Journal of Computational Biophysics and Chemistry

Research

Cite This: *J. Comput. Biophys. Chem.* **2023**, *22* (4), 503–511

Neighborhood Path Complex for the Quantitative Analysis of the Structure and Stability of Carboranes

Jian Liu[*,†], Dong Chen[‡,§], Feng Pan[‡,¶] and Jie Wu[†,¶]

[*]*School of Mathematical Sciences, Hebei Normal University, Heibei, 050024, P. R. China*
[†]*Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, P. R. China*
[‡]*School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, P. R. China*
[§]*Department of Mathematics, Michigan State University, MI, 48824, USA*

[¶]Corresponding authors. Emails: *panfeng@pkusz.edu.cn; wujie@bimsa.cn*

ABSTRACT: Thanks to the tremendous progress in data, computing power and algorithms, AI-based material mining and design have gained much attention. However, building high-performance AI models requires efficient material structure representation. In this work, we propose a structural characterization method based on the neighborhood path complex for the first time. Specifically, we use persistent neighborhood path homology to obtain the structural features by introducing a filtration. This approach preserves more elemental information, as well as the corresponding physicochemical information, through the directed edges of the neighborhood digraph. To validate our model, we perform cross-validation with the carborane structures. The Pearson coefficient for stability prediction is as high as 0.903, which is a 15.5% improvement compared to the traditional persistent homology method. In addition, we constructed a prediction model based on the neighborhood path complex, and the Pearson coefficients for the prediction of carboranes' HOMO, LUMO, and HOMO–LUMO gaps were 0.915, 0.946, and 0.941, respectively. The results show that our proposed method can effectively extract structural information and achieve accurate material property prediction.

KEYWORDS: Directed graphs; structure; neighborhood path complex; stability; carboranes.

1. INTRODUCTION

Feature engineering is of critical importance for artificial intelligence in areas such as materials design and new materials discovery. For these areas, the generalization ability of machine learning models is related to the high readings of the model inputs. For materials with physicochemical significance, extracting features with more intrinsic properties of the material can help machine learning models understand the intrinsic physical laws and thus produce more powerful models. Mathematical invariants extracted from frontier mathematics with topology, geometry, algebra, etc., can often describe the most intrinsic and fundamental properties of materials and structures. Feature engineering methods based on these advance mathematical tools have been successful in a variety of fields.[1,2]

A neighborhood complex,[3] roughly speaking, is a simplicial complex built on a graph with each simplex given by a collection of vertices that have a common neighbor. It can be used to study correlations between vertices with shared neighbors within a point set, and the neighborhood complex has also been applied in drug design-related fields through the persistent homology approach.[4] A variant of neighborhood complex, the neighborhood hypergraph was used to describe the structure and stability prediction of small fullerene molecules.[5] However, since the object studied in the theory of neighborhood complex is the vertex set and the elements within a vertex set are not different, the elemental information within materials can easily be weakened or even lost in the characterization process. Due to the path complex and path homology theory, named as GLMY theory, introduced by Grigor'yan *et al.*,[6–8] the relationship between different elements in a vertex set can be better handled, i.e. path, a kind of directed edge. The path complex

Received: 24 January 2023
Accepted: 8 February 2023
Published: 20 March 2023

DOI: 10.1142/S2737416523500229
J. Comput. Biophys. Chem. **2023**, *22* (4), 503–511

(which can usually be obtained from a directed graph) can characterize the difference between points by directed paths. The path complex shows application potential in molecular and materials sciences.[9] Therefore, as a generalization, neighborhood path complex has higher advantages for describing real material systems that contain different elements.

In this work, we propose a neighborhood path complex for the first time and extract structural features of materials using path homology theory, which is called neighborhood path homology. In addition, by introducing a filtration process, persistent neighborhood path homology can retain more geometric features, thus allowing the machine learning model to better predict the material function.

Original persistent homology has been successfully applied to the study of single-element clusters[10] and some multi-element clusters. However, in these studies, either the system has only one element, or the differences between different atoms are neglected during the analysis, and thus some critical physicochemical information is ignored. In this paper, we present a neighborhood path homology method and apply it for the first time to structural stability prediction and band gap prediction of closo-carboranes, which contain multiple elements.

As one of the stable cluster structures, carborane is a polyhedral cluster of borohydrides in which one or more borohydride B(H) vertices may be replaced by C(H) units.[11,12] The structure of such closo-carboboranes $C_2B_{n-2}H_n$ has been studied in both experimentally and theoretically.[13,14] In addition, the tendency of the stability of $C_2B_{n-2}H_n$ ($n = 5$–20) to vary with the increasing number of B atoms has attracted attention.[11,15–19] In this work, we use the $C_2B_{n-2}H_n$ ($n = 5$–20) cluster structure data set for the validation of the proposed NPH method. For carboranes, the basic idea is that both bonding and nonbonding in the structure can be used to estimate the orbital energy, and the relative stability of the structure can be directly expressed in terms of the relative energy. Therefore, in this work, the relative energy of the system is used to describe the stability of the structure,[15] and thus we achieve the prediction of structural stability by predicting the relative energy of the structure. Specifically, we used the persistent neighborhood path homology method to obtain the structural information of clusters, including zero- and one-dimensional topological information, and subsequently achieved accurate prediction of relative energy by linear regression model and boosted tree regression model, and successfully established the relationship between structural stability and path topological features. Through cross-validation, the method proposed in this work achieves accurate prediction of relative energy, HOMO, LOMO, and HOMO–LOMO gaps for the $C_2B_{n-2}H_n$ ($n = 5$–20) system, showing a very high correlation between the structural features extracted by PNPH and cluster properties.

2. THEORY AND METHODS

The functions and properties of molecules are determined by their structures. The simplicial complex has been used as a common mathematical model for portraying molecular structures. Another relatively new mathematical model for representing molecule is the digraph or the directed networks. In this work, we consider the neighborhood digraph of the structure, which indicates the interrelationship of atoms close to each other in a molecule. We adopt the path homology as the topological features for neighborhood digraphs to analyze the molecular structures. Furthermore, the topological persistence based on the path homology is applied to reduce the dimension of graph structure data.

2.1. The path homology theory

The GLMY theory introduced by Grigor'yan *et al.*[6–8] provides a tool for us to detect topological invariant on digraphs. The central idea for the GLMY theory is the path homology. We will briefly review the path homology for path complexes of digraphs. Let $G = (V, E)$ be a digraph, that is, a vertex set V equipped with an edge set $E \subseteq V \times V$. A simple digraph is a digraph with no loop or parallel edges, where a loop refers to the edge with the same starting and ending vertex, and parallel edges refer to those edges with the same direction between two vertices. In this paper, we will always consider the simple digraph.

An *elementary p-path* on V is a sequence of vertices $i_0 i_1 \dots i_p$ in V. A *path complex* on a finite set V is a collection $\mathcal{P}$ of elementary paths on V such that $i_0 i_1 \dots i_p \in \mathcal{P}$ implies $i_0 i_1 \dots i_{p-1}, i_1 i_2 \dots i_p \in \mathcal{P}$. An *allowed p-path* on G is an elementary p-path $i_0 i_1 \dots i_p$ on V such that $(i_{k-1}, i_k) \in E$ for $k = 1, 2, \dots, p$. Then the collection $\mathcal{P}(G)$ of allowed paths on G consists a path complex, called the path complex of the digraph G. From now on, the path complex considered is the path complex of a digraph.

Fix a field $\mathbb{K}$. Let $\Lambda_p = \Lambda_p(G; \mathbb{K})$ be the $\mathbb{K}$-linear space generated by all the elementary p-paths. We write the basis for $\Lambda_p(G; \mathbb{K})$ as $e_{i_0 i_1 \dots i_p}, i_0, i_1, \dots, i_p \in V$.

DOI: 10.1142/S2737416523500229

Then $(\Lambda_p)_{p\geq 0}$ can be regarded as a chain complex with the boundary operator

$$\partial_p e_{i_0 i_1 \ldots i_p} = \sum_{k=0}^{p} (-1)^k e_{i_0 i_1 \ldots \widehat{i_k} \ldots i_p}, \quad p \geq 1. \qquad (1)$$

Here, $\widehat{i_k}$ means that the index i_k is omitted. Moreover, we denote $\partial_0 e_{i_0} = 0$. Let $\mathcal{A}_p = \mathcal{A}_p(G; \mathbb{K})$ be the $\mathbb{K}$-linear space generated by the allowed p-path on G. In particular, $\mathcal{A}_0 = V$ and $\mathcal{A}_1 = E$. We have a chain complex $(\Omega_p)_{p\geq 0}$ with

$$\Omega_p = \partial_p^{-1}\mathcal{A}_{p-1} \cap \mathcal{A}_p = \{\sigma \in \mathcal{A}_p | \partial_p \sigma \in \mathcal{A}_{p-1}\}, \quad (2)$$

generated by the ∂-invariant paths. The *path homology* of G is

$$H_p(G; \mathbb{K}) := H_p(\Omega_*) = \frac{\ker \partial|_{\Omega_p}}{\mathrm{Im}\partial|_{\Omega_{p+1}}}, \quad p \geq 0. \qquad (3)$$

The pth Betti number β_p of a digraph G is defined as $\dim H_p(G; \mathbb{K})$. Roughly speaking, β_0 represents the number of connected components of G while β_1 gives the number of independent path cycles in G. The Betti numbers are the topological invariants for digraphs, which reveals the essential information of a graph. In this paper, our work is based on the digraph model for molecules, and the data features considered are built by the path homology of digraphs.

2.2. Neighborhood digraph and neighborhood path complex

The neighborhood complex[3] was first introduced by Lovász to deal with the Kneser's conjecture in 1978. For two reasons, we believe that neighborhood complex are useful for studying the structure of molecules. First, the neighborhood complex gives a description of the high-dimensional structure of graphs, which will help us understand the topology of a graph. Second, the neighborhood information reveals intrinsic connections between molecular bonds.

Let $G = (V, E)$ be a graph. The neighborhood complex[20] of G, denoted by $\mathcal{N}(G)$, is the abstract simplicial complex on V with simplices given by all the subsets of V that have a common neighbor. Mathematically, a simplex $\sigma \in \mathcal{N}(G)$ if and only if there exists a vertex $v \in V$ such that $(x, v) \in E$ for any $x \in \sigma$. As shown in Figs. 1(a) and 1(b), the neighborhood complex of a graph is a construction from a graph to a simplicial complex, which provides us with a topological view to analyze the properties of graphs.

In this paper, we consider the construction of digraphs such that the neighborhood path complex can be obtained. Let $G = (V, E)$ be a digraph. The *neighborhood digraph* $\mathcal{N}_D(G)$ of G is the digraph (V, E_D), where (i, j) is an edge in E_D if there exists a vertex $v \in V$ such that $(i, v), (v, j) \in E$. As shown in Fig. 1(c), from vertex 1, vertex 2 can be reached through vertex 3, then in neighborhood digraph (as in Fig. 1(d), vertex 1 and vertex 2 can be connected by a path. Note that a digraph is a generalization of a graph, a graph can be regarded as a digraph with bidirectional edges.

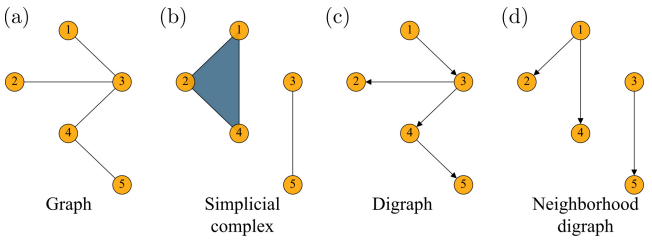

Fig. 1. (Color online) (a) The original graph G; (b) The neighborhood complex of G; (c) The directed graph (digraph) G'; (d) The neighborhood digraph of G'.

The following proposition shows that the neighborhood digraph is exactly the 1-skeleton of the neighborhood complex of the graph. This implies that the 0th Betti number of the neighborhood digraph coincides with the 0th Betti number of the neighborhood complex of a graph.

Proposition 1. *Let $G = (V, E)$ be a graph. Then the neighborhood digraph $\mathcal{N}_D(G)$ is the 1-skeleton $\mathrm{sk}_1\mathcal{N}(G)$ of the neighborhood complex of G.*

Indeed, if $\{i, j\} \in \mathrm{sk}_1\mathcal{N}(G)$, then there exists a vertex $v \in V$ such that $(i, v), (j, v) \in E$. Thus, there is a path ivj on G, which shows that $(i, j) \in E_D$. Here, E_D is the edge set of $\mathcal{N}_D(G)$. On the other hand, if $(i, j) \in E_D$, there exists a path ivj on G. This shows that i, j has the common neighbor. So we have $\{i, j\} \in \mathrm{sk}_1\mathcal{N}(G)$.

The neighborhood path complex is constructed on the neighborhood digraph of a graph. Precisely, let G be a digraph. The *neighborhood path complex* of G is the path complex $\mathcal{P}(\mathcal{N}_D(G))$. Moreover, by a straightforward verification, we have the following result.

Proposition 2. *Let G, H be digraphs. If $G \subseteq H$, then we have $\mathcal{N}_D(G) \subseteq \mathcal{N}_D(H)$ and $\mathcal{P}(\mathcal{N}_D(G)) \subseteq \mathcal{P}(\mathcal{N}_D(H))$.*

This result says that the construction of neighborhood path complexes preserves the filtrations of digraphs. The neighborhood path complex is essentially from the construction of the neighborhood complex. As we see, the construction of neighborhood path complex can turn a long path on a digraph into a

DOI: 10.1142/S2737416523500229
J. Comput. Biophys. Chem. 2023, 22 (4), 503–511

shorter path. Thus, we prefer to consider the path homology of the neighborhood digraph rather than the original digraph. This will enrich the information of low-dimensional path complex. It is very useful for us to have more low-dimensional information since we rarely calculate the higher-order path homology.

2.3. Topological persistence on neighborhood path complexes

Recall that the persistent homology is always obtained from a real-valued functions on a simplicial complex. The usual constructions of simplicial complexes are the Čech complex[21] and the Vietoris–Rips complex.[22] In this section, we obtain a filtration of digraphs from a data set embedded in a Euclidean space. Information about the position of atoms in a molecule can be inscribed, to a large extent, by Euclidean distances. Our method is based on the filtration of digraphs, which further provides us with topological persistence on neighborhood path complexes.

Let $G = (V, E)$ be a digraph embedded into a Euclidean space $\mathbb{R}^n$. We have a real-valued function $f : E \to \mathbb{R}$ on the edge set E of G given by

$$f(i,j) = \|i - j\|, \quad (i,j) \in E. \tag{4}$$

Here, $\|\cdot\|$ denote the Euclidean norm. For each positive number $t > 0$, we have a digraph $G_t = (V, E_t)$ parametrized by t, where $E_t = \{e \in E | f(e) \leq t\}$. For $a \leq b$, there is a natural inclusion $G_a \hookrightarrow G_b$. Thus, we obtain a filtration of digraphs $\{G_t\}_{t\geq 0}$. By Proposition 2, we have a new filtration of neighborhood digraphs $\{\mathcal{N}_D(G_t)\}_{t\geq 0}$. The (a, b)-persistent neighborhood path homology for the digraph G is given by

$$H_p^{a,b}(G;\mathbb{K}) = \mathrm{Im}(H_p^a(G;\mathbb{K}) \hookrightarrow H_p^b(G;\mathbb{K})), \quad a \leq b. \tag{5}$$

Here, $H_p^t(G;\mathbb{K})$ denotes the neighborhood path homology of G_t for $t \geq 0$. Moreover, the pth (a, b)-persistent Betti number $\beta_p^{a,b} = \dim H_p^{a,b}(G;\mathbb{K})$ is the dimension of the (a, b)-persistent path homology. For a specific molecule, such as a hydrocarbon molecule, we can obtain a digraph $G = (V, E)$ as follows. The vertex set V is given by the atoms. For every two atoms, we have directed edges given by the rule: $\mathrm{H} \to \mathrm{C}$, $\mathrm{C} \rightleftarrows \mathrm{C}$, $\mathrm{H} \rightleftarrows \mathrm{H}$. With this rule, we can obtain a digraph, and the atoms are canonical embedded into the usual three-dimensional Euclidean space. For example, consider the benzene ring C_6H_6. When the distance parameter is $t = 1.5$ Å, we have a digraph G given as Fig. 2(b) The corresponding neighborhood digraph of G is given in Fig. 2(c), on which the corresponding neighborhood path complex is built.

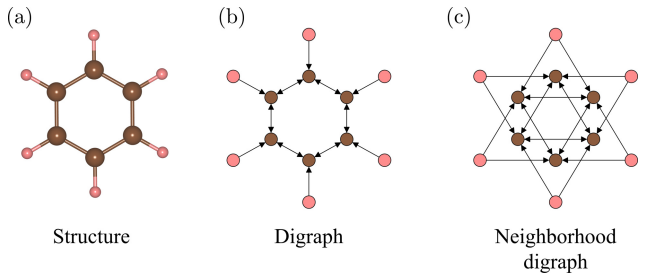

Fig. 2. (Color online) (a) The molecular structure of benzene; (b) The digraph representation of benzene; (c) The neighborhood digraph of benzene.

2.4. Euclidean-distance-based persistent neighborhood path homology

For a given structure, which contains a set of atoms, the pairwise distance in Euclidean space can be calculated and stored in a distance matrix $M_{n\times n}$, where n is the number of points. Pairwise atomic distance d_{ij} can be calculated as Eq. (4). As the filtration parameter d increases, a binary matrix $C_{n\times n}$ can be generated from the distance matrix $M_{n\times n}$ with the i-row and j-column elements are noted as p_{ij}. We denote $p_{ij} = 1$ when $d_{ij} \leq d$, indicating the existence of a path from node i to node j. On the other hand, we write $p_{ij} = 0$ when $d_{ij} > d$, indicating that no path exists. In addition, to exclude the loop path, we set $p_{ij} = 0$ when $i = j$.

To construct the digraph directly from the structure, we visualize the change of digraph during the growth of filtration parameter d by using a circle with vertex as the center and filtration parameter d as the radius. Figure 3 shows the persistent neighborhood path homology for benzene, when the circles overlap, i.e. $d_{ij} < d$, the path can be generated and the corresponding digraph and neighborhood digraph can be constructed. In this work, the direction of the path is determined by the electronegativity of the atom. The atom with small electronegativity points to the atom with large electronegativity, while when two atoms have the same electronegativity, they can be connected by two paths with opposite directions.

2.5. Structural stability prediction modeling

From the previous presentation, we can easily obtain information about the structure from the proposed persistent neighborhood path homology (PNPH), which includes topological information and geometric

DOI: 10.1142/S2737416523500229

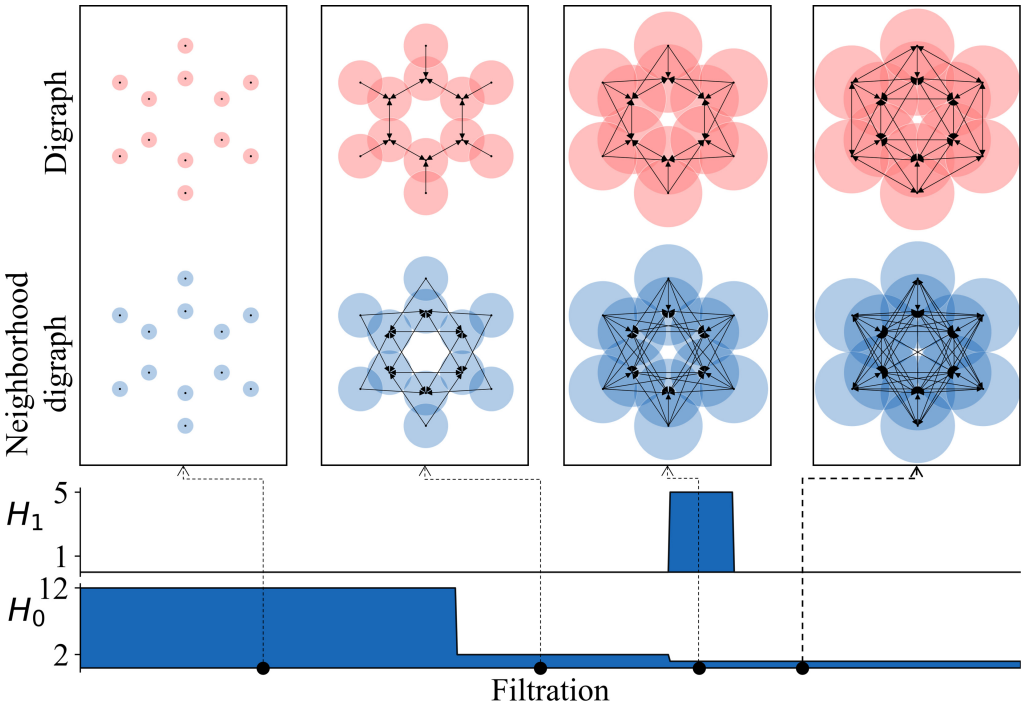

Fig. 3. (Color online) Persistent neighborhood path homology for benzene. As the filtration parameter increases, the circle with these points as the center grows and the overlap of the circles is formed with corresponding paths. The direction of the path is determined by the electronegativity of the element represented by each point. Note that the small electronegativity is defined to point to the large electronegativity.

information. Furthermore, since the molecular structure determines the function of the molecule, such as stability,[23–25] this provides the motivation to construct model to represent a structure–molecule function relationship.

In this work, we demonstrate the application of the proposed PNPH method in predicting the stability of *closo*-carboranes $C_2B_{n-2}H_n$ ($n = 5$–20, 16 structures in total). The relative energy, which indicates the stability of the structure, was used as the target. It is calculated based on the difference in single-point energies. The relative energy of $C_2B_{n-2}H_n$ is calculated by using the BH addition energy,[15] as in the following equation:

$$\Delta H_{\text{add}} = E(C_2B_{n-2}H_n) - E(C_2B_3H_5) - (n-5)E(\text{BH}), \quad (6)$$

where E(BH) is the difference in energy between B_6H_{10} and B_5H_9. The energy of $C_2B_3H_5$ is set as the baseline value. The most stable structure was used for this task and reoptimized by the DFT method. In addition, we performed data augmentation and randomly generated a total of 2193 $C_2B_{n-2}H_n$ ($n = 5$–20) structure data and calculated the related HOMO value, LUMO value, and the difference between them, i.e. HOMO–LUMO gap value, for each structure.

Specifically, the first task is to validate the proposed PNPH method using the most stable structure of $C_2B_{n-2}H_n$ ($n = 5$–20). To establish the relationship between the features generated by the PNPH method and the stability, here, we use the information of H_0. First, the feature is the value of the filtration parameter corresponding to the first decrease of the dimension of H_0, denoted as F_1, which is associated with the earliest 2-path in the structure. The second feature is the filtration parameter corresponding to the dimension of H_0 becomes 1, denoted as F_2. This value indicates that any two atoms in the structure can be linked by another atom, and the electronegativity of the neighbor atom will be between these two atoms. The third feature chosen is the average area per atom, which is the ratio of the area S enclosed by the H_0 curve, e.g. the blue-shaded area shown in Fig. 3, to the number of atoms N, denoted as F_3. To emphasize the relationship between structural stability and the PNPH features presented above, we performed leave-one-out cross-validation using linear regression,[26] where just three features, F_1, F_2, F_3, were used as inputs to the model. For the linear regression model, all the predictions will come out within 1 s.

The second task is to predict the HOMO, LUMO, and HOMO–LUMO gaps of the structure. Here, H_0

DOI: 10.1142/S2737416523500229
J. Comput. Biophys. Chem. **2023**, *22* (4), 503–511

and H_1 information are used for modeling. Specifically, we use the filtration parameters every 0.1 Å from 0 to 5 Å and calculate the corresponding zero- and one-dimensional path homology to produce a feature vector of dimension 100. We use the gradient-boosting regressions model for 5-fold cross-validation. In this work, the learning_rate of GBRT is set to 0.005, max_depth is set to 7, min_samples_leaf is set to 1, min_samples_split is set to 5, subsample is set to 0.4, and n_estimators is set to 10000. For the three prediction tasks HOMO, LUMO and HOMO–LUMO gap, the program runs on Intel(R) Core(TM) i7-8565U CPU and can be finished in 50 s, 46 s and 50 s, respectively.

To quantify the prediction accuracy of the model, in this paper, we use the Pearson correlation coefficient to measure the correlation between the predicted results and the results of the DFT calculation, defined as follows:

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (7)$$

where x_i, y_i are the prediction result and the quantum mechanic calculation result, respectively, and $\bar{x}, \bar{y}$ are the average predicted result and the average value of the quantum mechanic calculation result, respectively. In addition to this, mean absolute error (MAE) and root mean square error (RMSE)are also used to validate the performance.

2.6. Computational methods

All molecular structures used in this work are calculated by the Gaussian16 program.[27] The stable structures of closo-carboranes $C_2B_nH_{n+2}$ ($n = 5$–20) and closo-borane dianions were optimized at the B3LYP/6-311G(d,p) level of theory.[28] In addition, the single point energies are used to measure the relative energies of different structures. The VESTA software[29] is used for the presentation of molecular structures.

3. RESULTS AND DISCUSSION

3.1. PNPH-based stability analysis for closo-carboranes

In this work, both the persistent homology (PH) and the proposed PNPH method are used to predict the relative energies of closo-carboranes. For $C_2B_{n-2}H_n$ ($n = 5$–20), for each value of n, we have one most stable structures and its corresponding relative energy value. Thus, a total of 16 closo-carboranes structures were used to validate the PH method and the PNPH method. Specifically, for the PNPH method, as mentioned in the theory section, we use only three features extracted from H_0, namely, F_1, F_2, F_3, as inputs to the model. To demonstrate the extremely high correlation between the PNPH-based features and the structural stability, the model we utilize is a simple linear regression model, i.e. $y = c_0 + c_1F_1 + c_2F_2 + c_3F_3$, where $c_i, i = 0, 1, 2, 3$ are coefficients. For the PH method, we use the same workflow to extract three features from zero-dimensional homology. Thus, the first feature, denoted F_1^{PH}, is the value of the filtration parameter corresponding to when the dimension of H_0 is reduced for the first time. The second feature, denoted F_2^{PH}, is the filtration parameter corresponding to the time when the dimension of H_0 becomes 1. The third feature, denoted as F_3^{PH}, is the value of the area enclosed by the H_0 curve spread equally to each atom, i.e. the ratio of the area S enclosed by the H_0 curve to the number of atoms N. Furthermore, we utilize a leave-one-out cross-validation method to validate the PH and PNPH methods.

Figure 4(a) shows the leave-one-out cross-validation results of the PH-based model. The Pearson correlation coefficients of the predicted results and the true values are 0.782, RMSE is 0.893, and MAE is 0.660. Among all the structures, there are four clusters whose predicted values differ significantly from the true values, namely, $B_4C_2H_6$, $B_7C_2H_9$, $B_8C_2H_{10}$, and $B_{13}C_2H_{15}$. The structures of these four clusters are shown in Fig. 4(c). The difference between the predicted and true values of these four structures is greater than 0.85 eV. Figure 4(b) shows the results of the leave-one-out cross-validation based on the PNPH model. The final correlation coefficients of the PNPH-based linear model predictions with the true values are 0.903, RMSE 0.596, and MAE 0.487. Compared with the results of the PH method, the Pearson correlation coefficient improves by 15.5%, RMSE decreases by 33.2%, and MAE decreases by 26.2%. It is noteworthy that the four structures that were not well predicted by the PH method were well predicted by the PNPH method, as shown in Fig. 4(b). The digraph representation of these four structures and the neighborhood digraph representation are shown in Figs. 4(e) and 4(f). Compared with the original graph representation (Fig. 4(d), the digraph in which the differences between the vertices are encoded by path, and the neighborhood digraph constructs a richer relationship between the vertices. It can be found that in Fig. 4(f), the vertices representing carbon atoms are always isolated from their nearest neighboring vertices representing hydrogen atoms, which means the PNPH method obtains the information on the location of the key C elements. The above results show that the PNPH method can obtain information on different elements of multi-element

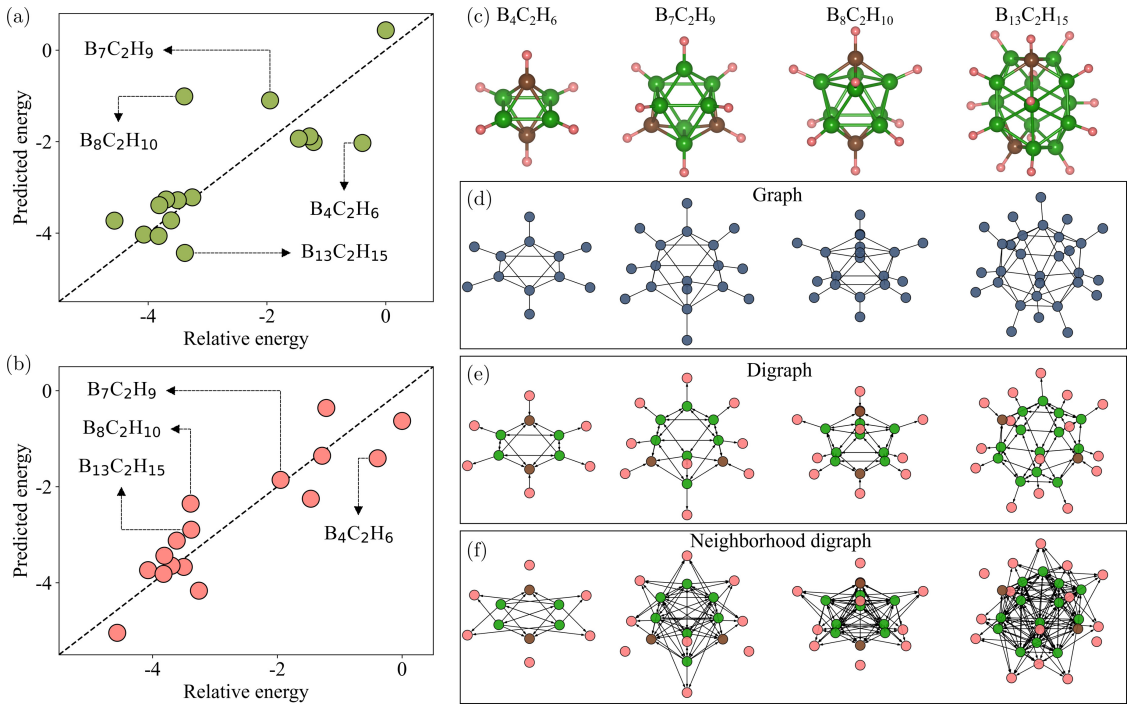

Fig. 4. (Color online) (a) Comparison between quantum mechanic method calculated relative energies (eV) and the persistent homology-based model predicted results. (b) Comparison between quantum mechanic method calculated relative energies (eV) and the PNPH-based model predicted results. (c) The structures of the four clusters $B_4C_2H_6$, $B_7C_2H_9$, $B_8C_2H_{10}$, and $B_{13}C_2H_{15}$. (d, e and f) The graph, digraph, and neighborhood digraph representations of four clusters $B_4C_2H_6$, $B_7C_2H_9$, $B_8C_2H_{10}$, and $B_{13}C_2H_{15}$.

clusters, which is superior to the conventional PH method. In addition, the results also demonstrate that the features based on the PNPH method are highly correlated with structural stability.

3.2. PNPH-based prediction of HOMO, LUMO, and HOMO–LUMO gap

To further validate the features based on the PNPH method, we predicted the HOMO, LUMO, and HOMO–LUMO gaps of closo-carboranes. First, we extracted the features of zero-dimension and one-dimension path homotopy of 2319 closo-carboranes structures using PNPH method and modeled them based on the gradient boosting tree algorithm's, and finally verified them by 5-fold cross-validation method, as shown in Fig. 5. In this work, we calculated the HOMO values, LUMO values, and HOMO–LUMO gaps for all structures using quantum mechanic methods and regard them as the true values. Figure 5(a) shows the

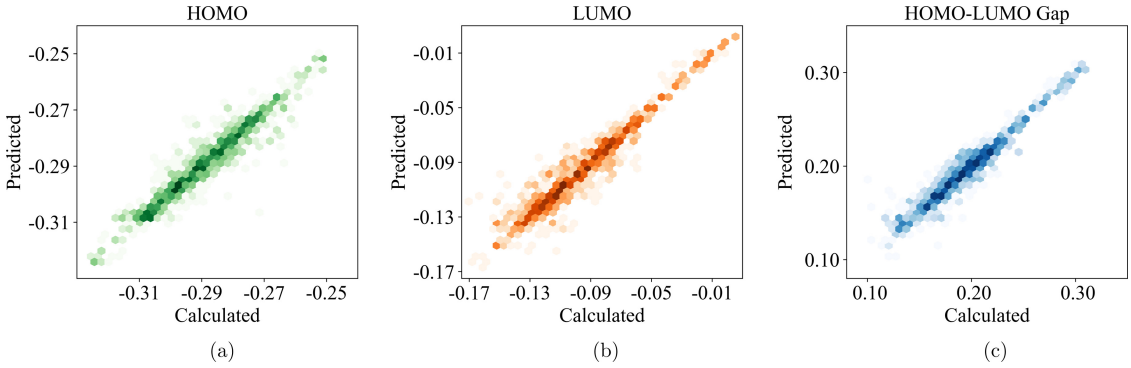

Fig. 5. (Color online) (a) Comparison of quantum mechanic-based method calculated HOMO values and PNPH-based model predicted HOMO. (b) Comparison of quantum mechanic-based method calculated LUMO values and PNPH-based model predicted LUMO. (c) Comparison of quantum mechanic-based method calculated HOMO–LUMO gaps and PNPH-based model predicted HOMO–LUMO gaps.

DOI: 10.1142/S2737416523500229
J. Comput. Biophys. Chem. **2023**, *22* (4), 503–511

comparison of the predicted HOMO values with the true values, which ends up with a Pearson correlation coefficient of 0.956, RMSE of 0.007, and MAE of 0.004. Figure 5(b) shows the comparison of the predicted LUMO values with the true values, which has a Pearson correlation coefficient of 0.973, RMSE of 0.007, and MAE of 0.004. And for the HOMO–LUMO differences, i.e. HOMO–LUMO gaps, which are usually used to characterize the conductive properties of the material. The final Pearson correlation coefficient of the PNPH-based method with the true value is 0.976, the RMSE is 0.002, and the MAE is 0.004. The above results show that the PNPH-based machine learning model exhibits excellent performance on the HOMO, LUMO, and HOMO–LUMO gaps predictions, illustrating the powerful capability of PNPH. In this work, we focus on demonstrating the proposed PNPH approach and its advantages over the original approach. The results of both tasks show that the PNPH method has great potential for material stability prediction, material function prediction, and can help drive the design and discovery of new materials.

4. CONCLUSION

In this work, we proposed a neighborhood path homology (NPH) method and introduced a filtration process to achieve the function of extracting structural features in multiple scales. In the method section, we introduced the concept and construction of NPH. To validate the proposed method, we apply the NPH method to the analysis of closo-carboranes structures. In the first task, we achieved a prediction accuracy of Pearson correlation coefficient of 0.903 using only three features extracted from the zero-dimensional path homology as input to the linear regression. As a comparison, we also followed the same procedure to validate the traditional persistent homology method, and the final prediction result was only 0.782. The prediction accuracy of the NPH method proposed in this work was improved by 15.5%. It shows that the inclusion of direction information in the path achieves the capture of information about different elements in the structure, thus improving the prediction results. In the second task, we also extracted the zero- and one-dimensional structural features by the PNPH method, and combined with the GBRT algorithm, we achieved the prediction of HOMO, LUMO, and HOMO–LUMO gaps for the closo-carboranes $C_2B_{n-2}H_n$ ($n = 5$–20) with Pearson correlation coefficients of 0.915, 0.946 and 0.941, respectively. In this work, we demonstrate the feasibility of the proposed PNPH method on the multi-element system with satisfactory prediction accuracy. In future work, by incorporating more sophisticated machine learning models, PNPH can be applied to more complex multi-element systems to achieve fast and accurate predictions, providing a feasible path for structure design and new material discovery.

FUNDING INFORMATION

The work of Wu and Liu was supported by Natural Science Foundation of China (NSFC grant no. 11971144), High-level Scientific Research Foundation of Hebei Province, the start-up research fund from BIMSA. The work of Pan and Chen was financially supported by the Shenzhen Science and Technology Research Grant (No. JCYJ20200109140416788), the Soft Science Research Project of Guangdong Province (No. 2017B030301013) and the Major Science and Technology Infrastructure Project of Material Genome Big-science Facilities Platform supported by Municipal Development and Reform Commission of Shenzhen.

CONFLICT OF INTEREST

The authors declare no competing interests.

References

1. Nguyen, D. D.; Cang, Z.; Wei, G.-W. A Review of Mathematical Representations of Biomolecular Data. *Phys. Chem. Chem. Phys.* **2020**, *22* (8), 4343–4367.
2. Chen, D.; Gao, K.; Nguyen, D. D.; Chen, X.; Jiang, Y.; Wei, G.-W.; Pan, F. Algebraic Graph-Assisted Bidirectional Transformers for Molecular Property Prediction. *Nat. Commun.* **2021**, *12* (1), 1–9.
3. Lovász, L. Kneser's Conjecture, Chromatic Number, and Homotopy. *J. Combin. Theory, Ser. A* **1978**, *25* (3), 319–324.
4. Liu, X.; Xia, K. Neighborhood Complex based Machine Learning (NCML) Models for Drug Design, *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*, Springer, 87–97, 2021.
5. Liu, J.; Chen, D.; Li, J.; Wu, J. Neighborhood Hypergraph Model for Topological Data Analysis. *Comput. Math. Biophys.* **2022**, *10* (1), 262–280.
6. Grigor' yan, A.; Lin, Y.; Muranov, Y. Y.; Yau, S.-T. Path Complexes and Their Homologies. *J. Math. Sci.* **2020**, *248* (5), 564–599.
7. Grigor'yan, A.; Lin, Y.; Muranov, Y.; Yau, S.-T. Homologies of Path Complexes and Digraphs. arXiv:1207.2834.

8. Grigor'yan, A.; Muranov, Y.; Yau, S. T. Homologies of Digraphs and Künneth Formulas. *Commun. Anal. Geom.* **2017**, *25* (5), 969–1018.
9. Chen, D.; Liu, J.; Wu, J.; Wei, G.-W.; Pan, F.; Yau, S.-T. Path Topology in Molecular and Materials Sciences. *J. Phys. Chem. Lett.* **2023**, *14*, 954–964.
10. Xia, K.; Feng, X.; Tong, Y.; Wei, G.-W. Persistent Homology for the Quantitative Prediction of Fullerene Stability. *J. Comput. Chem.* **2015**, *36* (6), 408–422.
11. Schleyer, P. V. R.; Najafian, K. Stability and Three-Dimensional Aromaticity of Closo-Monocarbaborane Anions, $CB_{n-1}H_{n-}$, and Closo-Dicarboranes, $C_2B_{n-2}H_n$. *Inorg. Chem.* **1998**, *37* (14), 3454–3470.
12. Williams, R. E. Early Carboranes and Their Structural Legacy. *Adv. Organometal. Chem.* **1994**, *36*, 1–55.
13. Zhang, J.; Xie, Z. Synthesis, Structure, and Reactivity of 13-and 14-Vertex Carboranes. *Acc. Chem. Res.* **2014**, *47* (5), 1623–1633.
14. Qiu, Z.; Ren, S.; Xie, Z. Transition Metal-Carboryne Complexes: Synthesis, Bonding, and Reactivity. *Acc. Chem. Res.* **2011**, *44* (4), 299–309.
15. Zhang, J.; Lin, Z.; Xie, Z. DFT Studies on Structures, Stabilities, and Electron Affinities of Closo-Supercarboranes $C_2B_{(n-2)}H_n$ ($n = 13$–20). *Organometallics* **2015**, *34* (23), 5576–5588.
16. Wales, D. J.; Michael, M. M.; Mingos, P.; Slee, T.; Lin, Z. Clusters in Inorganic and Molecular Beam Chemistry. Some Unifying Principles. *Acc. Chem. Res.* **1990**, *23* (1), 17–22.
17. Diaz, M.; Jaballas, J.; Tran, D.; Lee, H.; Arias, J.; Onak, T. Interaction of Trimethylamine and Closo-1, 6-$C_2B_7H_9$. Evidence for an "open" Cage $C_2B_7H_9$/Amine Adduct. *Inorg. Chem.* **1996**, *35* (16), 4536–4540.
18. Buhl, B.; Gauss, J.; Hofmann, M.; Schleyer, P. v. R. Decisive Electron Correlation Effects on Computed Boron-11 and Carbon-13 NMR Chemical Shifts. Application of the GIAO-MP2 Method to Boranes and Carbaboranes. *J. Am. Chem. Soc.* **1963**, *115* (26), 12385.
19. McKee, M. L. Theoretical Study of the Reaction of Acetylene with B4H8. A Proposed Mechanism of Carborane Formation 2. *J. Am. Chem. Soc.* **1996**, *118* (2), 421–428.
20. Kozlov, D. *Combinatorial Algebraic Topology*, Vol. 21, Springer Science & Business Media, 2008.
21. Carlsson, G. Topology and Data. *Bull. Am. Math. Soc.* **2009**, *46* (2), 255–308.
22. Vietoris, L. Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen. *Math. Ann.* **1927**, *97* (1), 454–472.
23. Xia, K.; Opron, K.; Wei, G.-W. Multiscale Multiphysics and Multidomain Models–Flexibility and Rigidity. *J. Chem. Phys.* **2013**, *139* (19), 11B614_1.
24. Xia, K.; Wei, G.-W. Stochastic Model for Protein Flexibility Analysis. *Phys. Rev. E* **2013**, *88* (6), 062709.
25. Xia, K.; Wei, G.-W. Molecular Nonlinear Dynamics and Protein Thermal Uncertainty Quantification. *Chaos* **2014**, *24* (1), 013103.
26. Wong, T.-T. Performance Evaluation of Classification Algorithms by k-Fold and Leave-One-Out Cross Validation. *Pattern Recogn.* **2015**, *48* (9), 2839–2846.
27. Frisch, M. E.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V. P. G. A.; Petersson, G. A.; Nakatsuji, H. J. R. A.; Li, X. Gaussian 16, **2016**.
28. Becke, A. *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design*. John Wiley & Sons, 2007.
29. Momma, K.; Izumi, F. Vesta 3 for Three-Dimensional Visualization of Crystal, Volumetric and Morphology Data. *J. Appl. Crystallogr.* **2011**, *44* (6), 1272–1276.

DOI: 10.1142/S2737416523500229
J. Comput. Biophys. Chem. **2023**, *22* (4), 503–511