Accelerating the prediction of large carbon clusters via structure search: evaluation of machine-learning and classical potentials

Bora Karasulu, Jean-Marc Leyssale, Patrick Rowe, Cedric Weber, Carla de Tomas

Accelerating the prediction of large carbon clusters via structure search: evaluation of machine-learning and classical potentials

Bora Karasulu,[1,2,*] Jean-Marc Leyssale,[3] Patrick Rowe,[1,4] Cedric Weber,[5] and Carla de Tomas[1,*]

[1] *Happy Electron Ltd., 32 Telford Way, London W3 7XS, United Kingdom*
[2] *Dept. of Chemistry, University of Warwick, Gibbet Hill Road, CV4 7AL Coventry, United Kingdom.*
[3] *University of Bordeaux, CNRS, Bordeaux INP, ISM, UMR 5255, F-33400 Talence, France*
[4] *Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*
[5] *King's College London, Theory and Simulation of Condensed Matter, The Strand, London WC2R 2LS, United Kingdom*

(Dated: December 27, 2021)

From as small as single carbon dimers up to giant fullerenes or amorphous nanometer-sized particles, the large family of carbon nanoclusters holds a complex structural variability that increases with cluster size. Capturing this variability and predicting stable allotropes remains a challenging modelling task, crucial to advance technological applications of these materials. While small cluster sizes are traditionally investigated with first-principles methods, a comprehensive study spanning larger sizes calls for a computationally efficient alternative. Here, we combine the stochastic *ab initio* random structure search algorithm (AIRSS) with geometry optimisations based on interatomic potentials to systematically predict the structure of carbon clusters spanning a wide range of sizes. We first test the transferability and predictive capability of seven widely used carbon potentials, including classical and machine-learning potentials. Results are compared against an analogous cluster dataset generated via AIRSS combined with density functional theory optimizations. The best performing potential, GAP-20, is then employed to predict larger clusters in the nanometre scale, overcoming the computational limits of first-principles approaches. Our complete cluster dataset describes the evolution of topological properties with cluster size, capturing the complex variability of the carbon cluster family. As such, the dataset includes ordered and disordered structures, reproducing well-known clusters, like fullerenes, and predicting novel isomers.

I. INTRODUCTION

The wide family of carbon nanoclusters develops under a range of natural and experimental conditions. Small molecular carbon clusters in the shape of chains were first observed in carbon vaporization experiments in the 80s. These pioneering experiments led to the breakthrough synthesis of a new form of carbon clusters with icosahedral symmetry, the C60 Buckminster-fullerene[1], well after its theoretical prediction. The in-lab synthesis of the buckyball, naturally present in interstellar media, sparked exhaustive research on the quest for other fullerenes with smaller and larger sizes, and other possible cluster morphologies beyond fullerenes. For this a number of experimental techniques have been employed, all based on vaporization of a carbon-rich target, such as arc discharge, laser deposition, supersonic beams and chemical-vapor deposition[2–5]. These techniques allow the formation of clusters via atom-by-atom aggregation in the plasma or carbon vapor, with *in situ* characterization of the existing clusters usually performed with mass spectrometry[3]. Spectrometry allows one to infer the relative stabilities, masses and energies of the existing clusters and thus size of the cluster in terms of number of atoms per cluster, with clusters as small as dimers and carbyne-chains up to larger clusters containing thousands of atoms[2,4,5]. Dense and non-hollow nanoclusters, typically with diamond-like features, have also been observed in other conditions, such as subproducts of industrial detonations[6] and as precursors of nano-onions (concentric fullerenes) in annealing processes[7].

However, experimental characterization cannot provide insight on the clusters atomic structure, and since the number of possible isomers for a given cluster size increases with the number of atoms, the landscape becomes extremely complex. To overcome experimental limitations, first-principles theoretical studies have shed light onto the microstructure, chemistry, nucleation and stability of carbon clusters, relying on highly predictive frameworks such as density-functional theory (DFT)[8–10], tight-binding DFT[11] and quantum Monte Carlo approaches[12]. However, first-principles methods are limited by their high computational cost and thus restricted to small cluster sizes of a few tens of atoms and small datasets.

The main challenges for accurate *in silico* prediction of carbon cluster structures and their isomers at the molecular level are two-fold: (1) accurate and efficient description of the interactions between carbon atoms, and (2) generation of reliable representative structural models; notwithstanding the overwhelmingly large conformational space that requires care to reduce the computational costs of demanding first-principles calculations. On this front, progress in the development of accurate carbon interatomic potentials has enabled substantial reduction of calculation overheads. However, a remaining bottleneck towards accurate large-scale predictions employing interatomic potentials, is the need for system-

atic benchmarking to ensure the reliability of the chosen potential. Since carbon potentials are usually developed with a given carbon material in mind, their transferability to other carbon materials is not guaranteed. This becomes particularly evident in disordered carbons and high-density carbon phases, where previous benchmarking works[13–16] revealed some potentials inaccurately describe the local coordination environments, indicated by underestimated sp^3 hybridization and unlikely ring topologies, as well as unphysical behaviours at densities above graphite density.

To overcome the limitations of classical interatomic potentials, machine-learning (ML) frameworks have recently opened an alternative path for carbon potentials development. ML-based potentials generally map a set of atomic environments to the numerical values for the corresponding energies and forces. This procedure involves the training of the energy and forces based on a large and accurate quantum-mechanical reference database, typically generated with DFT, and a subsequent interpolation to predict 'new' atomic environments using one of the many available ML algorithms. Upon a successful training, ML potentials can achieve a significantly higher accuracy than comparable classical models, often approaching that of quantum-mechanical methods at a reduced computational costs[17]. To date mainly three ML algorithms have been employed to generate ML potentials for carbon: the Gaussian approximation potential (GAP)[17–19], neural networks[20,21] and adaptive Bayesian inference (FLARE)[22] models. Of these, the most promising and carbon-focused approach is the GAP framework, with the last generation (GAP-20) being successfully applied to crystalline and amorphous carbon phases[19]. However its transferability to carbon clusters of different degree of order, morphologies and densities is yet to be tested.

Typically classical interatomic potentials have been employed in combination with molecular dynamics (MD) to study well-defined cluster structures, e.g. fullerenes and nano-onions[23,24], targeting their physical properties and behaviour under external conditions without targeting the generation of a carbon cluster energy landscape neither the evolution of geometries with cluster size. Only a few works have explored how to combine the potentials with high-throughput structure searching algorithms and energy optimization methods in order to i) improve over the limited size of DFT-only studies (in terms of number of atoms per cluster and size of datasets) and ii) reduce computational costs. Cai et al.[25] combined the Brenner potential[26] (aka REBO-I) with a global optimisation algorithm to produce clusters of several sizes containing up to 71 carbon atoms. Mauney et al.[27] also selected the REBO-I potential and combined it with Monte Carlo simulated annealing, basing-hopping and minima-hopping algorithms for structure searching of clusters containing up to 99 atoms, followed by a DFT-based energy optimization. Kosimov et al.[28,29] modified the REBO-I potential and combined it with an energy minimization based on the conjugate gradient algorithm to produce planar clusters of up to 55 atoms and applied the dataset to study defective graphene flakes.

Another popular global optimisation technique which has not yet been combined with carbon interatomic potentials to achieve large cluster datasets is the *ab initio* random structure search (AIRSS)[30], which relies on high-throughput stochastic structure generation followed by geometry relaxations. Unlike the other global optimisers, AIRSS benefits from a broad and uncorrelated sampling of configuration space, as it does not require to avoid the local minima on the corresponding potential energy surface. As such, AIRSS has proven useful in predicting new stable and metastable crystalline phases and cluster configurations for a range of materials (see Refs.[30,31] and references therein). AIRSS has hitherto been used jointly with *ab initio* methods to enforce the geometry optimisation of the predicted structures. However, after being interfaced with the widely-used large-scale molecular simulator software LAMMPS[32,33], AIRSS offers the flexibility to be used in combination with any interatomic potential (either classical or machine-learning) implemented in LAMMPS to perform the geometry optimisations. Nevertheless, employing AIRSS with carbon interatomic potentials calls for a rigorous benchmarking of the existing potentials because the choice of the potential can greatly affect the carbon structure and properties, especially in highly disordered systems. In addition, no potential was developed with isolated carbon clusters in mind, such as fullerenes, cyclo[n]carbons, chains and bowl-shaped clusters. Therefore, the characteristics of such carbon materials (highly curved surfaces, dangling bonds and different surface reconstruction to graphite or diamond) pose a challenge to test the transferability of the potentials.

Here, we predict a wide carbon clusters landscape combining AIRSS with an interatomic potential-driven geometry optimisation by enforcing a robust optimisation protocol in LAMMPS based on the FIRE optimiser[34]. Within this framework, we test the performance and transferability of common classical interatomic potentials for carbon (EDIP, $\text{ReaxFF}_{\text{C2013}}$, REBO-II, LCBOP-I, AIREBO, Tersoff) and a recently developed machine-learning-based potential (GAP-20) to generate a large dataset of carbon clusters. To benchmark these potentials, we analyse a set of structural properties in the predicted clusters and compare to those of our reference AIRSS+DFT generated dataset containing clusters of up to 200 atoms. In view of the benchmark results, to showcase the prediction and high-throughput capability of our methodology, we choose the best performing potential to predict the minimum-energy structure of larger carbon nanoclusters. Predictions reveal novel carbon allotropes for clusters sizes ranging from 200 to 720 atoms. Our methodology can be extended to predict even larger clusters and, potentially, carbon nano-powders and nanofoams where the particle size (typically <100 nm) is intractable with first-principles approaches.

II. METHODS

To explore the allotropes of different carbon cluster sizes and point group symmetries, we employed AIRSS[30,31] to generate the starting cluster atomic coordinates. These starting coordinates were then subjected to a geometry optimization to achieve stable structures. In our initial search, the energy optimization was performed with plane-wave density-functional theory (referred to as AIRSS+DFT method) to establish a referential clusters dataset. In our subsequent searches, the optimization was performed with an interatomic potential (referred to as AIRSS+potential method). Seven widely used carbon potentials were employed: a recently developed machine-learning-based potential (GAP-20)[19] and six common empirical potentials (Tersoff[35], C-EDIP[36], LCBOP-I[37], ReaxFF$_{C-2013}$[38], REBO-II[39] and AIREBO[40]).

A. Input structure generation using AIRSS

For each searching method (AIRSS+DFT and AIRSS+potential) two AIRSS searching approaches were used to account for different degrees of short-range order in the cluster microstructure: random search of atomic positions within a sphere of a given radius imposing (i) only C1 symmetry constraint to obtain disordered structures and (ii) up to 24 symmetry operations to obtain higher-symmetry structures. A minimum atomic separation (MINSEP command) of 1.4 Å was imposed for each C-C pair to prevent unphysical overlap of atoms. The radius of the initial sphere is read-in by the POSAMP command. By default AIRSS uses a push algorithm, which pushes atoms further away to fulfil the minimum atomic separation condition. This means that for structures with low number of atoms, the initial sphere is filled. When the number of atoms increases, it reaches a point where there is no space within the initial sphere to fulfil the imposed minimum atomic separation. Then, the algorithm starts pushing the atoms further outside of the region defined by the initial sphere, which originates denser clusters. A low value for the initial sphere of 3 Å is selected to ensure a smooth density increase with cluster size above cluster sizes 40 to 50 atoms (note C60 fullerene radius is 3.5 Å), where the cage-like geometry is already expected to prevail. Switching on the CLUSTER flag promotes the clusterization of the structure. All searches were performed over a cluster size interval of [4, 200] atoms. The subsequent optimisation of the input structures with either DFT or interatomic potentials was performed as follows: carbon atoms were placed in the centre of a sufficiently large periodic cubic box to prevent interactions with the periodic images, i.e. 50 Å per side for all interatomic potentials and 15 Å for DFT to minimise the computational cost. To discard any unwanted effects of using a smaller box in the DFT searches, we performed additional tests as discussed in the Supplementary (Section S2). To ensure a fair comparison across the AIRSS+DFT and AIRSS+potential methods for all the seven potentials studied here, the AIRSS search-related parameters in both approaches (with and without symmetry constraints) are kept identical, see the AIRSS scripts in the Supplementary (Section S1).

B. Structure optimisation with density-functional theory

Density-functional theory calculations were performed using the Vienna Ab initio Simulation Package (VASP v. 5.4.4)[41]. We employed the projector-augmented wave (PAW)[41] method jointly with the PBE version of the Generalized Gradient Approximation exchange–correlation potentials[42]. To limit computational cost, we use a smaller plane-wave energy cut-off of 300 eV in the AIRSS searches, whereas a higher cut-off of 520 eV is used for selected structures within 0.1 eV/atom of the corresponding minimum-energy. An electronic convergence criterion of 10^{-8} eV and a Gaussian smearing factor of 0.2 eV were adopted, the VASP default settings for the fast Fourier transform (FFT) grid and optimisation of the projection operators were employed. Due to the non-periodic nature of the isolated clusters, we only considered the centre of the reciprocal space (Γ point) when integrating the Brillouin zone. This also enabled us to benefit from the faster implementation in VASP designed for the Γ-only calculations, useful to achieve a large number of structures needed for the high-throughput AIRSS searches. During the geometry optimisations atomic positions were fully relaxed until all force components were below 50 meV/Å using the conjugate gradient (CG) algorithm[43].

C. Structure optimisation with interatomic potentials

Geometry optimisation using the classical and machine-learning interatomic potentials was performed in LAMMPS[32,33]. We used the standard implementations of the original carbon Tersoff[35], the 2015 reparametrization of the ReaxFF (ReaxFF$_{C-2013}$)[38] targeting solid carbon phases, the second generation of the Brenner potential (REBO-II)[39], the first generation of the Long-range carbon bond order potential (LCBOP-I)[37], the Adaptive intermolecular reactive empirical bond order potential (AIREBO)[40]; the carbon environment-dependent interaction potential (EDIP)[36] developed by Marks, with its LAMMPS implementation available upon request to the developer. The Gaussian approximation potential for carbon (GAP-20)[19] was employed via the routines implemented in the QUIP module within LAMMPS. For all the potentials the optimisation protocol was kept identical. The protocol consists of four consecutive geometry minimisations using the CG algo-

Method	Disordered clusters	Symmetric clusters
DFT	3,500 [5,000]	10,500 [13,000]
GAP-20	22,200 [24,000]	20,300 [24,000]
Tersoff	18,300 [24,000]	12,500 [24,000]
EDIP	22,200 [24,000]	18,300 [24,000]
REBO-II	20,700 [24,000]	15,400 [24,000]
AIREBO	18,800 [24,000]	20,000 [24,000]
$\text{ReaxFF}_{\text{C}-2013}$	20,200 [24,000]	19,800 [24,000]
LCBOP-I	22,300 [24,000]	16,600 [24,000]

TABLE I. Number of relaxed structures obtained for each curated dataset by AIRSS with different methods (avoiding structure duplicates, dissociated clusters and non-converging structures). Initial number of generated structures for each method is given in brackets for reference.

rithm, with energy and force tolerance criteria of 10^{-12} eV and 10^{-8} eV/Å, respectively, and a step size of 0.2 Å. Note that this is the default LAMMPS optimisation protocol distributed with AIRSS. However, this minimisation was not sufficient and resulted in a large number of structures with large pressures (>100 MPa). To prevent this, we implemented in the LAMMPS protocol an additional energy minimisation step using the FIRE[34] damped-dynamics followed by a Hessian-free truncated Newton algorithm procedure with a timestep of 1 fs, which was run until the pressure (internal stress) converged to ≈0 Pa, ensuring a negligible pressure contribution to the enthalpy ($\leq 10^{-5}$ eV). With this approach, only a small fraction of non-converging structures were observed and subsequently removed, optimising the dataset production.

D. Final datasets

Further to the removal of non-converging structures, all datasets (AIRSS+DFT and AIRSS+potential) were also screened to identify repeated structures and structures containing dissociated clusters. To filter out the repeated structures, we performed structural similarity analysis based on the radial distribution functions (RDF) as implemented in the MATADOR code[44]. To filter out the structures with dissociated clusters, we performed a cluster size analysis (using a 1.85 Å cut-off radius to determine bonded atoms) using OVITO[45]. Table I shows the final number of unique cluster structures comprising each dataset, which were used for the subsequent characterization analysis. Note that the AIRSS+DFT datasets initially contained fewer structures due to computational cost limitations. All datasets are available online.[46]

E. Analysis of the generated clusters datasets

To characterize the clusters microstructure we implemented the analysis tools from the OVITO package[45] in an in-house Python code and computed coordination numbers within a 1.85 Å cut-off radius, RDFs and density of cohesive energies by counting number of structures with the same energy within a bin width of 0.01 eV/atom. We used the R.I.N.G.S. code[47] combined with the Franzblau algorithm[48] to compute the ring statistics for each cluster. To limit the complexity in the resulting plots and to capture the relative contributions of each of the isomers of a given cluster size within the obtained subset of isomers, we show the Boltzmann-weighted average properties (unless stated otherwise). The Boltzmann weight of each structure with respect to the minimum-energy structure of the same stoichiometry was computed using the well-known relation

$$p_i = \frac{e^{-\epsilon_i/k_B T}}{\sum_{j=1}^{M} e^{-\epsilon_j/k_B T}} \quad (1)$$

where ϵ_i and ϵ_0 are the energy of the given structure and the minimum-energy structure respectively, k_B the Boltzmann constant and T =293 K. The cohesive energy (E_c) of a cluster of size n was computed using

$$E_c = (E_{tot} - nE_{ref})/n \quad (2)$$

where E_{tot} is the total energy of a cluster, predicted by DFT or an interatomic potential, and E_{ref} is the reference energy of a single carbon atom. Given that the pressure is negligible, the E_{tot} corresponds to the enthalpy of formation of the cluster, as reported by VASP and LAMMPS. By construction $E_{ref} = 0$ for the potentials, whereas for DFT results E_{ref} is the energy of an isolated carbon atom (–1.38 eV/atom).

F. Extrapolation of the methodology to search for arbitrarily large cluster sizes

The key to apply our method to larger clusters is simply to enlarge the cubic box size to prevent interactions with periodic images, and to increase the radius within which carbon atoms are added to the simulation cell to prevent extremely dense clusters. Thus, to target cluster sizes in the C200 to C720 range, we employ box sizes of 100 Å and an initial sphere radius of 9.5 Å. As known from previous studies[49,50], in this range size the fullerene/cage-like structures with high symmetry (T_h/T_d or I_h) are expected to form. Therefore we covered this possibility by enforcing all 24 symmetry operators while generating the input structures. The AIRSS input script is included in Supplementary Section S1C. The extra searches were performed in the [200,720] interval at selected cluster sizes (see script). This resulted in additional 4350 structures of which 1800 unique structures passed the filtering criteria described above. Additional intensive searches targeting only two cluster sizes (C240 and C540 searches) resulted in 840 and 410 new structures, respectively, of which 460 and 160 unique structures passed the filtering criteria.

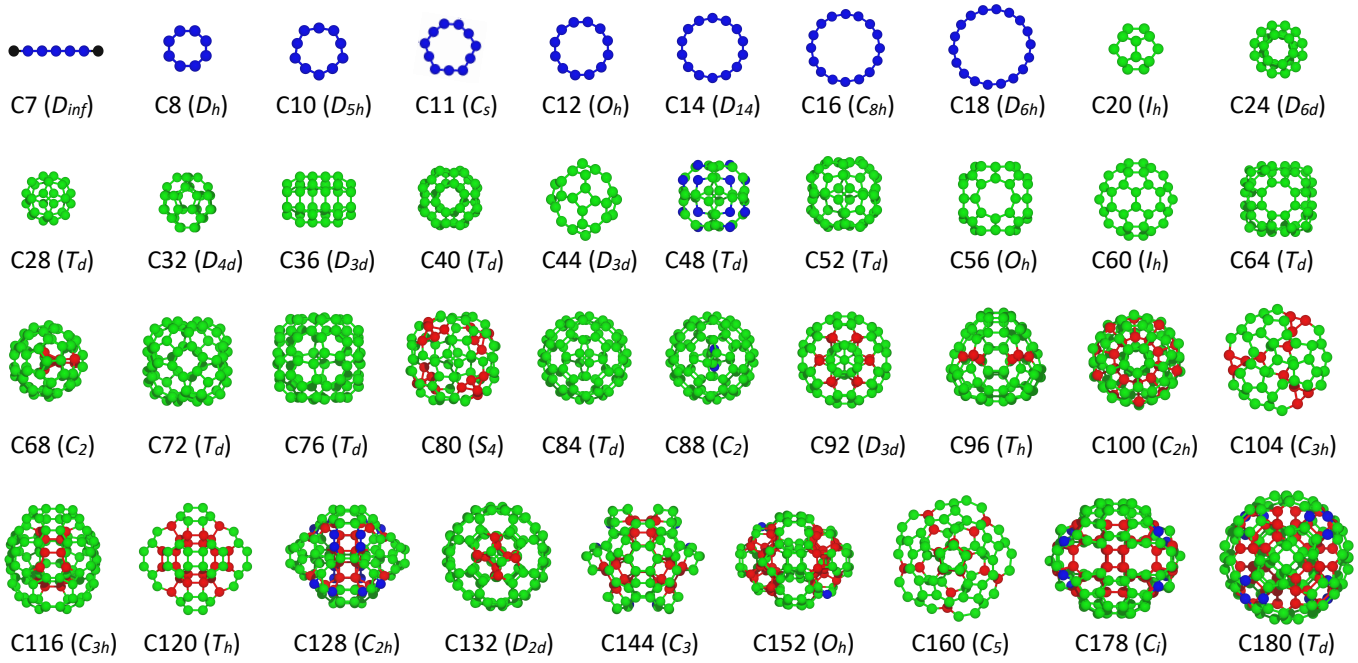

FIG. 1. Visualization of the minimum-energy structures in a selection of ordered carbon clusters generated with AIRSS+DFT. Clusters are labelled as Cn where n indicates the number of atoms. Point group symmetries are shown in parentheses. Carbon atoms are coloured according to their hybridization: blue, green and red correspond to sp, sp^2 and sp^3, respectively.

III. RESULTS AND DISCUSSION

We first present in subsection A the results of a global search using a wide cluster size range (C4 to C200); this is aimed at benchmarking the interatomic potentials against DFT. In subsection B we perform a single cluster size search on a well-known system (C60) as case study to further analyse the behaviour of the potentials compared to DFT. Finally, subsection C shows the application of the best performing potential to search for larger clusters unable to be studied with DFT.

A. Comparison of general structural trends as a function of cluster size in the range C4 to C200

To assess the performance of different interatomic potentials, we analyse energetic and structural properties of the predicted carbon clusters with AIRSS+potential and compare to the AIRSS+DFT benchmark. As detailed in the Methodology, for each of the AIRSS+potential and AIRSS+DFT searches we generate two datasets to account for symmetric and disordered structures.

1. *Minimum-Energy Structures*

Predicting minimum-energy (ME) structures in agreement with DFT-level theory is a key merit for any interatomic potential. Here we focus our discussion on the symmetric structures datasets for the sake of evaluating the point group symmetries (PGS) and cohesive energies of the ME structures obtained with the different AIRSS+potentials and compare with the AIRSS+DFT benchmark. A comparison table including cohesive energies E_c and PGS values for all the C4 to C200 ME structures is presented in the Supplementary (Table S1), while Fig. 1 collates some representative snapshots of selected cluster sizes.

Small carbon clusters typically present chain and ring morphologies competing for the most stable configuration. Upon reaching a critical mass of at least n=20-35 atoms the most stable configuration becomes a bowl-shaped or cage-like cluster, such as fullerenes and bucky-diamonds, with the C20 fullerene being the smallest closed cage-like cluster[51]. In line with this, a visual inspection of our AIRSS+DFT ME structures reveals that small clusters (C4 to C7 inclusive) form sp-hybridised carbon chains, whereas between C8 and C19, the sp-chains close into ring-like clusters, known as cyclo[n]carbons, where [n] denotes the number of atoms in the ring. At C20 we observe the expected change in morphology, obtaining the well-known smallest fullerene formed entirely by pentagons in a sp^2-hybridised spherical cage. As the cluster size increases up to C200, the structures densify and cage-like structures are the only observed ME structural form.

Our structures (Fig. 1 and Table S1) are in overall good agreement with those from Mauney et al.[27], who generated clusters up to C100 using global optimisation techniques (basin and minima hopping). In particular,

there is a good agreement in the DFT cohesive energy trends, in view of the gradual decrease from about -5 to -7.5 eV/atom with a consequent level-off upon going from C4 to C100. Mauney et al. also reported ME geometries similar to ours (Fig. 1). They showed that for the smallest carbon clusters (C4-C9) chains predominate except for C6, which presents a ring structure. For the cluster sizes C10 to C23, Mauney et al. found ring structures as the most stable form and did not predict the C20 fullerene, contrary to our AIRSS+DFT approach. As a consequence, they obtained the threshold for the transformation into cage-like structures at C23. For larger cluster sizes (Cn, n>60), we note that our structures have higher densities than corresponding fullerenes, and thus occur as filled spheres, as opposed to previous reports[27,49]. This is due to our structure generation strategy using AIRSS, which constrains atoms within a certain constant radius for the sake of simplifying the benchmark. However, AIRSS offers the flexibility to adopt diverse searching strategies, which can prioritize the generation of hollow fullerene-like structures rather than filled spheres. Varying the parameters in the searching protocol, we can recover the high-symmetry (hollow) fullerene structures of different sizes. We further discuss this in Section III B.

As for the interatomic potentials, the structures of carbon clusters with size up to C71 and their cohesive energies were previously reported using a global optimisation (GA) approach jointly with the first generation of REBO potential[25], enabling a comparison with the second generation REBO-II tested here. By inspecting the ME structures, we found a good agreement between our AIRSS+REBO-II structures and the GA-REBO structures[25], particularly considering the high symmetry structures of C20, C24, C28, C36, C50, C60 and C70 bearing the symmetry of I_h, D_{6d}, T_d, D_{6h}, D_{5h}, I_h and D_{5h}, respectively, whereas some discrepancies in the corresponding formation energies are also noted.

Among the potentials considered here, the AIRSS+GAP-20 searches gives overall the best match with the AIRSS+DFT dataset in terms of the predicted point group symmetries as well as the cohesive energies for the corresponding ME structures (see Supplementary Table S1). The only discrepancy for GAP-20 and DFT is noted for the small cluster range, where the conversion from chains into ring-like structures occurs at a somewhat larger size C10, rather than C8 in DFT. In contrast, the other potentials (except AIREBO) predict only ring-like clusters below C9. Note that for the C60 cluster size, all potentials predict the expected icosahedral fullerene as the ME structure, although discrepancies of ∼0.5 eV respect to the DFT value are observed, with GAP-20 giving the closest value to DFT. In Section III B we show additional details on intensive C60 searches to further assess the potentials performance in a well-known case study.

2. *Cohesive Energies*

Beyond minimum-energy structures, we analyse the cohesive energies of the entire structure datasets and compare the general trends in structural behaviour as a function of cluster size among potentials and DFT searches (see Fig. 2). Noteworthy is the initial decrease of cohesive energy for small cluster sizes up to ca. C40, followed by a levelling-off for larger clusters. The latter can be attributed to the well-known low stability of the smallest clusters (mostly chain or ring-like) and is predicted by all the potentials as well as DFT, albeit with a change of energy gradient across potentials.

Another trend observed for all interatomic potentials and DFT is the mismatch between symmetric and disordered distributions in the low energy region of the distribution: the minimum-energy and lower energy structures for each cluster size in the symmetric set (red) show lower energy values than in the disordered dataset (black), in line with chemical intuition, i.e. symmetry helping stabilise structures. This is observed in a lesser extent for the Tersoff potential, where only a few cluster sizes present lower minimum-energy values in the symmetric datasets.

Moreover, there is a wider energy distribution per cluster size in the symmetric sets (red datapoints) than in the disordered ones (black), as the spread datapoints indicate in Fig. 2. This is expected to follow from our methodology for creating the initial geometries. For the disordered clusters dataset, the constraint to place atoms in a 3 Å-radius sphere results in efficiently packed structures due to the lack of short-range order. However, this constraint results in an artificially large number of high-energy structures when symmetry operations need to be satisfied. This effect is most evident in the broadest distribution of ReaxFF$_{C2013}$, followed by AIREBO and Tersoff, indicating a poorer performance as compared to

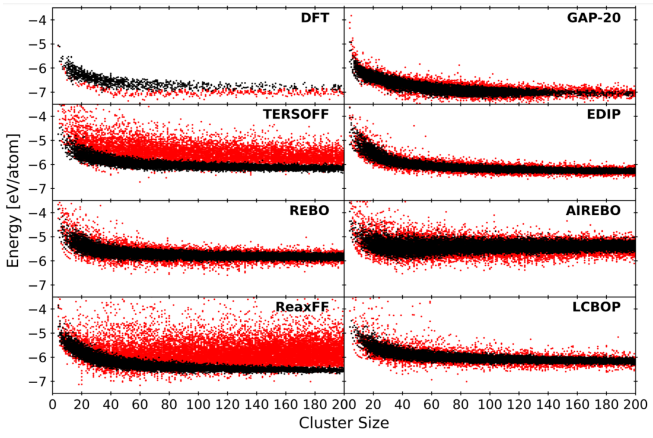

FIG. 2. Cohesive energy distributions of the predicted symmetric (red) and disordered (black) clusters with cluster size. Panels show the results given by the seven interatomic potentials tested in this work and the DFT benchmark.

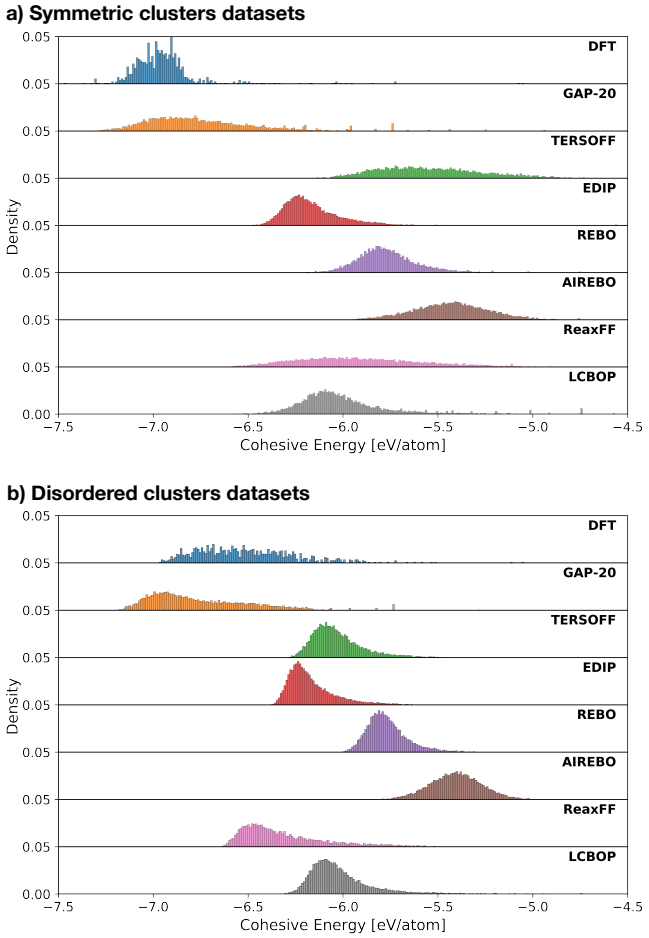

FIG. 3. Cohesive energy histogram for the (a) symmetric and (b) disordered structure datasets computed with different interatomic potentials and DFT considering the whole cluster size range [C4, C200].

GAP-20, EDIP and LCBOP-I. Beyond the width of the distribution, GAP-20 is the only potential that provides the lower energy values for the symmetric and disordered datasets, comparable to the DFT results and the graphite cohesive energy (-7.5 eV/atom). Note that the agreement is achieved despite GAP-20 was not trained using the same type of DFT theory.

For an alternative perspective on cohesive energies, Fig. 3 shows the histogram of cohesive energies for the symmetric and disordered datasets computed with AIRSS+potential and AIRSS+DFT for the whole cluster size range [C4,C200]. The plots reveal the high overlap between the DFT and GAP-20 energy distributions for both symmetric and disordered datasets (-7.5 to -6.0 eV/atom regions). In contrast, the classic potentials give cohesive energies in a higher regime (-6.5 to -4.5 eV/atom), suggesting a tendency to overestimate the clusters cohesive energies with respect to DFT.

3. *Coordination fractions*

Local coordination environments are commonly analysed to benchmark carbon interatomic potentials since it provides insight into the bond-making and -breaking capacity of the potentials[13,15]. To analyse the whole datasets beyond the minimum-energy structures, we compute the Boltzmann factors at 293 K based on the energies of individual structures for a given cluster size, relative to the associated minimum-energy state. The resulting Boltzmann-weighted (BW) ratios of sp, sp^2 and sp^3-hybridised carbons for the symmetric and disordered datasets are shown in Fig. 4. For reference, the raw data without averaging is included in the Supplementary Fig. S3. As a general note for all AIRSS+DFT and AIRSS+potentials results, the disordered clusters datasets (Fig. 4b) show average coordination fractions with smoother fluctuations than the symmetric datasets (Fig. 4a), despite having overall similar trends. This is due to less variability in the allotropes of a given cluster size in the disordered dataset, contrary to the higher symmetry case, where atoms are more spread due to symmetry constraints.

The AIRSS+DFT searches predict predominantly sp-hybridised clusters up to C20 in the symmetric and disordered dataset. This characteristic arises from the predominance of chain-like and ring-like clusters and agrees with previous DFT studies[8,27]. Additionally, in both datasets the sp and sp^2 curves intersect at C20 and clusters become predominantly sp^2-bonded with less than 20% sp^3 contribution in the C20-C200 range. Only for the largest clusters, above ca. 120 atoms in both datasets, the sp^3 fraction presents a small increase, due to the densification of the clusters. This densification is enforced by the choice of AIRSS parameters during the input structure generation: an increasing number of carbon atoms need to fit within the same confined space defined by our initial spherical domain for coordinate generation, which originates denser clusters. Note that the initial conditions can be tuned to target particular cluster densities. The sp^3 increase is more acute in the disordered dataset and correlates well with the experimentally observed linear increase of sp^3 fraction with density in bulk amorphous carbons[52]. Importantly, the high-density regime poses a challenge for the interatomic potentials since most potentials tend to underestimate sp^3 fractions in dense carbon phases[13,15].

Among the seven potentials employed in the AIRSS+potential searches, GAP-20, ReaxFF$_{C2013}$ and EDIP provide the best overall match to the AIRSS+DFT results for both symmetric and disordered datasets in all cluster size ranges. However over C100 the three potentials show small differences: GAP-20 shows a slightly larger sp^3 fraction for disordered clusters; while the DFT sp^2 fractions present a smooth decreasing trend with increasing cluster size, ReaxFF$_{C2013}$ and EDIP show sp^2 fractions plateauing at around 65% and 75% respectively in the disordered datasets, and around 70% and 80% re-

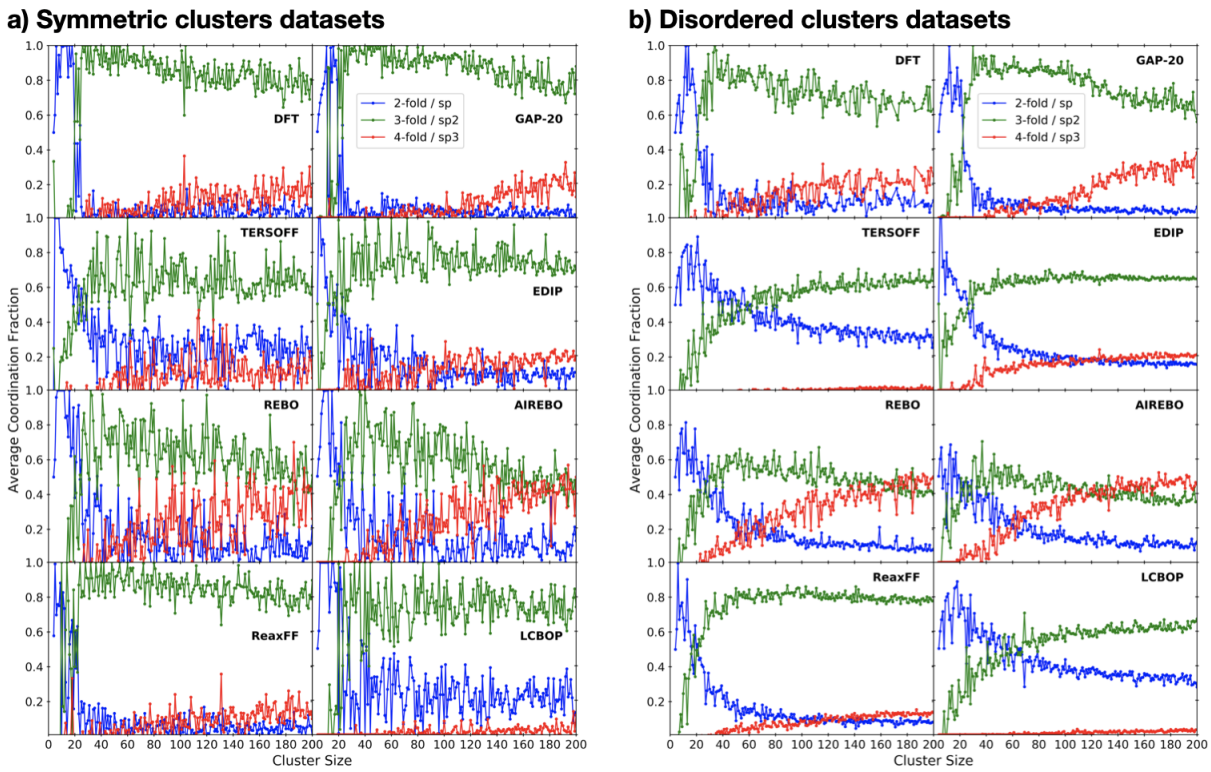

FIG. 4. Boltzmann-weighted coordination fractions as a function of cluster size for the datasets generated using different interatomic potentials and DFT searches. Panels show the symmetric (a) and disordered (b) clusters datasets.

spectively in the symmetric datasets. In contrast, the other potentials, viz. TERSOFF, REBO-II, AIREBO, and LCBOP-I, significantly deviate from the DFT benchmark in both symmetric and disordered clusters datasets. In particular, TERSOFF and LCBOP-I behave similarly and favour a high *sp* content (around 35%) for clusters between C30 and C200, comparable to the value obtained by Alonso et al.[53] at room temperature for bulk disordered carbon of density 1 g/cm^3. However this high *sp* fraction is not comparable to amorphous carbons with higher densities, as it is well-known that the *sp* fraction decreases with increasing density[13,53]. The origin of the high *sp* content in the TERSOFF and LCBOP-I clusters is revealed after visual inspection of minimum-energy structures, which contain highly *sp*-hybridised surface shells.

The REBO family of potentials (REBO-II and AIREBO) displays a different behaviour to the others. Both potentials produce disordered and symmetric clusters with a significantly lower sp^2 and higher sp^3 content with predominance of sp^3 as clusters densify, well-above the DFT benchmark. De Tomas et al.[13] also observed density-related issues with REBO-II, which produced unrealistically compressed graphitic bulk structures at densities well above graphite density. In the present case of clusters, this behaviour may be amplified by our choice of cluster generation methodology. The choice of a fixed radius while adding atoms during the structure generation in AIRSS, leads to a high density/close packing in the resulting carbon clusters, particularly over C100. While most of the interatomic potentials and DFT can handle these high-density input structures upon geometry optimisation and thus relax to less dense clusters, REBO-II and AIREBO fail at this and cannot drive away from these initial structures due to getting trapped in deep potential wells, and they hence converge to local minima even though we enforce zero internal pressure structures. Thus, to employ a potential of the REBO family in low-density and large cluster structure searches, we suggest an alternative structure generation procedure by setting a variable radius based on the cluster size to ensure a fixed target density in the inputs.

Computing the average coordination numbers for the symmetric and disordered datasets further supports the overall good agreement of GAP-20, EDIP and ReaxFF$_{\mathrm{C2013}}$ with DFT for both datasets (see Supplementary Fig. S4). In particular, GAP-20 performs relatively better than EDIP and ReaxFF$_{\mathrm{C2013}}$ in capturing the increase in coordination number in the small cluster size regime, i.e. C4-C30, where the *sp* to sp^2 conversion occurs. In contrast, REBO-II and AIREBO tend to slightly overestimate the mean coordination numbers for the larger clusters, in line with the dominant sp^3 character, while TERSOFF and LCBOP-I underestimate coordination numbers for the disordered clusters in line with the observed high *sp* fractions.

4. Ring Statistics

Characterizing carbon structures using ring statistics provides complementary information to the local coordination environments. To further assess the performance of different potentials against DFT, we computed the BW fractions of common carbon rings occurring within the clusters in the symmetric and disordered datasets. In the Supplementary Figs. S5a and S5b we show the fractions for the most relevant three- to ten-member rings (R3-R10) for each potential. To facilitate comparison with each potential, AIRSS+DFT results are shown in blue shade background in each panel.

The ring profiles of AIRSS+DFT clusters show significant differences between the symmetric and disordered datasets. Symmetric clusters above C20 contain predominantly hexagons (R6) and pentagons (R5), which drive positive curvature, resulting in cage-like clusters[54]. This behaviour is broadly well reproduced by all the potentials. Smaller clusters (below C20) only contain a small fraction of triangles and isolated peaks are observed for the ring size at the cluster size where DFT predicts stable cyclocarbons (see Fig.1), such as the peaks at C6, C8 and C10 clusters in the R6, R8 and R10 profiles respectively. Note that no peak is observed in the heptagons profile at C7, since the predicted minimum-energy structure is a chain-like cluster. The small symmetric clusters regime pose a challenge to all the potentials, since none of them provide an overall good agreement with DFT for all ring sizes. Disordered DFT clusters also show a change in ring profiles versus cluster size below and above C20. The smallest clusters present a predominance of triangles, which is broadly captured only by GAP-20, EDIP and AIREBO. However, EDIP, AIREBO, REBO-II and ReaxFF show a significant contribution of pentagons, which is absent in the DFT profile. Larger DFT clusters contain all types of rings, typical of disordered carbons, with most ring profiles given by the potentials broadly matching the DFT benchmark.

	Minimum-energy	2^{nd} lowest	3^{rd} lowest
DFT	-7.47 (I_h)	-7.23 (D_{5h})	-7.18 (D_{5d})
GAP-20	-7.57 (I_h)	-7.50 (D_5)	-7.45 (C_{6h})
Tersoff	-6.73 (I_h)	-6.66 (C_{6h})	-6.66 (D_{6h})
EDIP	-6.56 (I_h)	-6.54 (D_5)	-6.52 (C_{5v})
REBO-II	-6.84 (I_h)	-6.76 (C_{6h})	-6.74 (D_{6h})
AIREBO	-6.81 (I_h)	-6.77 (C_s)	-6.76 (D_{6d})
ReaxFF$_{C2013}$	-7.17 (I_h)	-7.12 (C_2)	-7.10 (D_5)
LCBOP-I	-6.93 (I_h)	-6.86 (C_{6h})	-6.66 (D_{4h})

TABLE II. Cohesive energy in (eV/atom) and point group symmetry of the three lowest-energy C60 clusters predicted by different interatomic potentials and compared to the reference DFT calculations.

B. Case study: C60 clusters

Having compared the general trends in the cluster properties as a function of cluster size, we now target a single cluster size to further test the AIRSS+potentials performance on a well-known system. For this we chose the most widely studied subfamily of carbon clusters, that is, clusters comprising 60 atoms. The ground-state of all possible C60 isomers is the Buckminster-fullerene and has been the object of exhaustive experimental and theoretical research for the last 30 years. The C60 fullerene is formed by 12 pentagonal (R5) and 20 hexagonal (R6) rings arranged into a hollow cage with icosahedral symmetry (I_h).

Analysis of the minimum-energy of C60 structures found in our wide cluster size (C4 to C200) searches revealed that REBO-II, AIREBO and GAP-20 did not include the C60-(I_h) fullerene. This can be due to either a lack of enough statistics for the particular cluster size or to higher density C60 clusters. Therefore, to enable fair comparison across all the potentials and discard statistical and density-related effects, we perform additional AIRSS+potentials searches targeting only C60 clusters. To target the particular fullerene morphology with optimised search efficiency, we adopt coordination constraints to rapidly locate the hollow icosahedral C60: atoms are not allowed to enter a spherical exclusion zone of radius 3.0 Å defined within the initial sphere while imposing 24 high-symmetry operations, a minimum bond angle of 91° and coordination number of 3, to avoid 4-fold rings (see script in Supplementary S1B). Since large datasets ensure better statistics in the AIRSS protocol, we then combine the structures produced in these new highly targeted searches (low density structures) with the C60 structures of the initial symmetric dataset. This way the dataset for C60 clusters spans a wide range of densities.

Analysis of the extended dataset reveals that now all of the AIRSS+potentials and AIRSS+DFT searches successfully predict the icosahedral Buckminster-fullerene as the minimum-energy structure (see Table II); however some discrepancies are observed in the formation energies values, with GAP-20 giving the closest value to the DFT benchmark, followed by ReaxFF$_{C2013}$. The energy of the other two most stable C60 isomers (2nd and 3rd lowest-energy structures) are also reported in Table II along with their corresponding symmetry group. For these structures again GAP-20, followed by ReaxFF$_{C2013}$, provides the closest energy values to the DFT reference, however the structures symmetry is widely varied, with no potential reproducing the DFT results. We note that the disagreement between the empirical models and DFT/GAP-20 may be affected by the choice of the energy of isolated atoms as reference, which would not be as prevalent if computing the cohesive energy relative to a condensed-phase state, such as graphite[19].

Figure 5a shows the radial distribution function predicted by each potential and DFT for the C60-(I_h)

minimum-energy structure. The DFT curve displays twin peaks at the first neighbours C-C distances of 1.41 Å and 1.46 Å, corresponding to the two bond lengths (pentagon-hexagon and hexagon-hexagon) featured in the fullerene structure. While most potentials reproduce the characteristic twin peaks, EDIP is the only potential that do not differentiate both bond lengths. This may be due to the bond-order term of EDIP's functional form being atom-centred, unlike the other classical potentials tested here where the bond-order term is bond-centred. This behaviour arises from the original focus on amorphous carbon when EDIP was developed, and may pose some transferability issues when applying it to isolated fullerene-like clusters. In contrast, GAP-20 provides a nearly identical radial distribution function to the DFT function; AIREBO and LCBOP closely match the DFT function while the rest of the potentials present an overall right shift of the functions, indicating slightly larger bond lengths. Further analysis of bond lengths is included in Supplementary Fig. S6. Beyond the minimum-energy fullerene structure, analysis of the whole C60 isomers datasets is presented in Fig. 5b, which shows the energy density distribution of the dataset as a function of the cluster cohesive energy. GAP-20, ReaxFF$_{C2013}$, REBO-II and LCBOP-I give an overall good agreement in the distributions widths and shapes compared to the DFT reference. However, the classical potentials present a right shift in the energy values with respect to the DFT benchmark and only GAP-20 gives the highest overlap. A similar good agreement among GAP-20 and DFT-level of theory was obtain by Aghajamali and Karton[55] in a study of isomerisation energies using a dataset of 1,812 C60 isomers.

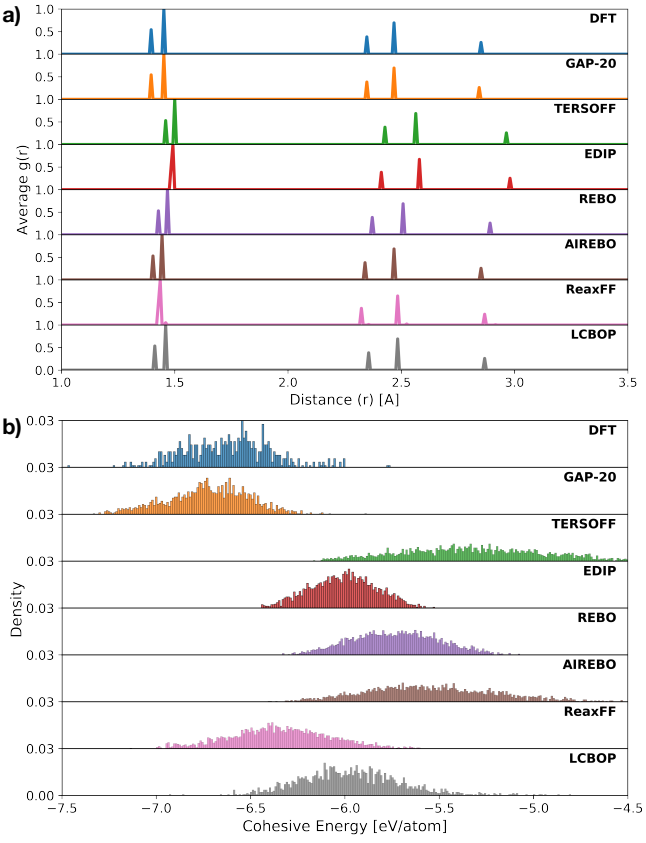

FIG. 5. Comparison of (a) radial distribution functions of the C60 (I_h) fullerene cluster and (b) energy distribution of the whole family of C60 isomers, as predicted by the potentials and DFT.

C. Applying the AIRSS+GAP-20 method to predict large clusters

To showcase the predictive capabilities of our AIRSS+potential approach, we select the best performing potential and run additional searches to find the low-energy structures of large carbon clusters, inaccessible with AIRSS+DFT due to the associated computational costs. According to our benchmarking results, GAP-20 gives the closest agreement with DFT over the cluster size range C4 to C200. Therefore, here we employ the AIRSS+GAP-20 method to target larger cluster sizes above C200. Computational details and key aspects to extrapolate our method to large structures are given in the Methodology. Here we note that the AIRSS+GAP-20 structure searches are not exhaustive and the dataset contains a limited number of structures. This is due to our intention to just give a flavour of the capabilities of our AIRSS+GAP-20 approach. A detailed study of the larger carbon clusters calls for a dedicated future work applying the methodology presented here.

1. *General trends observed for varying cluster sizes with high symmetry*

As a first predictive test, we expand our cluster size range above n=200 up to n=720 atoms and look at the evolution of apparent ground-state with increasing n within our initial searching constraints; i.e. increasing the cluster densities. Selected minimum-energy structures resulting from the general search are collated in Fig. 6. Visual inspection reveals that the degree of sp^3 hybridization increases with increasing cluster size, associated to the densification of the clusters. This behaviour is quantitatively shown in Supplementary Fig. S7c, where the computed coordination fractions indicate predominantly sp^2-bonded clusters up to ca. C320, correlated with hollow cage structures with spheroidal morphologies and octahedral/tetrahedral symmetries. This behaviour resembles that of the less dense icosahedral giant fullerenes occurring at specific cluster sizes, such as C240. Above C320 sp^3 fractions steadily increase and correlate with non-hollow cage-like structures. To accommodate a denser structure, external sp^2 shells with faceted octagonal morphologies seem to be preferred in order to encapsulate a predominantly sp^3 carbon core,

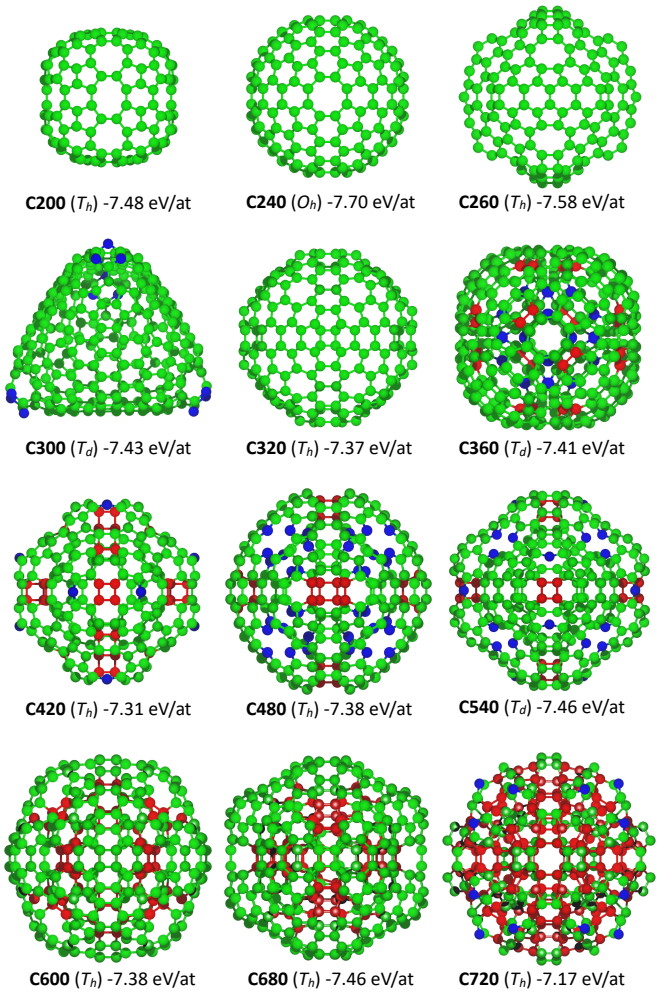

FIG. 6. Minimum-energy structures for large symmetric carbon clusters (Cn, n=atoms) of increasing density as predicted by AIRSS+GAP-20. Point group symmetries and cohesive energy per atom are are indicated. Carbon atoms are coloured by coordination: blue, green, and red spheres correspond to sp, sp^2 and sp^3-hybridised atoms, respectively.

similar to what has been experimentally observed in the thermal decomposition of nano-diamonds to carbon onions[56]. Furthermore, these structures seem to combine two behaviours independently observed in large carbon clusters: first the encapsulation capacity as observed in endohedral fullerenes able to encapsulate other clusters and chemical species[57], and second the faceted external shell as observed in large hollow clusters, where it is still under debate if alternative faceted pseudo-spherical morphology are more stable than the canonical icosahedral cages[58].

An alternative setup in the searches would be fixing the clusters density and allowing the volume to increase with the number of atoms. This would allow to explore other morphologies for a particular density. In particular, to target only large fullerene/cage-like hollow clusters as well as carbon onions, we suggest to adopt an dynamic initial sphere, so that its radius increases with the increasing cluster size, but this is beyond the scope of this work. A further aspect worth noting is the large variety of the predicted structures, including spherical and multi-faceted clusters as well as filled vs. hollow clusters, hinting at the advantage of using AIRSS over other methods for structure predictions due to the more balanced exploration of the topological search space.

2. *Structural Properties of C240 and C540 clusters*

Now we focus on two carbon clusters, viz. C240 and C540, selected after the availability of previous experimental and computational studies addressing their structure[23,49,59]. In particular, these clusters were shown by electron microscopy measurements to be building blocks of larger clusters or carbon onions[49] and the icosahedral (I_h) structures are presumed to be the ground-state for both C240 and C540. However, due to the large cluster size, multiple local minima are expected with energies close to the ground-state.

To elaborate on the structural variety of the C240 and C540 allotropes, we perform more comprehensive searches where clusters are generated with the symmetry of a randomly chosen point group with between two and 24 symmetry operations and without exclusion zone imposed in the center of the initial sphere. This way we account for both low and high symmetry structures without enforcing hollow icosahedral structures. Again, we note here that these structure searches are not exhaustive, and a fully-fledged computational study is thus needed to cover the vast allotropic space corresponding to these particular cluster sizes. By inspecting the predicted C240 clusters (Fig. 7a and Supplementary S11), the large topological variety becomes evident, which can be grouped into, e.g., high vs. low density, (quasi-)spherical vs. faceted, disordered vs. ordered, and classified by predominant hybridization (sp^3 vs. sp^2). The structural variety is accompanied by a large spread of formation energies. In this lot of structures, the fullerene-like (hollow) octahedron (O_h) geometry is predicted to be the lowest-energy one, with considerably lower cohesive energy (–7.57 eV/atom). In this structure, immediate C-C bond distances are predicted as 1.40, 1.44 and 1.49 Å, depending on the location of a given C-C bond (whether shared among pentagon-hexagon, hexagon-hexagon or hexagon-octagon) and the fullerenic radius is 7.2 Å.

Moreover, the AIRSS+GAP-20 searches also give some C240-I_h pseudo-fullerenes, presenting a hollow quasi-spherical fullerenes but containing other rings different to the canonical combination of hexagon, pentagons and heptagons. Interestingly, however, these I_h structures are predicted to have higher formation energies (-7.42 and -7.12 eV/atom) as compared to the C240-O_h structures (-7.56 eV/atom). This highlights the multiple local minima that can coexist at such large cluster sizes which hints at the need for intensive structure searches for a

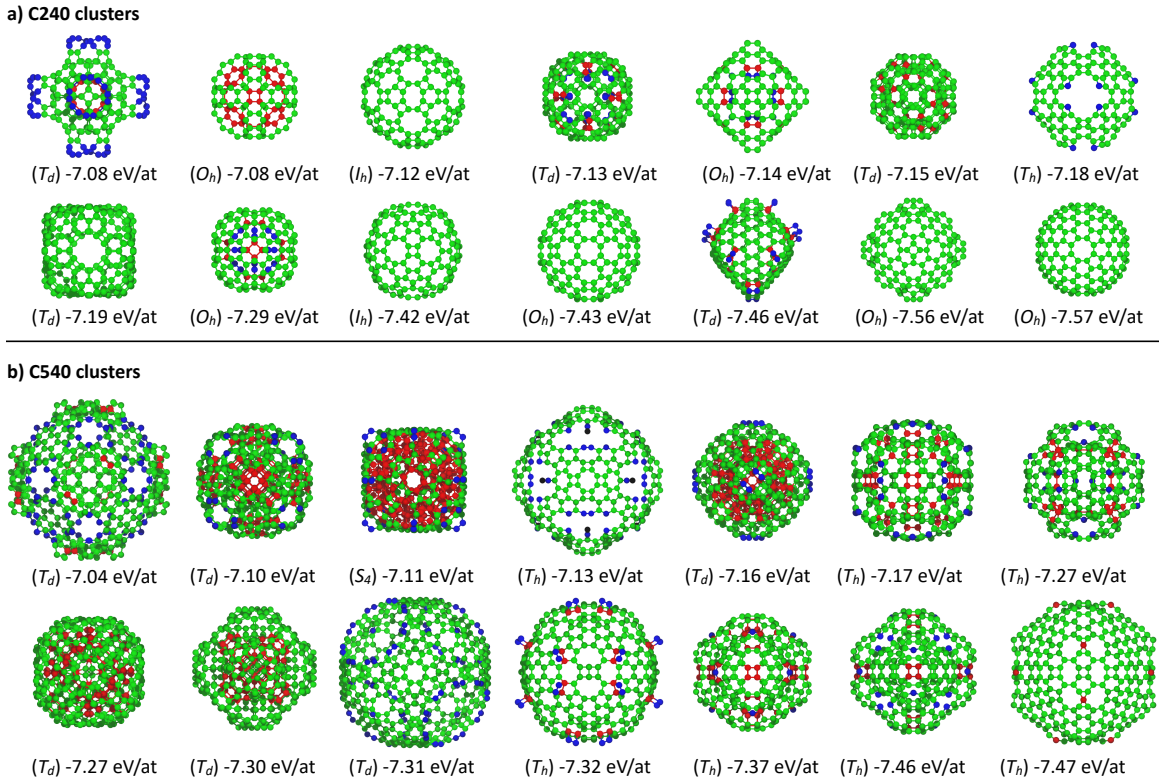

FIG. 7. Lowest-energy isomers of (a) C240 and (b) C540 clusters as predicted by AIRSS+GAP-20. Isomers are ordered by decreasing energy, with the minimum-energy structure shown on the bottom-left corner of each panel. Point group symmetries and cohesive energies per atom are indicated for each structure. Carbon atoms are colored by coordination: blue, green, and red spheres correspond to sp, sp^2 and sp^3-hybridised atoms, respectively. Additional higher-energy isomers of C240 and C540 are shown in Fig. S11

more complete understanding of the energy landscape beyond canonical fullerenes. Despite such study is out of the scope of this work, for the sake of completeness we performed an additional search targeting only hollow C240 clusters, i.e low-density, using an analogous searching protocol as described in Section II-B with a larger initial sphere (8 Å radius) and larger exclusion zone within the initial sphere (6 Å) with the aim to obtain the canonical C240 fullerene. A low-cost search with only 1000 isomers gives the canonical C240 (I_h) fullerene 8 times as the minimum-energy structure (-7.81 eV/atom), followed by hollow pseudo-fullerenes with O_h symmetry, similar to those found in the general search shown in Fig. 7-a, comprising energies between -7.77 and -7.57 eV/atom.

As for the C540 clusters (Fig. 7b and Supplementary S11), again a wide range of structures with different characteristic features have been generated by AIRSS. In particular, the minimum-energy structure is predicted to be a fullerene-like one with tetrahedral (T_d) symmetry, with the cohesive energy of -7.47 eV/atom. Moreover, the second lowest-energy structure with a comparable cohesive energy (within 10 meV/atom) is also of T_d symmetry; interestingly, however, it has a high-density geometry (as evident from the high sp^3 content). This finding suggests that allotropes other than fullerenes may compete with the canonical fullerene morphology and geometry, in line with recent findings pointing to octahedral fullerenes with energies comparable to icosahedral fullerenes when cluster sizes surpass the hundreds of atoms[58].

As a final remark we note that, although no carbon interatomic potential is perfect, developing new potentials and improving existing ones remains an active field of research which has seen increased impact and accuracy in the last years; particularly the use of machine-learning algorithms has fueled the race to achieve the quantum accuracy provided by DFT methods. Nevertheless, the low computational costs associated to classical potentials, such as REBO or Tersoff, are still tempting respect to machine-learning potentials. Therefore, introducing quantum corrections in classical potentials are desirable to achieve more accurate energies, specially when describing frustrated systems as in amorphous carbons[60]. Although such corrections are not straightforwardly applicable into the GAP formalism, extending and refining the training datasets to specifically deal with geometrical frustration will offer a solid path towards greater accuracy in the next-generation GAP potentials.

IV. CONCLUSIONS

We have evaluated the performance of widely used classical and machine-learning carbon interatomic potentials in high-throughput structure searches. For this, we generate large datasets of symmetric and disordered carbon clusters using the stochastic structure search algorithm AIRSS for the atomic coordinates generation combined with a geometry relaxation protocol driven by the potentials. The results of the AIRSS+potentials searches are evaluated against analogous AIRSS+DFT datasets, where the geometry relaxation is performed at the DFT-theory level and used as benchmark. For the examined cases, the machine-learning GAP-20 potential gives overall the best match with the DFT results for high-density carbon clusters (i.e. condensed carbon phases) with a wide range of sizes (Cn, n=4-200), at a lower computational cost. This conclusion was drawn based on an intensive dataset and a wide range of different metrics, including structural features like symmetry group, coordination fractions, cohesive energy distributions, ring statistics, and radial distribution functions. On the other hand, being developed for condensed carbon phases, $\text{ReaxFF}_{\text{C2013}}$ provides a promising cost-efficient alternative to DFT with a moderate accuracy trade-off. Our approach combining the GAP-20 potential with AIRSS provides a computationally cost-effective alternative, overcoming the DFT limitations due to large cluster sizes and big data. As such, our approach enables the production of large data sets with large cluster sizes, required to explore the vast energy space of carbon nanoclusters. Our method is also in particular timely with the rapidly increasing availability of data-driven numerical modelling methods, with valuable applications in structure prediction of crystals, surfaces and particles at the nanoscale.

ACKNOWLEDGMENTS

We thank Nigel A. Marks for providing the LAMMPS implementation of the Carbon EDIP interatomic potential and Georg Schusteritsch for helpful discussions regarding AIRSS. B.K. acknowledges The Engineering and Physical Sciences Research Council (EPSRC) Early-Career Fellowship (Grant no: EP/T026138/1).

* Corresponding authors: bora.karasulu@warwick.ac.uk and carla.detomas@he.co

[1] H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley. C60: Buckminsterfullerene. *Nature*, 318: 162–163, 1985. doi:10.1038/318162a0.

[2] E Barborini, P Piseri, A Li Bassi, A.C. Ferrari, C.E. Bottani, and P Milani. Synthesis of carbon films with controlled nanostructure by separation of neutral clusters in supersonic beams. *Chemical Physics Letters*, 300(5):633–638, 1999. doi:10.1016/S0009-2614(98)01449-3.

[3] C Lifshitz. Carbon clusters. *International Journal of Mass Spectrometry*, 200(1):423–442, 2000. ISSN 1387-3806. doi: 10.1016/S1387-3806(00)00350-X.

[4] C. S. Casari, A. Li Bassi, L. Ravagnan, F. Siviero, C. Lenardi, P. Piseri, G. Bongiorno, C. E. Bottani, and P. Milani. Chemical and thermal stability of carbyne-like structures in cluster-assembled carbon films. *Phys. Rev. B*, 69:075422, Feb 2004. doi:10.1103/PhysRevB.69.075422.

[5] A. Hu, Q.-B. Lu, W. W. Duley, and . Rybachuk. Spectroscopic characterization of carbon chains in nanostructured tetrahedral carbon films synthesized by femtosecond pulsed laser deposition. *The Journal of Chemical Physics*, 126(15):154705, 2007. doi:10.1063/1.2727450.

[6] J.A. Hammons, M.H. Nielsen, M Bagge-Hansen, S. Bastea, W.L. Shaw, J.R.I. Lee, J. Ilavsky, N. Sinclair, K. Fezzaa, L.M. Lauderbach, R.L. Hodgin, D.A. Orlikowski, L.E. Fried, and T.M. Willey. Resolving detonation nanodiamond size evolution and morphology at sub-microsecond timescales during high-explosive detonations. *The Journal of Physical Chemistry C*, 123(31):19153–19164, 2019. doi:10.1021/acs.jpcc.9b02692.

[7] S. Tomita, A. Burian, J.C Dore, D.LeBolloch, M. Fujii, and S. Hayashi. Diamond nanoparticles to carbon onions transformation: X-ray diffraction studies. *Carbon*, 40(9): 1469–1474, 2002. doi:10.1016/S0008-6223(01)00311-6.

[8] R. O. Jones. Density functional study of carbon clusters c2n (2n16). i. structure and bonding in the neutral clusters. *The Journal of Chemical Physics*, 110(11):5189–5200, 1999. doi:10.1063/1.478414.

[9] J.I. Martínez and J.A. Alonso. An improved descriptor of cluster stability: Application to small carbon clusters. *Physical Chemistry Chemical Physics*, 20(43):27368–27374, 2018. doi:10.1039/c8cp05059g.

[10] Lan Ting Shi, Zhao Qi Wang, Cui E. Hu, Yan Cheng, Jun Zhu, and Guang Fu Ji. Possible lower energy isomer of carbon clusters Cn (n=11, 12) via particle swarm optimization algorithm: Ab initio investigation. *Chemical Physics Letters*, 721:74–85, 2019. ISSN 00092614. doi: 10.1016/j.cplett.2019.02.028.

[11] T. W. Yen and S. K. Lai. Use of density functional theory method to calculate structures of neutral carbon clusters Cn (3 n 24) and study their variability of structural forms. *Journal of Chemical Physics*, 142(8), 2015. ISSN 00219606. doi:10.1063/1.4908561. URL `http://dx.doi.org/10.1063/1.4908561`.

[12] S. Sokolova, A. Lüchow, and J.B. Anderson. Energetics of carbon clusters C20 from all-electron quantum monte carlo calculations. *Chemical Physics Letters*, 323(3):229–233, 2000. doi:10.1016/S0009-2614(00)00554-6.

[13] 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. ISSN 0008-6223. doi:10.1016/j.carbon.2016.08.024.

[14] Alireza Aghajamali, Carla de Tomas, Irene Suarez-Martinez, and Nigel A Marks. Unphysical nucleation of diamond in the extended cutoff Tersoff po-

tential. *Mol. Simul.*, 44(2):164–171, 2018. doi: 10.1080/08927022.2017.1355555.

[15] 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. doi: 10.1016/j.carbon.2019.07.074.

[16] Longqiu Li, Ming Xu, Wenping Song, Andrey Ovcharenko, Guangyu Zhang, and Ding Jia. The effect of empirical potential functions on modeling of amorphous carbon using molecular dynamics method. *Appl. Sci. Res.*, 286:287–297, 2013. doi:10.1016/j.apsusc.2013.09.073.

[17] Volker L Deringer and Gábor Csányi. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B*, 95(9):094203–15, 2017. doi: 10.1103/PhysRevB.95.094203.

[18] Patrick Rowe, Gábor Csányi, Dario Alfè, and Angelos Michaelides. Development of a machine learning potential for graphene. *Phys. Rev. B*, 97:054303, 2018. doi: 10.1103/PhysRevB.97.054303.

[19] Patrick Rowe, Volker L. Deringer, Piero Gasparotto, Gábor Csányi, and Angelos Michaelides. An accurate and transferable machine learning potential for carbon. *Journal of Chemical Physics*, 153(3), 2020. ISSN 10897690. doi:10.1063/5.0005084. URL `https://doi.org/10.1063/5.0005084`.

[20] 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. *Phys. Rev. B*, 81:100103, 2010. doi:10.1103/PhysRevB.81.100103.

[21] Pilsun Yoo, Michael Sakano, Saaketh Desai, Md Mahbubul Islam, Peilin Liao, and Alejandro Strachan. Neural network reactive force field for c, h, n, and o systems. *npj Computational Materials*, 7(9), 2021. doi:10.1038/s41524-020-00484-3.

[22] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(20), 2020. doi:10.1038/s41524-020-0283-z.

[23] J.H. Los, N. Pineau, G. Chevrot, G. Vignoles, and J.-M. Leyssale. Formation of multiwall fullerenes from nanodiamonds studied by atomistic simulations. *Phys. Rev. B*, 80: 155420, 2009. doi:10.1103/PhysRevB.80.155420.

[24] Alexander S. Sinitsa, Irina V. Lebedeva, Andrey M. Popov, and Andrey A. Knizhnik. Transformation of Amorphous Carbon Clusters to Fullerenes. *Journal of Physical Chemistry C*, 121(24):13396–13404, 2017. ISSN 19327455. doi: 10.1021/acs.jpcc.7b04030.

[25] Wensheng Cai, Nan Shao, Xueguang Shao, and Zhongxiao Pan. Structural analysis of carbon clusters by using a global optimization algorithm with Brenner potential. *Journal of Molecular Structure: THEOCHEM*, 678(1-3):113–122, 2004. ISSN 01661280. doi: 10.1016/j.theochem.2004.03.017.

[26] Donald W. Brenner. Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films. *Phys. Rev. B*, 42(15):9458–9471, 1990. doi: 10.1103/PhysRevB.42.9458.

[27] C. Mauney, M.B. Nardelli, and D. Lazzati. Formation and properties of astrophysical carbonaceous dust. I. ab-initio calculations of the configuration and binding energies of small carbon clusters. *The Astrophysical Journal*, 800(1): 30, 2015. doi:10.1088/0004-637x/800/1/30.

[28] D. P. Kosimov, A. A. Dzhurakhalov, and F. M. Peeters. Theoretical study of the stable states of small carbon clusters C n (n=2-10). *Physical Review B - Condensed Matter and Materials Physics*, 78(23):1–8, 2008. ISSN 10980121. doi:10.1103/PhysRevB.78.235433.

[29] D. P. Kosimov, A. A. Dzhurakhalov, and F. M. Peeters. Carbon clusters: From ring structures to nanographene. *Physical Review B - Condensed Matter and Materials Physics*, 81(19):1–12, 2010. ISSN 10980121. doi: 10.1103/PhysRevB.81.195414.

[30] Chris J Pickard and R J Needs. Ab initiorandom structure searching. *Journal of Physics: Condensed Matter*, 23(5): 053201, jan 2011. doi:10.1088/0953-8984/23/5/053201.

[31] C.J. Pickard. Hyperspatial optimization of structures. *Phys. Rev. B*, 99:054102, 2019. doi: 10.1103/PhysRevB.99.054102.

[32] S Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Chem. Phys.*, 117(1):1–19, 1995. ISSN 0021-9991. doi:10.1006/jcph.1995.1039.

[33] LAMMPS. `http://lammps.sandia.gov`.

[34] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch. Structural relaxation made simple. *Phys. Rev. Lett.*, 97:170201, 2006. doi: 10.1103/PhysRevLett.97.170201.

[35] J. Tersoff. Empirical interatomic potential for carbon, with applications to amorphous carbon. *Phys. Rev. Lett.*, 61 (25):2879–2882, 1988. doi:10.1103/PhysRevLett.61.2879.

[36] N. A. Marks. Generalizing the environment-dependent interaction potential for carbon. *Phys. Rev. B*, 63(3):035401, 2000. doi:10.1103/PhysRevB.63.035401.

[37] J H Los and A Fasolino. Intrinsic long-range bond-order potential for carbon: Performance in Monte Carlo simulations of graphitization. *Phys. Rev. B*, 68(2):24107, 2003. doi:10.1103/PhysRevB.68.024107.

[38] S G Srinivasan, A C T van Duin, and P Ganesh. Development of a ReaxFF potential for carbon condensed phases and its application to the thermal fragmentation of a large fullerene. *J. Phys. Chem. A*, 119(4):571–580, 2015. doi: 10.1021/jp510274e.

[39] D. W. Brenner, O. A. Shenderova, J. A. Harrison, S. J. Stuart, B. Ni, and S. B. Sinnott. A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *J. Phys.: Condens. Matter*, 14 (4):783–802, 2002. doi:10.1088/0953-8984/14/4/312.

[40] Steven J Stuart, Alan B Tutein, and Judith A Harrison. A reactive potential for hydrocarbons with intermolecular interactions. *J. Chem. Phys.*, 112(14):6472–6486, 2000. doi:10.1063/1.481208.

[41] G. Kresse and J. Hafner. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B*, 47:558–561, 1993. doi: 10.1103/PhysRevB.47.558.

[42] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.

[43] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64:1045–1097, 1992. doi:10.1103/RevModPhys.64.1045.

[44] M. Evans, MATADOR: an aggregator, manipulator and runner of first-principles calculations (2019). `http://matador.science/`.

[45] A. Stukowski. Visualization and analysis of atomistic simulation data with OVITO: the Open Visualization Tool. *Modell. Simul. Mater. Sci. Eng.*, 18:015012, 2010. doi: 10.1088/0965-0393/18/1/015012.

[46] `https://github.com/bkarasulu/Carbon-PP-Benchmark-Paper-SI.git`.

[47] Ring statistics analysis of topological networks: New approach and application to amorphous ges2 and sio2 systems. *Computational Materials Science*, 49(1):70–83, 2010. doi:10.1016/j.commatsci.2010.04.023.

[48] D. S. Franzblau. Computation of ring statistics for network models of solids. *Phys. Rev. B*, 44(10):4925–4930, Sep 1991. doi:10.1103/PhysRevB.44.4925.

[49] H. Terrones and M. Terrones. The transformation of polyhedral particles into graphitic onions. *Journal of Physics and Chemistry of Solids*, 58(11):1789–1796, 1997. ISSN 00223697. doi:10.1016/S0022-3697(97)00067-X.

[50] C. John, Cheriyacheruvakkara Owais, Anto James, and Rotti Srinivasamurthy Swathi. Swarm intelligence steers a global minima search of clusters bound on carbon nanostructures. *The Journal of Physical Chemistry C*, 125(5): 2811–2823, 2021. doi:10.1021/acs.jpcc.0c09528.

[51] Peter R. Taylor, Eric Bylaska, John H. Weare, and Ryoichi Kawai. C20: fullerene, bowl or ring? new results from coupled-cluster calculations. *Chemical Physics Letters*, 235 (5):558–563, 1995. doi:10.1016/0009-2614(95)00161-V.

[52] J. Schwan, S. Ulrich, H. Roth, H. Ehrhardt, S. R. P. Silva, Robertson J., and *et al.* Tetrahedral amorphous carbon films prepared by magnetron sputtering and dc ion plating. *J. App. Phys.*, 79(3):1416–1422, 1996. doi: 10.1063/1.360979.

[53] Lydia Alonso, Julio A. Alonso, and María J. López. Computer simulations of the structure of nanoporous carbons and higher density phases of carbon. *Many-body Approaches at Different Scales: A Tribute to Norman H. March on the Occasion of his 90th Birthday*, pages 21–34, 2018. doi:10.1007/978-3-319-72374-7˙3.

[54] Q. L. Zhang, S. C. O'Brien, J. R. Heath, Y. Liu, R. F. Curl, H. W. Kroto, and R. E. Smalley. Reactivity of large carbon clusters: spheroidal carbon shells and their possible relevance to the formation and morphology of soot. *The Journal of Physical Chemistry*, 90(4):525–528, 1986. doi: 10.1021/j100276a001.

[55] Alireza Aghajamali and Amir Karton. Can force fields developed for carbon nanomaterials describe the isomerization energies of fullerenes? *Chemical Physics Letters*, 779:138853, 2021. doi:10.1016/j.cplett.2021.138853.

[56] L. Hawelek, A. Brodka, S. Tomita, J.C. Dore, V. Honkimäki, and A. Burian. Transformation of nano-diamonds to carbon nano-onions studied by x-ray diffraction and molecular dynamics. *Diamond and Related Materials*, 20(10):1333–1339, 2011. doi: https://doi.org/10.1016/j.diamond.2011.09.008.

[57] Alexey A. Popov, Shangfeng Yang, and Lothar Dunsch. Endohedral fullerenes. *Chemical Reviews*, 113(8):5989–6113, 2013. doi:10.1021/cr300297r.

[58] Tomas Lazauskas, Alexey A. Sokol, and Scott M. Woodley. Are octahedral clusters missing on the carbon energy landscape? *Nanoscale Adv.*, 1:89–93, 2019. doi: 10.1039/C8NA00013A.

[59] V. K. Dolmatov, P. Brewer, and S. T. Manson. Photoionization of atoms confined in giant single-walled and multiwalled fullerenes. *Phys. Rev. A*, 78:013415, 2008. doi: 10.1103/PhysRevA.78.013415.

[60] Zachary Bullard, Eduardo Costa Girão, Colin Daniels, Bobby G. Sumpter, and Vincent Meunier. Quantifying energetics of topological frustration in carbon nanostructures. *Phys. Rev. B*, 89:245425, Jun 2014. doi: 10.1103/PhysRevB.89.245425.