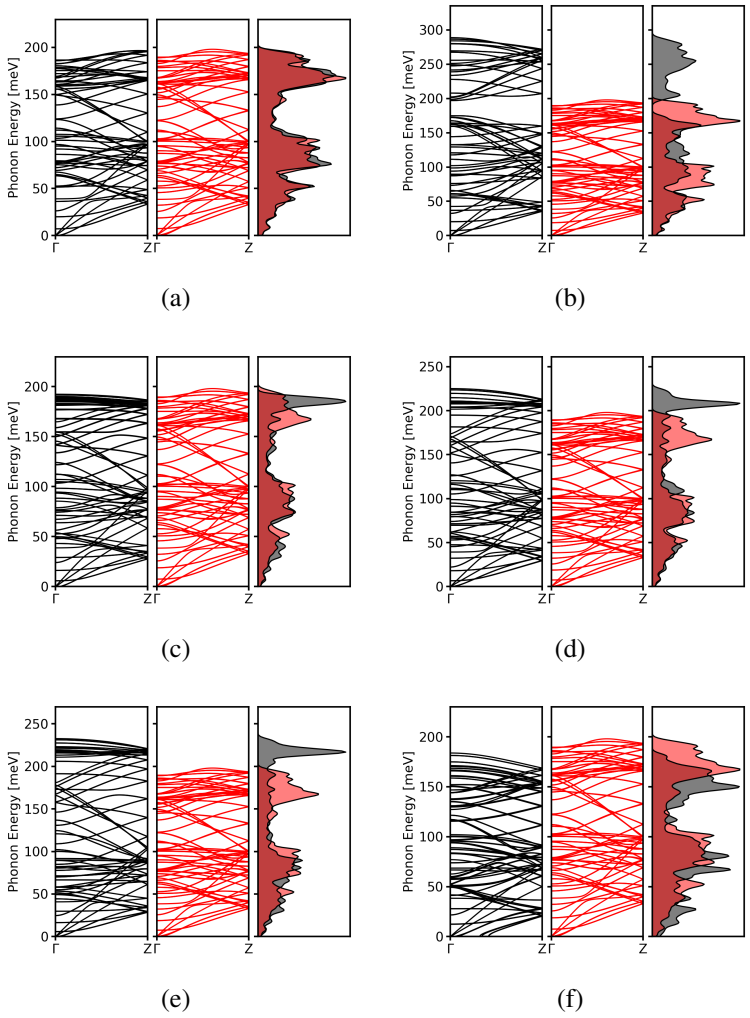

Figure B.5: Phonon dispersion curves and density of states for (9,0) index nanotube calculated with (A) GAP-20 (B) Tersoff Potential (C) LCBOP (D) REBO (E) AIREBO (F) GAP-17. DFT Reference data are shown in red, while model test data are shown in black. Note that the in this instance, we compare GAP-17 (trained using the LDA DFT functional) to optB88-vdW reference data, so some disagreement is inevitable. A comparison of the functional dependence of the phonon spectrum of graphene can be found in [107].

In figure B.5 we present the phonon dispersion curves and vibrational density of states for a (9, 0) index carbon nanotube as predicted by GAP-20, a number of empirical potentials and the GAP-17 model. As in the case of the (9, 9) index carbon nanotube, the Tersoff potential overestimates the energy of the high energy phonon modes by approx. 100 meV, while the REBO, LCBOP and AIREBO potentials have

a spurious peak in the vibrational density of states at high energies. In this instance, the GAP-17 model underestimates the energy of the phonon modes, but predicts the correct shape for the density of states. The GAP-20 model shows some broadening of the peaks in the vibrational density of states, but is generally accurate for both energies and dispersion shape.

B.7 Graphene bilayer separation energy

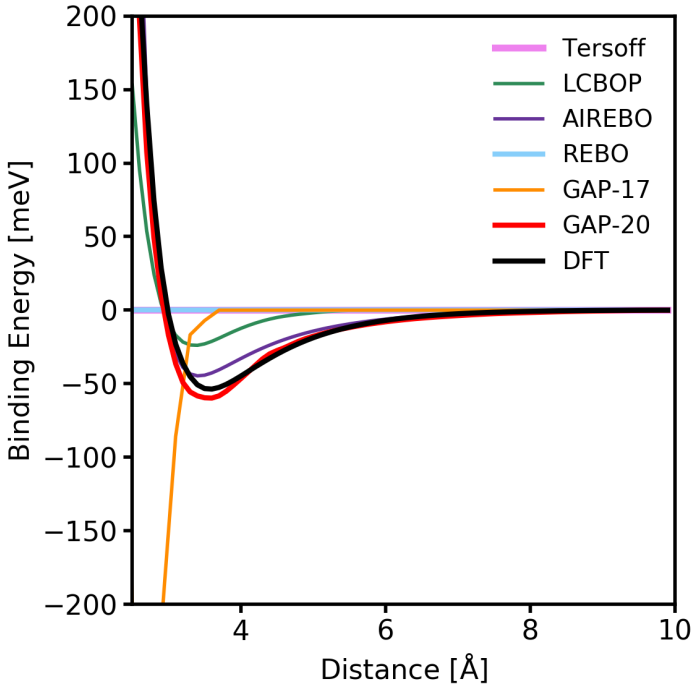

Figure B.6: Energy required to move a bound system of two graphene sheets out to large separation.

Figure B.6 shows the predicted energy as the distance between two graphene sheets is changed. It provides a useful test of how well a potential might capture the long-ranged Van der Waals interactions, which are important for modelling layered materials like graphite. GAP-20 accruately reproduces the shape and depth of the binding curve out to long separation as do both the AIREBO and LCBOP potentials, although the latter does underbind the two sheets. GAP-17 strongly overbinds the two sheets at very short distances. Due to their extremely short cutoffs, the REBO and Tersoff potentials predict no interaction between the graphene sheets even at very short distances.

B.8 Force and Energy Errors of the Potential

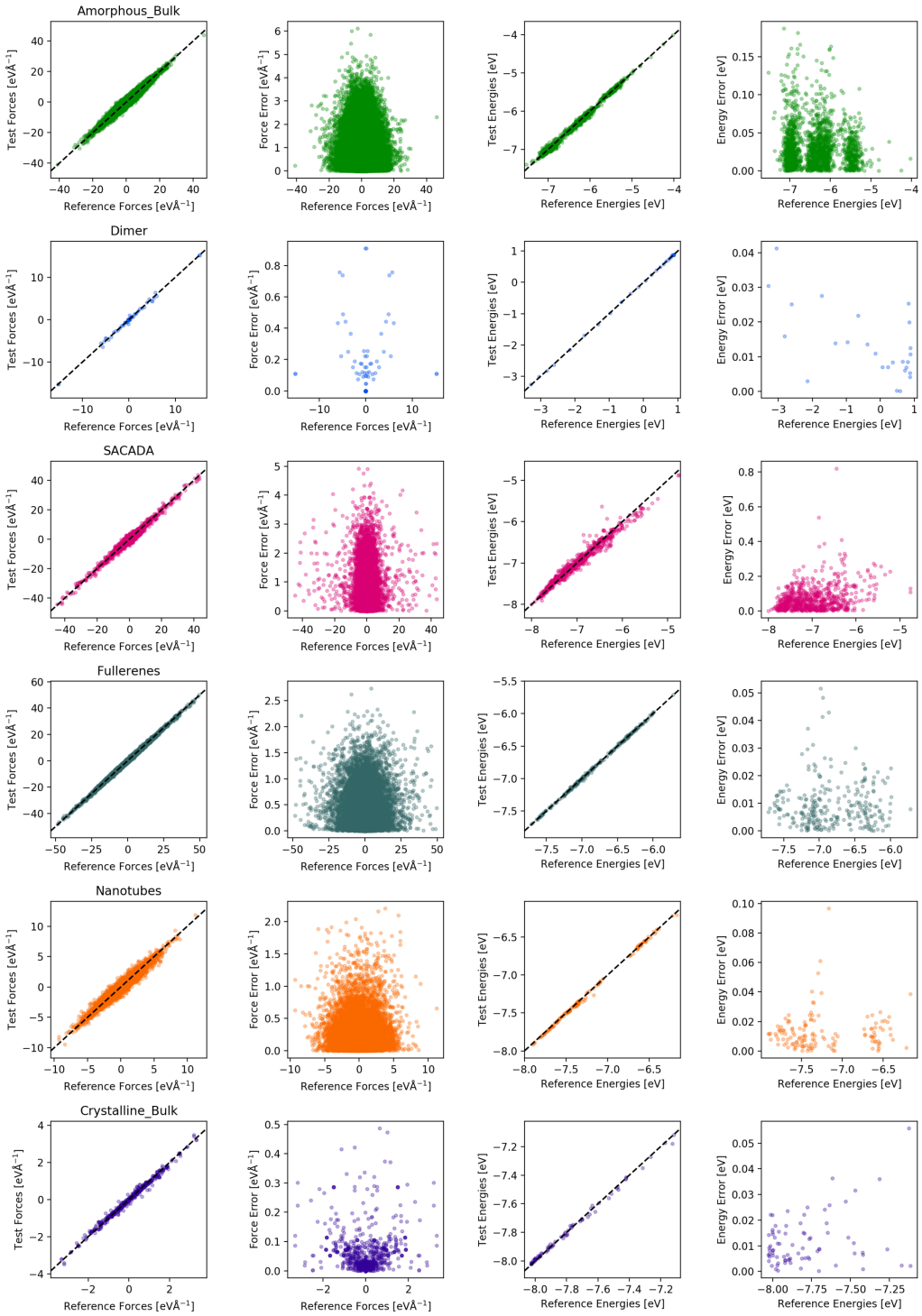

Figure B.7: Force Errors (1/3) (Left to right) Correlation between GAP-20 and DFT (optB88-vdW) forces, errors in force prediction for GAP-20, correlation between GAP-20 and DFT (optB88-vdW) energies and errors in energy prediction for GAP-20. (Top to bottom) data for bulk amorphous, dimer, SACADA [184] database, fullerenes, nanotubes and bulk crystalline structues.

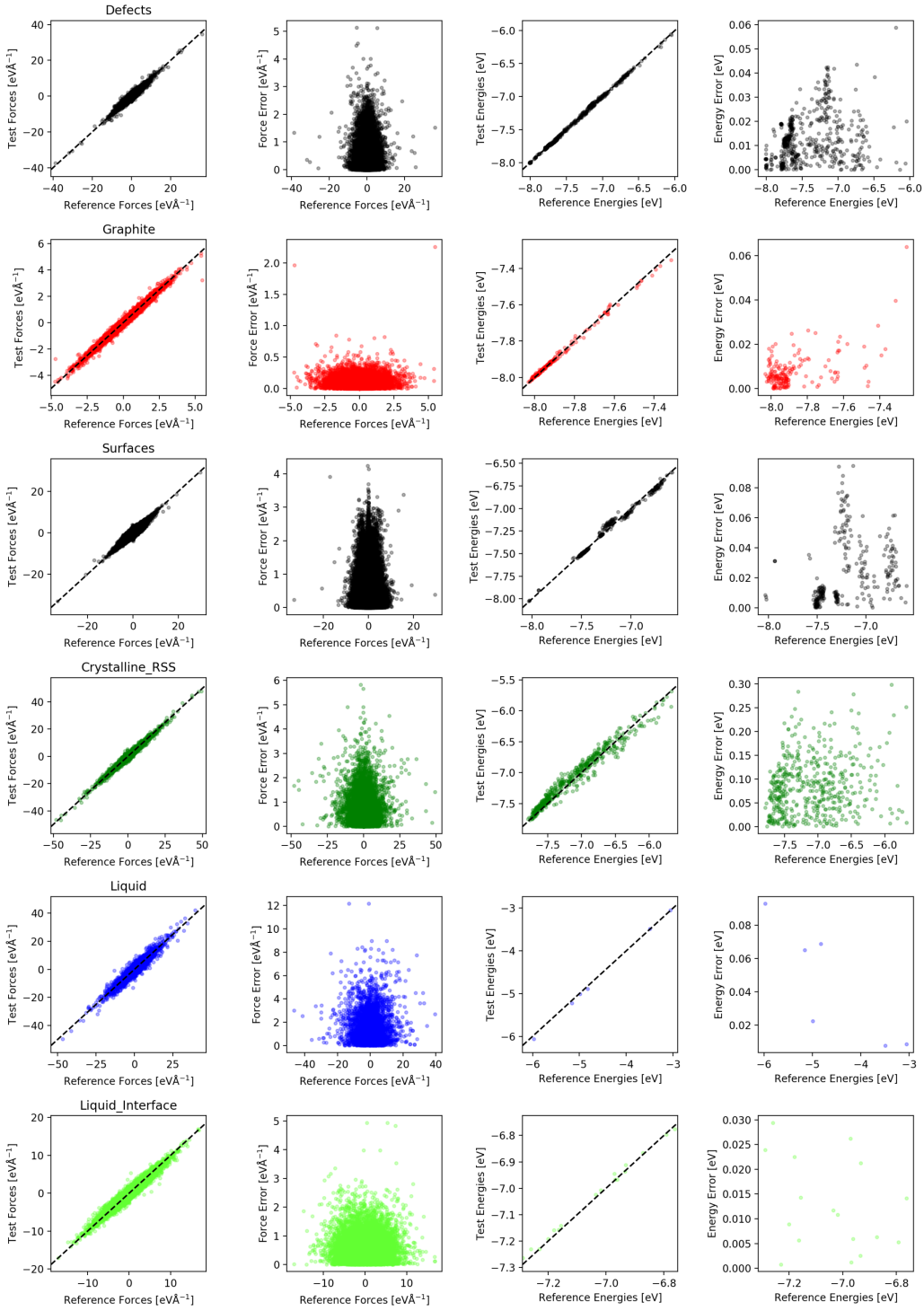

Figure B.8: Force Errors (2/3) (Left to right) Correlation between GAP-20 and DFT (optB88-vdW) forces, errors in force prediction for GAP-20, correlation between GAP-20 and DFT (optB88-vdW) energies and errors in energy prediction for GAP-20. (Top to bottom) data for defective carbon, graphite, carbon surfaces, crystalline RSS data [23], liquid carbon and liquid-solid interfaces.

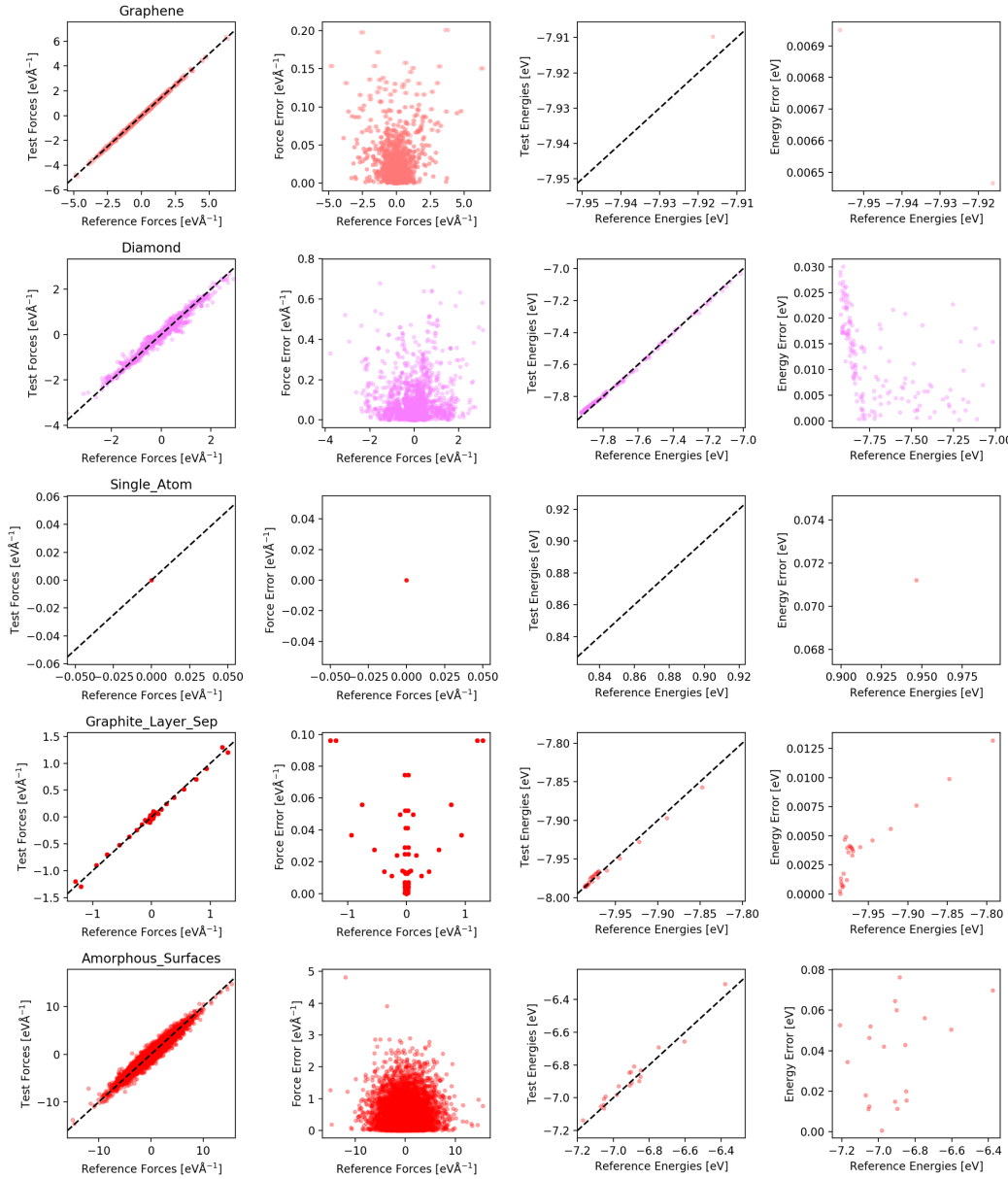

Figure B.9: Force Errors (3/3) (Left to right) Correlation between GAP-20 and DFT (optB88-vdW) forces, errors in force prediction for GAP-20, correlation between GAP-20 and DFT (optB88-vdW) energies and errors in energy prediction for GAP-20. (Top to bottom) data for graphene, diamond, the isolated gas-phase atom, bilayer graphene and amorphous carbon surfaces.

B.9 Optimisation of SOAP Descriptor

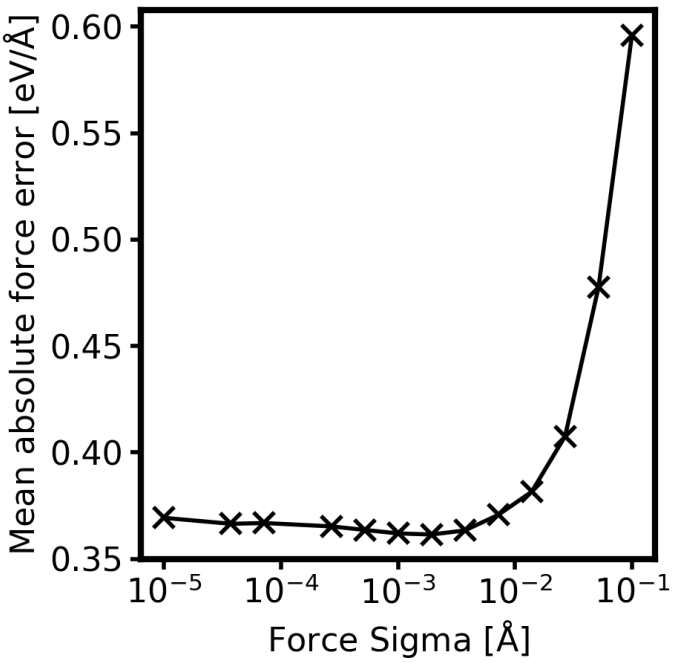

Figure B.10: Variation of mean absolute force errors on the test set for an l=8, n=8 SOAP descriptor with a 4.2 Åcutoff and 6000 sparse points with respect to the σ_{force} parameter used in GAP training. The minimum is present as a result of reaching an optimal balance between overfitting and underfitting.

Figure B.10 shows the convergence of the force error as a function of the selected value for σ_{force}. The accuracy of the potential is strongly affected if values larger than approx 0.01 eV Å^{-1} are used, and begin to increase again for very small values as over-fitting becomes problematic. In general, one wishes to use the largest value possible without negatively impacting the quality of the potential, thereby minimising the effects of over-fitting and maximising the resistance of the potential to noise or other possible artefacts in the training data.

B.10 Random Structure Search

In addition to showing the error of GAP-20 compared to DFT when computing the energies of the GAP-RSS structures, we here also plot in sketch-map representation (fig B.11 the energies of the structures themselves.

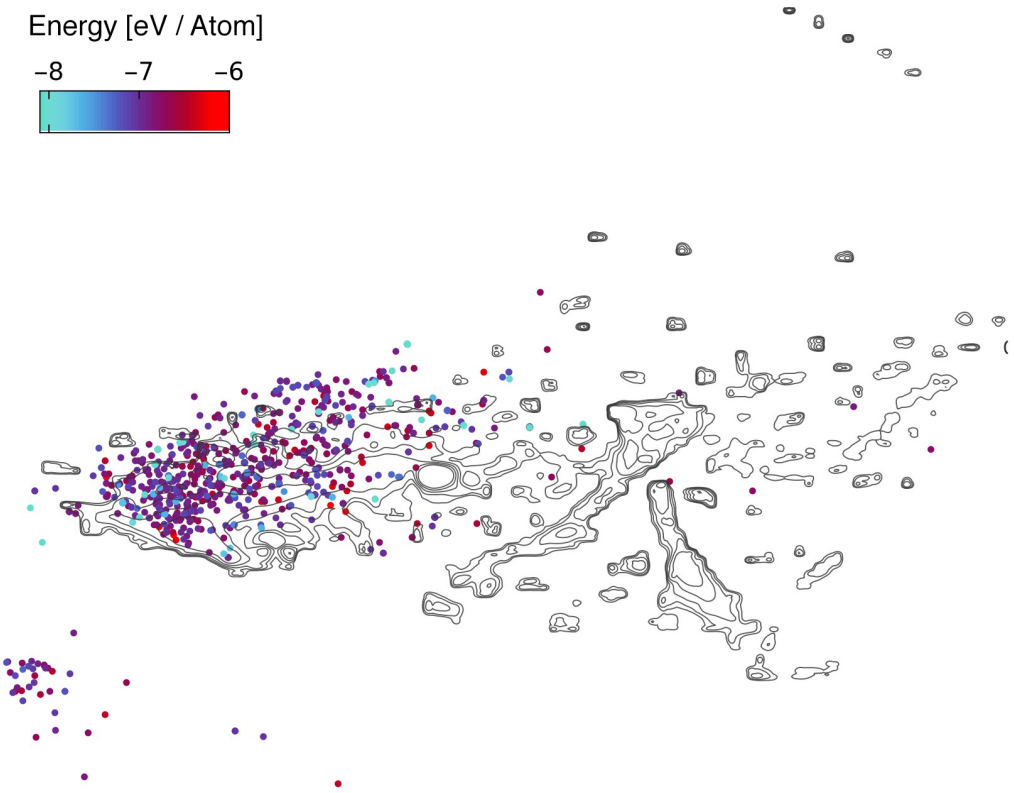

Figure B.11: Sketch-map representation of GAP-RSS structures, the energy of each structure is indicated by the colour of the point. The density of the population of structures in the GAP-20 training data are indicated by the black contour lines.

B.11 Cost of the Potential

The primary benefit to using a Gaussian approximation potential over direct *ab initio* simulation is the significant reduction in the cost of simulations. We have previously commented on the relative cost of direct *ab initio* simulation compared to GAP models, density functional tight binding and a number of empirical models for a small 200 atom graphene system [107]. Here, we compare the cost of our GAP model to that of direct *ab initio* simulation with VASP and to the LCBOP empirical many body potential for carbon. Although there are many empirical potentials for carbon available, for our illustrative purposes here they perform effectively the same; the LCBOP potential could be substituted for any number of other empirical potentials without affecting our conclusions here.

Simulations were performed in the NVT ensemble at 1000 K for diamond lat-

tices ranging from 8 to 5832 atoms. These simulations were performed using 72 cores spread over 3 nodes on the Thomas cluster, the UK National Tier 2 High Performance Computing Hub for Materials and Molecular Modelling.

In the case of GAP and LCBOP simulations, trajectories of 20,000 steps were generated, for VASP *ab initio* molecular dynamics simulations, trajectories of 10 steps were generated. *Ab initio* molecular dynamics simulations were performed at the gamma point, the electronic convergence criterion was set to 10^{-4} eV, all other settings were fixed as described for the generation of the training data. We note that for very small system sizes, the empirical models will spend a considerable portion of their time on communication rather than on calculation of the pair potential, however this is unavoidable if we wish to perform comparable simulations on larger cells with DFT.

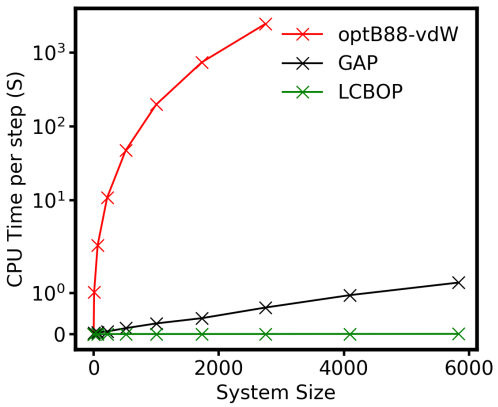

Figure B.12: Efficiency of the GAP model for simulations of various sizes. Note the logarithmic scale of the y-axis.

As can be seen from figure B.12, the GAP model is many orders of magnitude cheaper than *ab initio* molecular dynamics, even for relatively modest system sizes. Furthermore, since the scaling of the GAP model follows the approximately Nlog(N) scaling of empirical MD, while that of DFT scales with the cube of the number of electrons, the advantage to using the GAP model increases dramatically for larger systems. In fact, simulations using DFT for systems larger than 2744 atoms was not possible with the chosen computational setup. While GAP model does, have a much larger prefactor to this scaling than any of the empirical models

tested, it remains entirely feasible to use it to simulate systems of tens of thousands of atoms for nanoseconds at a time.

Bibliography

[1] Hugh Aldersey-Williams. *The Most Beautiful Molecule: The Discovery of the Buckyball*. John Wiley & Sons, New York, 1st ed edition, 1995.

[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] Nobel Media AB 2020. The Nobel Prize in Chemistry 1996, 1996.

[4] H. Kroto. C60: The celestial sphere which fell to earth. *Nanotechnology*, 3(3):111–112, 1992.

[5] Sumio Iijima, Tomonari Wakabayashi, and Yohji Achiba. Structures of carbon soot prepared by laser ablation. *Journal of Physical Chemistry*, 100(14):5839–5843, 1996.

[6] Sumio Iijima. Carbon nanotubes: Past, present, and future. *Physica B: Condensed Matter*, 323(1-4):1–5, 2002.

[7] Iskandar N. Kholmanov, Carl W. Magnuson, Richard Piner, Jin Young Kim, Ali E. Aliev, Cheng Tan, Tae Young Kim, Anvar A. Zakhidov, Giorgio Sberveglieri, Ray H. Baughman, and Rodney S. Ruoff. Optical, electrical, and electromechanical properties of hybrid graphene/carbon nanotube films. *Advanced Materials*, 27(19):3053–3059, 2015.

[8] Consol Farrera, Fernando Torres Andón, and Neus Feliu. Carbon Nanotubes as Optical Sensors in Biomedicine. *ACS Nano*, 11(11):10637–10643, 2017.

[9] Nobel Media AB 2020. The Nobel Prize in Physics 2010, 2010.

[10] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Griegorieva, and A. A. Firsov. Electric Field Effect in Atomically Thin Carbon Films. *Science*, 306(5696):666–669, 2004.

[11] A. H. Castro Neto, F. Guinea, N. M R Peres, K. S. Novoselov, and A. K. Geim. The electronic properties of graphene. *Reviews of Modern Physics*, 81(1):109–162, 2009.

[12] M. Mohr, J. Maultzsch, E. Dobardžić, S. Reich, I. Milošević, M. Damnjanović, A. Bosak, M. Krisch, and C. Thomsen. Phonon dispersion of graphite by inelastic x-ray scattering. *Physical Review B*, 76(3):035439, 2007.

[13] Jia An Yan, W. Y. Ruan, and M. Y. Chou. Phonon dispersions and vibrational properties of monolayer, bilayer, and trilayer graphene: Density-functional perturbation theory. *Physical Review B*, 77(12):125401, 2008.

[14] J Maultzsch, S Reich, C Thomsen, H Requardt, and P Ordejón. Phonon Dispersion in Graphite. *Physical Review Letters*, 92(7):075501, 2004.

[15] Yuanbo Zhang, YW Tan, HL Stormer, and Philip Kim. Experimental observation of the quantum Hall effect and Berry's phase in graphene. *Nature*, 438(November):201–204, 2005.

[16] Changgu Lee, Xiaoding Wei, Jeffrey W Kysar, and James Hone. Measurement of the elastic properties and intrinsic strength of monolayer graphene. *Science*, 321(5887):385–388, 2008.

[17] Phaedon Avouris, Zhihong Chen, and Vasili Perebeinos. Carbon-based electronics. *Nature Nanotechnology*, 2(10):605–15, 2007.

[18] F Bonaccorso, Z Sun, T Hasan, and A C Ferrari. Graphene Photonics and Optoelectronics. *Nature Photonics*, 4(9):611–622, 2010.

[19] Lars Pastewka, Stefan Moser, Peter Gumbsch, and Michael Moseler. Anisotropic mechanical amorphization drives wear in diamond. *Nature Materials*, 10(1):34–38, 2011.

[20] T. B. Shiell, 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. *Physical Review Letters*, 120(21):215701, 2018.

[21] Miguel Martinez-Canales and Chris J Pickard. Thermodynamically Stable Phases of Carbon at Multiterapascal Pressures. *Physical Review Letters*, 108(4):045704, 2012.

[22] R. C. Powles, N. A. Marks, and D. W M Lau. Self-assembly of sp2 -bonded carbon nanostructures from amorphous precursors. *Physical Review B - Condensed Matter and Materials Physics*, 79(7):1–11, 2009.

[23] Volker L. Deringer, Gábor Csányi, and Davide M. Proserpio. Extracting Crystal Chemistry from Amorphous Carbon Structures. *ChemPhysChem*, 18(8):873–877, 2017.

[24] David Tománek. *Guide Through the Nanocarbon Jungle*. Morgan & Claypool Publishers, San Rafael, 1 edition, 2014.

[25] Rasim Mirzayev, Kimmo Mustonen, Mohammad R. A. Monazam, Andreas Mittelberger, Timothy J. Pennycook, Clemens Mangler, Toma Susi, Jani Kotakoski, and Jannik C. Meyer. Buckyball sandwiches. *Science Advances*, 3(6):e1700176, 2017.

[26] M. H. Nazare and A. J Neves. *Properties Growth and Applications of Diamond*. INSPEC, London, 1 edition, 2001.

[27] Y Magnin, G D Förster, F Rabilloud, F Calvo, A Zappelli, and C Bichara. Thermal expansion of free-standing graphene: benchmarking semi-empirical potentials. *Journal of Physics: Condensed Matter*, 26(18):185401, 2014.

[28] A. K. Geim. Graphene: Status and prospects. *Science*, 324(5934):1530–1534, 2009.

[29] Monica Pozzo, Dario Alfè, Paolo Lacovig, Philip Hofmann, Silvano Lizzit, and Alessandro Baraldi. Thermal expansion of supported and freestanding graphene: Lattice constant versus interatomic distance. *Physical Review Letters*, 106(13):135501, 2011.

[30] K. V. Zakharchenko, M. I. Katsnelson, and A. Fasolino. Finite temperature lattice properties of graphene beyond the quasiharmonic approximation. *Physical Review Letters*, 102(4):046808, 2009.

[31] Carla de Tomas, Alireza Aghajamali, Jake L. Jones, Daniel J. Lim, Maria J. Lopez, Irene Suarez-Martinez, and Nigel A. Marks. Transferability in interatomic potentials for carbon. *Carbon*, 155:624–634, 2019.

[32] Carla de Tomas, Irene Suarez-Martinez, and Nigel A. Marks. Graphitization of amorphous carbons: A comparative study of interatomic potentials. *Carbon*, 109:681–693, 2016.

[33] Robert C. Sinclair, James L. Suter, and Peter V. Coveney. Graphene–Graphene Interactions: Friction, Superlubricity, and Exfoliation. *Advanced Materials*, 30(13):1705791, 2018.

[34] Ming Ma, Gabriele Tocci, Angelos Michaelides, and Gabriel Aeppli. Fast diffusion of water nanodroplets on graphene. *Nature Materials*, 15(1):66–71, 2016.

[35] Alexander Balandin. Thermal properties of graphene and nanostructured carbon materials. *Nature Materials*, 10(8):569–81, 2011.

[36] Vikas Varshney, Soumya S Patnaik, Ajit K Roy, George Froudakis, and Barry L Farmer. Modeling of Thermal Transport in Pillared-Graphene Architectures. *ACS Nano*, 4(2):1153–1161, 2010.

[37] F. Schwierz. Graphene transistors. *Nature Nanotechnology*, 5(7):487–496, 2010.

[38] Carlos P. Herrero and Rafael Ramírez. Vibrational properties and diffusion of hydrogen on graphene. *Physical Review B*, 79(11):115429, 2009.

[39] Carlos P. Herrero and Rafael Ramírez. Quantum effects in graphene monolayers: Path-integral simulations. *Journal of Chemical Physics*, 145(22):224701, 2016.

[40] J. Tersoff. Empirical interatomic potential for carbon, with applications to amorphous carbon. *Physical Review Letters*, 61(25):2879–2882, 1988.

[41] J. Tersoff. Modeling solid-state chemistry: Interatomic potentials for multicomponent systems. *Physical Review B*, 39(8):5566–5568, 1989.

[42] Donald W. Brenner. Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films. *Physical Review B*, 42(15):9458–9471, 1990.

[43] Steven J. Stuart, Alan B. Tutein, and Judith A. Harrison. A reactive potential for hydrocarbons with intermolecular interactions. *The Journal of Chemical Physics*, 112(2000):6472–6486, 2000.

[44] Thomas C. O'Connor, Jan Andzelm, and Mark O. Robbins. AIREBO-M: A reactive model for hydrocarbons at extreme pressures. *Journal of Chemical Physics*, 142(2):024903, 2015.

[45] J H Los and a Fasolino. Intrinsic long-range bond-order potential for carbon: Performance in Monte Carlo simulations of graphitization. *Physical Review B*, 68(2):24107, 2003.

[46] Luca M. Ghiringhelli, Jan H. Los, A. Fasolino, and Evert Jan Meijer. Improved long-range reactive bond-order potential for carbon. II. Molecular simulation of liquid carbon. *Physical Review B*, 72(21):214103, 2005.

[47] Jan H. Los, Luca M. Ghiringhelli, Evert Jan Meijer, and A. Fasolino. Improved long-range reactive bond-order potential for carbon. I. Construction. *Physical Review B*, 73(22):229901, 2006.

[48] N A Marks. Generalizing the environment-dependent interaction potential for carbon. *Physical Review B*, 63(December 2000):1–7, 2006.

[49] Lars Pastewka, Pablo Pou, Rubén Pérez, Peter Gumbsch, and Michael Moseler. Describing bond-breaking processes by reactive potentials: Importance of an environment-dependent interaction range. *Physical Review B - Condensed Matter and Materials Physics*, 78(16):161402, 2008.

[50] A C T van Duin, S Dasgupta, F Lorant, and W A Goddard III. ReaxFF: A Reactive Force Field for Hydrocarbons. *Journal of Physical Chemistry A*, 105(41):9396–9409, 2001.

[51] Sriram Goverapet Srinivasan, Adri 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. *Journal of Physical Chemistry A*, 119(4):571–580, 2015.

[52] D. Porezag, Th Frauenheim, Th Köhler, G. Seifert, and R. Kaschner. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Physical Review B*, 51(19):12947–12957, 1995.

[53] G. Seifert, D. Porezag, and Th. Frauenheim. Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme. *International Journal of Quantum Chemistry*, 58(2):185–192, 1996.

[54] Marcus Elstner and Gotthard Seifert. Density functional tight binding. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2011), 2014.

[55] Bobby G Sumpter, Coral Getino, and Donald W Noid. Theory and Applications of Neural Computing in Chemical Science. *Annual Reviews in Physical Chemistry*, 45:439–481, 1994.

[56] Dimitris K. Agrafiotis, Walter Cedeño, and Victor S. Lobanov. On the use of neural network ensembles in QSAR and QSPR. *Journal of Chemical Information and Computer Sciences*, 42(4):903–911, 2002.

[57] J. U. Thomsen and B. Meyer. Pattern recognition of the 1H NMR spectra of sugar alditols using a neural network. *Journal of Magnetic Resonance (1969)*, 84(1):212–217, 1989.

[58] Mehul M. Khimasia and Peter V. Coveney. Protein structure prediction as a hard optimization problem: The genetic algorithm approach. *Molecular Simulation*, 19(4):205–226, 1997.

[59] Peter V Coveney, Edward R Dougherty, Roger R Highfield, Dougherty Er, and Peter V Coveney. Big data need big theory too. *Philosophical Transactions of the Royal Society A*, 374:20160153, 2016.

[60] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403, 2010.

[61] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 2007.

[62] Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3:e1701816, 2017.

[63] Peter V. Coveney and Roger R. Highfield. From digital hype to analogue reality: Universal simulation beyond the quantum and exascale eras. *Journal of Computational Science*, 46:101093, 2020.

[64] Gordon M Moore. Cramming more components onto integrated circuits With unit cost. *Electronics*, 38(8):114, 1965.

[65] Matthias Rupp, Matthias R. Bauer, Rainer Wilcken, Andreas Lange, Michael Reutlinger, Frank M. Boeckler, and Gisbert Schneider. Machine Learning Estimates of Natural Product Conformational Energies. *PLoS Computational Biology*, 10(1):e1003400, 2014.

[66] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus Robert Müller, and O. Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15:095003, 2013.

[67] Katja Hansen, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O. Anatole Von Lilienfeld, Alexandre Tkatchenko, and Klaus Robert Müller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013.

[68] Matthias Rupp, Raghunathan Ramakrishnan, and O. Anatole von Lilienfeld. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *Journal of Physical Chemistry Letters*, 6(16):3309–3313, 2015.

[69] Alejandro Lopez-Bezanilla and O. Anatole Von Lilienfeld. Modeling Electronic Quantum Transport with Machine Learning. *Physical Review B*, 89(23):235411, 2014.

[70] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole Von Lilienfeld, Klaus Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *Journal of Physical Chemistry Letters*, 6(12):2326–2331, 2015.

[71] John C. Snyder, Matthias Rupp, Katja Hansen, Klaus Robert Müller, and Kieron Burke. Finding density functionals with machine learning. *Physical Review Letters*, 108(25):253002, 2012.

[72] Li Li, John C. Snyder, Isabelle M. Pelaschier, Jessica Huang, Uma Naresh Niranjan, Paul Duncan, Matthias Rupp, Klaus Robert Müller, and Kieron Burke. Understanding machine-learned density functionals. *International Journal of Quantum Chemistry*, 116(11):819–833, 2016.

[73] Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 115(16):1058–1073, 2015.

[74] Kevin Vu, John C. Snyder, Li Li, Matthias Rupp, Brandon F. Chen, Tarek Khelif, Klaus Robert Müller, and Kieron Burke. Understanding kernel ridge regression: Common behaviors from simple functions to density functionals. *International Journal of Quantum Chemistry*, 115(16):1115–1128, 2015.

[75] Jörg Behler. Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry*, 115(16):1032–1050, 2015.

[76] Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *Journal of Chemical Physics*, 145(17):170901, 2016.

[77] Jörg Behler. Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations. *Physical Chemistry Chemical Physics*, 13(40):17930–17955, 2011.

[78] Thuong T Nguyen, Eszter Székely, Giulio Imbalzano, Jörg Behler, Gábor Csányi, Michele Ceriotti, W Götz, Francesco Paesani, and Giulio Imbalzano. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *Journal of Chemical Physics*, 148:241725, 2018.

[79] Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyue Ping Ong. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *Journal of Physical Chemistry A*, 124(4):731–745, 2020.

[80] Tim Mueller, Alberto Hernandez, and Chuhong Wang. Machine learning for interatomic potential models. *Journal of Chemical Physics*, 152(5), 2020.

[81] Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Advanced Materials*, 31(46):1902765, 2019.

[82] F. Ercolesi and J. B. Adams. Interatomic potentials from first-principles calculations: The force-matching method. *Europhysics Letters*, 26(8):583–588, 1994.

[83] Xiang-Yang Liu, James B. Adams, Furio Ercolessi, and John A. Moriarty. Modelling and Simulation in Materials Science and Engineering Related content EAM potential for magnesium from quantum mechanical forces forces. *Modelling and Simulation in Materials Science and Engineering*, 4:293–303, 1996.

[84] Věra Kůrková. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5(3):501–506, 1992.

[85] Chris M. Handley and Paul L.A. Popelier. Potential energy surfaces fitted by Artificial Neural Networks. *Journal of Physical Chemistry A*, 114(10):3371–3383, 2010.

[86] Thomas B. Blank, Steven D. Brown, August W. Calhoun, and Douglas J. Doren. Neural network models of potential energy surfaces: Prototypical examples. *Journal of Chemical Theory and Computation*, 103:4129, 1995.

[87] Sönke Lorenz, Axel Groß, and Matthias Scheffler. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chemical Physics Letters*, 395(4-6):210–215, 2004.

[88] Jörg Behler, Roman Martoňák, Davide Donadio, and Michele Parrinello. Pressure-induced phase transitions in silicon studied by neural network-based metadynamics simulations. *Physica Status Solidi*, 245(12):2618–2629, 2008.

[89] Rustam Z. Khaliullin, Hagai Eshet, Thomas D. Kühne, Jörg Behler, and Michele Parrinello. Graphite-diamond phase coexistence study employing a neural-network mapping of the ab initio potential energy surface. *Physical Review B*, 81(10):100103, 2010.

[90] Rustam Z. Khaliullin, Hagai Eshet, Thomas D. Kühne, Jorg Behler, and Michele Parrinello. Nucleation mechanism for the direct graphite-to-diamond phase transition. *Nature Materials*, 10(9):693–697, 2011.

[91] Vanessa Quaranta, Matti Hellström, Jörg Behler, Jolla Kullgren, Pavlin D. Mitev, and Kersti Hermansson. Maximally resolved anharmonic OH vibrational spectrum of the water/ZnO(10-10) interface from a high-dimensional neural network potential. *Journal of Chemical Physics*, 148(24):241720, 2018.

[92] Gabriele C. Sosso, Giacomo Miceli, Sebastiano Caravati, Federico Giberti, Jörg Behler, and Marco Bernasconi. Fast crystallization of the phase change compound GeTe by large-scale molecular dynamics simulations. *Journal of Physical Chemistry Letters*, 4(24):4241–4246, 2013.

[93] Hagai Eshet, Rustam Z Khaliullin, Thomas D Kühne, Jörg Behler, and Michele Parrinello. Ab initio quality neural-network potential for sodium. *Physical Review B*, pages 1–8, 2010.

[94] Jörg Behler, Roman Martoňák, Davide Donadio, and Michele Parrinello. Metadynamics simulations of the high-pressure phases of silicon employ-

ing a high-dimensional neural network potential. *Physical Review Letters*, 100(18):1–4, 2008.

[95] Jörg Behler. RuNNer, A Program for Constructing High-Dimensional Neural Network.

[96] Andreas Singraber. n2p2 - A neural network package potential.

[97] Alireza Khorshidi and Andrew A. Peterson. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications*, 207:310–324, 2016.

[98] J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science*, 8(4):3192–3203, 2017.

[99] Yunqi Shao, Matti Hellström, Pavlin D. Mitev, Lisanne Knijff, and Chao Zhang. PiNN: A Python Library for Building Atomic Neural Networks of Molecules and Materials. *Journal of Chemical Information and Modeling*, 60(3):1184–1193, 2020.

[100] Farhad Arbabzadah, Stefan Chmiela, Klaus R Mu, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:13890, 2017.

[101] K T Schütt, H E Sauceda, P Kindermans, A Tkatchenko, K Müller, H E Sauceda, P Kindermans, A Tkatchenko, and K M. SchNet – A deep learning architecture for molecules and materials SchNet – A deep learning architecture for molecules and materials. *Journal of Chemical Physics*, 148:241722, 2018.

[102] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and E. Weinan. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Physical Review Letters*, 120(14):143001, 2018.

[103] Oliver T Unke and Markus Meuwly. PhysNet : A Neural Network for Predicting Energies , Forces , Dipole Moments , and Partial Charges. *Journal of Chemical Theory and Computation*, 15:3678–3693, 2019.

[104] Christoph Schran, Krystof Brezina, and Ondrej Marsalek. Committee neural network potentials control generalization errors and enable active learning. *Journal of Chemical Physics*, 104105(August), 2020.

[105] Carl Edward Rasmussen and K. I. Williams, Christopher. *Gaussian Processes for Machine Learning*. MIT Press, Massachusetts, 2nd edition, 2006.

[106] Volker L. Deringer and Gábor Csányi. Machine learning based interatomic potential for amorphous carbon. *Physical Review B*, 95(9):094203, 2017.

[107] Patrick Rowe, Gábor Csányi, Dario Alfè, and Angelos Michaelides. Development of a machine learning potential for graphene. *Physical Review B*, 97:054303, 2018.

[108] Miguel A Caro, Volker L Deringer, Jari Koskinen, Tomi Laurila, and Gábor Csányi. Growth Mechanism and Origin of High sp3 Content in Tetrahedral Amorphous Carbon. *Physical Review Letters*, 120(16):166101, 2018.

[109] Richard Jana, Daniele Savio, Volker L. Deringer, and Lars Pastewka. Structural and elastic properties of amorphous carbon from simulated quenching at low rates. *Modelling and Simulation in Materials Science and Engineering*, 27(8), 2019.

[110] Gabriele C Sosso, Volker L Deringer, Stephen R Elliott, and Gábor Csányi. Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials. *Molecular Simulation*, 44(11):866–880, 2018.

[111] Volker L. Deringer, Miguel A. Caro, Richard Jana, Anja Aarva, Stephen R. Elliott, Tomi Laurila, Gábor Csányi, and Lars Pastewka. Computational

Surface Chemistry of Tetrahedral Amorphous Carbon by Combining Machine Learning and Density Functional Theory. *Chemistry of Materials*, 30(21):7438–7445, 2018.

[112] Carla de Tomas, Alireza Aghajamali, Jake L. Jones, Daniel J. Lim, Maria J. López, Irene Suarez-Martinez, and Nigel A. Marks. Transferability in interatomic potentials for carbon. *Carbon*, 155:624–634, 2019.

[113] Volker L Deringer, Chris J Pickard, and Gábor Csányi. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Physical Review Letters*, 120(15):156001, 2018.

[114] Chris J Pickard and R J Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23:053201, 2011.

[115] Albert P Bartók, James Kermode, and Noam Bernstein. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Physical Review X*, 8(4):41048, 2018.

[116] Gabor Csányi, T. Albaret, M. C. Payne, and A. De Vita. "Learn on the fly": A hybrid classical and quantum-mechanical molecular dynamics simulation. *Physical Review Letters*, 93(17):175503, 2004.

[117] Zhenwei Li, James R. Kermode, and Alessandro De Vita. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Physical Review Letters*, 114(9):096405, 2015.

[118] Evgeny V. Podryabinkin and Alexander V. Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.

[119] Marco Caccin, Zhenwei Li, James R Kermode, and Alessandro De Vita. A Framework for Machine-Learning-Augmented Multiscale Atomistic Simulations on Parallel Supercomputers. *International Journal of Quantum Chemistry*, 115:1129–1139, 2015.

[120] Alexander V Shapeev. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Modelling and Simulation*, 14(3):1153–1173, 2016.

[121] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.

[122] Cas van der Oord, Geneviève Dusson, Gabor Csanyi, and Christoph Ortner. Regularised Atomic Body-Ordered Permutation-Invariant Polynomials for the Construction of Interatomic Potentials. *Machine Learning: Science and Technology*, 1:015004, 2020.

[123] Evgeny V. Podryabinkin, Evgeny V. Tikhonov, Alexander V. Shapeev, and Artem R. Oganov. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Physical Review B*, 99(6):064114, 2019.

[124] I. I. Novoselov, A. V. Yanilkin, A. V. Shapeev, and E. V. Podryabinkin. Moment tensor potentials as a promising tool to study diffusion processes. *Computational Materials Science*, 164(April):46–56, 2019.

[125] I. S. Novikov, Y. V. Suleimanov, and A. V. Shapeev. Automated calculation of thermal rate coefficients using ring polymer molecular dynamics and machine-learning interatomic potentials with active learning. *Physical Chemistry Chemical Physics*, 20(46):29503–29512, 2018.

[126] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016.

[127] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *34th International Conference on Machine Learning, ICML 2017*, 3:2053–2070, 2017.

[128] Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, 2018.

[129] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[130] W. Michael Brown, Aidan P. Thompson, and Peter A. Schultz. Efficient hybrid evolutionary optimization of interatomic potential models. *Journal of Chemical Physics*, 132(2), 2010.

[131] A Slepoy, M. D. Peters, and A. P Thompson. Searching for Globally Optimal Functional Forms for Interatomic Potentials Using Genetic Programming with Parallel Tempering. *Journal of computational chemistry*, 28:2465–2471, 2007.

[132] Alberto Hernandez, Adarsh Balasubramanian, Fenglin Yuan, Simon A.M. Mason, and Tim Mueller. Fast, accurate, and transferable many-body interatomic potentials by symbolic regression. *npj Computational Materials*, 5(1), 2019.

[133] Richard P. Feynman, R. B. Leighton, and Matthew Sands. *The Feynman Lectures on Physics, Volume 1*. Addison-Wesley Publishing Group Ltd, Boston, 1963.

[134] M. P Allen and D. J Tildesley. *Computer Simulation of LIquids*. Oxford University Press, Oxford, 2nd edition, 1991.

[135] J.E. Lennard-Jones. On the Determination o f Molecular Fields. — II. From the Equation of State of a Gas. *Proceedings of the Royal Society of London*, 106(738):463, 1924.

[136] J.E. Lennard-Jones. Cohesion. *Proceedings of the Physical Society*, 43(5):461–476, 1931.

[137] Joseph C.A. Prentice, Jolyon Aarons, James C. Womack, Alice E.A. Allen, Lampros Andrinopoulos, Lucian Anton, Robert A. Bell, Arihant Bhandari, Gabriel A. Bramley, Robert J. Charlton, Rebecca J. Clements, Daniel J. Cole, Gabriel Constantinescu, Fabiano Corsetti, Simon M.M. Dubois, Kevin K.B. Duff, José Mariá Escartín, Andrea Greco, Quintin Hill, Louis P. Lee, Edward Linscott, David D. O'Regan, Maximillian J.S. Phipps, Laura E. Ratcliff, Álvaro Ruiz Serrano, Edward W. Tait, Gilberto Teobaldi, Valerio Vitale, Nelson Yeung, Tim J. Zuehlsdorff, Jacek Dziedzic, Peter D. Haynes, Nicholas D.M. Hine, Arash A. Mostofi, Mike C. Payne, and Chris Kriton Skylaris. The ONETEP linear-scaling density functional theory program. *Journal of Chemical Physics*, 152(17), 2020.

[138] D. J Evans and B. L Holian. The Nose–Hoover thermostat. *Journal of Chemical Physics*, 83(8):4096, 1985.

[139] P Hohenberg and W Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, 1964.

[140] Jiří Klimeš, David R. Bowler, and Angelos Michaelides. Van der Waals density functionals applied to solids. *Physical Review B*, 83(19):195131, 2011.

[141] Jiří Klimeš, David R. Bowler, and Angelos Michaelides. Chemical accuracy for the van der Waals density functional. *Journal of Physics: Condensed Matter*, 22(2):022201, 2010.

[142] Gabriella Graziano, Jiří Klimeš, Felix Fernandez-Alonso, and Angelos Michaelides. Improved description of soft layered materials with van der Waals density functional theory. *Journal of Physics: Condensed Matter*, 24(42):424216, 2012.

[143] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996.

[144] Albert P. Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057, 2015.

[145] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 7th edition, 2005.

[146] Vollker L Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gabor Csanyi. Gaussian Process Regression for Materials and Molecules. *preprint*, 2021.

[147] Michael W. Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(3):697–702, 2009.

[148] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754, 2016.

[149] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *Journal of Chemical Physics*, 134(7):074106, 2011.

[150] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[151] O. Anatole Von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, 115(16):1084–1093, 2015.

[152] Vollker L Deringer, Miguel A Caro, and Gábor Csányi. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nature Communications*, 11:5461, 2020.

[153] Aldo Glielmo, Claudio Zeni, and Alessandro De Vita. Efficient nonparametric n -body force fields from machine learning. *Physical Review B*, 97(18):1–12, 2018.

[154] S. Plimpton. Fast Parallel Algorithms for Short – Range Molecular Dynamics. *Journal of Computational Physics*, 117(June 1994):1–19, 1995.

[155] Nicola Bonini, Michele Lazzeri, Nicola Marzari, and Francesco Mauri. Phonon anharmonicities in graphite and graphene. *Physical Review Letters*, 99(17):176802, 2007.

[156] Duhee Yoon, Young Woo Son, and Hyeonsik Cheong. Negative thermal expansion coefficient of graphene measured by raman spectroscopy. *Nano Letters*, 11(8):3227–3231, 2011.

[157] Vibhor Singh, Shamashis Sengupta, Hari S Solanki, Rohan Dhall, Adrien Allain, Sajal Dhara, Prita Pant, and Mandar M Deshmukh. Probing thermal expansion of graphene and modal dispersion at low-temperature using graphene nanoelectromechanical systems resonators. *Nanotechnology*, 21(16):165204, 2010.

[158] G. Kresse and J. Hafner. Ab initio molecular dynamics for liquid metals. *Physical Review B*, 47(1):558–561, 1993.

[159] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996.

[160] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist. Van der Waals density functional for general geometries. *Physical Review Letters*, 92(24):246401, 2004.

[161] G. Kresse. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59(3):1758–1775, 1999.

[162] Guillermo Román-Pérez and José M. Soler. Efficient implementation of a van der waals density functional: Application to double-wall carbon nanotubes. *Physical Review Letters*, 103(9):096102, 2009.

[163] H Monkhorst and J Pack. Special points for Brillouin zone integrations. *Physical Review B*, 13(12):5188–5192, 1976.

[164] Martin Fitzner, Gabriele C. Sosso, Stephen J. Cox, and Angelos Michaelides. The Many Faces of Heterogeneous Ice Nucleation: Interplay between Surface Morphology and Hydrophobicity. *Journal of the American Chemical Society*, 137(42):13658–13669, 2015.

[165] S. Linas, Y. Magnin, B. Poinsot, O. Boisron, G. D. Förster, V. Martinez, R. Fulcrand, F. Tournus, V. Dupuis, F. Rabilloud, L. Bardotti, Z. Han, D. Kalita, V. Bouchiat, and F. Calvo. Interplay between Raman shift and thermal expansion in graphene: Temperature-dependent measurements and analysis of substrate corrections. *Physical Review B*, 91(7):075426, 2015.

[166] Nicolas Mounet and Nicola Marzari. First-principles determination of the structural, vibrational and thermodynamic properties of diamond, graphite, and derivatives. *Physical Review B*, 71(20):205214, 2005.

[167] Jin Wu Jiang, Jian Sheng Wang, and Baowen Li. Thermal expansion in carbon nanotubes and graphene: nonequilibrium Green's function approach. *Physical Review B*, 80(20):205429–205436, 2009.

[168] E. R. Hernández, A. Rodriguez-Prieto, A. Bergara, and Dario Alfè. First-principles simulations of lithium melting: Stability of the bcc phase close to melting. *Physical Review Letters*, 104(18):185701, 2010.

[169] Dario Alfè. PHON: A program to calculate phonons using the small displacement method. *Computer Physics Communications*, 180(12):2622–2633, 2009.

[170] Ling Ti Kong. Phonon dispersion measured directly from molecular dynamics simulations. *Computer Physics Communications*, 182(10):2201–2207, 2011.

[171] Carlos Campañá and Martin H. Müser. Practical Green's function approach to the simulation of elastic semi-infinite solids. *Physical Review B*, 74(7):075420, 2006.

[172] L. Lindsay and D. A. Broido. Optimized Tersoff and Brenner empirical potential parameters for lattice dynamics and phonon thermal transport in carbon nanotubes and graphene. *Physical Review B*, 81(20):205441, 2010.

[173] Zhen Yao and L Kane. High-Field Electrical Transport in Single-Wall Carbon Nanotubes. *Physical Review Letters*, 84(13):2941–2944, 2000.

[174] T. M G Mohiuddin, A. Lombardo, R. R. Nair, A. Bonetti, G. Savini, R. Jalil, N. Bonini, D. M. Basko, C. Galiotis, N. Marzari, K. S. Novoselov, A. K. Geim, and A. C. Ferrari. Uniaxial strain in graphene by Raman spectroscopy: G peak splitting, Grüneisen parameters, and sample orientation. *Physical Review B*, 79(20):205433, 2009.

[175] Zhen Hua Ni, Ting Yu, Yun Hao Lu, Ying Ying Wang, Yuan Ping Feng, and Ze Xiang Shen. Uniaxial Strain on Graphene : Raman. *ACS Nano*, 2(11):2301–2305, 2008.

[176] Ming Ma, Gabriele Tocci, Angelos Michaelides, and Gabriel Aeppli. Fast diffusion of water nanodroplets on graphene. *Nature Materials*, 15(1):66–71, 2016.

[177] Miguel A. Caro, Anja Aarva, Volker L. Deringer, Gábor Csányi, and Tomi Laurila. Reactivity of Amorphous Carbon Surfaces: Rationalizing the Role of Structural Motifs in Functionalization Using Machine Learning. *Chemistry of Materials*, 30(21):7446–7455, 2018.

[178] Volker L. Deringer, Céline Merlet, Yuchen Hu, Tae Hoon Lee, John A. Kattirtzi, Oliver Pecher, Gábor Csányi, Stephen R. Elliott, and Clare P. Grey. Towards an atomistic understanding of disordered carbon electrode materials. *Chemical Communications*, 54(47):5988–5991, 2018.

[179] Jian Xing Huang, Gábor Csányi, Jin Bao Zhao, Jun Cheng, and Volker L. Deringer. First-principles study of alkali-metal intercalation in disordered carbon anode materials. *Journal of Materials Chemistry A*, 7(32):19070–19080, 2019.

[180] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, Cambridge, 1 edition, 2004.

[181] Piero Gasparotto, Robert Horst Meißner, and Michele Ceriotti. Recognizing Local and Global Structural Motifs at the Atomic Scale. *Journal of Chemical Theory and Computation*, 14:486–498, 2018.

[182] David J. Wales. Closed-shell structures and the building game. *Chemical Physics Letters*, 141(6):478–484, 1987.

[183] Kyuho Lee, Éamonn D. Murray, Lingzhu Kong, Bengt I. Lundqvist, and David C. Langreth. Higher-accuracy van der Waals density functional. *Physical Review B - Condensed Matter and Materials Physics*, 82(8):081101, 2010.

[184] Roald Hoffmann, Artyom A. Kabanov, Andrey A. Golov, and Davide M. Proserpio. Homo Citans and Carbon Allotropes: For an Ethics of Citation . *Angewandte Chemie International Edition*, 55(37):10962–10976, 2016.

[185] L. Li, S. Reich, and J. Robertson. Defect energies of graphite: Density-functional calculations. *Physical Review B - Condensed Matter and Materials Physics*, 72(18):184109, 2005.

[186] Tao Xu and Litao Sun. Structural defects in graphene. *Defects in Advanced Electronic Materials and Novel Low Dimensional Structures*, 5(1):137–160, 2018.

[187] J. C. Charlier. Defects in carbon nanotubes. *Accounts of Chemical Research*, 35(12):1063–1069, 2002.

[188] Stephen T Skowron, Irina V Lebedeva, Andrey M Popov, and Elena Bichoutskaia. Energetics of atomic scale structure changes in graphene. *Chemical Society Reviews*, 44(44):3143–3176, 2015.

[189] J. Ristein. Diamond surfaces : familiar and amazing. *Applied Physics A*, 384:377–384, 2006.

[190] G Kern and J Hanfer. Ab initio molecular-dynamics studies of the graphitization of flat and stepped diamond (111) surfaces. *Physical Review B*, 19:13167, 1998.

[191] Newton Ooi, Asit Rairkar, and James B. Adams. Density functional study of graphite bulk and surface properties. *Carbon*, 44(2):231–242, 2006.

[192] G Kern and J Hafner. Ab initio calculations of the atomic and electronic structure of diamond (111) surfaces with steps. *Physical Review B*, 58(4):2161–2169, 1998.

[193] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, 2011.

[194] Albert P. Bartók, Michael J. Gillan, Frederick R. Manby, and Gábor Csányi. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Physical Review B*, 88(5):054104, 2013.

[195] Daniele Dragoni, Thomas D. Daff, Gabor Csanyi, and Nicola Marzari. Achieving DFT accuracy with a machine-learning interatomic potential:

thermomechanics and defects in bcc ferromagnetic iron. *Physical Review Materials*, 2:013808, 2017.

[196] So Fujikake, Volker L Deringer, Tae Hoon Lee, Marcin Krynski, Stephen R Elliott, Gábor Csányi, So Fujikake, Volker L Deringer, Hoon Lee, and Marcin Krynski. Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures. *Journal of Chemical Physics*, 148:241714, 2018.

[197] Wojciech J. Szlachta, Albert P. Bartók, and Gábor Csányi. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Physical Review B*, 90(10):104018, 2014.

[198] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Physical Chemistry Chemical Physics*, 20(47):29661–29668, 2018.

[199] Felix A. Faber, Anders S. Christensen, Bing Huang, and O. Anatole Von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *Journal of Chemical Physics*, 148(24), 2018.

[200] Anders S. Christensen, Lars A. Bratholm, Felix A. Faber, and O. Anatole Von Lilienfeld. FCHL revisited: Faster and more accurate quantum machine learning. *Journal of Chemical Physics*, 152(4):044107, 2020.

[201] Atsushi Togo and Isao Tanaka. First principles phonon calculations in materials science. *Scripta Materialia*, 108:1–5, 2015.

[202] G Kern and J Hafner. Ab initio calculations of the atomic and electronic structure of clean and hydrogenated diamond (110) surfaces. *Physical Review B*, 56(7):4203–4210, 1997.

[203] S Thinius, M. M. Islam, and T Bredow. Reconstruction of low-index graphite surfaces. *Surface Science*, 649(1):60–65, 2016.

[204] M. I. J. Probert and M. C. Payne. Improving the convergence of defect calculations in supercells: An ab initio study of the neutral silicon vacancy. *Physical Review B*, 67(7):075204, 2003.

[205] D. Hunt, D. Twitchen, M. Newton, J. Baker, and T. Anthony. Identification of the neutral carbon (100)-split interstitial in diamond. *Physical Review B - Condensed Matter and Materials Physics*, 61(6):3863–3876, 2000.

[206] A. J. Stone and D. J. Wales. Theoretical Studies of Icosahedral C60 and Some Related Species. *Chemical Physics Letters*, 128(5):501–503, 1986.

[207] J. 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. *Physical Review B - Condensed Matter and Materials Physics*, 83(24):245420, 2011.

[208] Jie Ma, Dario Alfè, Angelos Michaelides, and Enge Wang. Stone-Wales defects in graphene and other planar sp2 -bonded materials. *Physical Review B*, 80(4):033407, 2009.

[209] Alfredo A. Correa, Stanimir A. Bonev, and Giulia Galli. Carbon under extreme conditions: Phase boundaries and electronic properties from first-principles theory. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1204–1208, 2006.

[210] Monica Pozzo, Chris Davies, David Gubbins, and Dario Alfè. Transport properties for liquid silicon-oxygen-iron mixtures at Earth's core conditions. *Physical Review B - Condensed Matter and Materials Physics*, 87(1):014110, 2013.

[211] H M Strong and R E Hanneman. Crystallization of Diamond and Graphite. *Journal of Chemical Physics*, 46(9):3668–3676, 1967.

[212] A Sorkin, Joan Adler, and R Kalish. Nucleation of diamond from liquid carbon under extreme pressures : Atomistic simulation. *Physical Review B*, 74(6):064115, 2006.

[213] W. H. Gust. Phase transition and shock-compression parameters to 120 GPa for 3 types of graphite. *Physical Review B*, 22(10):4744–4756, 1980.

[214] Christopher J. Mundy, Alessandro Curioni, Nir Goldman, I. F.Will Kuo, Evan J. Reed, Laurence E. Fried, and Marcella Ianuzzi. Ultrafast transformation of graphite to diamond: An ab initio study of graphite under shock compression. *Journal of Chemical Physics*, 128(18), 2008.

[215] Volker L. Deringer, Davide M. Proserpio, Gábor Csányi, and Chris J. Pickard. Data-driven learning and prediction of inorganic crystal structures. *Faraday Discussions*, 211:45–59, 2018.

[216] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97(17):170201, 2006.

[217] Chris J Pickard. Predicted Pressure-Induced s -Band Ferromagnetism in Alkali Metals. *Physical Review Letters*, 107(8):087201, 2011.

[218] Richard J. Needs and Chris J. Pickard. Perspective: Role of structure prediction in materials discovery and design. *APL Materials*, 4(5), 2016.

[219] Yuko Kumeda and David J. Wales. Ab initio study of rearrangements between C60 fullerenes. *Chemical Physics Letters*, 374(1-2):125–131, 2003.

[220] David J Wales. Global Optimization of Clusters, Crystals, and Biomolecules. *Science*, 285(5432):1368–1372, 1999.

[221] P. J.F. Harris and S. C. Tsang. High-resolution electron microscopy studies of non- graphitizing carbons. *Philosophical Magazine A: Physics of Condensed Matter, Structure, Defects and Mechanical Properties*, 76(3):667–677, 1997.

[222] Peter J.F. Harris. New perspectives on the structure of graphitic carbons. *Critical Reviews in Solid State and Materials Sciences*, 30(4):235–253, 2005.

[223] P. J. F. Harris. Structure of non-graphitising carbons. *International Materials Reviews*, 42(5):206–218, 2014.

[224] Xinwei Dou, Ivana Hasa, Damien Saurel, Christoph Vaalma, Liming Wu, Daniel Buchholz, Dominic Bresser, Shinichi Komaba, and Stefano Passerini. Hard carbons for sodium-ion batteries: Structure, analysis, sustainability, and electrochemistry. *Materials Today*, 23(March):87–104, 2019.

[225] Fabian L. Thiemann, Patrick Rowe, Erich A. Müller, and Angelos Michaelides. Machine Learning Potential for Hexagonal Boron Nitride Applied to Thermally and Mechanically Induced Rippling. *The Journal of Physical Chemistry C*, 124(40), 2020.

[226] Andrea Zen, Jan Gerit Brandenburg, Angelos Michaelides, and Dario Alfè. A new scheme for fixed node diffusion quantum Monte Carlo with pseudopotentials: Improving reproducibility and reducing the trial-wave-function bias. *Journal of Chemical Physics*, 151(13), 2019.

[227] Andrea Zen, Jan Gerit Brandenburg, Jiří Klimeš, Alexandre Tkatchenko, Dario Alfè, and Angelos Michaelides. Fast and accurate quantum Monte Carlo for molecular crystals. *Proceedings of the National Academy of Sciences of the United States of America*, 115(8):1724–1729, 2018.

[228] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *Journal of Chemical Physics*, 132(15):154104, 2010.

[229] Stefan Grimme. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *Journal of Computational Chemistry*, 27(15):1787–1799, 2009.