Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: B. Liu, J. Jin and M. Liu, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00241A.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D5DD00241A

Extrapolating Beyond C_{60}: Advancing Prediction of Fullerene Isomers with FullereneNet

Bin Liu[1,2†], Jirui Jin[1,2†], and Mingjie Liu[*1,2*]

[1]Department of Chemistry, University of Florida, Gainesville, FL 32611, United States.
[2]Quantum Theory Project, University of Florida, Gainesville, FL 32611, United States.
[†]These authors contributed equally to this work.

November 5, 2025

Abstract

Fullerenes, carbon-based nanomaterials with sp^2-hybridized carbon atoms arranged in polyhedral cages, exhibit diverse isomeric structures with promising applications in optoelectronics, solar cells, and medicine. However, the vast number of possible fullerene isomers complicates efficient property prediction. In this study, we introduce FullereneNet, a graph neural network-based model that predicts fundamental properties of fullerenes using topological features derived solely from unoptimized structures, eliminating the need for computationally expensive quantum chemistry optimizations. The model leverages topological representations based on the chemical environments of pentagon and hexagon rings, enabling efficient capture of local structural details. We show that this approach yields superior performance in predicting the C–C binding energy for a wide range of fullerene sizes, achieving mean absolute errors of 3 meV/atom for C_{60}, 4 meV/atom for C_{70}, and 6 meV/atom for C_{72}–C_{100}, surpassing state-of-the-art machine learning interatomic potentials GAP-20. Additionally, the FullereneNet model accurately predicts 11 other properties, including HOMO-LUMO gap and solvation free energy, demonstrating robustness and transferability across fullerene types. This work provides a computationally efficient framework for high-throughput screening of fullerene candidates and establishes a foundation for future data-driven studies in fullerene chemistry.

[*]Corresponding author.
E-mail address: mingjieliu@ufl.edu (Mingjie Liu)

View Article Online
DOI: 10.1039/D5DD00241A

1 Introduction

Fullerenes are carbon-based nanomaterials composed of sp^2 hybridized carbons with the general formula C_{20+2n} ($n \geq 0,\ n \neq 1$). [1, 2] These carbon atoms form pentagon and hexagon rings on the spherical polyhedral cages. The different arrangements of these rings lead to numerous isomeric structures for each fullerene size, with the number of possible isomers increasing at a rate of $O(N^9)$ for $N = 20 + 2n$ carbon atoms, resulting in a diverse family of fullerenes. [3] Thanks to their unique spherical cage structure and remarkable physicochemical properties, various fullerenes have been synthesized and applied in optoelectronics, solar cells, gas storage and separation, biology, and medicine. [4–11] For example, fullerene C_{60}, C_{70}, C_{84} and their derivatives, which conform to the isolated pentagon rule (IPR) with 12 pentagons separated by hexagons to minimize steric strain and ensure high stability, are used as electron acceptors in organic solar cells. [12, 13] Pristine non-IPR fullerenes, such as C_{20} and C_{36}, which are unstable and highly reactive due to their condensed pentagons and increased strain, have been successfully synthesized and isolated. [14, 15] Additionally, non-IPR fullerenes with endohedral functionalization (encapsulation of metal and/or nonmetal atoms, e.g., $Sc_2@C_{66}$ and $Sc_3N@C_{68}$), or exohedral functionalization (anchoring functional groups on the cage surface, e.g., chlorinated $C_{50}Cl_{10}$, $C_{60}Cl_{12}$, and $C_{60}Cl_8$), exhibit unusual electronic, magnetic, and mechanical properties. [16–19] So far, only a small portion of fullerenes have been investigated for real-world applications through laborious trial-and-error approach. To fully harness their capabilities in customizing electronic properties and functionalization for practical uses, it is crucial to accurately and efficiently predict the fundamental properties, including stability, electronic properties, and solubility for high-throughput screening potential fullerene candidates. However, the expansive fullerene family—with its range of sizes and numerous isomers for each size (e.g., C_{84} has 51,592 isomers)—not only poses significant challenges to mathematical enumeration and topological analysis [20], but also surpasses the computational capacity of current high-performance resources for quantum chemistry calculations.

To circumvent this computational bottleneck, early studies explored connections between the graph-theoretical properties of fullerenes and their physical characteristics. Schwerdtfeger et al.'s review [21] provides an extensive account of such topological indicators, demonstrating how graph-based indices derived from adjacency matrices can qualitatively predict properties such as the HOMO–LUMO gap through approaches like Hückel theory, and even anticipate phenomena such as Jahn-Teller distortions. However, these traditional approaches rely heavily on human-derived heuristics and predefined physical approximations, which constrain their predictive power to the limits of established chemical intuition

Data-driven machine learning (ML) techniques have become powerful tools for predicting properties of fullerenes and their functionalized derivatives. For example, Pablo et al. [22] trained various conventional ML models to predict the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), and gap renormalization energies using a dataset of 163 fullerenes and their derivatives, ranging from C_{28} to C_{180}. They tested numerous variables as input representations, including density functional

View Article Online
DOI: 10.1039/D5DD00241A

theory (DFT)-calculated electronic properties, geometric features, phonon features, bond length features, and bond order features. Liu et al. [23] trained a SchNet model-based neutral network potential (NNP), which uses molecule coordinates as input representations, and a dataset of 120k non-isomorphic $C_{2n}Cl_m$ chlorofullerene isomers generated from 500 pristine cages of C_{40}, C_{50}, C_{60}, and C_{70}. The NNP achieved a mean absolute error (MAE) of 0.37 eV for relative energy prediction with respect to DFT benchmarks and demonstrated excellent transferability to other exohedral functionalized fullerenes $C_{2n}X_m$ (X=H, F, I, Cl, Br, OH, CF_3, CH_3) across different functional groups, number of addends, and cage sizes. However, previous studies have relied on DFT-calculated features or those derived from DFT-optimized geometry for ML training, which requires significant computational cost, making it unsuitable for high-throughput screening of fullerene candidates. Machine Learning Interatomic Potentials (MLIPs) have also been employed for fullerene systems. Aghajamali and Karton [24] applied the Gaussian Approximation Potential (GAP-20) [25] force field to investigate the isomerization energies and thermal stabilities of C_{40} fullerenes, while Karasulu et al. [26] used GAP-20 to predict large carbon clusters and search for novel isomers, overcoming the computational limits of first-principles approaches. However, GAP-20 is limited to energy-potential surface prediction and cannot predict other fundamental properties which are crucial for real-world applications, such as molecular orbital levels, HOMO-LUMO gaps, and solubilities. More importantly, employing MLIPs requires geometry optimization to determine ground-state energies, resulting in additional computational costs.

Graph neural networks (GNNs) have gained considerable attention in chemistry due to their exceptional performance in molecular and materials science. [27–30] Their effectiveness stems from the ability to represent molecules and materials as graphs, where nodes correspond to atoms and edges to bonds. For example, TensorNet [31], which utilizes Cartesian tensor representations, demonstrates state-of-the-art (SOTA) performance on the small organic molecule dataset QM9 [32]. Matfomer [33], incorporating a transformer architecture for learning periodic graph representations, is considered a leading GNN model for predicting crystal properties such as formation energy and band gap. However, these models require optimized structures as inputs, since precise structural information is essential for training, significantly increasing computational costs. Moreover, fullerenes, consisting solely of carbon, present unique challenges due to identical node features, complicating the design of effective GNN models for this system. To the best of our knowledge, no existing GNN model can accurately and efficiently predict a wide range of fundamental properties for fullerenes of any size without relying on quantum chemistry-optimized structures as inputs.

In this work, we developed a GNN-based model to predict the fundamental properties of fullerenes using two sets of topological features based on the arrangement of pentagons and hexagons over the cage surface, as proposed in our previous study. [34] These topological features rely solely on carbon atom connectivity and can be efficiently extracted from unoptimized geometries, eliminating the need for quantum chemistry-based geometry optimization. Our results show that these features can effectively capture the intricate

View Article Online
DOI: 10.1039/D5DD00241A

local structural environments of carbon atoms in fullerene, enabling excellent accuracy and extrapolation capability for predicting the C-C binding energy beyond C_{60}. Notably, our GNN model trained with a C_{20}–C_{58} dataset achieved MAEs of 3, 4, and 6 meV/atom for test sets C_{60}, C_{70}, and C_{72}–C_{100}, respectively, surpassing the accuracy of the SOTA MLIP GAP-20. More profoundly, with the same topological features as structural representations, our retrained GNN models demonstrated high prediction accuracy for 11 other properties, including HOMO, LUMO, gap, solvation free energies, and logP. Our study highlights the superior capability of topological features for predicting various fundamental properties of fullerenes with excellent accuracy and transferability. Our study lays a fundamental basis for future data-driven research in fullerene chemistry.

2 Methodology

2.1 Dataset construction

In this study, we constructed three datasets for model training and testing. The C_{20}–C_{60} dataset, consisting of 5770 structures, was adopted from our previous work [34].This dataset provides a complete set of C_{20}–C_{60} structures, with properties such as the HOMO-LUMO gap, formation energy, and IP-EA computed using DFT (with the same methods applied in this work). The C_{70} dataset comprising 8149 isomers, were generated using the FULLERENE program (version 4.5) and subsequently optimized with a harmonic oscillator approximation-based force field [35], which provides a good compromise between computational efficiency and accuracy in reproducing fullerene geometries. The C_{72}–C_{100} dataset (except C_{88}) with 1171 IPR-conforming isomers were obtained from the Fullerene Library created by M. Yoshida. [36] Table S1 and S2 lists numbers of possible isomers, C-C bonds, hexagon rings for each fullerene size in each set.

2.2 Computational details

We applied Gaussian 16 package [37] for all the DFT calculations. We used B3LYP hybrid functional [38–40] with D3 dispersion correction [41] and the 6-31G* basis set for geometry optimization. The maximum force threshold is 0.02 eV/Å. Then, we conducted a single-point calculation at singlet state with B3LYP functional and 6-311G* basis set to compute energies and electronic properties of the isomers. The solvation energies were calculated using the Solvation Model based on Density (SMD).[42] The dielectric constants for water and ODCB solvents were set at 78 and 10, respectively.

Geometry optimization with GAP-20 potential were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package [43], compiled with the Quantum mechanics and Interatomic Potentials (QUIP) library [44, 45] including GAP routines. Geometry relaxation and energy minimization were conducted with the conjugate gradient method. The convergence criteria were set to 10^{-4} eV for the total energy and 10^{-6} eV/Å for the atomic forces.

View Article Online
DOI: 10.1039/D5DD00241A

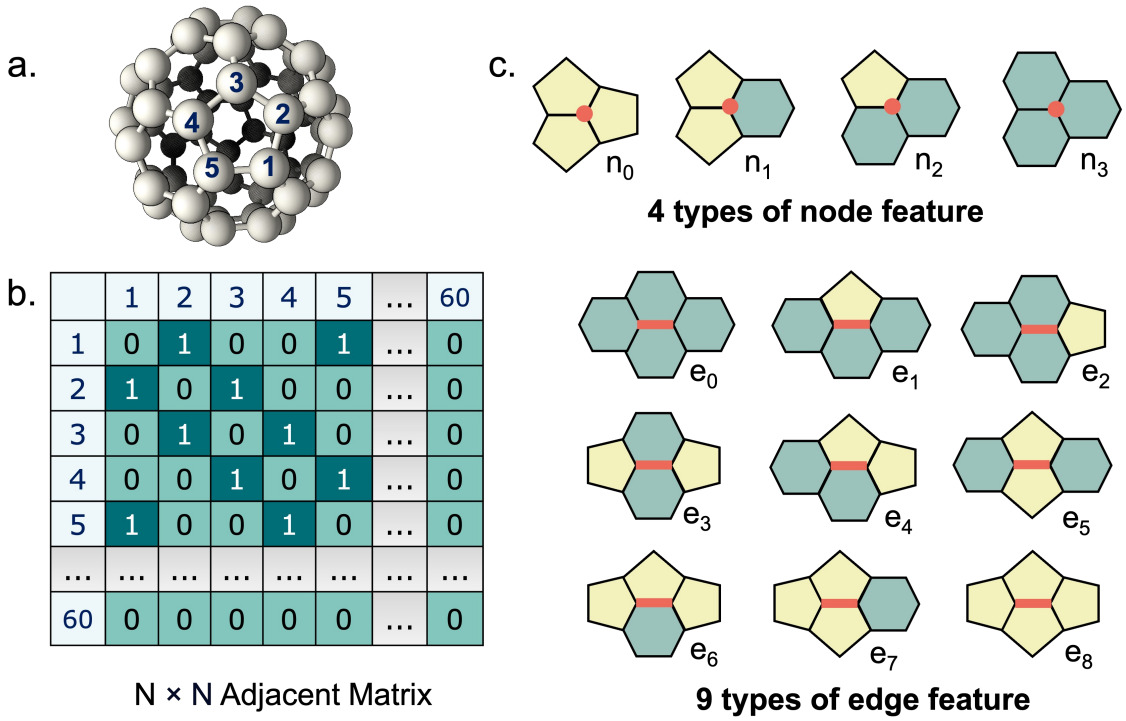

Figure 1: **Feature Construction. (a)** Take C_{60} as an example, each atom in the fullerene is assigned a unique sequence number. **(b)** An adjacency matrix is generated, where entries of 1 indicate a connection between atoms, and entries of 0 indicate no connection. **(c)** Topological node and edge features are derived from the adjacency matrix.

2.3 Feature space construction

We used atom and bond representations from our previous work [34] as node and edge features for GNN model training. In fullerene molecules, each carbon atom is fused to three adjacent polygonal rings, and each C-C bond is surrounded by four polygonal rings, either pentagons or hexagons. Atom features categorize carbon atoms into four types $\{v_i|i = 0, 1, 2, 3\}$, based on the number of adjacent pentagons (ranging from 0 to 3). The node feature is one-hot encoded with four digits to represent the local geometry of each carbon atom. Similarly, bond features categorize C-C bonds into nine types $\{e_i|i = 0, ..., 8\}$, based on the number and arrangement of the surrounding polygonal rings. The edge feature is one-hot encoded with nine digits to indicate the local geometry of each C-C bond. The detailed atom and bond features are shown in Figure 1.

Beyond the canonical fullerenes studied in this work, our methodology may also be applied to non-canonical fullerenes that include other polygons such as heptagons and octagons. While the present study focused exclusively on canonical fullerenes, we propose an extension of the current methodology to address non-canonical cases (see Section S1), which will be carefully investigated in future work.

View Article Online
DOI: 10.1039/D5DD00241A

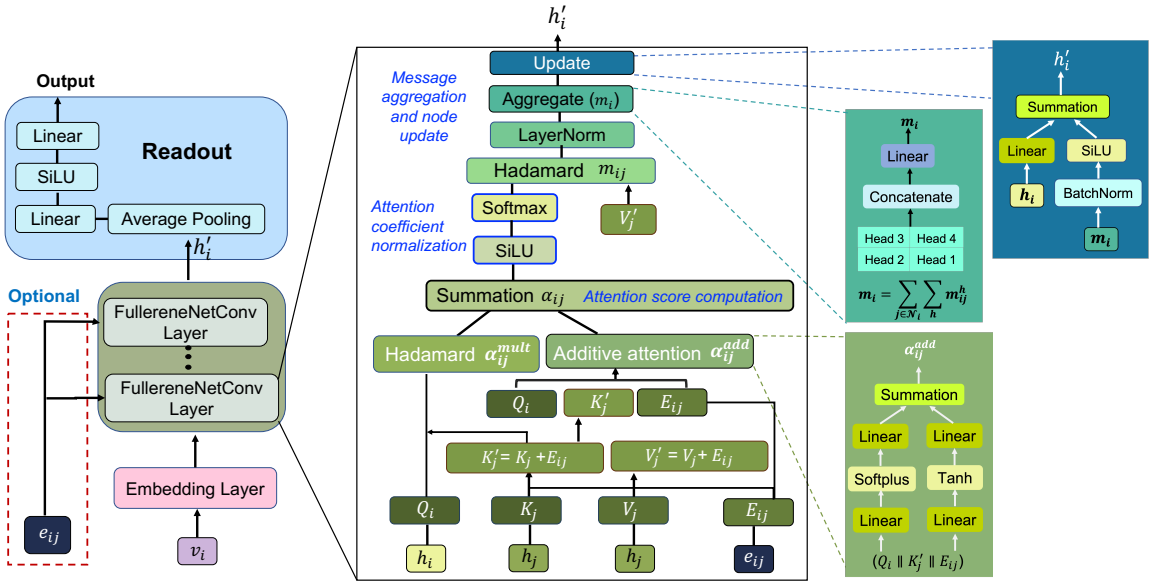

Figure 2: The detailed architecture of FullereneNet.

2.4 GNN Model architecture

Here, we present our model named FullereneNet, inspired by the transformer-based architecture utilized in Matformer [33] (model details are in section S2). This model is designed to eliminate the need for quantum mechanical optimization of fullerene structures, as illustrated in Figure 2. We define the molecular graph as $G = (V, E)$, where V and E represent the set of atoms and bonds in a molecule, respectively. Let $v_i \in V$ denote the node feature of atom i, and $e_{ij} \in E$ represent the edge feature between atoms i and its neighbor atom j. The overall model architecture consists of an embedding layer, multiple convolutional layers, and a readout function. The embedding layer projects the atomic feature v_i to a higher-dimensional vector, denoted as h_i. The convolutional layers are the core of the GNN model, where message passing occurs [46, 47]. This process is essential for capturing interactions between atoms. The readout function aggregates the updated atomic representations into a molecular-level feature, which is then used for property prediction via a pooling function.

Our proposed message passing scheme is composed of three steps: attention score computation, attention coefficient normalization, and message computation with node information updating. In the first step, query Q_i, key K_j, and value V_j are calculated following the regular attention mechanism [48]. These vectors serve to evaluate how relevant each neighboring node is to the node being updated. By comparing the query of a node with the keys of its neighbors, the model calculates attention scores that determine the influence each neighbor should have. Specifically,

$$Q_i = \mathrm{LN}_Q(h_i), \quad K_j = \mathrm{LN}_K(h_j), \quad V_j = \mathrm{LN}_V(h_j), \quad E_{ij} = \mathrm{LN}_E(e_{ij}) \tag{1}$$

View Article Online
DOI: 10.1039/D5DD00241A

where LN_Q, LN_K, LN_V, and LN_E denote the linear transformations to compute query, key, value, and edge embedding. Then K'_j and V'_j are obtained by summing K_j or V_j with E_{ij}, respectively. Next, we compute the attention score using both additive and multiplicative components. The additive attention computes attention scores by processing the concatenation of Q_i, K'_j, and E_{ij} through two separate multi-layer perceptrons (MLPs) with different activation functions:

$$\alpha_{ij}^{\text{add1}} = \mathrm{MLP}_{\text{add1}}(Q_i \| K'_j \| E_{ij}), \tag{2}$$

$$\alpha_{ij}^{\text{add2}} = \mathrm{MLP}_{\text{add2}}(Q_i \| K'_j \| E_{ij}), \tag{3}$$

$$\alpha_{ij}^{\text{add}} = \alpha_{ij}^{\text{add1}} + \alpha_{ij}^{\text{add2}}, \tag{4}$$

where $\|$ denotes concatenation, $\mathrm{MLP}_{\text{add}}$ consists of linear layers with an activation function (e.g., Tanh, Softplus), and α_{ij}^{add} is the additive attention score. Another part is multiplicative attention, which is computed by using the scaled Hadamard product:

$$\alpha_{ij}^{\text{mult}} = \frac{Q_i \odot K'_j}{\sqrt{d}} \tag{5}$$

where $\odot$ represents the Hadamard product, and d is the dimensionality of the key vectors. The overall attention score α_{ij} is obtained by summing the α_{ij}^{add} and $\alpha_{ij}^{\text{mult}}$ components, followed by an activation function.

After obtaining the attention score α_{ij}, we then move to the second part. A non-linear activation function, such as SiLU, is first applied to the combined attention scores α_{ij}. To ensure that the attention coefficients are properly normalized, a softmax function is applied over the neighbors of atom i:

$$\alpha_{ij} = \text{softmax}(\alpha_{ij}) = \frac{\exp(\alpha_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(\alpha_{ik})} \tag{6}$$

where $\mathcal{N}(i)$ denotes the set of neighbor nodes of atom i.

Finally, in the third part, messages from neighbors are aggregated and updated to the original atom representations. The messages from neighbors are first computed by weighting the value vectors V'_j with the attention coefficients α_{ij} using the Hadamard product:

$$m_{ij} = \alpha_{ij} \odot V'_j \tag{7}$$

The processed messages are then passed through a layer normalization function. Subsequently, the messages from each node and attention head are aggregated using a summation method:

$$m_i = \sum_{j \in \mathcal{N}_i} \sum_h m_{ij}^h \tag{8}$$

The last step is the node update with a residual connection:

View Article Online
DOI: 10.1039/D5DD00241A

$$h_i' = \mathrm{LN}_{Lin}(h_i) + \sigma(\mathrm{BN}(m_i)) \tag{9}$$

where LN_{Lin} denotes linear transformations, σ is an activation function, and BN indicates batch normalization. The updated node representation is obtained by adding the original node features with the message m_i.

Matformer introduces an effective transformer variant tailored for crystal graph learning. However, our proposed model, FullereneNet, diverges from Matformer in terms of feature construction and message passing schemes. While Matformer leverages elemental diversity to extract a broad range of atomic features, it struggles with fullerenes, which consist solely of carbon and exhibit uniform atomic properties. To overcome this limitation, we designed topological atomic features that distinguish pentagon and hexagon arrangements around carbon atoms, effectively capturing structural variations and enhancing representation. Additionally, Matformer takes a DFT-optimized structure to derive bond distances and angles, whereas FullereneNet, with its bespoke feature design, bypasses the need for geometry optimized via *ab initio* methods as input, significantly reducing computational costs. Regarding message passing, Matformer combines multiple information streams through a complex triple concatenation process. In comparison, FullereneNet employs a streamlined method that separately processes atomic interactions through two complementary pathways—one capturing linear relationships and another capturing non-linear relationships between connected atoms. This dual-pathway design eliminates redundant computational steps while better preserving the essential topological information that determines fullerene stability. Given the uniform node degree (*i.e.*, three) in fullerenes, FullereneNet does not encounter the challenge to distinguish nodes with different degrees as Matformer [33]. Instead, it benefits from applying softmax normalization, which enhances training stability and model focus. The hyperparameters search range for FullereneNet is summarized in Table S3.

2.5 Model training strategy for extrapolation

In this study, we aim to develop a GNN model with strong extrapolation capabilities for predicting the stability of larger fullerenes. To achieve this, we trained and validated our model on a dataset of small fullerenes ranging from C_{20} to C_{58}, including 3958 distinct isomers. Three larger fullerene groups were utilized as test sets: the complete C_{60} dataset containing 1812 isomers, a C_{70} dataset consisting of 100 randomly selected non-IPR isomers from a total of 8149 entries, and a C_{72}–C_{100} dataset comprising 1171 IPR-conforming isomers. The random selection of C_{70} isomers ensures unbiased sampling across the energy landscape while avoiding the prohibitive computational cost of evaluating all 8149 isomers, while the curated C_{72}–C_{100} dataset provides validation against established structural benchmarks. Moreover, the inclusion of both IPR and non-IPR fullerenes in the test sets allows us to demonstrate the model's reliable extrapolation capability.

Previous studies [49, 50] have shown that different data-splitting strategies can enhance a model's extrapolation ability. To explore their impact on improving the robustness and

View Article Online
DOI: 10.1039/D5DD00241A

effectiveness of extrapolation, we applied four distinct strategies to partition the training and validation sets (Figure 3):

a. Leave-One-Group-Out (LOGO): Following the approach described by Zhao et al. [50], the dataset was partitioned based on fullerene cage size (*i.e.*, total number of carbon atoms). For instance, C_{58} was used as the validation set while the fullerenes ranging from C_{20} to C_{56} constituted the training set. By rotating the validation set across fullerene sizes from C_{20} to C_{58}, five GNN models were generated. Notably, C_{20}–C_{50} fullerenes were grouped as a single validation set since their isomer counts are relatively small, and their combined total is comparable to the number of isomers found in each of the larger cages (C_{52}, C_{54}, C_{56}, and C_{58})

b. Leave-One-Cluster-Out (LOCO): In contrast to LOGO, where the dataset is categorized by fullerene cage size, the LOCO method, as proposed by Meredig et al. [49], employs the k-means clustering algorithm [51] to cluster the data into five distinct groups. Each cluster was sequentially used as the validation set while the remaining clusters were used for training, producing five GNN models.

c. Five-Fold Cross-Validation (5-fold CV): The C_{20}–C_{58} dataset was divided into training and validation sets using 5-fold cross-validation, resulting in the training of five GNN models.

d. Random Split: The C_{20}–C_{58} dataset was randomly partitioned into training and validation sets with a 4:1 ratio. Five GNN models were trained, each using a different random seed.

The extrapolation performance of each strategy was evaluated by averaging the prediction outcomes of the corresponding five GNN models across the three test sets: the C_{60}, C_{70}, and C_{72}–C_{100} datasets.

3 Results and Discussion

3.1 Stability prediction

We evaluated the extrapolation performance of FullereneNet trained using the four strategies discussed in Section 2.5. The performance metrics for predicting the binding energies of fullerene in three test sets — C_{60}, C_{70}, and C_{72}–C_{100} — are summarized in Table 1. For each strategy, the final prediction value was obtained by averaging the prediction values of the five corresponding GNN models on the test set. As shown in Table 1, all FullereneNet models trained on the C_{20}–C_{58} dataset, irrespective of the training strategies employed, consistently demonstrate high accuracy in predicting binding energies of the C_{60}, C_{70}, and C_{72}–C_{100} datasets as indicated by MAE and root mean squared error (RMSE) values. Notably, these models achieve small MAE values ranging from 3 meV/atom to 7 meV/atom.

View Article Online
DOI: 10.1039/D5DD00241A

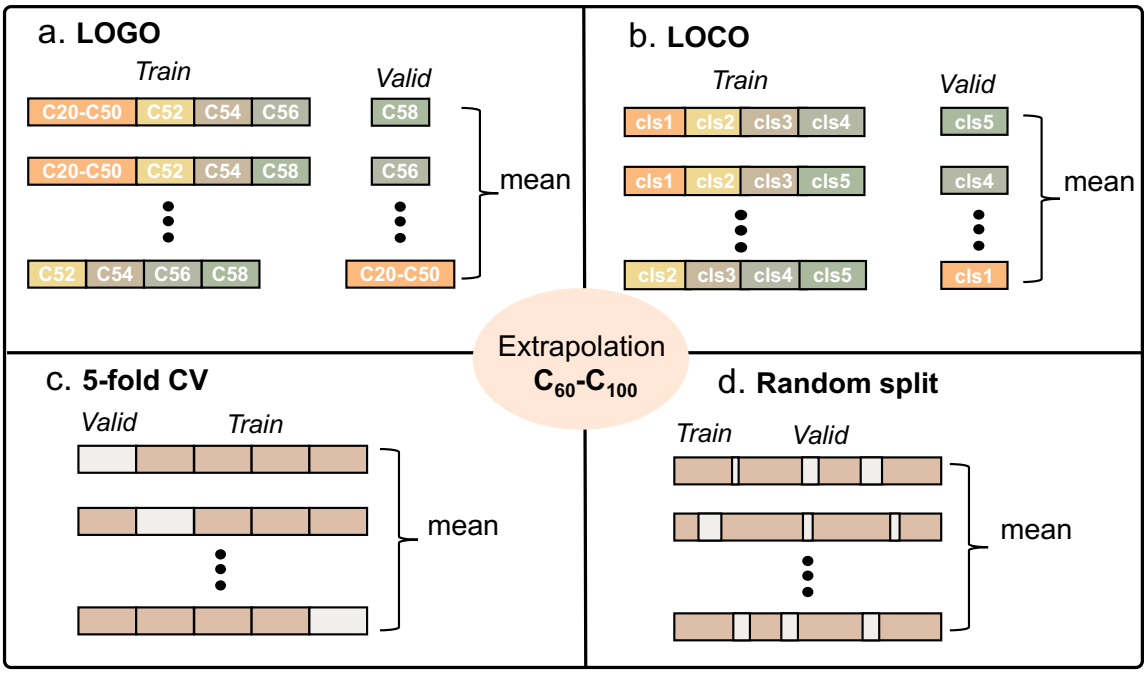

Figure 3: Four different training strategy for extrapolation, including Leave-One-Group-Out (LOGO), Leave-One-Cluster-Out (LOCO), Five-fold Cross-Validation (5-fold CV), and Random Split, respectively.

The strong extrapolation capabilities of FullereneNet highlight the effectiveness of atom and bond features as input representations for predicting fullerene's stability. Based on our results, the four different data split strategies did not exhibit a statistically significant difference in their predictive performance for large-sized fullerenes. Furthermore, the prediction error for the C_{72}–C_{100} test set was consistently larger than that for the other two test sets, regardless of the training strategy employed. For instance, using the random split method, the FullereneNet model achieved a coefficient of determination (R^2) of 0.563 and an MAE of 5 meV/atom for C_{72}–C_{100}, compared to 0.989 and 3 meV/atom for C_{60}, and 0.974 and 4 meV/atom for C_{70}. This slightly lower accuracy for the C_{72}–C_{100} test set can be attributed to the fact that the binding energy distribution of the C_{72}–C_{100} set is markedly different from that of the training set (C_{20}–C_{58}), as illustrated in Figure S1.

View Article Online
DOI: 10.1039/D5DD00241A

Table 1: Performance metrics (R^2, MAE, and RMSE) across fullerene datasets using four different validation methods, including Leave-One-Group-Out (LOGO), Leave-One-Cluster-Out (LOCO), Five-fold Cross-Validation (5-fold CV), and Random Split, respectively.

Dataset	Method	R^2	MAE	RMSE
C_{60}	LOGO	0.988	0.003	0.004
	LOCO	0.988	0.003	0.004
	5-fold CV	0.989	0.003	0.004
	Random Split	**0.989**	**0.003**	**0.003**
C_{70} (non-IPR)	LOGO	0.969	0.004	0.005
	LOCO	0.972	0.004	0.005
	5-fold CV	0.973	0.004	0.005
	Random Split	**0.974**	**0.004**	**0.005**
C_{72}–C_{100} (IPR)	LOGO	0.304	0.007	0.009
	LOCO	0.431	0.006	0.008
	5-fold CV	0.391	0.006	0.008
	Random Split	**0.563**	**0.005**	**0.007**

As detailed in section 2, the FullereneNet incorporates node and edge features extracted from unoptimized fullerene structures as input. In contrast, Matformer, the predecessor of FullereneNet, relies on atomic attributes such as atomic volume, valence electron count, and bond distances calculated from optimized Cartesian coordinates. However, since fullerene consists exclusively of carbon atoms, these conventional atomic descriptors become uniform across the structure, limiting their discriminative power for property prediction. Here, we evaluated the performance of Matformer, trained on bond distance obtained from atomic coordinates of unoptimized fullerene structures, against FullereneNet in predicting the binding energies of fullerenes ranging from C_{60} to C_{100}. As shown in Figure S2, Matformer performs poorly with unoptimized structures, whereas optimizing fullerene structures significantly improves prediction accuracy. For example, in the C_{70} dataset, utilizing optimized structures to train Matformer increases the R^2 value from 0.542 to 0.977 and reduces the MAE from 19 meV/atom to 4 meV/atom, as illustrated in Figure S3.

The state-of-the-art MLIP, GAP-20, has been developed to accurately and efficiently predict isomerization energies, assess thermal stability, and identify new carbon clusters and fullerene isomers [26]. We evaluated the performance of GAP-20 on DTF optimized structures by predicting the relative binding energies for the C_{60}, C_{70}, and C_{72}–C_{100} datasets, using the binding energy of C_{60}-isomer-1 as the reference (Figure S4). We emphasize that GAP-20 exhibits excellent accuracy in geometry optimization when benchmarked against the DFT method. For the C_{60} dataset, we confirmed a strong correlations between relative binding energies calculated using GAP-20 with both DFT-optimized and GAP-20-optimized structures, achieving an impressive R^2 value of 0.97 and a low MAE of 5

View Article Online
DOI: 10.1039/D5DD00241A

meV/atom (Figure S5), consistent with previous studies. [52]

Figure 4 presents a comprehensive comparison of binding energy prediction performance among FullereneNet (using unoptimized structures), Matformer, and GAP-20 (both using optimized structures). Our results demonstrate that FullereneNet consistently outperforms the other two models across all three datasets. Matformer exhibits poor performance on the C_{72}–C_{100} test sets, achieving an MAE of 0.020 eV/atom. Similarly, GAP-20 shows suboptimal performance on the C_{60} test set, with an an MAE of 0.016 eV/atom. Detailed MAE and R^2 values for all methods are summarized in Table S4.

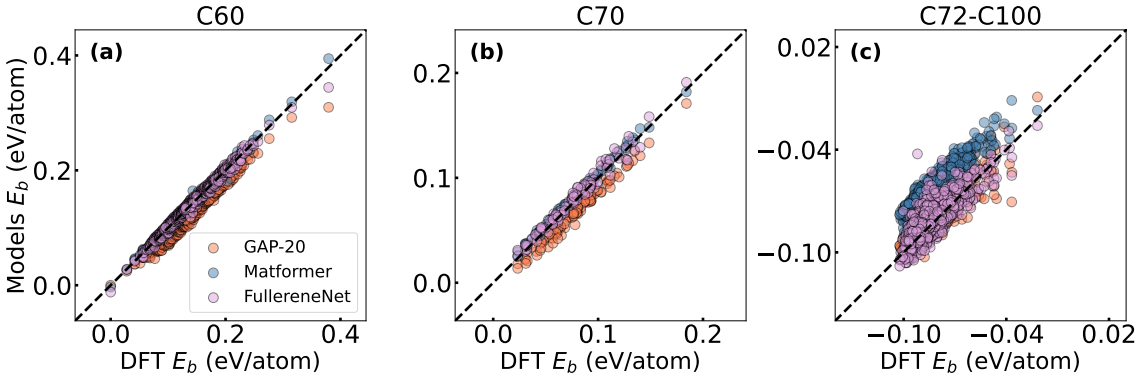

Figure 4: Parity plots comparing the prediction results of FullereneNet (unoptimized structures), Matformer, and GAP-20 (optimized structures) against DFT values, using C_{60}-isomer-1 as the reference. (a) Performance of the three models on the C_{60} test set. (b) Performance on the C_{70} test set. (c) Performance on the C_{72}–C_{100} test set.

Besides the accuracy, one significant benefit of our model is that it can avoid the huge computational cost associated with computing large-size fullerenes. Take the one C_{720} isomer [36] as an example, which represents a computationally challenging system due to its substantial size. To quantify the computational advantages, we performed a detailed cost analysis comparing FullereneNet with DFT and MLIP approaches. The DFT calculations (using Gaussian 16 [37] on 48 CPU cores), geometry optimization with B3LYP/6-31G* required 33 hours 56 minutes of wall time, followed by an additional 6 hours 4 minutes for single-point energy calculations with B3LYP/6-311G*, totaling approximately 40 hours of computation time. In contrast, GAP-20 calculation (using LAMMPS with convergence criteria of 10^{-4} eV for total energy and 10^{-6} eV/Å for atomic forces) completed geometry optimization in approximately 1 minute on a single CPU core. Remarkably, FullereneNet prediction required less than 5 seconds with one Nvidia L4 GPU, representing a tremendous speedup compared to DFT calculations. This dramatic computational acceleration arises from FullereneNet's ability to directly predict binding energies based on the arrangement of polygon rings, eliminating the need to compute energies from optimized geometry using DFT or MLIP.

In summary, our results demonstrate that FullereneNet effectively leverages topological features from unoptimized structures to accurately predict binding energies, showcasing

View Article Online
DOI: 10.1039/D5DD00241A

strong extrapolation capabilities for larger fullerenes. In contrast, Matformer and GAP-20 rely heavily on optimized structures for accurate prediction. Given GAP-20's exceptional ability to generate optimized geometries closely matching DFT results, integrating it with FullereneNet could enhance predictions of additional properties, such as ionization potential and electron affinity. This combined approach will be explored further in Section 3.3.

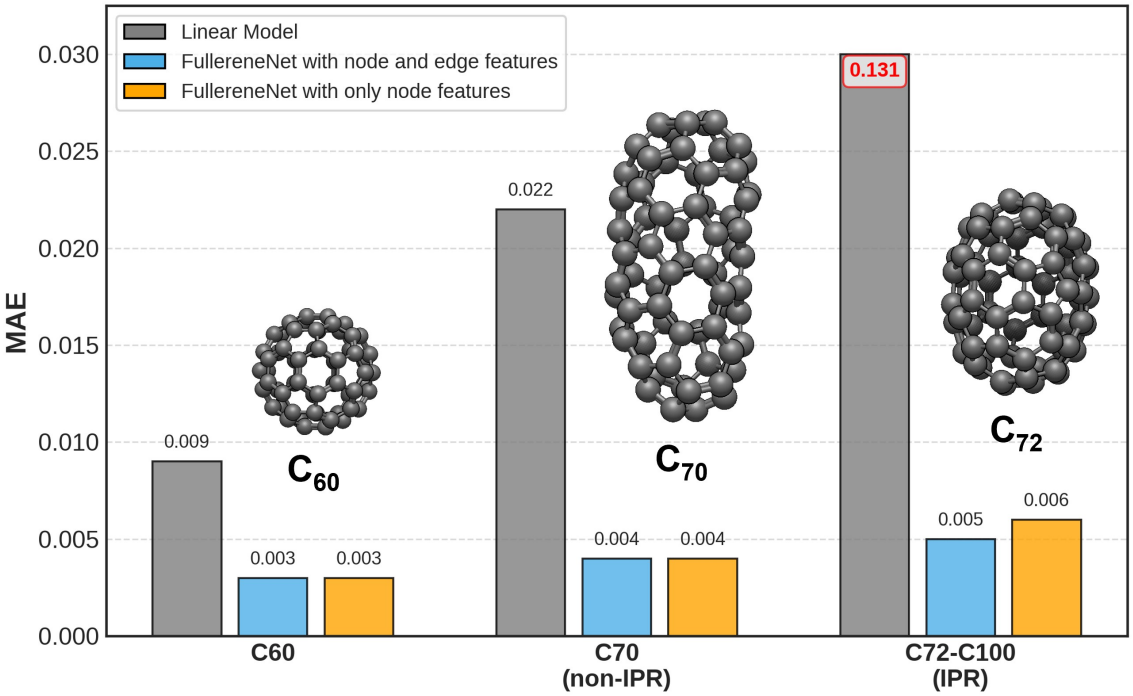

Figure 5: Comparison of binding energy prediction performance across diverse fullerene datasets. Three models are evaluated: a linear regression model (with notably high error values, shown in red), FullereneNet utilizing both node and edge features, and FullereneNet employing only node features.Inset molecular structures depict three representative fullerenes from our dataset: C_{60} isomer 1 (left), C_{70} isomer 11 (center), and C_{72} isomer 1 (right).

3.2 Feature analysis

Feature engineering remains a fundamental challenge in materials and molecules discovery, where the choice of structural representations significantly impacts model performance. For conventional GNN models designed for molecules, researchers typically utilize atom features such as element type and hybridization, along with bond features including bond types and distances [53]. However, these conventional features become inadequate for fullerene modeling, as they consist exclusively of sp^2-hybridized carbon atoms, resulting in uniform node features across the structure. In this study, we developed topologically-

View Article Online
DOI: 10.1039/D5DD00241A

informed node and edge features based on pentagon and hexagon arrangements and investigated whether these features can capture the structural nuances of fullerene systems composed of chemically identical atoms.

As detailed in Section 2.3, all node and edge features were derived from the adjacency matrix, whose elements reflect the connectivity between pairs of carbon atoms within a fullerene molecule (see Figure 1). Given that the connectivity among carbon atoms varies across different fullerene structures, each yields a distinctive adjacency matrix, thereby enabling a unique representation of each structure. However, when using only the adjacency matrix, along with Gaussian-random-sampled node and edge features as inputs, the GNN model demonstrates extremely low predictive capability (see Figure S6). These findings indicate that, while the node and edge features are derived from the adjacency matrix, they offer distinct structural dimensions crucial for interpreting structure–property relationships. Specifically, the adjacency matrix records atom-pair connectivity, capturing local connectivity within a fullerene molecule. In contrast, node features specify the types of three rings each atom shares, and edge features detail the types of four rings shared by each bond, providing semi-local structural information that cannot be effectively inferred by the GNN model but must be incorporated through human domain expertise. This strategic integration of chemical knowledge about ring types and arrangements enables our model to differentiate between carbon atoms that would otherwise appear indistinguishable, establishing a hierarchical representation that captures both local connectivity and higher-order topological patterns essential for stability prediction. Our findings highlight the crucial role of human domain knowledge in feature extraction and representation design, contributing to the development of more robust, reliable, and accurate ML models.

The manually derived node and edge features enable the GNN-based model, FullereneNet, to achieve superior performance in binding energy predictions, as demonstrated above. Since both node and edge features are derived from a single adjacency matrix to capture the semi-local chemical environment of pentagons and hexagons, we further evaluated the necessity of incorporating both feature types. To this end, we retrained the FullereneNet model using only node features to predict C–C binding energies across three test sets. The corresponding performance metrics are presented in Figure 4 and Table S5. Similar to the model trained with both node and edge features, the model utilizing only node features demonstrates excellent extrapolation performance across four training strategies, yielding an average R^2 of 0.989 and an MAE of 3 meV/atom for C_{60}, 0.974 and 4 meV/atom for C_{70}, and 0.364 and 6 meV/atom for C_{72}–C_{100}. The slightly reduced accuracy for the C_{72}–C_{100} test set can be attributed to its binding energy distribution, which falls outside the range of the training set (see Figure S1). These results suggest that since both node and edge features originate from the adjacency matrix, the inclusion of edge features offers minimal additional advantage in enhancing the extrapolation performance of GNN models when node features are already incorporated.

View Article Online
DOI: 10.1039/D5DD00241A

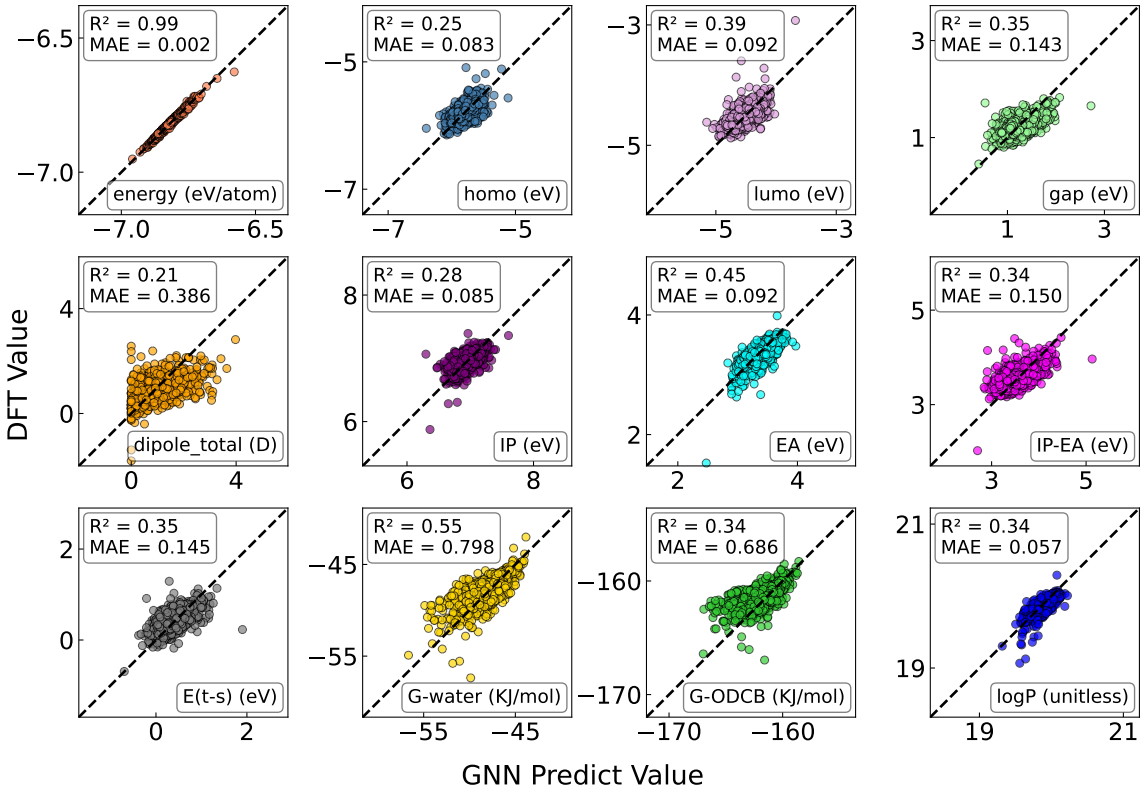

Figure 6: Extrapolation performance of FullereneNet on 12 properties in the C_{60} dataset, trained on data from C_{20} to C_{58}. X-axis represents the predicted values of various properties obtained using FullereneNet, while the Y-axis shows the corresponding DFT-calculated values. The labels' meaning for each subplot are summarized in Table S6.

3.3 Further Discussion

One of the key challenges in the computational design of molecules and materials is the high computational cost associated with structural optimization [54, 55]. In previous sections, we demonstrated that the FullereneNet achieves superior performance in predicting fullerene stability through carefully designed topological features rather than relying on optimized 3D Cartesian coordinates. This approach offers a substantial advantage by eliminating the computational cost associated with geometry optimization, which can be prohibitively expensive for high-throughput screening. To reinforce this advantage, we evaluated FullereneNet's predictive capability across a broader range of properties beyond stability. We tested FullereneNet on 11 other essential properties relevant for practical applications [56], including various electronic characteristics and solubility metrics, as summarized in Table S5.

First, we benchmarked the Matformer model using both unoptimized and optimized structural data as input. It is important to note that GAP-20 is limited to predicting the stability of fullerenes and cannot forecast other fundamental properties. As shown in Fig-

View Article Online
DOI: 10.1039/D5DD00241A

ure S7, the Matformer model struggles to accurately capture the structure–property relationships when using unoptimized structures on 11 properties. In contrast, structure optimization leads to significant improvements in both R^2 and MAE values. For example, in predicting the HOMO-LUMO gap, the R^2 value increased from -0.64 to 0.51, while the MAE decreased from 0.23 eV to 0.12 eV (Figures S7 and S8).

Subsequently, we retrained the FullereneNet model using both node and edge features derived from unoptimized structures and applied the model to extrapolate predictions for the C_{60} dataset. As illustrated in Figure 5 and Figure S8, FullereneNet achieves comparable predictions to Matformer with optimized structures for electronic properties while outperforming Matformer in solubility-related properties predictions with higher average R^2 values and lower MAE. Specifically, when predicting free solvation energies in water (ΔG_{sol}(water)) and 1,2-dichlorobenzene (ΔG_{sol}(ODCB)), and the 1,2-dichlorobenzene-water partition coefficient (logP), FullereneNet achieves MAE values of 0.80 kJ/mol, 0.69 kJ/mol, and 0.06, respectively. In contrast, the Matformer model trained on optimized structures yielded MAE values of 0.85 kJ/mol, 1.74 kJ/mol, and 0.26 (Figure S8). These results highlight the effectiveness of our feature design in capturing the chemical characteristics of fullerene systems, thereby enhancing the transferability of the FullereneNet model in predicting a diverse range of fundamental properties of fullerenes.

It is important to note that while FullereneNet achieves exceptional accuracy in predicting binding energies (R^2 = 0.99), its performance varies across other properties. For instance, electronic properties such as the HOMO-LUMO gap and electron affinity exhibit moderate predictive accuracy (R^2 = 0.35 and 0.45, respectively), indicating that these quantum mechanical properties are influenced by factors beyond the current topological descriptors. This observation is consistent with established chemical principles, as electronic properties often require more sophisticated quantum mechanical descriptors to accurately capture electron density distributions and orbital interactions [57]. Nonetheless, FullereneNet provides reliable predictions across multiple properties without costly geometric optimization, marking a significant advancement in high-throughput fullerene screening. By effectively balancing computational efficiency and predictive accuracy, our model enables the rapid identification of promising candidates for further computational or experimental validation. These findings also highlight the limitations of topological descriptors in capturing complex electronic properties while underscore the broader applicability of FullereneNet in efficient property prediction.

Conclusion

In this work, we developed a graph neural network (GNN)-based model, FullereneNet, to predict a wide range of fundamental properties of fullerenes using topological features derived from unoptimized structures. By leveraging the chemical environments of pentagons and hexagons within the fullerene cage, we demonstrated that these topological features efficiently capture the local structural details of fullerenes, enabling accurate prop-

View Article Online
DOI: 10.1039/D5DD00241A

erty predictions without the need for computationally expensive quantum chemistry optimizations. Our model significantly outperforms existing machine learning interatomic potentials GAP-20 and Matformer, achieving superior accuracy in predicting C–C binding energies across various fullerene sizes. Additionally, FullereneNet exhibits robust performance in predicting 11 other properties, including HOMO-LUMO gaps, solvation free energies, and partition coefficients, demonstrating its versatility and transferability. This study provides a computationally efficient framework for high-throughput screening of fullerene candidates, offering a valuable tool for advancing the exploration and application of fullerenes in various fields, from optoelectronics to materials science.

Author contributions

B.L. and M.L. initiated this study. B.L. conducted the theoretical calculations, while J.J. trained the machine learning models. All authors contributed to data analysis and manuscript writing.

Conflicts of interest

There are no conflicts to declare.

Data availability

The fullerene structures, including both unoptimized and optimized geometries, are available in the Zenodo repository [58].

Code availability: The Python implementation of FullereneNet is openly accessible on GitHub and Zenodo for long-term availability and reproducibility. [59]

Acknowledgements

This work was financially supported by the University of Florida's new faculty start-up funding. 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.

View Article Online
DOI: 10.1039/D5DD00241A

References

1. Chen M, Guan R, and Yang S. Hybrids of fullerenes and 2D nanomaterials. Advanced Science 2019;6:1800941.
2. Gaur M, Misra C, Yadav AB, Swaroop S, Maolmhuaidh FÓ, Bechelany M, et al. Biomedical applications of carbon nanomaterials: fullerenes, quantum dots, nanotubes, nanofibers, and graphene. Materials 2021;14:5978.
3. Thurston WP. Shapes of polyhedra and triangulations of the sphere. arXiv preprint math/9801088 1998.
4. Weaver J and Poirier D. Solid state properties of fullerenes and fullerene-based materials. In: *Solid State Physics*. Vol. 48. Elsevier, 1994:1–108.
5. Guldi DM and Martin N. Fullerenes: from synthesis to optoelectronic properties. Vol. 4. Springer Science & Business Media, 2002.
6. Kausar A. Breakthroughs of fullerene in optoelectronic devices—an overview. Hybrid Advances 2024:100233.
7. Thompson BC and Fréchet JM. Polymer–fullerene composite solar cells. Angewandte chemie international edition 2008;47:58–77.
8. Adisa OO, Cox BJ, and Hill JM. Methane storage in spherical fullerenes. Journal of Nanotechnology in Engineering and Medicine 2012;3:041002.
9. Mao D, Wang X, Zhou G, Chen L, Chen J, and Zeng S. Fullerene-intercalated graphene nanocontainers for gas storage and sustained release. Journal of Molecular Modeling 2020;26:1–6.
10. Goodarzi S, Da Ros T, Conde J, Sefat F, and Mozafari M. Fullerene: Biomedical engineers get to revisit an old friend. Materials Today 2017;20:460–80.
11. Castro E, Garcia AH, Zavala G, and Echegoyen L. Fullerenes in biology and medicine. Journal of Materials Chemistry B 2017;5:6523–35.
12. Vandewal K, Tvingstedt K, Gadisa A, Inganäs O, and Manca JV. On the origin of the open-circuit voltage of polymer–fullerene solar cells. Nature materials 2009;8:904–9.
13. Lee HKH, Telford AM, Röhr JA, Wyatt MF, Rice B, Wu J, et al. The role of fullerenes in the environmental stability of polymer: fullerene solar cells. Energy & environmental science 2018;11:417–28.
14. Prinzbach H, Weiler A, Landenberger P, Wahl F, Wörth J, Scott LT, et al. Gas-phase production and photoelectron spectroscopy of the smallest fullerene, C20. Nature 2000;407:60–3.
15. Piskoti C, Yarger J, and Zettl A. C36, a new carbon solid. Nature 1998;393:771–4.
16. Omer AM, Emara HM, Zaghloul RA, Abdel-Monem MO, and Dawwam GE. Research Journal of Pharmaceutical, Biological and Chemical Sciences. 2016.

View Article Online
DOI: 10.1039/D5DD00241A

17. Puente Santiago AR, Sanad MF, Moreno-Vicente A, Ahsan MA, Cerón MR, Yao YR, et al. A new class of molecular electrocatalysts for hydrogen evolution: Catalytic activity of M3N@ C2 n (2 n= 68, 78, and 80) fullerenes. Journal of the American Chemical Society 2021;143:6037–42.

18. Xie SY, Gao F, Lu X, Huang RB, Wang CR, Zhang X, et al. Capturing the labile fullerene [50] as C50Cl10. Science 2004;304:699–9.

19. Tan YZ, Liao ZJ, Qian ZZ, Chen RT, Wu X, Liang H, et al. Two I h-symmetry-breaking C60 isomers stabilized by chlorination. Nature Materials 2008;7:790–4.

20. Bille A, Buchstaber V, and Spodarev E. Some open mathematical problems on fullerenes. Journal of Chemical Information and Modeling 2025;65:2911–23.

21. Schwerdtfeger P, Wirz LN, and Avery J. The topology of fullerenes. Wiley Interdisciplinary Reviews: Computational Molecular Science 2015;5:96–145.

22. García-Risueño P, Armengol E, García-Cerdaña À, Garcia-Lastra JM, and Carrasco-Busturia D. Electron-vibrational renormalization in fullerenes through ab initio and machine learning methods. Physical Chemistry Chemical Physics 2024.

23. Liu M, Han Y, Cheng Y, Zhao X, and Zheng H. Exploring exohedral functionalization of fullerene with automation and Neural Network Potential. Carbon 2023;213:118180.

24. Aghajamali A and Karton A. Correlation between the energetic and thermal properties of C40 fullerene isomers: an accurate machine-learning force field study. Micro and Nano Engineering 2022;14:100105.

25. Rowe P, Deringer VL, Gasparotto P, Csányi G, and Michaelides A. An accurate and transferable machine learning potential for carbon. The Journal of Chemical Physics 2020;153.

26. Karasulu B, Leyssale JM, Rowe P, Weber C, and Tomas C de. Accelerating the prediction of large carbon clusters via structure search: Evaluation of machine-learning and classical potentials. Carbon 2022;191:255–66.

27. Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, et al. Geometry-enhanced molecular representation learning for property prediction. Nature Machine Intelligence 2022;4:127–34.

28. Ziaee SS, Rahmani H, Tabatabaei M, Vlot AH, and Bender A. DCGG: drug combination prediction using GNN and GAE. Progress in Artificial Intelligence 2024;13:17–30.

29. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Communications Materials 2022;3:93.

30. Choudhary K and DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Computational Materials 2021;7:185.

View Article Online
DOI: 10.1039/D5DD00241A

31. Simeon G and De Fabritiis G. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. Advances in Neural Information Processing Systems 2024;36.

32. Ramakrishnan R, Dral PO, Rupp M, and Lilienfeld OA von. Quantum chemistry structures and properties of 134 kilo molecules. Scientific Data 2014;1.

33. Yan K, Liu Y, Lin Y, and Ji S. Periodic graph transformers for crystal material property prediction. Advances in Neural Information Processing Systems 2022;35:15066–80.

34. Liu B, Jin J, and Liu M. Mapping structure-property relationships in fullerene systems: a computational study from C20 to C60. npj Computational Materials 2024;10:227.

35. Schwerdtfeger P, Wirz L, and Avery J. Program fullerene: a software package for constructing and analyzing structures of regular fullerenes. Journal of computational chemistry 2013;34:1508–26.

36. Yoshida M. C_n Fullerenes. Accessed: 2025-02-24. 2025. URL: `https://nanotube.msu.edu/fullerene/fullerene-isomers.html`.

37. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian˜16 Revision C.01. Gaussian Inc. Wallingford CT. 2016.

38. Becke AD. A new mixing of Hartree-Fock and local density-functional theories. Journal of chemical Physics 1993;98:1372–7.

39. Stephens PJ, Devlin FJ, Chabalowski CF, and Frisch MJ. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. The Journal of physical chemistry 1994;98:11623–7.

40. Lee C, Yang W, and Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Physical review B 1988;37:785.

41. Grimme S, Antony J, Ehrlich S, and Krieg H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. The Journal of chemical physics 2010;132.

42. Marenich AV, Cramer CJ, and Truhlar DG. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. The Journal of Physical Chemistry B 2009;113:6378–96.

43. Thompson AP, Aktulga HM, Berger R, Bolintineanu DS, Brown WM, Crozier PS, et al. LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. Computer physics communications 2022;271:108171.

44. Csányi G, Winfield S, Kermode JR, De Vita A, Comisso A, Bernstein N, et al. Expressive Programming for Computational Physics in Fortran 95+. IoP Comput. Phys. Newsletter 2007:Spring 2007.

View Article Online
DOI: 10.1039/D5DD00241A

45. Bartók AP, Payne MC, Kondor R, and Csányi G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys. Rev. Lett. 2010;104:136403.

46. Daigavane A, Ravindran B, and Aggarwal G. Understanding convolutions on graphs. Distill 2021;6:e32.

47. Sanchez-Lengeling B, Reif E, Pearce A, and Wiltschko AB. A gentle introduction to graph neural networks. Distill 2021;6:e33.

48. Liu Y, Yuan H, Wang Z, and Ji S. Global pixel transformers for virtual staining of microscopy images. IEEE Transactions on Medical Imaging 2020;39:2256–66.

49. Meredig B, Antono E, Church C, Hutchinson M, Ling J, Paradiso S, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Molecular Systems Design & Engineering 2018;3:819–25.

50. Zhao ZW, Del Cueto M, and Troisi A. Limitations of machine learning models when predicting compounds with completely new chemistries: possible improvements applied to the discovery of new non-fullerene acceptors. Digital Discovery 2022;1:266–76.

51. Wu J. Advances in K-means clustering: a data mining thinking. Springer Science & Business Media, 2012.

52. Aghajamali A and Karton A. Can force fields developed for carbon nanomaterials describe the isomerization energies of fullerenes? Chemical Physics Letters 2021;779:138853.

53. Chen C, Ye W, Zuo Y, Zheng C, and Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials 2019;31:3564–72.

54. Packwood D, Kermode J, Mones L, Bernstein N, Woolley J, Gould N, et al. A universal preconditioner for simulating condensed phase materials. The Journal of Chemical Physics 2016;144.

55. Chen M, Baer R, and Rabani E. Structure optimization with stochastic density functional theory. The Journal of Chemical Physics 2023;158.

56. Nelson J. Polymer: fullerene bulk heterojunction solar cells. Materials today 2011;14:462–70.

57. Li SC, Wu H, Menon A, Spiekermann KA, Li YP, and Green WH. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? Journal of the American Chemical Society 2024;146:23103–20.

58. Liu B, Jin J, and Liu M. Extrapolating Beyond C60: Advancing Prediction of Fullerene Isomers with FullereneNet. Zenodo 2025. DOI: 10.5281/zenodo.17400608.

View Article Online
DOI: 10.1039/D5DD00241A

59. Liu B, Jin J, and Liu M. Extrapolating Beyond C60: Advancing Prediction of Fullerene Isomers with FullereneNet. Version v1.0.0. 2025. DOI: 10.5281/zenodo.17426461. URL: https://doi.org/10.5281/zenodo.17426461.

View Article Online
DOI: 10.1039/D5DD00241A

Data availability statement

The fullerene structures, including both unoptimized and optimized geometries, are available in the Zenodo repository https://zenodo.org/records/17400608

Code availability: The Python implementation of FullereneNet is openly accessible on GitHub for long-term availability and reproducibility. https://github.com/Liu-Group-UF/FullereneNet and Zenodo: https://zenodo.org/records/17426461