PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2024, **26**, 22346

Transferable machine learning interatomic potential for carbon hydrogen systems†

Somayeh Faraji and Mingjie Liu *

Received 6th June 2024,
Accepted 2nd August 2024

DOI: 10.1039/d4cp02300e

rsc.li/pccp

In this study, we developed a machine learning interatomic potential based on artificial neural networks (ANN) to model carbon–hydrogen (C–H) systems. The ANN potential was trained on a dataset of C–H clusters obtained through density functional theory (DFT) calculations. Through comprehensive evaluations against DFT results, including predictions of geometries and formation energies across 0D–3D systems comprising C and C–H, as well as modeling various chemical processes, the ANN potential demonstrated exceptional accuracy and transferability. Its capability to accurately predict lattice dynamics, crucial for stability assessment in crystal structure prediction, was also verified through phonon dispersion analysis. Notably, its accuracy and computational efficiency in calculating force constants facilitated the exploration of complex energy landscapes, leading to the discovery of a novel C polymorph. These results underscore the robustness and versatility of the ANN potential, highlighting its efficacy in advancing computational materials science by conducting precise atomistic simulations on a wide range of C–H materials.

1. Introduction

Carbon (C), one of the most abundant elements in nature, displays various types of hybridized bonds, contributing to a rich energy landscape and a diverse range of properties across different structural phases.[1–3] This inherent versatility, combined with boundless possibilities of its combination with hydrogen (H) leads to a plethora of structures and chemical environments, ranging from simple hydrocarbons like CH_4 to complex organic molecules like carotenes ($C_{40}H_{56}$).[4–9] The investigation of the hydrocarbons and other C–H systems at atomistic level is crucial for understanding chemical interactions and advancing materials design, thereby, attracting significant attention from researchers. For instance, the advancements in C-based materials have revolutionized fields like hydrogen storage[10–15] and the capture of polycyclic aromatic hydrocarbons pollutants.[16–18]

Recent advancements in theoretical and computational methodologies, particularly those based on quantum mechanics (QM), such as density functional theory (DFT)[19,20] have significantly enhanced our ability to study and explore materials at the atomic scale. While QM methods provide accurate understanding into material behavior, the computational cost of them increases with the system size,[21,22] hindering their applications for exploring extensive energy landscapes or large-scale simulations. Therefore, we need a trade-off between accuracy and computational cost in modeling materials at atomic scale. Machine learning interatomic potentials (MLIPs), as computationally efficient alternatives to QM-based methods, have gained attention for their ability to capture complex atomic interactions and predict material properties with remarkable precision, enabling the exploration of extensive chemical spaces and previously inaccessible molecular dynamics (MD). There have been numerous efforts to develop MLIPs specifically for pure C.[23–34] These studies aim to improve the accuracy and transferability of the potential by training on dataset covering a broad spectrum of atomic environments and configurations, such as MD trajectories at different temperatures and pressures[23,24] or including 0D–3D systems to have diverse boundary conditions[25,26] to capture the bond diversity. Based on specific applications, ongoing efforts aim to improve MLIPs by presenting different versions. For instance, the Gaussian approximation potential (GAP)[30,31] was first developed to study the behavior of liquid and amorphous C,[32] later improved to encompass van der Waals corrections for C_{60} fullerene and nanoporous C structures,[33,34] and later ordered graphite configurations with different stacking patterns were added to its training dataset for exploring the graphitic energy landscape of C.[29]

The existence of MLIPs specifically tailored for pure C highlights the difficulty and challenges in modeling such systems. Pure C itself presents significant training challenges; incorporating H to develop accurate MLIPs for C–H systems adds further complexity. This includes effectively capturing

Department of Chemistry, University of Florida, Gainesville, FL 32611, USA. E-mail: mingjieliu@ufl.edu

† Electronic supplementary information (ESI) available: Impact of training data energy span on potential accuracy, structures and other details of the investigated 0D–3D systems, structural details and properties of the novel carbon polymorph. See DOI: https://doi.org/10.1039/d4cp02300e

bond variations and intramolecular interactions, such as hydrogen bonding. Achieving transferability of the MLIP across systems with various C/H ratios and different boundary conditions is more demanding than for pure C systems. This is due to the greater complexity and diversity of C–H compounds, necessitating the generation of a larger and more diverse training dataset. Efforts have been made to train MLIPs that include C and H for specific applications, such as predicting CH stretching modes in small molecules,[35] bond dissociation energy prediction in drug-like molecules,[36] C–C bond breaking in small molecules,[37] C–H bond activation of CH_4 on Pt(111),[38] and constructing the potential energy surface (PES) of CH_4 and study its vibrational levels.[39] Some other efforts have aimed to go beyond specific system and provide a general descriptions for all organic molecules, rather than for specific system.[40–42] However, these methods are trained for specific applications and/or lack sufficient accuracy for different systems that are not essentially close to their equilibrium state.[43]

The complexities and challenges in developing accurate MLIPs for pure C, compounded by the additional difficulties when H is incorporated into the system, highlight the lack of sufficiently accurate and transferable potentials for C–H systems. Motivated to address this gap in the current research landscape, we focused on creating an MLIP specifically tailored for C–H systems. Therefore, in this study, we present an MLIP based on artificial neural networks (ANNs) for C–H systems. To enhance the diversity of the training dataset to represent the complexities of the systems, we train the ANN potential on cluster systems, rather than including different boundary conditions. This approach provides a broader representation of the system's behavior. We demonstrate that our trained potential, solely based on cluster C–H systems, can be applied not only to C–H systems but also to pure C systems under various boundary conditions. Furthermore, our potential's accuracy and versatility enable it to be used in diverse contexts, including reactivity and lattice dynamics. The trained ANN potential is also utilized for crystal structure prediction (CSP) and has identified a novel C polymorph, showcasing the potential's practical applications.

2. Methods

2.1. Feed-forward ANN

In this study, we utilized high-dimensional ANN proposed by Behler[44,45] for potential training. Such ANNs commonly operate in a feed-forward manner, transmitting signals in one direction through the layers. The ANN structure consists of interconnected nodes linked by weights, arranged in layers including input, hidden, and output layers. Firstly, atomic coordinate representations are fed into the input layer by converting each atomic position into a set of atomic symmetry functions $\{\mathbf{G}_i\}$, describing the chemical environment of the atom. We employed radial ($\mathbf{G}^2$) and angular ($\mathbf{G}^5$) symmetry functions,[45] totaling 70 symmetry functions (16 radial and 54 angular), as parameterized in previous work.[46] For these symmetry functions, we chose a cutoff radius of 6 Å based on testing different values conducted in our work. Secondly, the appropriate level of network complexity to accurately represent the underlying physics without overfitting the data was determined by testing various number of hidden layers and node counts in each of them. In this work, we explored ANN models with two and three hidden layers, each with varying numbers of nodes. We omitted single hidden layer models due to their inadequacy in learning such a complex problem. Additionally, we did not go beyond three hidden layers due to the computational cost as well as the increased risk of overfitting. After training the ANN models, it is crucial to evaluate their performance on an unseen dataset. This helps assess their generalization capability and performance across various scenarios. Based on our evaluations, it was found that a network configuration with two hidden layers, each containing 17 nodes, reduced the root mean square error (RMSE) to below 22 meV/atom. Finally, the output layer, comprising a single node, yields the energy of atom in the system. For our ANN with two hidden layers and 17 nodes in each hidden layer and 70 symmetry function in the input layer, the total energy of the atom is obtained as

$$E = f\left(b_1^3 + \sum_{k=1}^{17} a_{k1}^{23} \cdot f\left(b_k^2 + \sum_{j=1}^{17} a_{jk}^{12} \cdot f\left(b_j^1 + \sum_{i=1}^{70} G_i \cdot a_{ij}^{01}\right)\right)\right) \quad (1)$$

where f is the activation function, b_i^j are the bias weights. Each node i in each layer k is connected to the nodes j in the next layer $l = k + 1$ by weights a_{ij}^{kl}.[45] The total energy (E_{tot}) of the system is the collective sum of these atomic energies, each computed *via* an individual ANN process. The force F acting on each atom is subsequently computed from the negative gradients of the total energy with respect to its atomic coordinates according to $F_n = -\nabla_n E_{tot}$ ($n = x, y, z$). In this scenario, a direct relationship is absent, attributed to the conversion of atomic Cartesian coordinates into the symmetry functions. Consequently, to compute the force components acting on each atom, the chain rule must be employed.[45]

2.2. Training dataset preparation

The training dataset used for constructing the ANN potential consisted of C–H cluster structures, varying in size from 10 to 71 atoms. Employing cluster structures allowed for a broad sampling of atomic configurations within C–H systems, thereby enhancing the transferability of the ANN potential. The initial dataset was constructed by about 7000 fully optimized defective graphene nanoflakes obtained from our local database (initial training data generation in Fig. 1). These structures were generated by introducing 1–24 C vacancies in a zigzag graphene flake.[47] The original pristine graphene flake contains 54 C atoms, with 18H atoms passivating the edges. The geometries of these defective structures underwent optimization using DFT implemented in the Gaussian 16 package.[48] Generally, the creation of vacancies induces structural instability, leading to a transition toward amorphization at higher vacancy concentrations.[49,50] Consequently, these defective structures

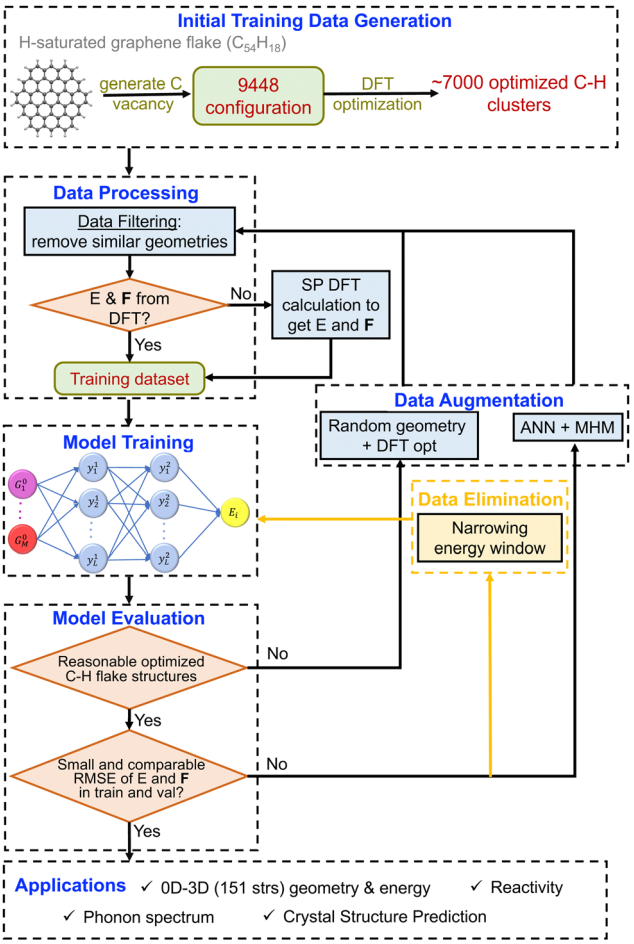

Fig. 1 A schematic workflow for training the ANN potential for C–H materials. It is divided into six blocks: initial training data generation, data processing, model training, model evaluation, data augmentation, and applications. To improve the RMSE in last training iteration, data elimination was applied.

exhibited significant structural reconstructions after geometry optimization, resembling amorphous phases. Given the possibility of similarities among structures for specific number of vacancies, we screened the initial dataset to ensure the structural diversity. To achieve this, we employed distances of atomic environment descriptors[51,52] to identify and eliminate configurations that were similar to each other. This initial data-filtering process results in an initial dataset comprising 4629 optimized defective flake structures. Given our interest in applying the ANN potential to systems with diverse boundary conditions, the energy and forces of these structures were recalculated by employing Vienna Ab initio Simulation Package (VASP version 6.4.1)[53–55] as described in Section 2.4.

Training potential on well-optimized structures generally limits its comprehension of non-equilibrium behavior, impeding its ability to precisely predict non-zero forces as atoms deviate from their ideal positions. In order to capture non-zero forces, the training dataset was gradually augmented by randomly selection of structures from the initial dataset and were subjected to random atomic position displacements, each with an amplitude of 0.05 Å, in subsequent training cycles. Additional structures were randomly generated and optimized using DFT to expand the dataset (data augmentation in Fig. 1). This was necessary because the initial generation of the ANN potential failed in geometry optimization, resulting in widely dispersed or collapsed structures. After improving the ANN potential to handle reasonable geometry optimization without such issues, additional structures were generated and used as initial guesses to explore low-energy regions of the energy landscape. This exploration was carried out using the minima hopping global geometry optimization method (MHM) and the enhanced ANN potential.[56,57] The resulting structures were filtered based on the fingerprint method discussed in the previous paragraph. This filtering ensured the diversity of the training data. Then, single point (SP) calculations with DFT was done for the selected structures to get the energy (E) and forces (F). In this way, we generated a dataset consisting of 26 731 structures, from which 14 664 structures were included in training the final ANN potential due to the energy filtering that will be discussed in Section 2.3. Fig. 2 displays some structures from our dataset, revealing their disordered nature similar to amorphous solids.

2.3. Training process

The training process was conducted iteratively, starting from the well-optimized structures of defective graphene nanoflakes from DFT. After structural filtering, the FLAME code[52] was utilized to train the ANN potential. The code incorporates tools to convert geometries into symmetry functions fed into the ANN, along with the extended Kalman filter algorithm[58] to train the feed-forward ANNs. During training, the entire dataset was randomly partitioned into training and validation sets,

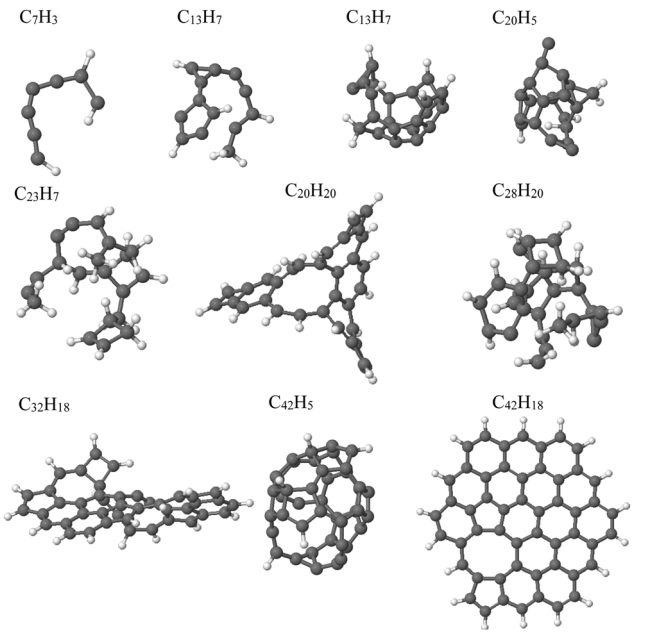

Fig. 2 Representative C–H cluster structures used as training data points. The gray and white spheres represent C and H atoms, respectively.

constituting 70% and 30% of the data, respectively. To test the potentials during the training, we also prepared a small test dataset containing C–H flakes with various edge types (the first condition in Model Evaluation block of Fig. 1). Starting with 4629 data points, we found that the obtained potentials had RMSE less than 3 meV per atom for training and validation data. However, the error in our tests was large. The accuracy of the potential improved after six iterations of training, increasing the dataset size, and capturing non-zero forces.

Despite increasing the training dataset size, we observed minimal improvement in the accuracy of the trained potential during the last training cycle. We hypothesized that this could be attributed to the complexity and significant diversity in energy among the data points (Fig. S1, ESI†), which varied widely across a range of 4.52 eV per atom. After conducting multiple training sessions at different energy range values (as explained in Section S1 and shown in Fig. S2 and S3, ESI†), we identified the optimal dataset with an energy range of 2.0 eV per atom. Narrowing our attention to this energy range and further refining the data, we eliminated training data with a final dataset size of 14 664. The detailed analysis of this dataset's composition, illustrating the distribution of data across various C/H ratios and the count of C atoms, is summarized in Fig. 3 and Fig. S4, S5 (ESI†). The Fig. 3(a) also highlights the absence of pure C systems and systems with a C/H ratio less than 1, as well as an uneven distribution across the available ratios. By training ANN potentials with this data, we noticed an improvement in the accuracy (Fig. S4, ESI†) and transferability of the candidate potentials when applied to test cases. Based on

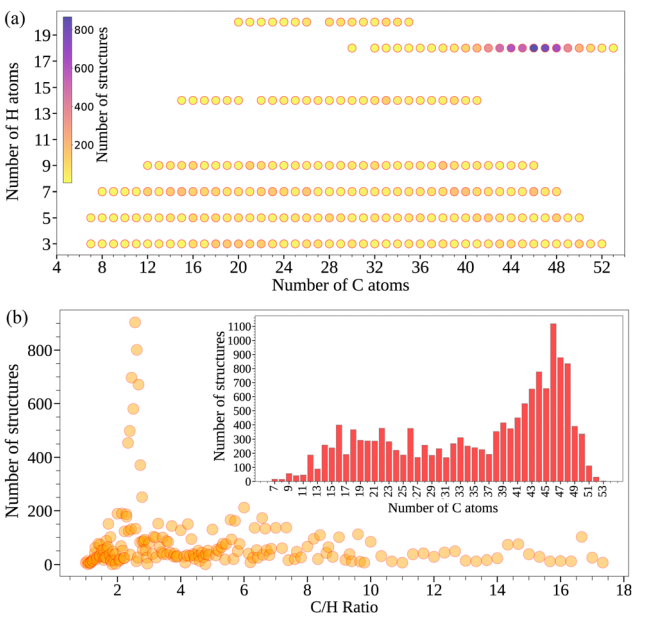

Fig. 3 The configuration analysis of data within the energy range of 2.0 eV per atom. (a) The frequency of structures for distinct C–H ratios, depicted by a colormap (color intensity) representing the count of structures within each ratio. (b) The distribution of structures *vs.* the C/H ratio. The inset plot illustrates the distribution of data based on the count of C atoms.

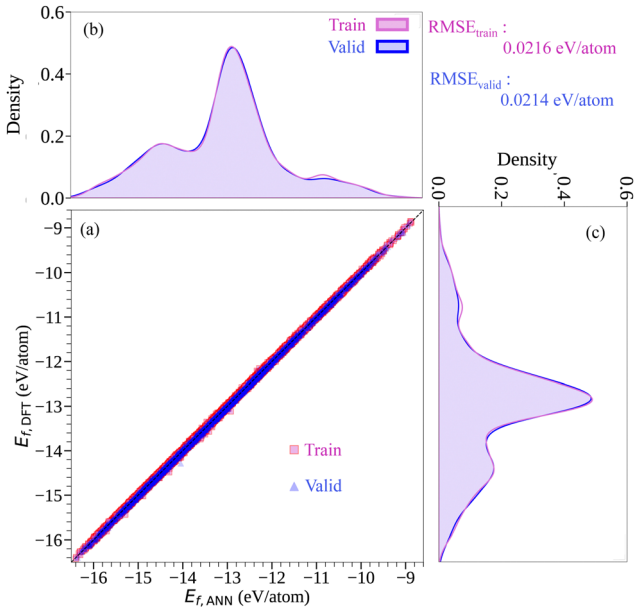

Fig. 4 The reference formation energies obtained from DFT calculations *versus* the predicted values by ANN potential are shown in panel (a), alongside probability density plots in panels (b) and (c), which illustrate the distribution of configurations in the training and validation datasets across different energy values. Pink and blue colors represent the training and validation datasets, respectively.

their accuracy in test cases, we identified a potential with an RMSE of 0.0216 and 0.0214 eV per atom in energy for the training and validation sets, respectively. Fig. 4 exhibits the DFT total energies *versus* the ANN potential predicted values. The figure also provides energy distribution of the structures within training and validation datasets. Both histograms exhibit similar distributions and have consistency in peaks and tails, demonstrating that the potential is not suffering from overfitting.

2.4. DFT calculations

The DFT calculations in this study were performed by employing VASP version 6.4.1. The Perdew–Burke–Ernzerhof (PBE)[59] functional within the generalized gradient approximation (GGA) was adopted to treat the exchange–correlation interactions and the projector-augmented wave basis set with a 500 eV cutoff was used. The convergence thresholds for energy and force during structural relaxation were set to 10^{-4} eV and 0.01 eV Å^{-1}, respectively. All the calculations were performed in a non-spin-polarized manner. As our training dataset consists of both closed- and open-shell clusters, we show in the detailed discussion in Section S2 of the ESI† that ignoring spin in open-shell calculations results in negligible errors in energies and forces. For non-periodic systems, a Monkhorst–Pack mesh of $1 \times 1 \times 1$ *k*-points was used to sample the Brillouin zone. For periodic systems, the smallest allowed spacing between *k*-points was set to 0.40 Å^{-1}. To prevent interactions between images, a vacuum of 10 Å was selected for the aperiodic directions of the systems.

3. Result and discussion

3.1. Geometry and energy comparison of 0D–3D systems

We first examine the transferability of the trained ANN potential, which was exclusively trained on a dataset of C–H clusters. We apply this trained potential to C–H systems across various dimensions: 0D, 1D, 2D, and 3D. This comprehensive analysis elucidates the accuracy and transferability of our ANN potential in predicting the energetics and structural characteristics of diverse systems. Notably, our test cases here encompass pure C systems as well as the structures where the number of H atoms exceeds the number of C atoms, a distinctive inclusion given that the training dataset lacked pure C configurations and those with C/H $\leq$ 1. These intentional inclusions allow us to assess the extrapolative capability of the ANN potential in scenarios absent from its original training data.

3.1.1. 0D systems. We studied 87 non-periodic systems, spanning five chemical groups: alkanes, alkenes, alkynes, aromatic rings, and fullerenes, detailed in Table S2 (ESI†). Each structure underwent geometry optimization using both DFT and the ANN potential. In the following, we compare the optimized geometries and then formation energies (E_f).

The geometry comparison was conducted using V_sim software.[60] Color-coded representations of bonds were employed, reflecting the varying bond lengths within the molecules, as shown in Fig. S7–S26 (ESI†). By visualization, we found that except three cases (cyclooctatetraene, propadiene-12, and C_{11}-008), the predicted geometries by ANN potential are similar to DFT, however, some bond lengths are not identical. To quantify the discrepancy in bond and angles, we did bond and angle analysis by employing cheminformatics library RDKit.[61] Based on this analysis, we found that the C–C–C and C–C–H angles and C–C bond lengths obtained from the ANN potential are generally in agreement, and the C–C bonds are slightly underestimated by the ANN potential. However, for H–C–H angles and C–H bonds, no correlation is observed despite similar geometries from DFT and the ANN potential (Fig. S27, ESI†). This may be due to the sensitivity of bond lengths and angles to small deviations in atomic positions. This validation underscores the robustness and reliability of the ANN potential in representing the intricate bonding patterns exhibited by such molecular structures.

For each 0D group, we conducted an energy analysis by obtaining the E_f values. The reference energy of H was taken as 1/2 of the H_2 molecule energy in the gas phase from DFT, and the reference energy of a C atom was taken as the C atom in cubic diamond from DFT. The RMSE, maximum absolute error (MAE), and mean percentage error (MPE) of E_f were obtained, as presented in Fig. 5. Based on these three metrics, alkynes showed the largest deviation from the DFT results, with an RMSE of 0.126 eV per atom for E_f and an MPE of 0.62%. The MAE, primarily from C_2H_2, is 0.25 eV per atom. The results suggest that the model performs relatively well in predicting energy for these groups. Another notable point is that the MPEs are positive, indicating that the E_f predicted by the ANN potential tend to be overestimated, albeit by less than 1% on average.

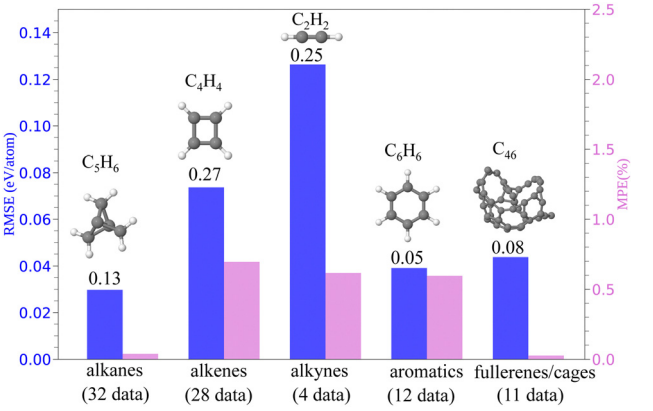

Fig. 5 Comparison of RMSE and MPE for E_f across different chemical groups in our 0D test set (entries in parenthesis next to the chemical group's names are the number of investigated structures for this group). The MAE (in eV) values and the associated structure are displayed atop each bar.

In summary, our analysis of 0D systems evaluated by both DFT and the ANN potential demonstrates the effectiveness of the ANN potential in providing reasonable predictions for molecular geometries, capturing intricate bonding patterns, and predicting energies that are close to DFT results with errors of less than 1%. However, there are a few cases where the optimized geometries from DFT and the ANN potential differ, *e.g.* cyclooctateraene and propadiene-12. Additionally, the RMSE of E_f for all the studied systems is 0.057 eV per atom, which, while slightly larger than chemical accuracy, is also not unreasonably large. It is worth noting that the majority of compositions in this test set were not included in our ANN potential training dataset, particularly those with H/C larger than 1 and pure C systems. Despite this, applying the ANN potential on them did not result in unreasonable results.

3.1.2. 1D and 2D systems. To assess the transferability of the trained ANN potential to boundary conditions that were not included during training, we examined 11 1D and 2D systems, as depicted in Fig. S28 (ESI†). The 1D systems with periodicity along the *z*-direction comprise three 10-atom C-chains including the pure C-chain and its one- and two-side H-saturated configurations, two pristine and two fully H-saturated single-wall carbon nanotubes (SWCNT) with chiralities (4,4) and (8,0). The four 2D systems include graphene and graphyne-X (X = 1, 2, 3) with periodicity in the *xy*-plane.

Firstly, we compared the optimized lattice (a) constants and C–C bond lengths (d_{C-C}) from DFT and the ANN potential, as summarized in Table 1. This comparison revealed that the ANN potential generally underestimates the lattice constants and bond lengths compared to DFT results. Quantitatively, the MAE of SWCNT's diameter (D), a, and d_{C-C} in the SWCNTs are 0.63 Å, 0.17 Å, and 0.10 Å, respectively. For 10-atom C chains, the H-saturated configurations were obtained by adding the 10 H atoms in two ways: either by placing all of them on one side of the chain or by alternating their placement on both sides in a repeating up-and-down sequence. The geometry analysis of the periodic 10-atom C chains shows that when H atoms are added

Table 1 Geometrical and E_f results obtained by DFT and the ANN potential for the 1D and 2D systems. For SWCNT, D is the diameter of the tube and the entries written in parentheses next to the D values are the lattice constant of the SWCNT's unit cell along the nanotube (z-axis). For 10-atom C-chains, graphene, and graphyne, a is the optimized lattice constant of the unit cell along the periodic directions (z-axis). d_{C-C} is the bond length between C (for graphynes, r indicates the C–C bond in the ring and c indicates the C–C bond along chains). All the D, a, and d values are in Å and the energies are in eV per atom

System	D_{DFT} (a_{DFT})	D_{ANN} (a_{ANN})	ΔD	$d_{C-C,DFT}$	$d_{C-C,ANN}$	$E_{f,DFT}$	$E_{f,ANN}$	ΔE_f
SWCNT(4,4)	5.55 (2.46)	5.48 (2.37)	0.07 (0.09)	1.43, 1.44	1.39, 1.40	0.138	0.039	0.099
SWCNT(8,0)	6.37 (4.27)	6.30 (4.10)	0.07 (0.17)	1.43	1.40	0.073	−0.012	0.085
H-SWCNT(4,4)	6.28 (2.59)	5.90 (2.43)	0.38 (0.16)	1.56, 1.58	1.46, 1.49	0.135	0.102	0.033
H-SWCNT(8,0)	7.36 (4.43)	6.73 (4.47)	0.63 (−0.04)	1.57	1.52	0.181	0.154	0.027

System	a_{DFT}	a_{ANN}	Δa	$d_{C-C,DFT}$	$d_{C-C,ANN}$	$E_{f,DFT}$	$E_{f,ANN}$	ΔE_f
C-chain	12.84	12.45	0.39	1.28	1.24	0.914	0.712	0.202
H–C chain (one-side)	15.09	13.79	1.30	1.51	1.38	1.094	1.157	−0.063
H–C chain (two-side)	12.36	12.12	0.24	1.40	1.37	−0.033	−0.006	−0.027
Graphene	2.44	2.39	0.05	1.41	1.38	−0.123	−0.119	−0.004
Graphyne-1	6.89	6.72	0.17	1.43r, 1.35c	1.42r, 1.34c	0.503	0.502	0.001
Graphyne-2	9.46	9.18	0.28	1.43r, 1.40c, 1.23c	1.43r 1.32c, 1.21c	0.647	0.586	0.061
Graphyne-3	12.03	11.67	0.36	1.43r, 1.40c, 1.23c, 1.34c 1.24c	1.43r, 1.32c 1.21c, 1.25c 1.23c	0.701	0.619	0.082

to one side, all C atoms align in a straight line with H atoms oriented perpendicularly to the chain. In contrast, the distribution of atoms on both sides, result in a zigzag structure (similar to *trans*-polyacetylene). Both the ANN potential and DFT give similar H-saturated C-chains patterns. The comparison of d_{C-C} obtained from DFT and the ANN potential for both H-saturated chains show that the C–C bonds are underestimated by the ANN potential. For 2D systems, the MAE of lattice constant a and d_{C-C} is 0.36 Å and 0.09 Å, respectively, smaller than those in 1D systems. Despite these geometrical differences in 1D and 2D systems, the ANN potential predictions remain qualitatively consistent with DFT results. For instance, in 1D systems, both the ANN potential and DFT results show that adding H atoms to the SWCNTs increases D and d_{C-C}. In 2D graphyne-X, the C–C bond lengths obtained by the ANN potential exhibit a similar trend compared to DFT when X increases from 1 to 3: the d_{C-C} in the hexagonal ring are larger than those along the chains and the presence of the C–C bond lengths alternations along these chains. The presence of various bond lengths denote different bond hybridization in hexagonal ring and in connecting chains (acetylenic linkages) as discussed in details in literature.[62,63]

For energy analysis by comparing E_f values, except graphene and H-saturated C-chains, there is a slight overestimation with the ANN potential, *i.e.*, the absolute values of the ANN potential's results are smaller than DFT values. However, the trends of stability are successfully captured by the ANN potential. For example, the SWCNT with chirality (8,0) is relatively more stable than that with chirality (4,4), however, this stability order changes after adding H. For C chains, two-side H-saturated is more stable than pure and one-side H-saturated chains. Similarly, for 2D systems, both the ANN potential and DFT consistently rank the structures, with graphene being more stable than all graphyne-X structures. Among the graphynes, graphyne-1 is the most stable, followed by graphyne-2 and then graphyne-3.

In summary, our analysis of 1D and 2D systems indicates that the trained ANN potential can effectively address boundary conditions that were not encountered during training. Our ANN potential overall underestimates the geometrical parameters and slightly overestimates the energies. Nonetheless, it demonstrates promising performance in predicting these properties of 1D and 2D systems.

3.1.3. 3D systems. Materials with 3D periodicity represent a significant increase in complexity compared to the lower dimensions previously discussed, necessitating a more intricate geometry optimization process.[64] In these materials, the geometry optimization extends to include the stress tensor, requiring the simultaneous optimization of lattice constants, lattice angles, and atomic positions. Here we have considered 53 bulk materials, including 8 structures with C–H and 45 structures with only C. The 45 pure C systems can be categorized into four groups: molecular crystals composed of either flakes/clusters (group I), fullerenes bulks (group II), layered structures (group III), and normal crystal (group IV). The geometries and details such as chemical formula, space group, energy and geometric parameters, for the 53 systems, are summarized in Fig. S29–S31 and Tables S3, S4 (ESI†).

Geometrical analysis of the optimized structures from the ANN potential and comparison with DFT revealed several key findings. Notably, from the 53 bulk phases, only the layered C–H system optimized to an unreasonable structure using the ANN potential, with C–C atoms too close together and desorbed H atoms. Specifically, the ANN potential optimization of this structure resulted in significantly smaller C–C bond lengths compared to DFT counterparts ($d_{C-C,ANN}$ = 0.95 Å and $d_{C-C,DFT}$ = 1.54 Å). Due to this discrepancy and its significant impact on geometric properties and energy comparisons, this C–H layered system was excluded from subsequent analysis. Furthermore, errors in lattice constants for C–H systems were found to be smaller than those for pure C structures: the maximum RMSE

Table 2 Maximum absolute error (MAE, in Å), and RMSE (in Å) of lattice constants *a*, *b*, and *c* in the 3D systems. The metrics for C–H* are after excluding the layered system. The RMSE values of E_f are in eV per atom

	C–H	C–H*	C-I	C-II	C-III	C-IV
MAE_a	0.760	0.430	1.156	1.494	1.876	2.352
MAE_b	0.760	0.430	1.605	1.444	1.876	0.593
MAE_c	0.540	0.540	1.013	0.515	0.079	9.219
$RMSE_a$	0.340	0.222	0.516	0.855	0.557	0.782
$RMSE_b$	0.365	0.264	0.744	0.890	0.547	0.192
$RMSE_c$	0.301	0.318	0.446	0.288	0.026	2.078
MAE_{E_f}	18.524	0.060	0.199	0.046	0.520	0.761
$RMSE_{E_f}$	6.549	0.032	0.097	0.027	0.231	0.236

and maximum MAE in lattice constants for C–H systems were 0.318 Å and 0.540 Å; in contrast, for C systems, the corresponding maximum RMSEs and maximum MAEs were 2.078 Å and 9.219 Å, respectively, for groups II–IV. Table 2 summarized the MAE and RMSE of geometrical parameters. The MPE values can be found in Table S5 (ESI†).

Based on the E_f analysis, summarized in Table 2, we found that the RMSE and MAE of C–H systems and C-II systems are notably smaller than other 3D systems. In contrast, the largest errors are identified for the C-III and C-IV groups in the pure C systems. Additionally, within each group, the comparison of E_f values reveals that the ANN potential correctly predicts the energy ordering of various structures for C–H, C groups I and II. However, for groups III and IV, the ANN potential fails to provide accurate energy ordering, particularly for the layered C systems (group III) which the values are identical as documented in Table S3 (ESI†), which could be a consequence of inaccuracy of the ANN potential for lattice parameters in these systems.

In summary, despite being trained on C–H cluster systems, the ANN potential demonstrates applicability to periodic C–H systems, successfully identifying the energy ordering and capturing correct morphologies, except for layered structures. For pure C molecular crystals (group I and II), the potential can give the right energy ordering for the stability of the configurations based on E_f analysis. However, for the other systems, especially for the layered ones and the low-symmetry C systems with space group *P*1, the accuracy of the potential needs to be improved.

3.1.4. Overall evaluation across dimensions. Based on the evaluations of 0D–3D C–H and pure C materials using the ANN potential trained on C–H clusters, we identified notable strengths and areas for improvement in the ANN potential.

Firstly, geometry optimization using the ANN potential generally provided reasonable geometries for 0D–3D systems, in terms of having similar patterns, bond angles/lengths, and lattice constants close to DFT results. The comparison with DFT values indicates an overall underestimation of geometrical parameters. Furthermore, while the ANN potential works for most of these systems, there were some discrepancies. For instance, it provided planar geometry for non-planar cyclooctateraene (C_8H_8) and propadiene-12 (C_3H_4) molecules and yielded an unreasonable layered C–H structure (C_4H_4) in 3D systems, where the C–C bonds were approximately 0.9 Å smaller than DFT values with desorbed H atoms.

Secondly, the energy investigations show a slight overestimation, meaning E_f from the ANN potential are more negative than DFT values. The comparison of E_f values shows that the largest MAE belongs to 3D pure C systems, particularly for group III (layered systems) and those from group IV with lower structural symmetries (*i.e.* space group *P*1).

These geometrical and energy investigations demonstrate overall robust performance of the ANN potential across various boundary conditions, despite being trained on clusters. However, according to the deviations encountered in analyses of different 0D–3D systems, refinement is required to improve its accuracy. This refinement necessitates the incorporation of training data points consisting of pure C, layered flakes, and clusters with other C/H ratios, particularly those ≤ 1. Additionally, it is important to highlight that our training dataset included open-shell systems, while all the 0D–3D test cases were closed-shell systems. Due to potential issues with using the PBE functional and ignoring spin during our data preparation, the current version of the ANN potential may not be accurate for open-shell systems. In the later improved version, we will consider using hybrid functionals and including spin in preparing the training datasets.

3.2. Reactivity comparison

In this section, we assess the performance of the ANN potential in modeling various chemical processes. We evaluate potential energy curves for C–C bond dissociation in a carbon dimer and three hydrocarbons, comparing the results to DFT calculations to determine how well the ANN potential captures the PES across different bond lengths. Then, we explore the adsorption of $CH_X(X = 1–4)$ and H atom on graphene, analyzing the accuracy of the ANN potential in predicting adsorption sites and adsorption energies. Finally, we study the fully hydrogenation of both periodic and non-periodic 10-atom C-chains to assess the ability of our ANN potential in capturing the favorable hydrogenated configuration and energies.

3.2.1. C–C PES. Capturing the PES is crucial for obtaining reasonable structure during geometry optimization because the PES represents the energy landscape of the system, mapping out how energy changes with variations in atomic positions. The inability of a trained MLIP to accurately capture the PES could result in missing low-energy configurations of the system or giving the structures that are physically not meaningful. Capturing correct PES across various regions is not an easy task. For instance, a recent study on the performance of MLIPs in capturing the PES of C dimer (C_2) revealed that despite correct representation around the equilibrium region, they encountered problems at C–C distances smaller/larger than equilibrium, resulting in overstabilized collapsed/dissociated structures.[65] To evaluate how our ANN potential behaves across various regions of PES, we examined the potential energy curves associated with the C–C bond dissociation process in C_2, ethyne (C_2H_2), ethene (C_2H_4), and ethane (C_2H_6) and compared the results with DFT calculations.

Comparison of the potential energy curves obtained from our ANN potential and DFT reveals a close resemblance in curve

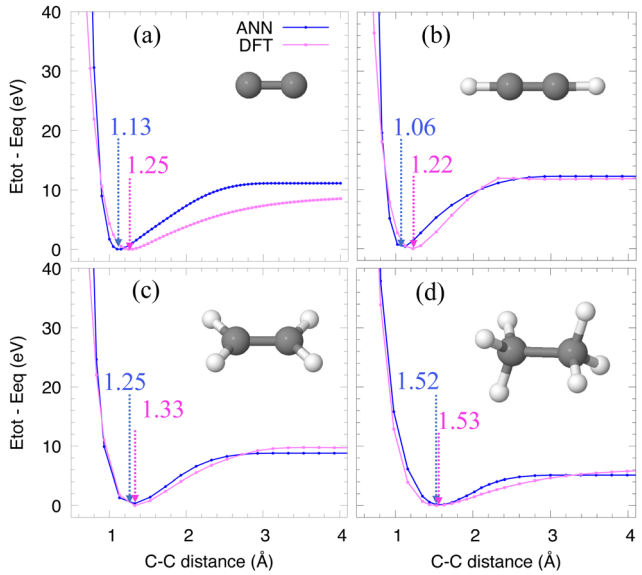

Fig. 6 Potential energy curves for (a) C dimer, (b) ethyne, (c) ethene, and (d) ethane calculated by DFT and ANN potential. The equilibrium bond lengths from the ANN potential and DFT are shown by arrows and the values (in Å) are written in blue and pink, respectively.

shape and energy variation across all four molecules (Fig. 6). This suggests that the ANN potential accurately captures the overall trend of the PES, performing well in predicting reasonable structures at near- and far-equilibrium regions. Additionally, analyzing energy values indicates that the ANN potential predicts energies for the C_2 dimer that are higher than DFT across various bond lengths. However, the discrepancy improves as the H content increases, likely because breaking unsaturated C–C bonds in C_2, C_2H_2, and C_2H_4 introduces complexities in their electronic structure calculations with DFT due to the need for more than one Slater determinant, leading to less accurate predictions. Furthermore, moving from C–C dimer to C_2H_2, C_2H_4, and then to C_2H_6, both the ANN potential and DFT consistently show an increase in equilibrium bond lengths. This trend and the obtained values (Fig. 6) also align with the recent values proposed from bond orders and populations (BEBOP) model.[66]

To conclude, our evaluation of C–C PES shows consistent trends across C_2H_{2X} (X = 0, 1, 2, 3), suggesting that the ANN potential captures the general behavior of the energy *vs.* bond length relationship well. The discrepancy between the ANN potential and DFT results are primarily observed in the C–C dimer system. This can be attributed to the fact that our training dataset does not include pure C systems.

3.2.2. Surface adsorption. We next evaluate the accuracy of the ANN potential by predicting the configuration and surface adsorption energy of CH_X (X = 1–4) and H atom on graphene. The 5 × 5 graphene supercell, consisting of 50 C atoms, was constructed from the DFT and the ANN potential optimized primitive cells of graphene. The in-plane lattice parameters are 12.33 Å and 11.96 Å from DFT and the ANN potential, respectively. This large lattice size, combined with a vacuum of 25 Å perpendicular to the graphene plane (z direction), prevent interactions between periodic images. For each adsorbate, we examined different adsorption sites, including the top of a C atom (top), the midpoint of a C–C bond (bridge), and the center of a hexagonal C ring (hollow), as shown in Fig. 7(a).

Based on the total energies obtained after geometry optimizations, we found that the adsorption sites predicted by the ANN potential agreed with DFT results. Geometrical analysis revealed that most adsorbates exhibited similar molecular orientations, with the exception of CH adsorption. In the case of CH adsorption, the ANN potential predicted CH to be perpendicular to graphene, resulting in a H–C–C bond angle of 180°, while DFT indicated a tilted CH orientation toward the substrate with a H–C–C bond angle of 120° (Fig. 7(e)). The values of C–C bond lengths (d) between the C atom in CH_x and the C atom in the graphene substrate to which CH_x is bonded, calculated from both DFT and the ANN potential were in close agreement. The maximum deviation 0.11 Å of d was observed for the CH case, while the others were smaller than 0.06 Å. For the case of H adsorption, d is the C–H bond length between the adsorbed H and the C atom in graphene obtained from the ANN potential is close to DFT value. The obtained d values of the adsorbed CH_X and H atom on graphene, along with the reported DFT values in literature, are summarized in Table 3. Comparing our results with other reported values in literature shows there is agreement between our findings and those values.

After geometrical analysis, we evaluated the adsorption energies (E_{ads}) of each species using the equation:

$$E_{ads} = E_{X/gr} - E_{gr} - E_X, \quad (2)$$

where, $E_{X/gr}$ is the total energy of the optimized graphene systems with the X species (H or CH_x) adsorbed, E_{gr} is the energy of pristine graphene, and E_X is the energy of the isolated X species. For H atom, the E_X is the energy of single atom H

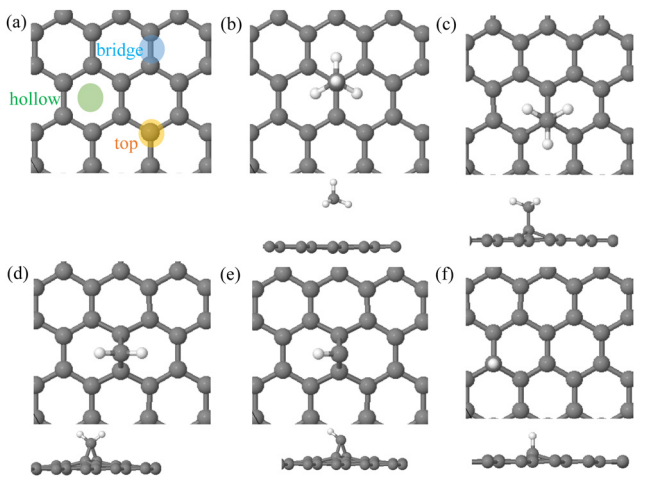

Fig. 7 (a) Schematic illustration of various adsorption sites on graphene. (b)–(f) DFT-predicted top and side views of the most stable configurations of CH_4, CH_3, CH_2, CH_1, and H adsorbed on graphene, respectively. C and H atoms are presented in gray and white, respectively.

Table 3 The adsorption site, the calculated C–C bond lengths (d in Å) of adsorbed CH_X on the graphene, and the calculated adsorption energies (E_{ads} in eV). For H adsorption, d is the C–H bond length. The literature values for d are reported in column five and the corresponding E_{ads} are in the last column

	Site	d_{DFT}	d_{ANN}	d	E_{ads}^{DFT}	E_{ads}^{ANN}	E_{ads}
CH_4	Top	3.33	3.35	3.358[67]	0.03	0.00	−0.33[67]
CH_3	Top	1.59	1.56	1.585[67]	−0.76	−0.50	−0.46[67]
CH_2	Bridge	1.51	1.46	1.515[67]	−2.62	−3.22	−2.94[67]
CH_1	Bridge	1.48	1.37	1.482[67]	−2.10	−4.04	−2.20[67]
H	Top	1.13	1.08	1.128[67]	−0.76	−2.28	−1.52[67]
				1.12[68]			−0.82[68]
							−0.69 to −0.87[69]
							−0.94[70]
							−0.80[71]
							−0.84[72]

from DFT with spin-polarized calculations. Our results for CH_x adsorptions show that, except for the case of CH, where the geometry predicted by the ANN potential is not consistent with DFT, the ANN potential effectively captures the relative adsorption strength of CH_x species despite differences in E_{ads} between DFT and the ANN potential. Both methods indicate an increase in E_{ads} as the number of H atoms in CH_x decreases which is in agreement with other DFT studies. However, for H adsorption, there is a notable discrepancy in the E_{ads} predicted by the ANN potential, which is significantly more exothermic than those from our DFT calculation and literature values. This discrepancy could be due to the low accuracy of the ANN potential for pure 2D systems and predicting C–H bond lengths, as discussed in 0D systems. Therefore, the accuracy of the ANN potential needs to be improved to address accurately the binding strengths of the adsorbed H atom(s). It is also worth noting that accurately describing the interaction of H with graphene to obtain its E_{ads} is a general challenge in modeling. As shown in Table 3, the summarized literature values for this system exhibit variability, ranging from −1.52 eV to −0.69 eV, depending on different models, DFT functionals, and basis sets.[67–72] This variability indicates inherent challenges in achieving consistent results for H adsorption on graphene.

These results highlight the robustness and reliability of the ANN potential in modeling interactions between molecules and materials, such as the adsorption of CH_x and H on graphene. The ANN potential effectively identifies correct adsorption sites and predicts the relative adsorption strength which are close to DFT values. It suggests that the ANN potential is a promising tool for studying catalysis, which requires identifying correct adsorption sites and configurations. However, further improvements is needed to enhance the accuracy, particularly in predicting E_{ads} and C–H bond lengths.

3.2.3. C chain hydrogenation. We next investigate the hydrogenation of periodic 10-atom C chains. In Section 3.1.2, as explained in detail, the hydrogenation process was done by adding one and two-side H atom to each C atom in the chain. Our geometry analysis showed the predicted d_{C-C} bonds are underestimated by the ANN potential. Here, we compare the hydrogenation energy (E_{hydro}) of the chain which represents the energy change associated with adding H atoms to C. E_{hydro} was calculated as the total energy difference between the hydrogenated (saturated) and non-hydrogenated (pristine) states of the C-chain:

$$E_{hyrdo} = (E_{sat} - E_{pr} - 10.0 \times E_H)/N_C. \quad (3)$$

Here, E_{sat} and E_{pr} are the total energies of the chain with and without 10 H atoms, E_H is half the energy of the H_2 molecule, and N_C is the number of C atoms in the chain. The E_{hydro} obtained by DFT for one- and two-side hydrogenated chains were 1.273 eV per atom and −0.980 eV per atom, respectively. These DFT results indicate that two-side hydrogenation is energetically 2.253 eV per atom more favorable than one-side saturation. Similarly, the ANN potential results indicate that the two-side hydrogenated chain is 2.326 eV per atom more favorable than the one-side configuration, with E_{hydro} values of 1.601 eV per atom for the one-side and −0.725 eV per atom for the two-side configurations, respectively. Therefore, despite absolute value differences between DFT and the ANN potential for E_{hydro}, the physical trends are consistent.

3.3. Lattice dynamics

We further assess the accuracy of the ANN potential by studying the lattice dynamics, which requires higher order derivatives of the PES, to build the dynamical matrices in order to obtain properties such as phonon dispersions. We employed the finite displacement method as implemented in Phonopy[73] package to obtain the interatomic force constants. These calculations involve creating supercell structures with an optimal number of displacements based on the structural symmetry. We considered the phonon dispersions of two experimentally synthesized crystalline phases of pure C, namely cubic diamond $Fd\bar{3}m$ (C_8, mp-66) and hexagonal diamond $P6_3/mmc$ (C_4, mp-47), and one C–H system with space group $I2_13$ (C_4H_4, mp-1079612) by making supercells with sizes $4 \times 4 \times 4$, $4 \times 4 \times 2$, and $4 \times 4 \times 4$, respectively. The reason for selecting these structures is that they are non-molecular crystals which provide non flat curves, making them suitable for our comparative analysis.

The phonon dispersions along different high-symmetry points in the Brillouin zone were obtained using DFT and the ANN potential. The comparison between DFT and the ANN potential results revealed a clear similarity in the overall emerging patterns, as depicted in Fig. 8. The ANN potential accurately captures the essential features of the phonon spectra, particularly in the case of acoustic modes, where frequency discrepancies are minimal. However, a more pronounced deviation is observed for optical modes, indicating a higher error in reproducing their frequencies. This observed discrepancy may arise from the inherently challenging nature of capturing intricate details associated with higher-energy optical vibrations.[74,75] To assess the magnitude of errors in comparison with experimental data, we specifically consider cubic diamond's longitudinal optical (LO) modes for which reliable experimental measurements are available. The experimentally measured LO frequencies at Γ and L points are reported as 1314.69 cm^{-1} and 1250.17 cm^{-1}, respectively.[74] Our DFT

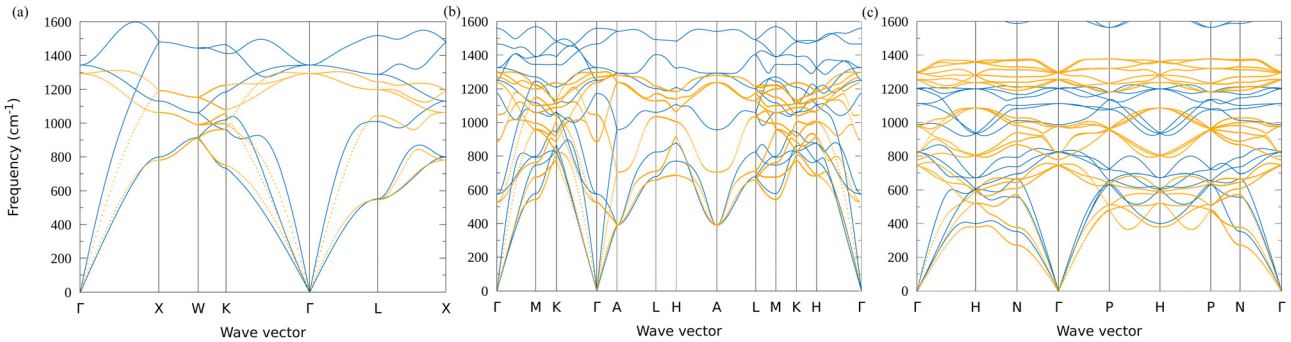

Fig. 8 The comparison of phonon band dispersions obtained from the ANN potential (blue lines) and DFT (orange dotted lines) is conducted for three structures: (a) mp-66 (diamond) (b) mp-47 (c) mp-1079612.

calculations closely align with these values, yielding 1292.69 cm^{-1} and 1244.81 cm^{-1}. However, the ANN potential predictions exhibit a slight deviation, with LO frequencies of 1345.79 cm^{-1} at Γ and 1291.20 cm^{-1} at L. Given these differences, it is noteworthy that the absolute discrepancies of 31.10 cm^{-1} and 41.03 cm^{-1} from ANN potential are relatively small. Despite quantitative disparities, the qualitative agreement in the phonon dispersions underscores the reliability of the ANN potential in capturing the fundamental characteristics of lattice vibrations. Additionally, in practical applications, it is important to highlight that acoustic modes play a crucial role in influencing thermal properties due to their lower energies and significant contributions to heat conduction. Consequently, the accurate representation of acoustic modes by the ANN potential underscores its reliability in predicting key material properties.

3.4. Uncovering a novel C polymorph through ANN-guided structural exploration

Phonon dispersion serves as a theoretical tool for assessing the stability of structures by confirming the absence of imaginary frequencies, an important aspect of CSP. However, for larger systems, phonon dispersion calculations for candidate structures can be demanding using DFT. These calculations typically require creating a supercell, densely sampling the Brillouin zone with k-points to accurately capture phonon modes, and calculating force constants by solving the dynamical matrix equations for each atom pair and phonon mode. Taking advantage of fast energy and force evaluations *via* the ANN potential, we assessed the efficiency and applicability of our trained ANN potential for CSP. Employing the MHM at zero pressure with our ANN potential and utilizing the cubic diamond structure as the initial configuration, several structures within the energy range of 1.0 eV per atom above the cubic diamond were revealed. Comparing them with the known phases in SACADA database collected by Hoffman *et al.*,[76] we confirmed that one of the structures is a new C polymoprh. In the following, we provide some DFT-calculated properties of the discovered C polymorph.

This novel monoclinic polymorph of C with the space group $C2/m$ (No. 12) contains eight C atoms in its primitive cell, as shown in Fig. S32 (ESI†). The atomic positions and lattice

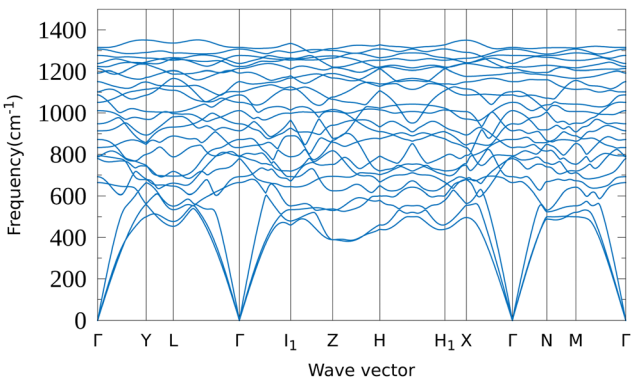

Fig. 9 The phonon band dispersion of the novel C polymorph.

constants are summarized in Table S6 (ESI†). The bond lengths are within 1.50–1.69 Å and the atoms are three- and four-coordinated, reflecting sp^2–sp^3 bonded hybridization, and contains only 6-member C rings. Its dynamic stability was confirmed by the absence of negative frequency modes through the entire Brillouin zone as depicted in Fig. 9. The thermodynamic stability was examined by taking the energy of C atom in diamond structure as reference energy. The calculated E_f of $C2/m$ is 0.203 eV per atom which is energetically more favorable than experimentally synthesized T-carbon[77,78] (space group $Fd\bar{3}m$) and other carbon allotropes such as orthorhobmic carbon o$C20$[79] (space group $Cmcm$) which have been reported.[78,80] Its mechanical stability has also been validated by twelve mechanical stability criteria for orthorhombic structures.[81] Other properties, such as mechanical properties as well as thermal and electronic band structure are summarized in Table S7 (ESI†) and represented in Fig. S33 and S34 (ESI†).

4. Conclusion

In this study, we developed a MLIP based on ANN for modeling C–H systems, achieving an impressive RMSE in energy less than 22 meV per atom. The potential was trained on a diverse dataset of C–H clusters ranging in size from 10 to 71 atoms and C/H > 1. Extensive evaluations against DFT results demonstrated

its accuracy and transferability across a range of scenarios. We examined its performance in geometry optimizations and formation energy predictions across 0D to 3D C–H structures, including systems with C/H ratios excluded in training. While deviations were most pronounced in pure C systems, particularly 3D layered structures, as well as C–H systems with C/H $\leq$ 1, the overall performance was robust. We further assessed the reactivity accuracy of the ANN potential through investigations of potential energy curves for C–C bond dissociation in C_2 dimer and other small hydrocarbons, as well as predictions of adsorption sites and adsorption energies, and C chain hydrogenation. The results indicated its ability to accurately capture the PESs, adsorption sites, and hydrogenation energies. All these evaluations highlight that training on clusters provided diverse environment, resulting in the ANN potential's versatility in addressing diverse system properties. Furthermore, we evaluated the potential's accuracy in lattice dynamics for both pure C and C–H systems. Despite quantitative differences in optical modes, the phonon dispersions exhibited qualitative agreement with DFT results. Finally, by taking advantage of the computational efficiency and accuracy of the ANN potential in predicting energy, forces, and obtaining phonon dispersions, we conducted a rapid structural search, leading to the discovery of a novel C polymorph that is energetically more favorable than other experimentally reported C allotropes. To enhance the accuracy of the current version of the ANN potential and address identified limitations in layered and 2D systems, particularly in pure C and H-rich clusters, additional data points should be incorporated. This includes data encompassing pure C and C/H ratio less than or equal to 1, as well as layered flakes. These efforts constitute the scope of our forthcoming research endeavors.

Author contributions

Somayeh Faraji: data curation, formal analysis, software, methodology, visualization, writing original draft, writing – review and editing; Mingjie Liu: conceptualization, supervision, writing – review and editing.

Data availability

The dataset used for training the ANN potential, which includes 14 664 C–H clusters, and the trained ANN potential are available at **https://github.com/Liu-Group-UF/MLIP-CH-Dataset**. Additional data, including the impact of the training data energy span on potential accuracy, the configurational analysis of training and validation data, details of the 0D–3D systems used for the evaluation of the ANN potential, and figures comparing the 0D geometries optimized by DFT and the ANN potential, are included in the ESI.† The ESI† also contains the structural parameters, mechanical properties, and electronic band structure of the novel C polymorph.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was financially supported by the University of Floridas new faculty start-up funding. We would like to express our sincere gratitude to Dr Alireza Ghasemi for his invaluable assistance in guiding us through the installation process of the FLAME package on HiperGator supercomputer. The authors acknowledge the University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication.

Notes and references

1 S. K. Tiwari, V. Kumar, A. Huczko, R. Oraon, A. D. Adhikari and G. Nayak, *Crit. Rev. Solid State Mater. Sci.*, 2016, **41**, 257–317.
2 Z. Wang, F. Dong, B. Shen, R. Zhang, Y. Zheng, L. Chen, S. Wang, C. Wang, K. Ho and Y.-J. Fan, *et al.*, *Carbon*, 2016, **101**, 77–85.
3 V. Thapliyal, M. E. Alabdulkarim, D. R. Whelan, B. Mainali and J. L. Maxwell, *Diam. Relat. Mater.*, 2022, **127**, 109180.
4 S. E. Stein and A. Fahr, *J. Phys. Chem.*, 1985, **89**, 3714–3725.
5 A. V. Plyasunov and E. L. Shock, *Geochim. Cosmochim. Acta*, 2000, **64**, 439–468.
6 A. Ricca, C. W. Bauschlicher, C. Boersma, A. G. Tielens and L. J. Allamandola, *Astrophys. J.*, 2012, **754**, 75.
7 X. Yang, A. Li, R. Glaser and J. Zhong, *Astrophys. J.*, 2016, **825**, 22.
8 G. Laudadio, Y. Deng, K. van der Waals, D. Ravelli, M. Nuño, M. Fagnoni, D. Guthrie, Y. Sun and T. Noël, *Science*, 2020, **369**, 92–96.
9 A. Hamadi, W. Sun, S. Abid, N. Chaumeix and A. Comandini, *Combust. Flame*, 2022, **237**, 111858.
10 S. u Rather, *Int. J. Hydrogen Energy*, 2020, **45**, 4653–4672.
11 A. Salehabadi, M. F. Umar, A. Ahmad, M. I. Ahmad, N. Ismail and M. Rafatullah, *Int. J. Energy Res.*, 2020, **44**, 11044–11058.
12 I. Hassan, H. S. Ramadan, M. A. Saleh and D. Hissel, *Renewable Sustainable Energy Rev.*, 2021, **149**, 111311.
13 M. Etesami, M. T. Nguyen, T. Yonezawa, A. Tuantranont, A. Somwangthanaroj and S. Kheawhom, *Chem. Eng. J.*, 2022, **446**, 137190.
14 E. Biehler, Q. Quach and T. M. Abdel-Fattah, *ECS J. Solid State Sci. Technol.*, 2023, **12**, 081002.
15 C. Liu, Y. Fan, M. Liu, H. Cong, H. Cheng and M. S. Dresselhaus, *Science*, 1999, **286**, 1127–1129.
16 B. Li, P. Ou, Y. Wei, X. Zhang and J. Song, *Materials*, 2018, **11**, 726.
17 V. Mehmeti and M. Sadiku, *Computation*, 2022, **10**, 68.
18 V. Nagarajan, R. Bhuvaneswari and R. Chandiramouli, *Comput. Theor. Chem.*, 2023, **1230**, 114391.
19 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864.

20 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133.
21 P. Verma and D. G. Truhlar, *Trends Chem.*, 2020, **2**, 302–318.
22 M. Bursch, J.-M. Mewes, A. Hansen and S. Grimme, *Angew. Chem., Int. Ed.*, 2022, **61**, e202205735.
23 J. Wang, H. Shen, R. Yang, K. Xie, C. Zhang, L. Chen, K.-M. Ho, C.-Z. Wang and S. Wang, *Carbon*, 2022, **186**, 1–8.
24 J. T. Willman, K. Nguyen-Cong, A. S. Williams, A. B. Belonoshko, S. G. Moore, A. P. Thompson, M. A. Wood and I. I. Oleynik, *Phys. Rev. B*, 2022, **106**, L180101.
25 M. Qamar, M. Mrovec, Y. Lysogorskiy, A. Bochkarev and R. Drautz, *J. Chem. Theory Comput.*, 2023, **19**, 5151–5167.
26 Y. Li and J.-W. Jiang, *Phys. Chem. Chem. Phys.*, 2023, **25**, 25629–25638.
27 V. Zaverkin, D. Holzmüller, L. Bonfirraro and J. Kästner, *Phys. Chem. Chem. Phys.*, 2023, **25**, 5383–5396.
28 A. Singh and Y. Li, *Comput. Mater. Sci.*, 2023, **227**, 112272.
29 G. A. Marchant, M. A. Caro, B. Karasulu and L. B. Pártay, *npj Comput. Mater.*, 2023, **9**, 131.
30 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.
31 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2017, **96**, 019902.
32 V. L. Deringer and G. Csányi, *Phys. Rev. B*, 2017, **95**, 094203.
33 H. Muhli, X. Chen, A. P. Bartók, P. Hernández-León, G. Csányi, T. Ala-Nissila and M. A. Caro, *Phys. Rev. B*, 2021, **104**, 054106.
34 Y. Wang, Z. Fan, P. Qian, T. Ala-Nissila and M. A. Caro, *Chem. Mater.*, 2022, **34**, 617–628.
35 F. Lu, L. Cheng, R. J. DiRisio, J. M. Finney, M. A. Boyer, P. Moonkaen, J. Sun, S. J. Lee, J. E. Deustua and T. F. Miller III, *et al.*, *J. Phys. Chem. A*, 2022, **126**, 4013–4024.
36 E. Gelzinyte, M. Öeren, M. D. Segall and G. Csányi, *J. Chem. Theory Comput.*, 2023, **20**, 164–177.
37 Q. H. Hu, A. M. Johannesen, D. S. Graham and J. D. Goodpaster, *Dig. Discovery*, 2023, **2**, 1058–1069.
38 S. Lee, H. W. Kim, J. Im, S. K. Kim, Y. T. Kim, H. Chang, T.-W. Ko, J. Lee and Y. Hyon, *J. Korean Phys. Soc.*, 2020, **77**, 680–688.
39 A. Owens, S. N. Yurchenko, A. Yachmenev, J. Tennyson and W. Thiel, *J. Chem. Phys.*, 2016, **145**, 104305.
40 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.
41 X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith and A. E. Roitberg, *J. Chem. Inf. Model.*, 2020, **60**, 3408–3415.
42 H. Zhang, S. Liu, J. You, C. Liu, S. Zheng, Z. Lu, T. Wang, N. Zheng and B. Shao, *Nat. Comput. Sci.*, 2024, 1–14.
43 S. Zhang, M. Z. Makos, R. B. Jadrich, E. Kraka, K. Barros, B. T. Nebgen, S. Tretiak, O. Isayev, N. Lubbers and R. A. Messerly, *et al.*, *Nat. Chem.*, 2024, 1–8.
44 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
45 J. Behler, *J. Chem. Phys.*, 2011, **134**, 074106.
46 N. Artrith and A. Urban, *Comput. Mater. Sci.*, 2016, **114**, 135–150.
47 J. Jin, S. Faraji, Z. Wang, M. Ross, M. U. U. Abdullaev, J. D. Strikowski and M. Liu, manuscript in preparation.
48 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision C.01*, Gaussian Inc., Wallingford, CT, 2016.
49 C. Carpenter, A. Ramasubramaniam and D. Maroudas, *Appl. Phys. Lett.*, 2012, **100**, 203105.
50 J. Kotakoski, A. Krasheninnikov, U. Kaiser and J. Meyer, *Phys. Rev. Lett.*, 2011, **106**, 105505.
51 L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte and C. Wolverton, *et al.*, *J. Chem. Phys.*, 2016, **144**, 034203.
52 M. Amsler, S. Rostami, H. Tahmasbi, E. R. Khajehpasha, S. Faraji, R. Rasoulkhani and S. A. Ghasemi, *Comput. Phys. Commun.*, 2020, **256**, 107415.
53 G. Kresse and J. Hafner, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **47**, 558–561.
54 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.
55 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 169.
56 S. Goedecker, *J. Chem. Phys.*, 2004, **120**, 9911–9917.
57 M. Amsler and S. Goedecker, *J. Chem. Phys.*, 2010, **133**, 224104.
58 I. Rivals and L. Personnaz, *Neurocomputing*, 1998, **20**, 279–294.
59 J. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1998, **80**, 891.
60 L. Sim, *ν_sim*, 2016, **https://gitlab.com/l_sim/v_sim/**.
61 G. Landrum, **https://www.rdkit.org**, 2006.
62 N. Narita, S. Nagai, S. Suzuki and K. Nakao, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **58**, 11009.
63 N. B. Singh, B. Bhattacharya and U. Sarkar, *Struct. Chem.*, 2014, **25**, 1695–1710.
64 L. Wu and T. Li, *J. Mech. Phys. Solids*, 2024, 105639.
65 N. V. Tkachenko, A. A. Tkachenko, B. Nebgen, S. Tretiak and A. I. Boldyrev, *Phys. Chem. Chem. Phys.*, 2023, **25**, 21173–21182.
66 B. Zulueta, S. V. Tulyani, P. R. Westmoreland, M. J. Frisch, E. J. Petersson, G. A. Petersson and J. A. Keith, *J. Chem. Theory Comput.*, 2022, **18**, 4774–4794.
67 K. Li, H. Li, N. Yan, T. Wang and Z. Zhao, *Appl. Surf. Sci.*, 2018, **459**, 693–699.
68 L. Chen, A. C. Cooper, G. P. Pez and H. Cheng, *J. Phys. Chem. C*, 2007, **111**, 18995–19000.

69 A. Dumi, S. Upadhyay, L. Bernasconi, H. Shin, A. Benali and K. D. Jordan, *J. Chem. Phys.*, 2022, **156**, 144702.
70 V. Ivanovskaya, A. Zobelli, D. Teillet-Billy, N. Rougeau, V. Sidis and P. Briddon, *Eur. Phys. J. B*, 2010, **76**, 481–486.
71 Ž. Šljivancanin, E. Rauls, L. Hornekær, W. Xu, F. Besenbacher and B. Hammer, *J. Chem. Phys.*, 2009, **131**, 084706.
72 S. Casolo, O. M. Løvvik, R. Martinazzo and G. F. Tantardini, *J. Chem. Phys.*, 2009, **130**, 054704.
73 A. Togo and I. Tanaka, *Scr. Mater.*, 2015, **108**, 1–5.
74 M. Schwoerer-Böhning, A. Macrander and D. Arms, *Phys. Rev. Lett.*, 1998, **80**, 5572.
75 K. Nakano, T. Morresi, M. Casula, R. Maezono and S. Sorella, *Phys. Rev. B*, 2021, **103**, L121110.
76 R. Hoffmann, A. A. Kabanov, A. A. Golov and D. M. Proserpio, *Angew. Chem., Int. Ed.*, 2016, **55**, 10962–10976.
77 X.-L. Sheng, Q.-B. Yan, F. Ye, Q.-R. Zheng and G. Su, *Phys. Rev. Lett.*, 2011, **106**, 155703.
78 J. Zhang, R. Wang, X. Zhu, A. Pan, C. Han, X. Li, D. Zhao, C. Ma, W. Wang and H. Su, *et al.*, *Nat. Commun.*, 2017, **8**, 683.
79 Q. Wei, C. Zhao, M. Zhang, H. Yan, Y. Zhou and R. Yao, *Phys. Lett. A*, 2018, **382**, 1685–1689.
80 Q. Wei, X. Yang, B. Wei, M. Hu, W. Tong, R. Yang, H. Yan, M. Zhang, X. Zhu and R. Yao, *Solid State Commun.*, 2020, **319**, 113994.
81 F. Mouhat and F.-X. Coudert, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **90**, 224104.

