IOP Publishing Modelling Simul. Mater. Sci. Eng. 34 (2026) 025004 https://doi.org/10.1088/1361-651X/ae3e04

Modelling and Simulation in Materials Science and Engineering

PAPER

Transferability of carbon potentials for novel carbon polymorphs

Zuzanna Malinowska-Trzmielak[1,*], **Nicole Grobert**[2] and **Mark Wilson**[1]

[1] Physical and Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford, Oxford OX1 3QZ, United Kingdom
[2] Department of Materials, University of Oxford, Oxford OX5 1PF, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: zuzanna.trzmielak@chem.ox.ac.uk, nicole.grobert@materials.ox.ac.uk and mark.wilson@chem.ox.ac.uk

Keywords: diaphites, diamond-graphite nanocomposites, carbon polymorphs, molecular dynamics, potentials, AIMD

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
27 October 2025

REVISED
16 December 2025

ACCEPTED FOR PUBLICATION
27 January 2026

PUBLISHED
5 February 2026

Abstract

Choosing a suitable potential model to study dynamic processes in novel structures is an ambitious task often relying on chemical intuition. This paper addresses this challenge through a case study of diaphites, diamond-graphite nanocomposites, that are the only naturally occurring crystalline form of carbon featuring both sp^3 and sp^2 hybridized atoms. Since their synthesis is expensive and difficult to control, molecular dynamics (MD) simulations of their formation would be highly valuable. However, none of the available carbon potentials explicitly includes diaphites in their parameterization. Here, we benchmark several well-established carbon potentials (Tersoff 1989, Tersoff 1994, REBO-II, LCBOP-I, AIREBO, AIREBO-M, GAP-20, ACE) against *ab initio* MD (AIMD) at the PBE+D2 level of theory. Comparison of structural labeling disqualified Tersoff 1989, Tersoff 1994, REBO-II, AIREBO, and AIREBO-M. To enable long-timescale simulations on systems of a few thousand atoms, an machine-learning (ML)-AIMD model was developed using AIMD acceleration with an on-the-fly Gaussian approximation potential (GAP). ML-AIMD accurately reproduced AIMD results and was therefore used as a benchmark. Extended testing revealed that ACE is the most transferable and computationally efficient potential for MD simulations of diaphites, reproducing the sp^2 fraction across all temperatures at a cost at least four times lower than GAP-20. LCBOP-I performed comparably below 2000 K and remains preferable when computational resources are limited. The presented benchmarking framework efficiently identifies the most suitable potentials and provides a general strategy for selecting MD models for novel materials.

1. Introduction

Elemental carbon is renowned for its exceptional structural and topological diversity, forming a wide range of allotropes such as nanotubes, graphene sheets, fullerenes, graphite, cubic and hexagonal diamond, as well as amorphous carbon. In bulk crystalline polymorphs, the diamond and graphitic forms each exhibit a single type of hybridization—sp^3 and sp^2, respectively, corresponding to coordination numbers $n_C^C = 4$ and 3 [1]. In contrast, amorphous carbon has both sp^2 and sp^3 sites and shows no ordering in those sites at either short- or long-range, leading to combined topological and site disorder.

Diaphites are crystalline structures that show a mixture of diamond- and graphite-like local coordination environments and have recently been observed in meteorite impact samples [2]. In some sense, these represent an 'intermediate' phase between pure diamond or graphite crystals and the amorphous states. For example, diaphites show both graphite- and diamond-like local coordination environments, but these appear to be arranged in *domains* rather than randomly distributed across the structure. As a result, they have a highly ordered (crystalline) structure at short range length scales but may have disordered structure over long range as the domains may be arranged randomly (figure 1).

Special attention was drawn to the graphite-diamond grain boundary, called gradia, which is an inherent consequence of these two polymorphs being present in a single solid material. Figure 2 shows

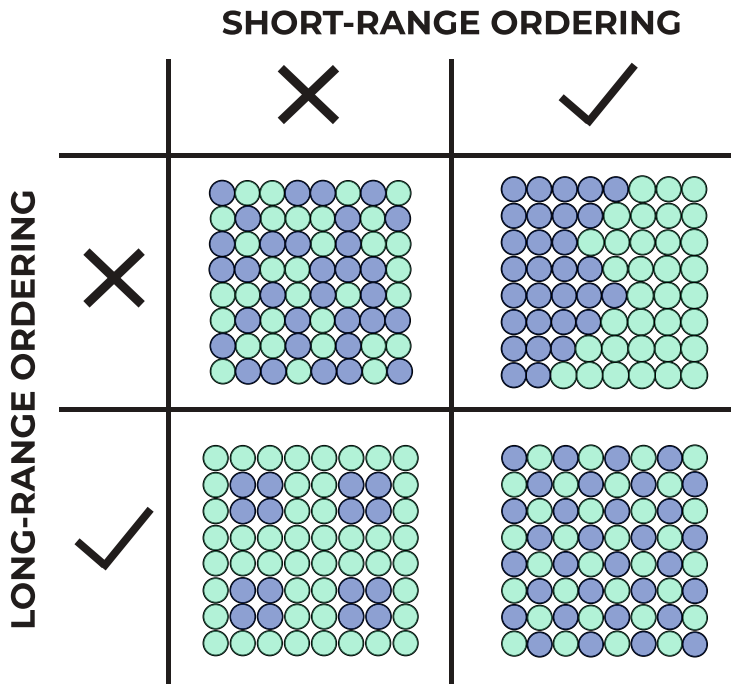


Figure 1. Examples of structures with ordering in short-range, long-range, both or neither. Diaphites are materials with short-range ordering (due to crystalline domains) but without long range ordering (due to an imperfect distribution of domains), corresponding to the top right corner in the matrix.

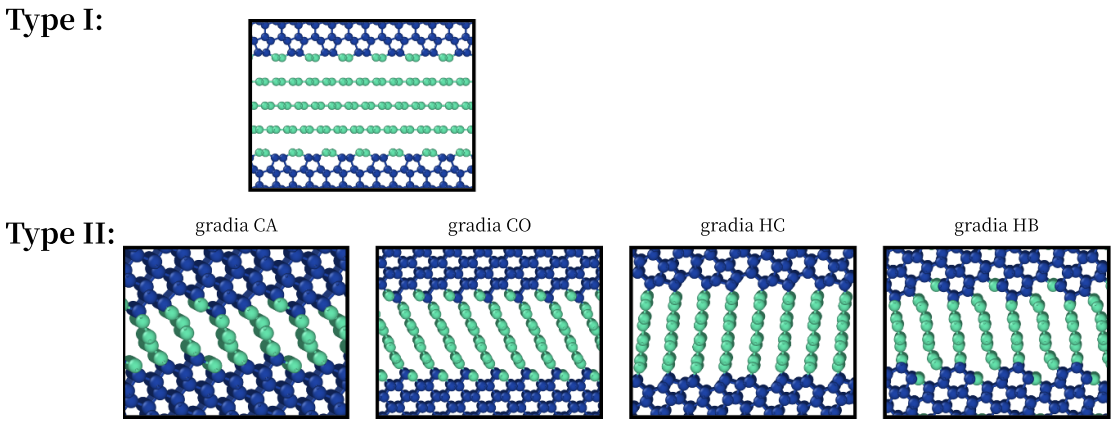


Figure 2. Example representations of different types of diaphites and graphite-diamond interface (known as 'gradia'). Graphite-like, 3-coordinate atoms are marked in green; diamond-like, 4-coordinate atoms are marked in blue. The first letter of the gradia name label refers to the type of diamond at the interface (*C*—cubic; *H*—hexagonal); the second letter refers to either the angle at which graphitic layer is bonded (*A*—acute; *O*—obtuse) or conformation of hexagonal building blocks of diamond (*C*—chair; *B*—boat) [3].

examples of diaphite structures with different types of grain boundary. The graphitic domain can be sandwiched between diamond domains, giving rise to type I diaphite, or connected to it, giving rise to type II diaphite. Within type II diaphite, four types of gradia were identified and labeled with two-letter acronyms (see figure 2). The first character specifies the diamond form at the interface: 'C' for 'cubic' and 'H' for 'hexagonal'. The second letter refers to the angle/confirmation at which the graphitic layer is attached: 'A' for 'acute' and 'O' for 'obtuse' in case of cubic diamond, while 'C' for 'chair' and 'B' for 'boat' in case of hexagonal diamond [3].

This intermediate structural behavior leads to a range of intriguing predicted properties that arise from the nanocomposite nature of diaphites. Such materials could combine the advantages of both diamond and graphite while mitigating their individual limitations. For example, the resulting composite could exhibit diamond-like hardness while retaining electrical conductivity through embedded graphitic domains, as demonstrated by Zou *et al* [4] Theoretical studies have also suggested the possibility of

high-temperature superconductivity driven by strong electron-phonon coupling at the diamond-graphite interfaces [5]. Moreover, diamond-graphite nanocomposites are expected to show enhanced elongation under tensile stress, as coherent interfaces enable sequential structural transformations rather than catastrophic fracture [6]. The presence of graphitic domains within the diamond matrix may further contribute to crack energy dissipation by promoting local graphitization instead of brittle failure. Finally, due to the contrasting electronic properties of the two polymorphs, controlling the diamond-to-graphite ratio offers a route to carbon-based semiconductors with tunable band gaps [7, 8].

The implication is that precise (atomistic) control may allow these key properties to be manipulated as required. Detailed empirical experimental observations are problematic due to the extreme and therefore costly conditions under which diaphites are formed (typically $T > 1000$ K, $p \sim 10$ GPa) [8]. As a result, computational investigations, in which the system energy can be deconstructed into fundamental interactions, would appear to be a promising avenue.

High-level methods (*ab initio* or DFT) can accurately identify local minima, estimate the relative energy of given structures, and hence highlight their relative stability with high precision. However, since they are based on iteratively solving the Kohn–Sham equation, it puts strict restrictions on the simulation length- and time-scales. Performing molecular dynamics (MD) simulations in which the atoms are systematically moved along their lines of force to explore the phase space requires evaluation of the system energy (and related forces) at each time step. Furthermore, while linear-scaling DFT approaches exist [9], DFT methods most often display much less favorable computational cost scaling with increasing number of atoms (N), often N^3 [10]. In the case of diaphites, the presence of well-defined domains implies that relatively large simulation cells will be required (i.e. N of the order of magnitude of thousands of atoms), making *ab inito* MD (AIMD) an unviable choice.

An alternative is to use potential models. In contrast to DFT and AIMD, instead of solving the Kohn–Sham equation, they use Newtonian laws of motion and account for quantum behaviors by using optimized parameters. The key interactions are therefore expressed as relatively simple functions of the atomic coordinates. The effective 'bypassing' of the underlying quantum mechanics makes these methods potentially orders of magnitude more rapid. As a result, due to their computational efficiency, potential model methods can facilitate the study of dynamic processes such as heating, quenching, pressurization, or the simple evolution of structure at the given temperature.

Although potential models are the solution to the problem of AIMD scalability, the cost of that may be a loss in transferability. AIMD or DFT may be applied to any given configuration (though the quality of the output could vary depending on the functional and the system). However, potential models are generally developed by parameterizing with respect to a subset of known structures. Only recently, with the rapid development of artificial intelligence, have potentials become available which include an extremely broad range of structures, and hence atom environments [11, 12]. However, most potentials have been developed with specific structures in mind. Carbon potentials can be divided into two categories: empirical and machine-learning (ML) potentials. Empirical potentials usually come down to relatively simple functions of the atom coordinates, in which each parameter has a clear physical meaning. However, they are limited in terms of the quantity of data used for parametrization and the process is, in general, not automated. With ML potentials, the situation is reversed in some sense: ML model can be optimized for a massive amount of data points, but the process of training a model is outsourced to the computer and, after setting up hyperparameters, happens without outside interference. Despite the development of at least 40 empirical carbon potentials and the rapid expansion of ML potentials, to date, none of the carbon potentials explicitly included diaphites, nor many mixed sp^3 and sp^2 phases except amorphous carbon. Therefore, one can only speculate on their accuracy when simulating these structures.

With the growing role of computational methods in materials discovery and design, a standardized framework for comparing interatomic potentials is essential to ensure reliability and reproducibility. Benchmarks currently reported in the literature are often inconsistent-or, in the case of MD simulations, overlooked entirely. To address this gap, the present work introduces a systematic approach for selecting suitable MD potentials for novel structures. As a demonstration, we benchmarked several widely used carbon potentials against high-level DFT calculations for diaphites.

Benchmarking, in this context, refers to the comparison of multiple computational methods against a reference of established accuracy, allowing researchers to identify the most appropriate approach for their system of interest. Here, we assess how well general-purpose carbon potentials reproduce DFT-level behavior in diamond-graphite nanocomposites by analyzing force distributions, total energies, sp^2/sp^3 ratios, and structural topology throughout and at the end of the simulations.

This paper aims to serve as a guide for anyone interested in performing MD on diaphites but most importantly as a general example of how much MD results could depend on the choice of potential and how to do it effectively.

The paper is organized as follows: section 2 outlines the overall computational workflow; section 3 presents the main findings from each stage of the study; section 4 discusses the rationale and assumptions underlying each step, details the development of the ML-AIMD model, and places the results in the broader context of emerging ML-based interatomic potentials and their transferability; finally, section 5 summarizes the key results and highlights the importance of assessing potential transferability when investigating novel structures.

2. Methods

The overall benchmarking workflow is summarized in figure 3 and detailed below. Example input and coordinate files for all structures are provided in the supplementary information (SI).

2.1. Parent structure generation

Representative diaphite unit cells were constructed using Wyckoff coordinates reported by Li *et al* [13], implemented via ASE [14]. The diamond surfaces in type I diaphites were modeled using the Pandey (2×1) reconstruction [1, 15]. By systematically varying the number of graphitic and diamond layers across types and gradia, a total of 23 parent structures were generated (see table S1). To balance diversity and computational feasibility, structures contained 20–32 atoms, with at least one distinct graphitic and one diamond layer (see figures S1–S5).

2.2. Preliminary screening

Each parent structure was subjected to small random displacements, producing 230 configurations in total. These were used to benchmark empirical and ML interatomic potentials against reference DFT calculations. DFT-based, single-point calculations were performed using CASTEP [16] with the PBE functional [17] and D2 dispersion correction [18]. Computational parameters (*k*-point spacing, plane-wave cutoff, and energy tolerance) were optimized for accuracy and efficiency (details in SI). Equivalent 0 K single-point evaluations were carried out using several empirical and ML potentials in LAMMPS: [19] Tersoff (1989, 1994), REBO [20], LCBOP-I [21], AIREBO [22], AIREBO-M [23], GAP-20, [24] and ACE [25]. Comparison of energies and forces with DFT-based results guided potential selection for subsequent simulations.

2.3. ML-AIMD model development

To enable larger-scale simulations with near-DFT accuracy, we employed accelerated AIMD with an on-the-fly Gaussian approximation potential (GAP), following Stenczel *et al* [26] The adaptive model parameters and implementation details are given in the SI. Each parent structure was simulated at $T = 1500$, 2000, 3000, and 4000 K for 1 ps (1 fs timestep) in the NVT ensemble using both AIMD and the on-the-fly GAP-AIMD approach. Energies and sp^2 content evolution were compared to verify model fidelity. All AIMD snapshots were pooled into a database (≈15 000 configurations) to train the final ML-AIMD potential, validated via mean absolute errors (MAE) on energies and forces.

2.4. Extended testing

The validated potentials-Tersoff 1989, LCBOP-I, GAP-20, ACE, and ML-AIMD-were tested on the 23 diaphitic structures across four temperatures (1500–4000 K) and two system sizes: moderate (≈1000 atoms) and large (≈8000 atoms), generated by replicating the cells from the Preliminary Screening. Simulations (52 ps, NVT ensemble) were performed in LAMMPS with periodic boundaries. Results were benchmarked against ML-AIMD in terms of total potential energies, temporal evolution of sp^2 fraction, and final atomic topology. Additional analysis and potential performance discussion are presented in section 4.

3. Results

The preliminary stage of a benchmarking (section 3.1) analyzes discrepancies between DFT and potential-derived energies and forces in order to eliminate some of the potentials without the need for expensive and extensive MD simulations. The next stage (section 3.2) aims to develop a potential of

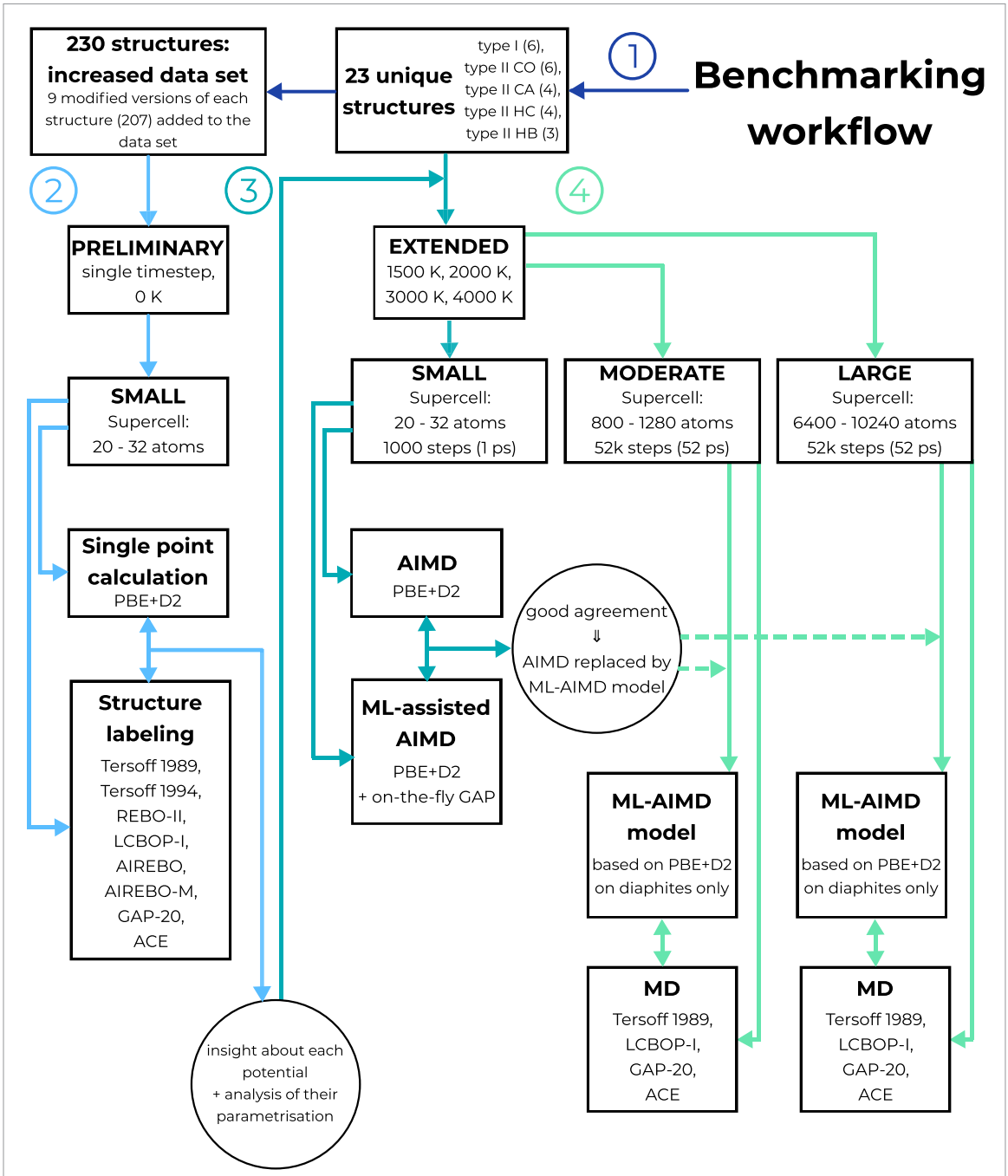


Figure 3. Carbon potential benchmarking workflow. Numbers in circles and like-colored arrows signify consecutive stages of benchmarking: (1) Structure Building, (2) Preliminary Screening based on single snapshot calculations, (3) ML-AIMD model development to replace classical AIMD, (4) Extended Testing in which long-scale simulation with moderate and large supercells developed with MD potentials were compared to those developed with ML-AIMD model. Double sided arrow indicates comparison between resulting datasets. Bigger circles with black circumference mean selection or thought process that influenced the next stage.

DFT/AMID accuracy for simulations on diaphites under the conditions of interest (NVT ensemble, T = [1500 K, 2000 K, 3000 K, 4000 K]). The suitability of the model architecture was determined by comparing the energies derived from accelerated AIMD with the on-the-fly GAP model with those obtained from AIMD, along with changes in the percentage of sp^2 environments. The model, referred to as the ML-AIMD model, was then trained and validated using AIMD structures by looking at the MAE of the forces and energies.

The final stage (section 3.3) involved analysis of the performance of the potential in relatively long ($t \sim$52 ps) simulations at four temperatures. The potential energies of the final structures and the changes in the percentage of sp^2 local environments were compared with the results derived from ML-AIMD. In some cases, the final snapshot topology was also assessed. An additional discussion of the decisions made throughout can be found in section 4.

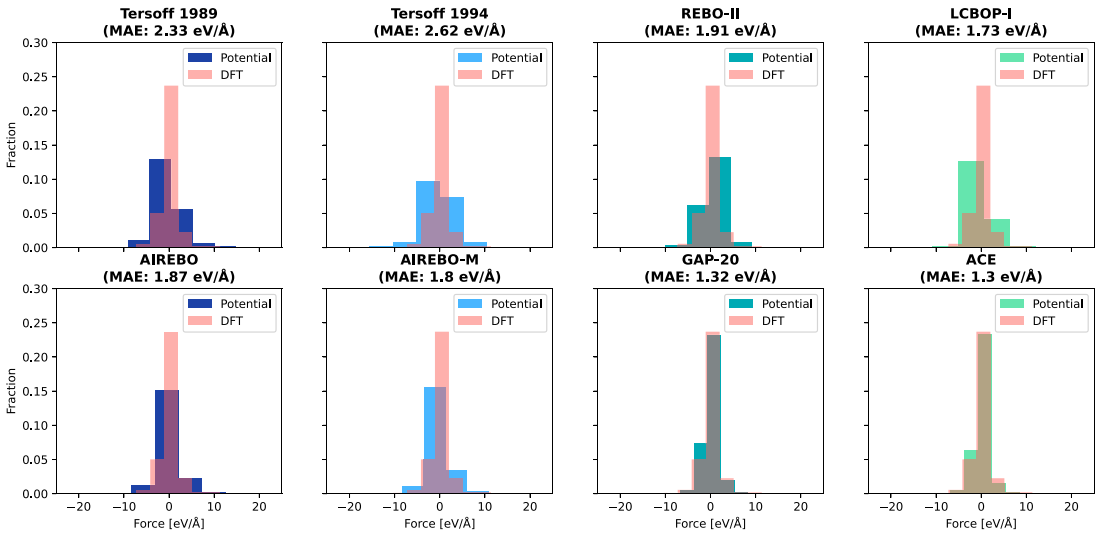


Figure 4. Distribution of forces acting on all atoms across the whole dataset (230 structures), using different carbon potentials, along with stated MAE value. The distribution of DFT forces is superimposed (viridian green) on each histogram for comparison.

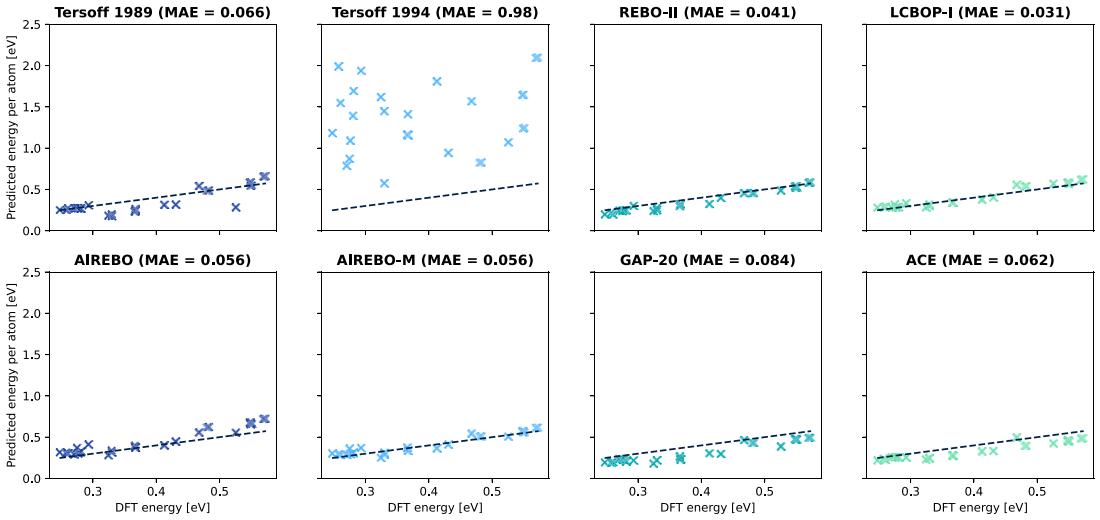


Figure 5. The difference in energy per atom between diamond and diaphite structures predicted using various carbon potentials shown against the DFT derived energy. The dashed line shows the result of perfect agreement with the DFT derived results. All are shown on the same scale to aid comparison.

3.1. Preliminary screening analysis

The preliminary screening analysis focused on the distribution of the forces acting on the atoms (figure 4) and the predicted energy per atom (figure 5) comparing the single-point results obtained from the potentials with those obtained from DFT.

Figure 4 shows the distributions of forces for all atoms (showing all three Cartesian components). The superimposed distribution of forces derived from DFT, along with the resultant MAEs are shown. A plot of the potential-derived forces against the corresponding DFT-derived forces can be found in the SI (figure S6). The ACE and GAP-20 potentials show the most similar forces distribution compared to those derived from the DFT, with the lowest MAEs among the tested potentials. The Tersoff 1989 and Tersoff 1994 potentials show a relatively poor overlap with the DFT forces and correspondingly have the highest MAEs. The MAEs of the force distributions obtained by the REBO-II, LCBOP-I, AIREBO, and AIREBO-M potentials are of a similar order of magnitude. Interestingly, despite AIREBO showing better overlap with the reference distribution, its MAE is higher than for LCBOP-I, which suggests a mismatch between the AIREBO and DFT-derived corresponding forces, i.e. similar distribution is a result of fortunate coincidence rather than accuracy of the potential.

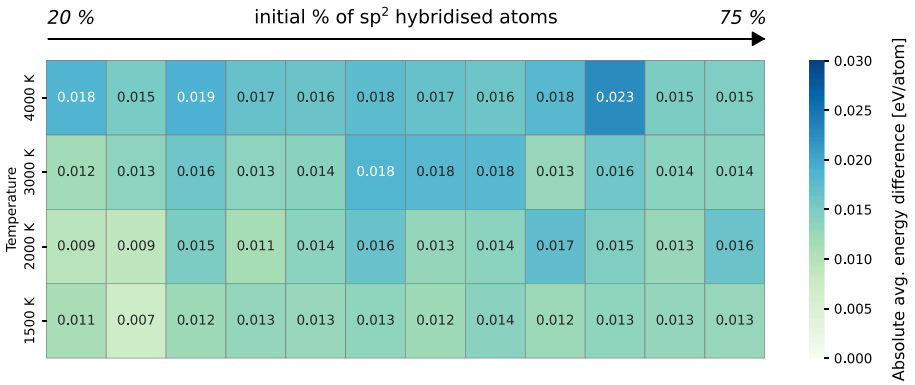


Figure 6. Absolute average energy difference in eV/atom (see color scale on the right) between the on-the-fly GAP model and AIMD derived results with respect to the temperature of the simulation (abscissa on the left) and the percentage of the initial sp^2 content (ordinate at the top). Values corresponds to averaged differences from all performed checks throughout accelerated AIMD simulations.

Figure 5 shows the energies per atom obtained from the potential models against DFT-derived energies for the same structures. As the source of parametrization/training data differs between potentials, the energy is stated with respect to the energy of the diamond crystal. This ensures fair comparison, taking into account relative energies, rather than absolute energies. Absolute energies and further explanation can be found in the SI.

All potentials, except Tersoff 1994, show a linear trend, good overlap with DFT-based reference data, and small MAE. Differences of $\Delta\text{MAE} = 0.05$, e.g. between GAP-20 and LCBOP-I, can generally be regarded as negligible and did not influence the further exclusion of some potentials as explained in section 4.

Based on the above results, three potentials were chosen for the 'Extended' stage of benchmarking: LCBOP-I, GAP-20, and ACE, along with Tersoff 1989 as a so-called sanity check. The detailed analysis of potentials' parametrization in the context of the above results is presented in section 4. It also outlines the reasoning behind the selection of the potentials for the next stage of benchmarking.

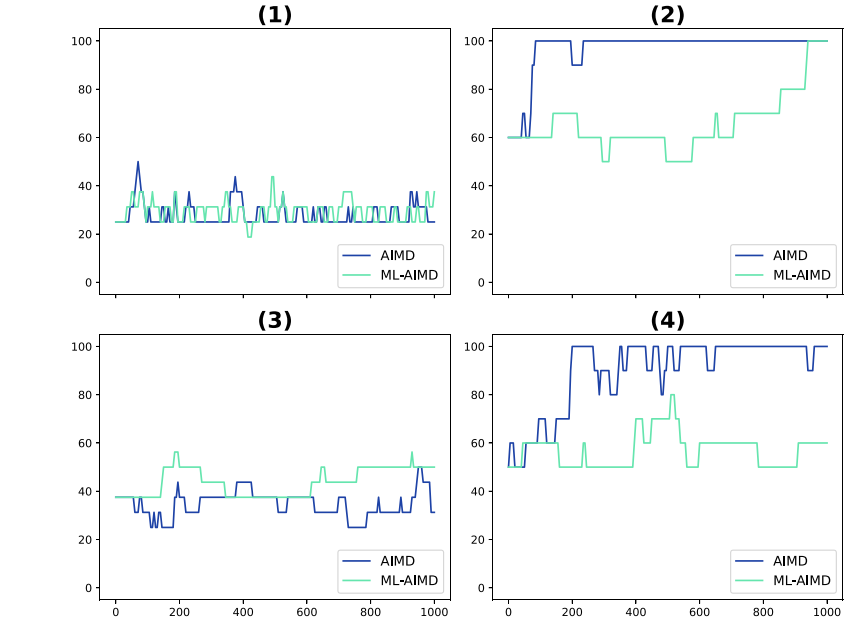
3.2. AIMD and ML-assisted AIMD comparison

3.2.1. Architecture testing

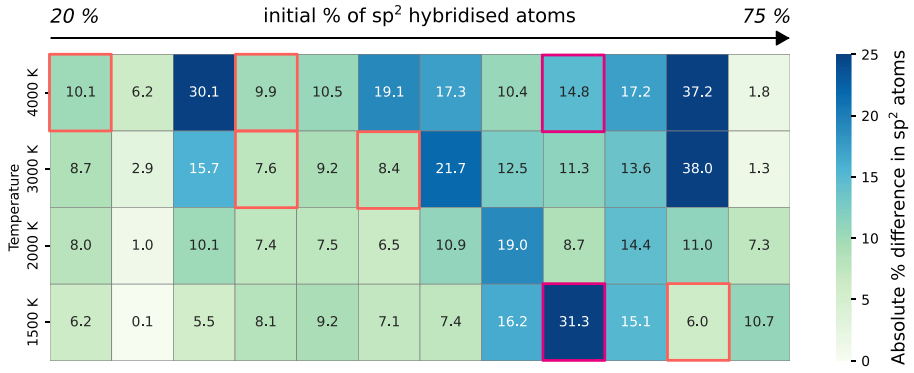
The suitability of the ML model architecture was determined by considering the energy differences between the on-the-fly GAP model and AIMD (figure 6), the average absolute difference in the percentage of sp^2 local environments over the course of the simulation (figure 7(b)), and the absolute difference in the sp^2 percentage difference between the final snapshots of the AIMD and ML-accelerated simulation (figure 7(c)). For those unfamiliar with the work of Stenczel *et al* [26], an accessible explanation of the process that allowed for this comparison is provided in section 4.

The energy difference (figure 6) was generally below 0.02 eV per atom, which corresponds to approximately 2.0 kJ mol^{-1}. Taking into account the Boltzmann law (k_bT), the meaningful difference in energy at the tested temperatures is $\gtrsim$12.5 kJ mol^{-1}, $\simeq$6 times higher than observed values. On average, differences in estimated energies increase with increasing temperature of the simulation. However, as the differences stayed well below the limit designated by the Boltzmann law, those differences can be regarded as negligible.

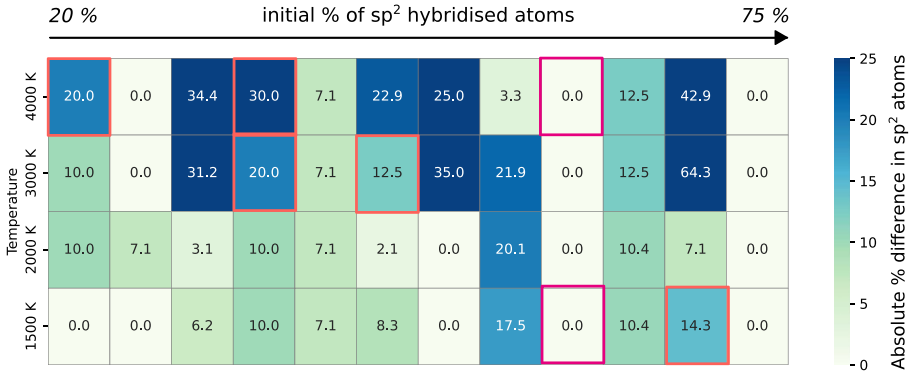
When it comes to the analysis of figure 7, they are a measure of how similar the simulations obtained with AIMD were and accelerated AIMD with the on-the-fly GAP model, whose architecture was to be used to develop the final ML-AIMD model. They yield essentially four types of results: with good agreement with AIMD both throughout and at the end of the simulation, with good agreement in only one of them (see: orange and pink frames in figures 7(b) and (c)) or poor agreement in both. These cases are visually represented in figure 7(a). The so-called 'good agreement' is defined as the result that does not exceed the 12.5% difference, as this corresponds, for the unit cells in question, to the change in coordination for fewer than 4 atoms. Most of the data points in figure 7 fall into the first category, when it is safe to assume that the AIMD results were correctly reproduced by the accelerated AIMD with the on-the-fly GAP model.



(a)



(b)



(c)

Figure 7. (a) Example of four possible courses of simulations: with good agreement between on-the-fly GAP model and AIMD both throughout and at the end of simulation (1), with good agreement in either of them (2 and 3) or poor agreement in both (4). (b) Absolute, percentage difference in sp^2 content between AIMD and on-the-fly GAP results throughout the simulation. Percentage difference of sp^2 content with respect to AIMD results was calculated for every 5th step and then averaged. (c) Absolute, percentage difference in sp^2 content between AIMD and on-the-fly GAP model results for the last snapshot from the simulation. Absolute percentage difference of sp^2 content at the 1000th was calculated with respect to AIMD value. Orange frames point out the data points that displayed good agreement between AIMD and on-the-fly GAP model throughout the simulation but final sp^2 content differ significantly. Pink frames point out the data points that displayed poor agreement between AIMD and on-the-fly GAP model throughout the simulation but had similar or identical final sp^2 content.

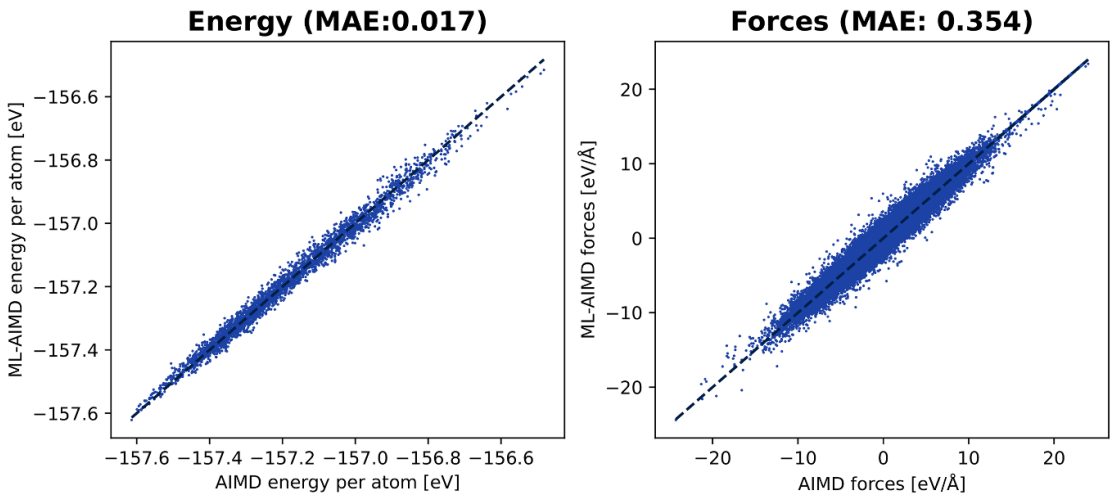


Figure 8. Energies and forces on the validation set derived from ML-AIMD model against respective energies and forces from AIMD snapshots. Dashed line is the trend line for AIMD results, i.e. data resembling AIMD results perfectly would fall into that trend line.

If the percentage difference throughout the simulation is low but high for the last snapshot (see orange frames in figures 7(b) and (c)), the simulation diverged just at the end, but was likely reproduced with high accuracy for the majority of timesteps taken.

If the percentage difference throughout the simulation is high but low for the last snapshot (see pink frames in figures 7(b) and (c)), the simulation obtained with accelerated AIMD with the on-the-fly GAP method likely corrected its course and eventually reproduced the AIMD results. Another possibility is that they corrected their course only seemingly and the obtained structure has the same amount of sp^2 content but differs in topology.

The on-the-fly GAP model performance should be analyzed in the context of the working principles of accelerated AIMD developed by Stenczel *et al* [26]. and simulation settings, as they can address its shortcomings in the presented comparison. All this can be found in the section 4.

In conclusion, considering negligible differences in energy estimations and mostly good agreement in the percentage of sp^2 content throughout and at the end of the simulation, it was agreed that the architecture is suitable to build a model replicating the results of AIMD, later called the ML-AIMD model. More details are presented in section 4.

3.2.2. ML-AIMD model validation

Figure 8 presents the energies and forces calculated by the ML-AIMD model for structures in the validation set (*ergo* structures not seen by the model during training), shown against the energies and forces obtained from the AIMD snapshots. The MAE for both quantities is exceptionally low, compared even to those presented in figures 5(b) and 4 for a single snapshot.

In general, the newly developed ML-AIMD model was deemed suitable to be used as a benchmark in the next stage. A full discussion of the reasoning behind this decision, the model's strengths and limitations can be found in section 4. The estimated energies and forces on the training set (*ergo* structures seen by the model during training) shown against the energies and forces from AIMD snapshots can be found in the SI (figure S8).

3.3. Extended testing

Extended benchmarking involved simulations on two systems' sizes: moderate (800 to 1280 atoms) and large (6400 to 10 240 atoms). The reason for that was to check how system size would influence the models' performance with respect to ML-AIMD. Since no major difference in the general trend was observed, the following section focuses solely on results for large supercells. Results for moderate supercells can be found in the SI (figures S8 and S9).

As described in section 2.4, the analysis involved comparison of final structures' energies, percentage difference in 3-coordinated atoms at the end and throughout the simulation, and in some cases topology analysis, with statistical topology analysis being attempted.

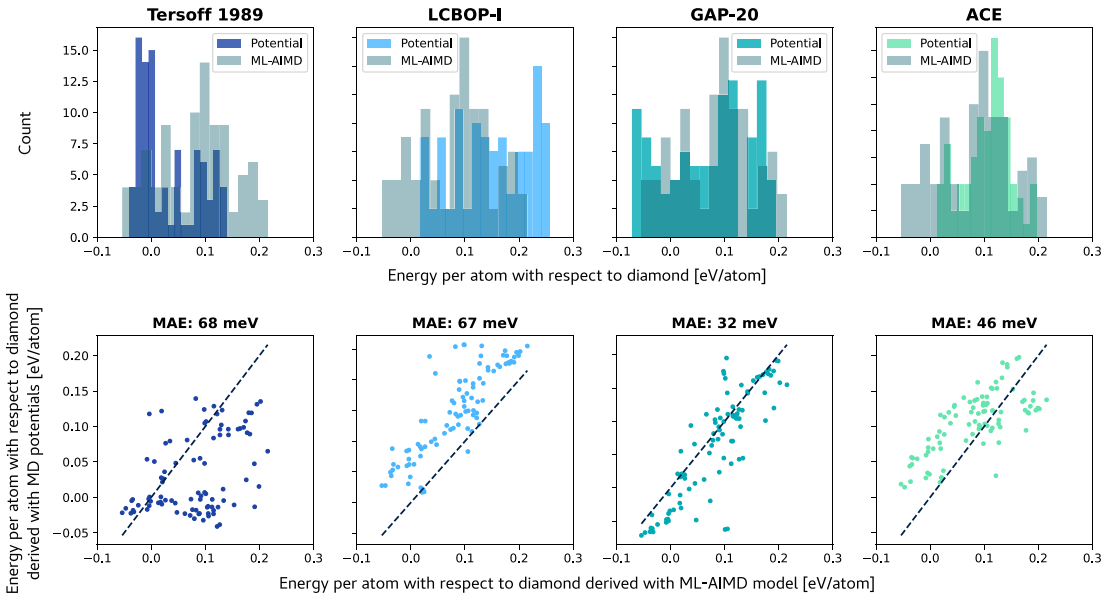


Figure 9. Distribution of Final Structures' energies with respect to energy of diamond lattice at the same temperature (top row) and direct comparison of those energies with ML-AIMD derived results for the same structure (bottom row). The superimposed, gray distributions correspond to results obtained with ML-AIMD potential, while dashed line is the trend line for ML-AIMD results, i.e. results resembling the benchmark perfectly would fall into that trend line.

3.3.1. Final structures' energies

Distributions of final structures' potential energies obtained in the simulations with respect to diamond on large scales are presented in figure 9 along with direct comparison to energies for the same structures obtained with ML-AIMD potential. Values above zero corresponds to structures less thermodynamically stable than diamond.

GAP-20 results shows both lowest MAE in direct comparison and most similar value range in the histogram. As expected, Tersoff 1989's energy distribution is neither similar in shape nor in values, which is clear form its scatter plot and highest MAE. Both LCBOP-I and ACE do not predict any of the structures to be more stable than diamond. This contradicts with topology analysis presented in section 3.3.3 and absolute percentage of 3-coordinated atoms in the SI (figure S11). It is clear that some of the initial diaphitic structures reached either substantial or full graphitization, and therefore their energy should be lower than that of a diamond lattice. Additionally, most energies estimated with these two potentials are systematically overestimated compared to reference values, similarly to those in figure 5(a). It is visible in both the histograms and scatter plots (figure 9) that, upon correction, ML-AIMD trend would be much better reproduced.

The reason for this lays in the Energy-Volume curves for each of these potentials (figure 13) explored in more details in section 4.2. Figure 13(b) shows that energy-volume curves, when varying equilibrium volume of graphite and diamond, are much closer in energy for LCBOP-I and ACE than for GAP-20. Therefore, even small compression of graphitic layers could lead to the estimated energy being higher than that of equilibrium diamond for LCBOP-I and ACE.

Overall, GAP-20 performed best in this task. However, LCBOP-I and ACE systematic increase in energy should be attributed to encoded difference in graphite and diamond energy differences rather than poor potentials' performance, as relative energies of the diaphitic structures can still be reproduced well.

3.3.2. Change in % of 3-coordinated atoms

Figure 10 shows percentage difference in the number of 3-coordinated atoms throughout and at the end of the simulations performed with different MD potentials compared to analogous simulations performed with ML-AIMD. Data are presented with respect to the initial percentage of 3 coordinated atoms (x-axis) and temperature of the simulation (y-axis). Darker colors indicate greater deviation from ML-AIMD.

In contrast to figure 7, the percentage difference in 3-coordinated atoms throughout the simulations and at the end is consistent in most cases, meaning that there were no sudden changes in structures towards the end of the simulations. It is likely related to the length of the simulation (1 ps vs 52 ps)

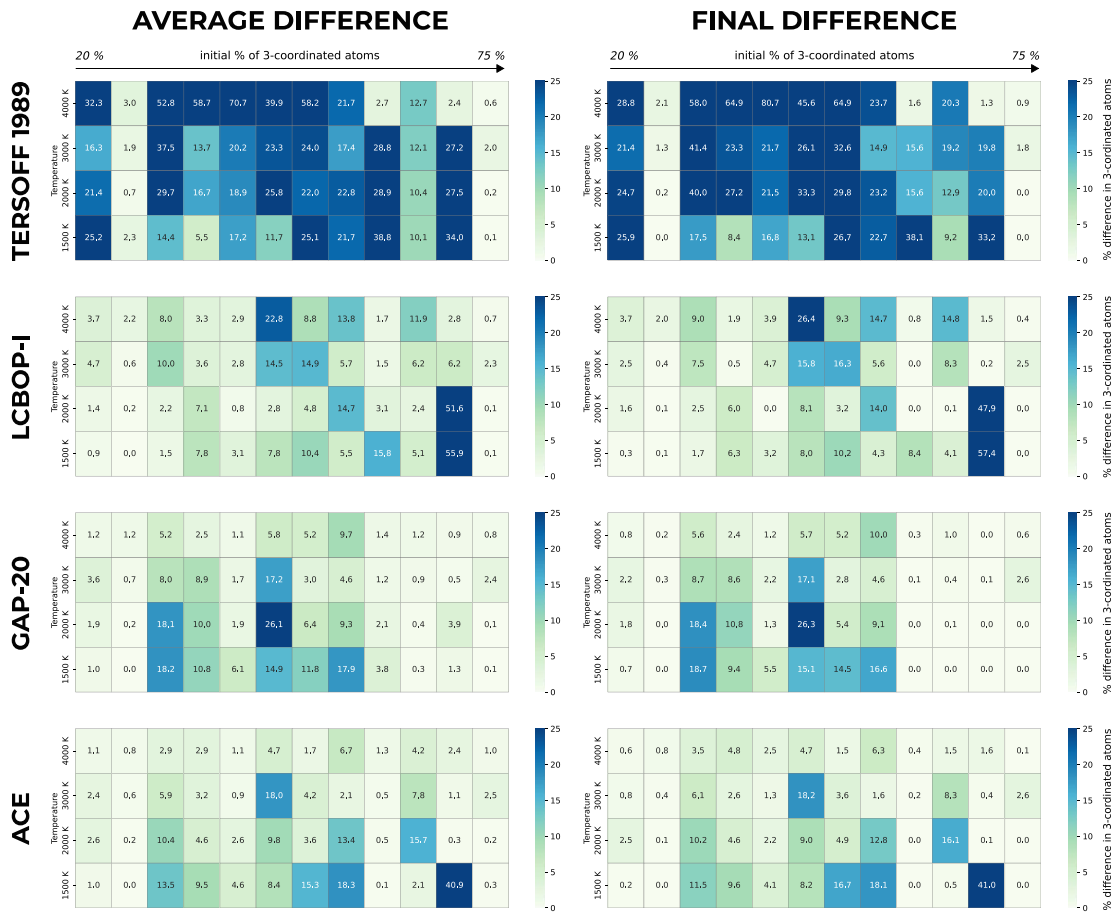


Figure 10. Difference in percentage 3-coordinated atoms content throughout (left) and at the end of the simulations (right) compared to ML-AIMD results (color scale) and with respect to temperature (*y*-axis) and initial percentage of 3-coordinated atoms (*x*-axis). The darker the color, the greater the difference.

Table 1. The average percentage difference for all the simulations throughout their course *(at the end)* for each potential compared to ML-AIMD, presented with respect to temperature of the simulation and averaged over all temperatures. The best results in each series are in **bold**.

TEMP.	Tersoff 1989	LCBOP-I	GAP-20	ACE
4000 K	29.58% *(32.40%)*	10.32% *(11.40%)*	4.75% *(4.65%)*	**3.77% *(3.36%)***
3000 K	19.28% *(19.96%)*	7.29% *(7.14%)*	5.34% *(5.19%)*	**5.29% *(5.08%)***
2000 K	20.12% *(22.89%)*	**7.91% *(7.65%)***	8.88% *(8.48%)*	8.51% *(8.35%)*
1500 K	17.28% *(18.01%)*	**7.79% *(7.02%)***	10.28% *(9.87%)*	11.11% *(10.67%)*
Average	21.57% *(23.31%)*	8.33% *(8.30%)*	7.31% *(7.05%)*	**7.17% *(6.87%)***

and its stability. The agreement between ML-AIMD and the potentials increases down the figure: the less intensively colored squares are observed. The average percentage difference for all the simulations throughout their course *(at the end)* for each potential compared to ML-AIMD (both overall and stated with respect to temperature of the simulations) are presented in table 1.

Both GAP-20 and ACE tend to perform the best at higher temperatures, while LCBOP-I yields the most accurate results at lower temperatures. The structures developed with the former two potentials tend to graphitize at lower temperatures. Also, average variation in % of 3-coordinated atoms was found to be higher for diaphite type I and type II with CO gradia. Heatmaps showing the absolute percentage of 3-coordinated atoms for simulations with each potential can be found in the SI. (figure S10)

Overall, the differences between LCBOP-I, GAP-20 and ACE are not dramatic, but the accuracy might be structure and temperature dependent.

3.3.3. Topology analysis

Statistical topology analysis was attempted with Ring Statistics and Persistent Homology (PH). The former was expected to amplify the differences in the amount and type of interfaces present. However,

heat maps presenting cosine distances between structures developed with ML-AIMD and the potentials followed the same patterns as those in figures 10 and S9. As they did not provide any new context to the results presented above, they have been included in the SI (figure S11).

PH as implemented in both the `ripser` and `gudhi` python packages, turned out to be too computationally intensive for system of those size and available resources. While authors recognize the potential application of PH in diaphites' feature classification (e.g. the type of interface present), it is deemed to assume that the evidence presented is sufficient to draw conclusions on the transferability of carbon potentials. The diaphites' medium-range order statistical analysis with PH in a form of persistence diagrams will be of interest in future work.

Since examining the topology of a final structure case-by-case would be unpractical, the following section focuses on two structured: D-5-G-3-CO (figure 11) and D-3-G-4-CA (figure 12). This choice was dictated by the closest to average difference in percentage 3-coordinated atom content across all three candidate potentials (LCBOP-I, GAP-20, and ACE). It allowed to gain an insight on how dissimilar to the benchmark (ML-AIMD) an average case can be and whether it is a significant difference or not.

The ideal potential would reproduce ML-AIMD results perfectly. It is clearly not the case for snapshots presented in figures 11 and 12. In case of D-5-G-3-CO, LCBOP-I somewhat agrees with the benchmark for simulations at 2000 K and 4000 K, with slightly more graphitized regions. For D-3-G-4-CA it performs better at lower temperatures. While results obtained with GAP-20 replicate ML-AIMD ones relatively well for D-3-G-4-CA (except at 3000 K), it fails when it come to D-5-G-3-CO, not accounting for any change in domains shape and orientation at all. ACE-predicted structures for D-5-G-3-CO resemble the benchmark quite well. For D-3-G-4-CA structures, it changed the shape of domains much more than ML-AIMD and favored greater degree of graphitization for lower temperatures but it was not worse than other potentials.

It did not escape our attention that the simulated domain structures produced by all tested potentials qualitatively resemble those observed in HRTEM and HAADF-STEM images [1, 3, 27]. Although such agreement cannot be used to rank their quantitative accuracy, it suggests that the potentials reproduce the essential structural features observed experimentally. Thorough examination of these features—like their formation mechanism, relative stability and transformations–will be a subject of future work.

The comparison above and ring statistics presented in the SI (figure S12) suggest a strong, yet expected, correlation between difference in 3-coordinated atoms and topology. Therefore, it suggest that the potential which agrees the most in percentage of 3-coordinated atoms with ML-AIMD is also the one that captures topology most accurately: namely ACE.

3.4. Final result

The results above strongly supports the conclusion that the most transferable carbon potential overall for diaphites is ACE. Its backed up by good prediction in energy trends, lowest overall difference in percentage of 3-coordinated atoms and reasonable structure predictions even for structures that were expected to deviate from the benchmark by quite a lot. While it is quite comparable with GAP-20 (except exact energy predictions in which GAP-20 surpasses) it is much more computationally efficient. It was found to yield results approximately four times faster than GAP-20, while other claim it to be even eighty times faster [25]. Interestingly, LCBOP-I was not far off from more complex ML-based potentials. Contradictorily, it outperformed ACE at lower temperatures as evidenced in table 1, figures 10, and 12. Therefore, it might be of use for simulations that require lower temperature ranges and greater computational efficiency. It is worth pointing out that the differences between the potentials filtered in the Preliminary Screening were generally quite small. It shows the effectiveness on Preliminary Screening in separating 'the good' from 'the bad' but also demonstrates how hard it might be to make an executive decision regarding MD potentials.

4. Discussion

4.1. Choice of DFT exchange-correlation functional

The PBE functional was chosen as a result of its computational efficiency, good performance in benchmarks for extended solids, and generality, since it is not in a semi-empirical functional, no bias is introduced [28, 29]. However, it is known to describe dispersion interactions relatively poorly [30] and so it was combined with the Grimme dispersion correction [18, 31, 32]. The simplest D2 correction was chosen deliberately due to planned AIMD, which is relatively computationally intensive [18]. Employing more accurate D3 [31] or D3BJ [32] dispersion corrections would substantially increase computational time with little added potential benefit.

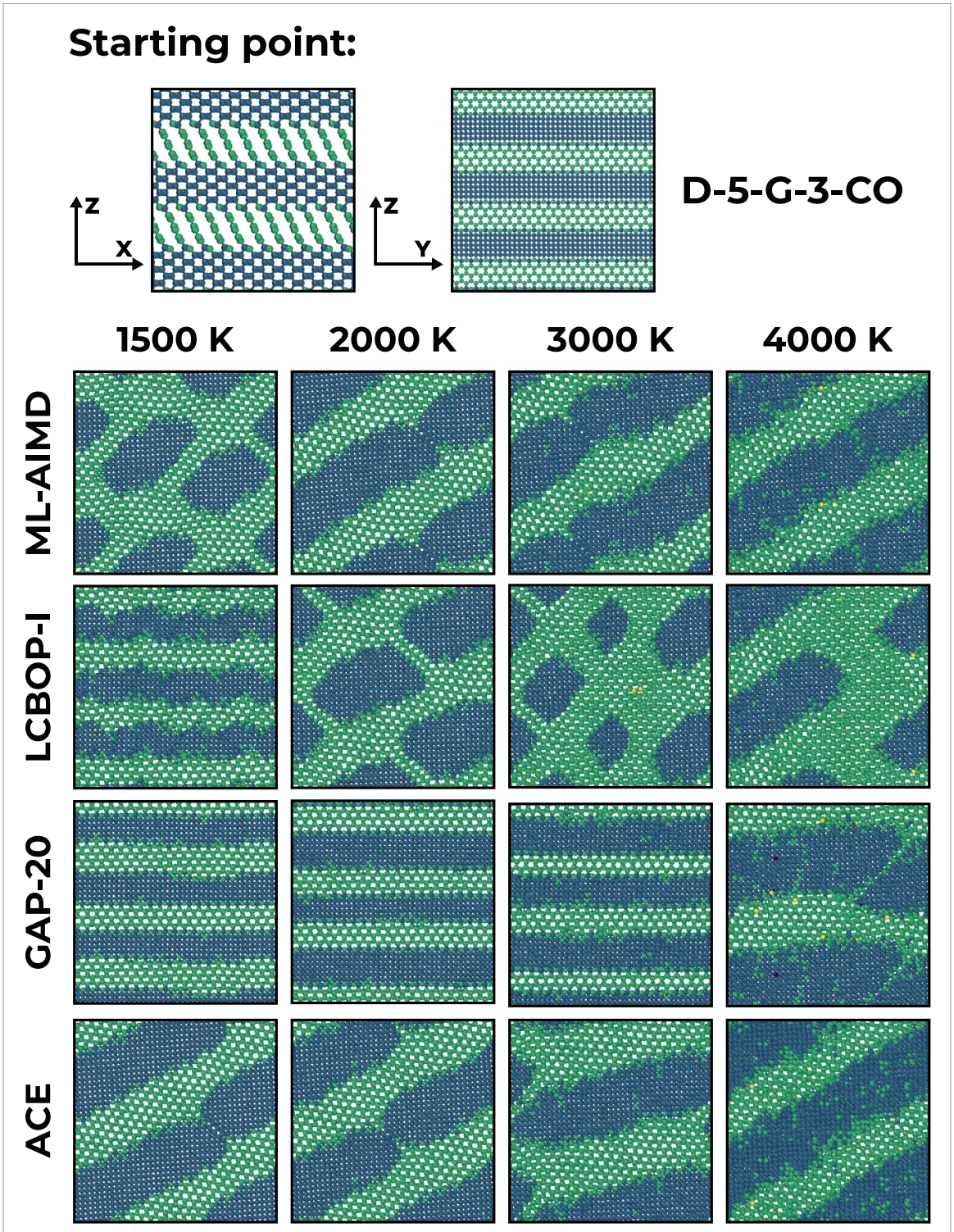


Figure 11. Final snapshots form the Extended simulations (52 ps, with time step of 1 ps) at four temperatures for D-5-G-3-CO diaphite obtained using ML-AIMD (the benchmark) and LCBOP-I, GAP-20, and ACE potentials. Final snapshots presented along *x*-axis.

Some might argue that, similarly to the MD potential, the DFT exchange-correction functional used in such a test should also be chosen via benchmarking, but against e.g. CCSD(T)+CBS instead. In this case, it was decided to refer to the literature as DFT functionals are inherently more general than MD potentials. In addition, the systems in question consist of only carbon atoms, leaving any kind of strong bond polarization or inner core electron effects out of the picture.

4.2. Connection between preliminary screening results and parametrization

While the performance of the potentials should be the main concern when benchmarking, it is best to understand why certain potentials perform better than others. It may be related to the underlying

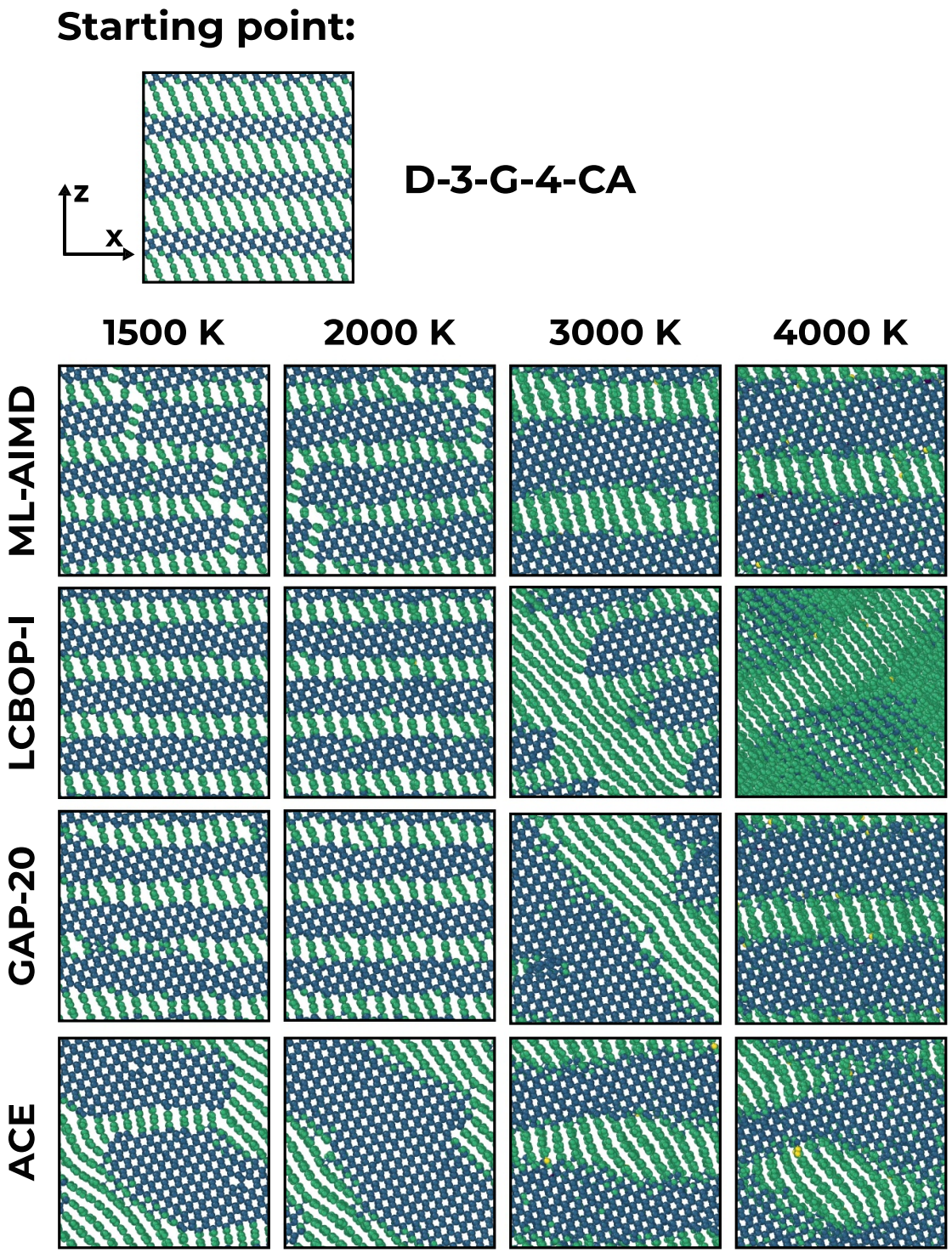


Figure 12. Final snapshots form the Extended simulations (52 ps, with time step of 1 ps) at four temperatures for D-3-G-4-CA diaphite obtained using ML-AIMD (the benchmark) and LCBOP-I, GAP-20, and ACE potentials. Final snapshots presented along *y*-axis.

parametrization, in case of empirical potentials, or training data, in case of ML potentials. To develop this point, figure 13 shows the energy/volume curves for the diamond and graphite crystal structures, obtained using eight models.

The following provides a concise summary of the main characteristics of each potential in the context of diaphites modeling. The Tersoff 1989 potential was optimized for short-range interactions, having a cut-off distance for considering atom-atom interactions of 2.7 Å, which is smaller than layer separation in graphite (3.1 Å) [35]. As a result, it is intrinsically incapable of predicting the correct ground state. This is confirmed by figure 13(b) which shows that the Tersoff 1989 potential predicts that diamond is

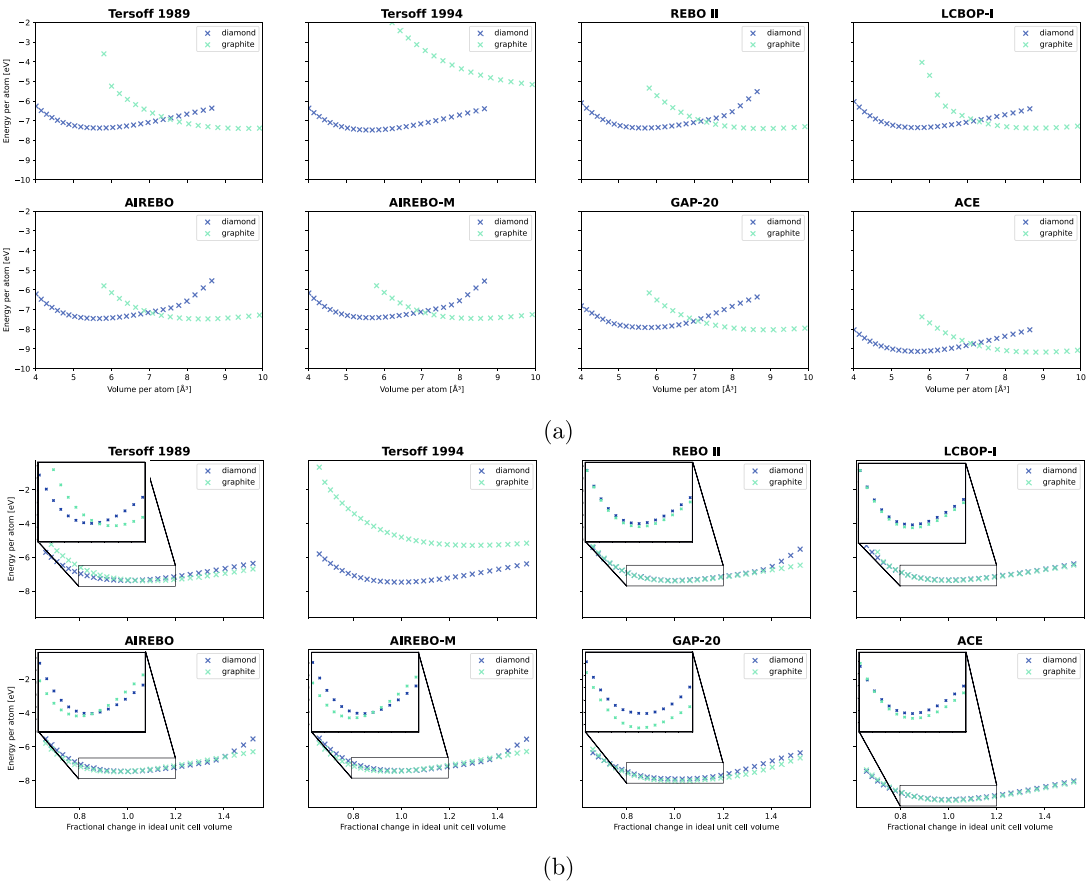


Figure 13. Energy-Volume curves for diamond (navy '*x*') and graphite (celadon green '*x*') estimated with tested potentials. (a) Energy per atom plotted against volume per atom. (b) Energies per atom plotted against the fractional change in the volume of an experimentally-derived, equilibrium unit cells [33, 34]. The fragments of a curves for fractional change in equilibrium unit cell volume between 0.8 and 1.2 were zoomed in for the potentials where graphite and diamond curves overlaps at smaller magnification.

more stable than graphite. Furthermore, while it can estimate the energy of some diamond surfaces, it performs poorly for the Pandey surface present in diaphite type I [36].

Similarly, the Tersoff 1994 potential was optimized with respect to short-range interactions in crystalline solids [37]. In this case, the target compound was SiC, which does not form graphite-like sheets. Therefore, a concomitant reduction in performance was observed (figure 5). Figure 13 indicate that the energy-volume curves for the Tersoff 1994 potential are entirely different from the others.

The REBO-II model expands on the empirical bond order in the Tersoff 1989 model by explicitly including local coordination on the atom and bond angles [20]. However, it still does not consider any non-bonding interactions, focusing on network solids such as diamond or small molecular fragments, making it too short-ranged [21].

LCBOP-I accommodates long-range interactions using a Morse potential, which is known to have a softer repulsion region compared to Lennard–Jones. As a result, it is considered more transferable when it comes to pressure-induced structural changes [21, 23], which might be responsible for the good agreement with the distribution of DFT forces (figure 4). In addition, the parametrization was focused on the diamond graphitization process—a process that diaphites are susceptible to [21]. In addition, the fitting also included the Pandey surface (present in type I diaphites) and clusters. This shows a strong focus on solid carbon polymorphs.

AIREBO and AIREBO-M are twin potentials, using Lennard–Jones and Morse potentials, respectively, for treating long-range interactions. Both potentials were created to consider generic hydrocarbons, [22, 23] to model reactions of liquid hydrocarbons, building on the REBO-II model adding non-bonded interactions. As a result, these models do not focus specifically on covering different carbon polymorphs. Instead, they are optimized for reactions and interactions between hydrocarbons in the condensed phase, whilst perturbing potential performance on the former as little as possible [22]. Their energy-volume curves reflect that since only AIREBO and AIREBO-M potentials predict diamond to be more stable than graphite upon expansion of their unit cell from the equilibrium.

GAP-20 is one of the ML potentials considered in this benchmark study and one of the best-performing potentials in the Preliminary stage (section 3.1). It was trained on a broad dataset of approximately 5000 carbon structures, including crystalline, amorphous, high-pressure, and theoretically predicted phases, as well as 1D and 2D materials, clusters, and defects. Training data were derived using DFT with the optB88-vdW functional, which is a dispersion-inclusive DFT functional [38], which means it should not require additional dispersion corrections, such as D2. In terms of DFT accuracy, optB88-vdW is positioned above standard GGA functionals, like PBE, as it includes non-local dispersion interactions. However, those higher-level DFT functionals are typically based on a GGA functional. However, in this case, the employed GGA functional was B88 rather than PBE, which explains the systematic discrepancies in the energies of diaphitic structures that disappear after correction [39] (figure 5(a)).

ACE, the other potential with nearly perfect agreement with DFT (section 3.1), was trained on a similar scope of structures, but its training set contained over 17 000 entries [25]. Since the PBE functional alone is unable to accurately estimate dispersion forces present in some carbon polymorphs [17], ACE incorporates dispersion corrections in a postprocessing manner, where an additional energy term is included, rather than modifying the functional itself, similar in principle to how Grimme's dispersion corrections are applied [25, 31]. The alignment of the functional with a Grimme-like dispersion correction resulted in near-perfect agreement between ACE-predicted forces and energies and the DFT reference data in section 3.1, even before applying any corrections.

In general, broad and extensive training datasets likely contribute to the superior performance of ML potentials compared to empirical ones, making both GAP-20 and ACE more robust and transferable when dealing with novel structures such as diaphites.

Despite some superficial similarities between these two ML potentials (e.g. extensive training data and the scope of covered structures), their underlying methodologies are entirely different and worth noting. GAP-20 is based on Gaussian process regression with SOAP and other descriptors, while ACE uses Polynomial Basis Functions with ACE basis functions [40–43]. In the case of GAP, SOAP descriptors encode the local atomic environment using spherical harmonics and radial basis functions, which are evaluated through a kernel-based, nonlinear approach [42]. In contrast, ACE represents atomic environments using an expansion in orthogonal polynomials based on both radial and angular coordinates, with parameters fitted using linear regression [43]. This transition from kernel-based evaluation to linear regression drastically reduces computational cost, making ACE orders of magnitude faster than GAP-20 while maintaining comparable accuracy, as demonstrated in section 3.1 [25].

In summary, an analysis of the parametrization involved in developing the models provides insight into why certain potentials performed better than others. The Tersoff 1989, Tersoff 1994 and REBO-II potentials appear too short-range to account for all interactions present in diaphites, especially those between graphite sheets. As a result, they were deemed insufficient for accurate simulations on diaphites and excluded from the Extended benchmarking, with the exemption of Tersoff 1989. The latter was chosen to serve as a 'sanity check' to demonstrate how an unsuitable but seemingly efficient and general potential could perform. AIREBO and AIREBO-M were not intended for solid-state carbon materials, which is reflected in the results shown in figure 13(b). The trend presented for energy change with respect to volume deviations is visibly different from other potentials that performed better in the Preliminary benchmarking (like ACE or GAP-20). Consequently, they were also discarded. LCBOP-I was specifically parametrized to capture changes in solid-state carbon polymorphs. This advantage is reflected in slightly better results in the Preliminary stage compared to AIREBO or AIREBO-M. As a result, it was chosen as the only empirical potential tested in the Extended stage. Last but not least, both GAP-20 and ACE were selected. This was supported by excellent agreement with the DFT data (upon relevant corrections) and their extensive training data. At this stage, computational expense was not considered as a disqualifying factor. However, it is taken into account in the discussion of the Extended stage, as it plays a much more significant role.

4.3. ML-AIMD model development

4.3.1. Working principle

In this section, we explain the basics of creating an AIMD-based ML model using accelerated AIMD with the on-the-fly GAP model. More detailed and technical explanations can be found in the paper by Stenczel *et al* [26].

As diaphites are regarded as nanomaterials, to represent them accurately in dynamic simulations, sufficiently large supercells are required (i.e. of the order of thousands of atoms). However, this is far beyond current AIMD capabilities and hence there are no sensible benchmarks for large-scale simulations. The answer was to develop this benchmark ourselves using accelerated AIMD with on-the-fly GAP, which works in the following way.

Classical AIMD is performed for the first few steps of the simulations (ten in this work). The results from those steps are fed to the ML-GAP model [44] (see SI for example input files). The newly trained GAP model carries on the simulation for a specified number of steps, in this work set to two. AIMD results are generated for the same configuration, and the results are compared with respect to tolerance argument (here, the tolerance was based on energy difference and set to 0.01 eV per atom). If the tolerances conditions are met, the ML model carries on the simulation until the next check, when the comparison starts again. If the tolerances conditions are not met, then the model undergoes refitting and from now on the refitted model continues the simulation until the next check.

The frequency of the checks depends on the chosen mode: fixed or adaptive. In the fixed mode, checks are performed every *n*th step regardless of whether tolerance conditions were met or not. In adaptive mode, when tolerance conditions are met, the checks become less frequent; similarly, when tolerance conditions are not met, the checks become more frequent. This is governed by the scaling factor, set to two in this work. If the checking criteria for AIMD and on-the-fly GAP model comparison are met, the next check is scheduled to take place after twice as many time steps. If the criteria are not met, the next check is to occur after half as many steps.

Each simulation performed with accelerated-AIMD with the on-the-fly GAP model results in three important outputs:

- tolerances values (in this case, energy difference), which show how well the model performed during checks; it is a direct measure of how well the model's architecture fits the given task.
- the simulation, which includes steps derived both by the AIMD and on-the-fly GAP model; its outcome can be compared with the results of pure AIMD simulations.
- file with all AIMD steps extracted from the simulations; they served as a training data set for the on-the-fly GAP model.

The first two of these were used to validate the model's architecture (see section 3.2 and figures 6, 7). When it was deemed suitable (see section 3.2 and the next section), all AIMD steps form accelerated-MD simulation (last relevant output mentioned above) were included into the training data for the final ML-AIMD model.

4.3.2. Overcoming the limitations of the on-the-fly GAP model

We first address the choice of the model architecture adapted from Stenczel *et al* [26] The model architecture was repurposed from that intended to study carbon graphitization at 3000 K. As this model was specifically constructed for carbon, for processes that can occur in diaphites and for similar temperatures, it seemed reasonable to reuse it, and the validation results in section 3.2.2 confirmed this assumption. However, being aware that ML is often treated like a so-called 'black-box', the reasoning behind the choice of this specific architecture is provided below.

Here, the model architecture encompasses a number of key aspects: mathematical basis, optimization algorithm, training data, descriptors, hyperparameters, noise control, and memory management. When choosing to train a GAP model, we have already committed to Gaussian process as a mathematical basis and Gaussian Process Regression as a optimization algorithm. Training data were provided on-the-fly by AIMD performed by CASTEP. Noise control is more related to mitigating overfitting when performing a Gaussian Process rather than the fact that it is a carbon material. Memory management does not play a vital role in model's accuracy. Therefore, the following paragraph focuses on the descriptors and hyperparameters and why they are suitable for a diaphite-specific model.

The model architecture described by Stenczel *et al* [26]. includes three descriptors: two-body, three-body and smooth overlap of atomic positions (SOAP), which is a highly-dimensional many-body descriptor [42]. The interaction cut-off distances are set to 3.7, 3.0 and 3.7 Å, respectively, which is appropriate to capture interactions between graphitic layers. The use of all three types of descriptor ensured an accurate depiction of different bonding environments, including the diamond-graphite interface. In the case of SOAP, its accuracy is based on the number of sparse points or representative points extracted from the region around each atom within the cut-off distance. It was set to 200 which is relatively low. For comparison, GAP-17 used more than 4000 sparse points [45]. This choice increased computational efficiency, at the cost of accuracy. It was partially mitigated by algorithm that chooses the most representative sparse points. Then, when training the ML-AIMD model on all AIMD results, the loss of accuracy due to the low number of sparse points was outweighed by the large training data set.

Hyperparameters govern the expected error in the target properties: energy, forces, and stress tensor (set to 0.002 eV atom^{-1}, 0.2 eV Å^{-1}, and 0.2 eV/Å°, respectively). The values make the model sensitive to

local changes in structures (e.g. at the grain boundary), which is of importance in meta-stable structures like diaphites.

The above arguments were deemed convincing to assume that an ML model with this architecture would be suitable to create a diaphite-specific ML-AIMD model.

However, the validation presented in section 3.2.1 may appear unconvincing given the near-ideal agreement between the AIMD and ML-AIMD models in section 3.2.2 and the non-ideal reproducibility of simulation outcomes as measured by the percentage of sp^2 content (section 3.2.1, figure 7(b)). How is it possible that the 'mediocre' on-the-fly GAP model turned into extremely accurate ML-AIMD? The answer lies in the training data set. The on-the-fly GAP model has access to AIMD calculations from a single run, that is, to the first ten initial AIMD steps (section 4.3.1) and other AIMD calculations done for retraining/checking purposes. In comparison, the final ML-AIMD model used nearly 12 000 AIMD calculations for training obtained in accelerated-AIMD simulations.

Additional factors that could influence the differences between AIMD and accelerated-AIMD included the small system size and simulation time. Because of the former, even small differences in atomic positions could have an effect on the forces calculated in the next step, and collectively those small changes led to massively different results. As the cut-off distance for interactions, programmed as GAP's hyperparameter is 3.7 Å to account for interactions between graphitic layers, there is a possibility that atoms interact directly with images of themselves. However, the same problem could arise in the classical AIMD, as the periodic boundary conditions and size of unit cells were the same in both cases. Another factor mentioned earlier was the simulation time. It is possible that some systems did not fully equilibrate [46], resulting in discrepancies between AIMD and accelerated-AIMD. Although increasing the system size and/or simulation length could improve that, it was not viable due to the massive increase in computational cost (i.e. scaling as N^3, where N is the number of atoms).

An increasing amount of training data was hoped to compensate for those shortcomings. Looking at incredibility low MAE for the energy and forces estimations in figure 8, it can be said that the aim was achieved.

Essentially, the whole process yielded an ML-AIMD model later used as a benchmark in the "extended" benchmarking using long simulations on the larger supercells. It is worth reiterating that the ML-AIMD model is, in fact, a GAP potential, differing from GAP-20 in its hyperparameters and training data. The phrase 'AIMD' in the name simply emphasizes the fact that it was derived from AIMD simulations.

In the face of that, one might cast doubt regarding the sense of the extended stage of the benchmarking: after all, the ML-AIMD model would be as cheap as GAP-20 and nearly as accurate as AIMD, so why not skip the effort and just use this? Although there is some truth to this analysis, it overlooks a few factors that make the ML-AIMD model suitable as a benchmark but poor as a general carbon potential. The ML-AIMD model was trained solely on AIMD simulations on diaphites at four specific temperatures in the NVT ensemble. The validation of the model (section 3.2.2) proved the accuracy of the model under those conditions and for those structures only. As a result, there is no guarantee that the model is transferable e.g.: for pressurized systems or when other carbon polymorphs are embedded into the structure. While some AIMD snapshots used for training contained graphitized, diamondized or amorphous phases, as the simulation led to those structures and they were present during checks, it is not enough to talk about general transferability of the model. Above all, the main aim of this work was to determine which widely-available general carbon potential is the most suitable for simulations on diaphites. ML-AIMD model used in this benchamarking does not meet this criterion. However, development of a model that includes diaphites in its training/parametrization data might be an interesting path to take. It would be interesting to see how inclusion of diaphites changes parameters in empirical potentials, as well as whether and how it improves the accuracy of models with different architectures, like GAP, ACE, or MACE.

4.4. Other MD potentials

It is worth mentioning that the presented transferability testing is not meant to be exhaustive, as there are much many empirical and ML-based MD potentials available for carbon [47]. Since the paper serves mostly as a demonstrative example for the transferability workflow, we believe that it can be extended to other carbon potentials—whether by comparing results obtained with the potential of interest with the presented work or by analogy. Such an analogy can be made, for example, when considering the growing field of ML potentials.

The field of ML-based MD potentials grows and develops dynamically. Models with more complex architecture are becoming more and more accessible and widely used. Simultaneously, a great

deal of effort is put into creating foundational ML potentials that aim to cover large parts (if not the entire scope) of the periodic table and their chemical space [11, 12, 48]. This poses questions about the upcoming redundancy of single-type-of-atom potentials (like carbon-only potentials). Considering all the above, it is important to contextualize the presented work with respect to other potentials, specifically ML-based ones.

Currently, one of the most widely employed models architecture for MD is Multilayered Atomic Cluster Expansion, shortened to MACE [48]. It is considered the-state-of-the-art in the machine-learning force-fields. [49] Not including the MACE carbon potential in this work could be considered disqualifying for some. However, MACE is, in fact, a flexible nonlinear extension of the ACE potential [48, 50]. Both models evaluate the descriptors in the same way; while ACE fitting is based on a linear model, MACE uses an equivariant message-passing graph neural network (GNN) [50]. It allows for greater flexibility, as each layer of the GNN can be adjusted to fine-tune the model. It results in generally improved accuracy compared to GAP or ACE models, keeping in mind that errors in the later models for physical properties are in the order of a few percents, deemed acceptable [50]. 'The elephant in the room' to address is hardware requirements. For efficient calculations with MACE, as is generally the case for GNN, GPU units are required or at least highly recommended [51]. Therefore, it is less accessible than CPU-based GAP and ACE, not to mention even cheaper empirical potentials.

When it comes to foundational models, there is no surprise that the most efficient of them are based on the MACE architecture, as it is designed to deal with large data quantities in parallel. Although models like MACE-MP0 [11] or MACE-OFF23 [12] are able to generate stable dynamics, they often yield approximate results and, in some cases, are trained on equilibrium configurations only [50, 52]. Their accuracy is still quite far from ML-based MD potentials focused on one type or a narrow set of atoms [53].

Despite MACE models being slightly more accurate than GAP or ACE potential, they might still not be a viable option for many due to GPU requirement. Additionally, the foundational models do not live up to the accuracy expectations laid down by atom-specific potentials, or at least not just yet. In light of these, it is valid to stick to CPU-friendly potentials, as done in this work, since the accuracy loss is minimal. However, MACE models are surely an option to explore in the future as GPU units become more widely accessible.

4.5. General workflow for future transferability testing

One of the first questions that should be asked before running MD simulations is what potential to choose. The answer is usually straightforward when one works with a known system, included in parametrization or training data set of the potential. When it comes to novel polymorphs, it becomes more complicated. In such a case, the choice is usually based on: the informed chemical intuition (facilitated by a literature review), the efficiency of the potential or the novelty of the potential, and the presumed superiority. Recently, with the increased ease of training ML-driven MD potentials, one might also be tempted to create their own potential. While all of these choices are valid if supported by right arguments, a more quantitative way to assess the potential suitability to run simulations for a given structures is highly valuable. It makes extended computational research more impactful and ensures the connection between high-level theory, simulations, and chemical reality. Although the work presented in this paper focuses specifically on the transferability of carbon potentials, the workflow presented here is applicable to other novel materials whose dynamics are of interest. It consists of the following stages:

- Single frame energy and forces analysis.
- Development of benchmark potential based on short AIMD simulations with small systems.
- Comparison of long MD simulations for larger systems obtained with readily available potential and developed benchmark.

The first stage can easily eliminate unsuitable potentials with little computational costs. In some cases, if the results are extremely conclusive, it might be the only stage needed.

The development of a benchmark potential allows access to long simulations on large systems with the accuracy of AIMD. CASTEP with the on-the-fly model [26] integrates the whole process, making it accessible. The conditions chosen for developing these should match exactly the conditions employed in the final stage of this workflow (ensemble, temperature ranges, pressure ranges, etc). Preferably, they should also match the conditions of interest in the actual computational investigation.

Finally, a comparison between long MD simulations should lead to a comprehensive understanding of potentials suitability and their limitations. The required data analysis might vary from case to case,

but it can involve looking at: coordination change, ring statistics, angle distribution, potential energy change, RDF, and forces distribution. It is worth noting that many-particle correlation functions might be preferred as averaged quantities, like RDFs or even coordination statistic can be misleading—it highly depends on the system of interest. It is also recommended to take into account the computational cost, as it can vary substantially from potential to potential [25].

Following the above workflow should provide a clearer answer to the question of what available potential to choose (if any). Finding out in the early stage of computational research that the method employed is nonphysical or disproportionately expensive to accuracy increase can save time, effort, and misleading results from being published.

5. Conclusions

This paper presented a transferability test of MD potentials with a case study of diaphites—a novel polymorph of carbon, not included in any parametrization or training data. The benchmarking consisted of three stages. The simplified stage involved the comparison of energy and forces for a single snapshot and DFT-derived single-point calculations. Together with a thorough literature review and contextualization of the results, it eliminated five potentials from further testing, leaving three promising candidates. To perform extended molecular simulations on supercells with >1000 atoms, the benchmark ML-AIMD potential was developed in a way that replicates AIMD results with high accuracy. The extended testing concluded that ACE is the most transferable carbon potential across all temperatures and can be used for the simulation of diaphites. Although GAP-20 achieves similar accuracy, it does so at much higher computational cost. LCBOP-I achieved slightly lower accuracy in final energy estimation and average 3-coordinated atom content throughout the simulation. However, it performed slightly better than the other two models in simulations at lower temperatures. It is therefore a computationally efficient option when it comes to simulations below 2000 K.

Similar accuracies of all three potentials prove that the preliminary step, involving a single snapshot comparison, is an effective way to filter more suitable potentials. The extended stage, enabled by Accelerated AIMD with on-the-fly GAP model, allows for determination of the most transferable potential and a cheaper, yet accurate, alternative which, without this structured process, could be discarded. In general, the presented framework has been shown to be successful in determining the most transferable model and can be applied to other novel systems.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Supplementary Data available at https://doi.org/10.1088/1361-651X/ae3e04/data1.

Acknowledgments

Special thanks to The Deringer Group, especially Zakariya El-Machachi and Daniel Thomas de Toit, for sharing their expertise and support. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work https://doi.org/10.5281/zenodo.22558.

Funding

Funding from the EPSRC Centre for Doctoral Training in Inorganic Chemistry for Future Manufacturing (OxICFM), EP/S023828/1.

Conflict of interest

Authors have no conflict of interest to declare.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used ChatGPT (OpenAI) in order to improve the clarity, readability, and phrasing of some part of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Author contributions

Zuzanna Malinowska-Trzmielak 0009-0003-0929-0294
Formal analysis (equal), Investigation (equal), Methodology (equal), Visualization (equal), Writing – original draft (equal)

Nicole Grobert 0000-0002-8499-8749
Funding acquisition (equal), Supervision (equal), Writing – review & editing (equal)

Mark Wilson 0000-0003-4599-7943
Conceptualization (equal), Funding acquisition (equal), Resources (equal), Supervision (equal), Writing – review & editing (equal)

References

[1] Németh P *et al* 2020 *Nano Lett.* **20** 3611–9
[2] Németh P, Garvie L A J and Salzmann C G 2023 *Phil. Trans. R. Soc. A* **381** 20220344
[3] Luo K *et al* 2022 *Nature* **607** 486–91
[4] Zou Y *et al* 2026 *J. Mater. Sci. Technol.* **240** 80–86
[5] Ge Y *et al* 2022 *Mater. Today Phys.* **23** 100630
[6] Li B, Luo K, Ge Y, Zhang Y, Tong K, Liu B, Yang G, Zhao Z, Xu B and Tian Y 2023 *Carbon* **203** 357–62
[7] Németh P, McColl K, Garvie L A J, Salzmann C G, Murri M and McMillan P F 2020 *Nat. Mater.* **19** 1126–31
[8] Zhai Z, Zhang C, Chen B, Xiong Y, Liang Y, Liu L, Yang B, Yang N, Jiang X and Huang N 2024 *Adv. Funct. Mater.* **n/a** 2401949–60
[9] Prentice J C A *et al* 2020 *J. Chem. Phys.* **152** 174111
[10] Pan J 2021 *Nat. Comput. Sci.* **1** 95–95
[11] Batatia I, *et al* 2024 A foundation model for atomistic materials chemistry (arXiv:2401.00096) [physics]
[12] Kovács D P *et al* 2025 MACE-OFF: Transferable Short Range Machine Learning Force Fields for Organic Molecules (arXiv:2312.15211) [physics]
[13] Li B, Liu B, Luo K, Tong K, Zhao Z and Tian Y 2024 *Acc. Mater. Res.* **5** 614–24
[14] Larsen A H *et al* 2017 *J. Phys.: Condens. Matter* **29** 273002
[15] Pandey K C 1982 *Phys. Rev. B* **25** 4338–41
[16] Clark S J, Segall M D, Pickard C J, Hasnip P J, Probert M I J, Refson K and Payne M C 2005 *Z. Kristallogr. Cryst. Mater.* **220** 567–70
[17] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865–8
[18] Grimme S 2006 *J. Comput. Chem.* **27** 1787–99
[19] Thompson A P *et al* 2022 *Comput. Phys. Commun.* **271** 108171
[20] Brenner D W, Shenderova O A, Harrison J A, Stuart S J, Ni B and Sinnott S B 2002 *J. Phys.: Condens. Matter* **14** 783
[21] Los J H and Fasolino A 2003 *Phys. Rev. B* **68** 024107
[22] Stuart S J, Tutein A B and Harrison J A 2000 *J. Chem. Phys.* **112** 6472–86
[23] O'Connor T C, Andzelm J and Robbins M O 2015 *J. Chem. Phys.* **142** 024903
[24] Rowe P, Deringer V L, Gasparotto P, Csányi G and Michaelides A 2022 *J. Chem. Phys.* **156** 159901
[25] Qamar M, Mrovec M, Lysogorskiy Y, Bochkarev A and Drautz R 2023 *J. Chem. Theory Comput.* **19** 5151–67
[26] Stenczel T K, El-Machachi Z, Liepuoniute G, Morrow J D, Bartók A P, Probert M I J, Csányi G and Deringer V L 2023 *J. Chem. Phys.* **159** 044803
[27] Li Z *et al* 2023 *Nat. Mater.* **22** 42–49
[28] Fabin 2020 https://mattermodeling.stackexchange.com/users/295/fabian, What makes PBE the most preferred functional over other GGA functionals? (available at: https://mattermodeling.stackexchange.com/questions/266/what-makes-pbe-the-most-preferred-functional-over-other-gga-functionals)
[29] Rappoport D, Crawford N R M, Furche F and Burke K 2009 *Encyclopedia of Inorganic Chemistry* (Wiley)
[30] Mardirossian N and Head-Gordon M 2017 *Mol. Phys.* **115** 2315–72
[31] Grimme S, Antony J, Ehrlich S and Krieg H 2010 *J. Chem. Phys.* **132** 154104
[32] Grimme S, Ehrlich S and Goerigk L 2011 *J. Comput. Chem.* **32** 1456–65
[33] Ergun S 1973 *Nat. Phys. Sci.* **241** 65–67
[34] Koike J, Parkin D M and Mitchell T E 1992 *Appl. Phys. Lett.* **60** 1450–2
[35] Tersoff J 1989 *Phys. Rev. B* **39** 5566–8
[36] Tersoff J 1988 *Phys. Rev. B* **37** 6991–7000
[37] Tersoff J 1994 *Phys. Rev. B* **49** 16349–52
[38] Dion M, Rydberg H, Schröder E, Langreth D C and Lundqvist B I 2004 *Phys. Rev. Lett.* **92** 246401
[39] Grau-Crespo R 2016 How to use optB88-vdW functional correctly in VASP?
[40] Rowe P, Deringer V L, Gasparotto P, Csányi G and Michaelides A 2020 *J. Chem. Phys.* **153** 034702
[41] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 *Chem. Rev.* **121** 10073–141
[42] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev. B* **87** 184115
[43] Drautz R 2019 *Phys. Rev. B* **99** 014104

[44] Bartók A P and Csányi G 2015 *Int. J. Quantum Chem.* **115** 1051–7
[45] Deringer V L and Csányi G 2017 *Phys. Rev. B* **95** 094203
[46] Ormeño F and General I J 2024 *Commun. Chem.* **7** 1–11
[47] Hale L 2016 NIST Interatomic potentials repository https://data.nist.gov/od/id/EBC9DB05EDF05B0EE043065706812DF87
[48] Batatia I, Kovács D P, Simm G N C, Ortner C and Csányi G 2023 MACE: higher order equivariant message passing neural networks for fast and accurate force fields (arXiv:2206.07697) [stat]
[49] Kovács D P, Batatia I, Arany E S and Csányi G 2023 *J. Chem. Phys.* **159** 044118
[50] Bernstein N 2024 *KR* **2** 1–10
[51] Pandey M, Fernandez M, Gentile F, Isayev O, Tropsha A, Stern A C and Cherkasov A 2022 *Nat. Mach. Intell.* **4** 211–21
[52] Choi J, Nam G, Choi J and Jung Y 2025 *JACS Au* **5** 1499–518
[53] Liu X, Zeng K, Wang Y and Zhao T 2025 A study on the fine-tuning performance of universal machine-learned interatomic potentials (U-MLIPs) (arXiv:2506.07401) [physics]