Journal of Computational Biophysics and Chemistry

Research

Cite This: *J. Comput. Biophys. Chem.* **2025**, *24* (1), 1–11

Persistent Path Homology for Quantitative Analysis of Carboranes

Bingxu Wang, Mingzheng Zhang and Feng Pan*

School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, P. R. China

*Corresponding author. E-mail: *panfeng@pkusz.edu.cn*

ABSTRACT: Persistent path homology represents an advanced mathematical approach explicitly tailored for directed systems. With its remarkable ability to characterize unbalanced or asymmetrical relationships within data, this method demonstrates great promise in qualitative analysis of the intrinsic topological features present in materials and molecules. In this work, we introduce persistent path homology at the first time for carborane analysis. Intrinsic path topological features are used to predict the stability of closo-carboranes. We qualitatively explain the connection between path topological features and properties on o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$. The correlation coefficients between linear predictions based on persistent path homology and thermodynamic stability are higher than 0.95, and that for chemical stability are about 0.85. While the correlation coefficients based on nonlinear models are increased to 0.99 and 0.95, respectively. These results indicate that persistent path homology shows excellent capabilities in structural and stability analysis of multi-element cluster physics.

KEYWORDS: Filtration; topology; mathematical methods; cluster structure; multi-element system.

1. INTRODUCTION

Carborane is a class of polyhedral borohydride clusters, in which one or more borohydride B(H) vertices are replaced by C(H) units.[1] The bonding theories and structure–energy relationship of carborane clusters, specifically carborane derivatives,[2,3] have shown great potential in such diverse areas as catalysis (including metal catalyst[4] and C–F, C–H and C–C functionalizations[5]), medicine (including the diagnosis of cancer[6] and neutron capture therapy[7]), nonlinear optical materials,[8] conducting organic polymers (including solution-based and solid-state electrolytes[5]), coordination polymers,[9] and other diverse fields.[10,11] The quantitative analysis of the stability of the ground-state structure of carborane clusters is currently a subject of intensive research in cluster physics. For example, the stabilities of closo-carboranes were compared and analyzed employing comprehensive ab initio calculations.[10] With the aid of DFT calculations, the difficulty of the experimental synthesis of closo-carboranes was explained by the value of the HOMO-LUMO gaps.[12] Although these quantum computations are classical tools in the field of cluster physics, they still have many limitations such as time consumption and high computational cost, and they become difficult to handle as the number of atoms in the system increases. Closo-carboranes, $C_2B_nH_{n+2}$, the first kind of carboranes discovered by humans,[11] have been favored by scholars both experimentally and theoretically.[13,14] Moreover, the stability of closo-carboranes has attracted extensive attention with the increase in the number of B atoms.[10,12,15]

An emerging family of data analysis methods called Topological Data Analysis (TDA) enables the simplification of complex data by combining ideas from algebraic topology[16] with other mathematical tools, such as neighborhood complex,[17] Morse theory[18] and hypergraph[19] and has yielded compelling results on problems in biological systems. Among them, persistent homology has been developed as a powerful tool to capture geometric and topological information

Received: 6 June 2024
Accepted: 6 July 2024
Published: 22 November 2024

DOI: 10.1142/S2737416524500339
J. Comput. Biophys. Chem. **2025**, *24* (1), 1–11

of data in changing scales[20–22] and has been applied in various fields.[23–28] Especially in the field of materials science, persistent homology has been extensively employed for property prediction of amorphous solids,[29] halide perovskites[30] and other substances.[31]

In recent years, it has been successful in the quantitative analysis of cluster structures, for example, by obtaining topological fingerprints of different structural configurations analyzed to quantitatively predict the stability of single-element clusters such as fullerenes.[32] Another result is the use of persistent homology methods to obtain topological fingerprints of a large number of different cluster configurations, which reveal the hidden structure–energy relationships in single-element lithium clusters by combining them with machine learning models.[33] Persistent homology has become a powerful method for analyzing single-element cluster structures.

Despite its successful application in various fields, persistent homology theory still cannot directly deal with multi-element systems. For example, carboranes are heteronuclear molecules, which can be formed by different elements. Since different types of atoms have different properties, polarizability becomes more and more apparent. However, the classical theory of persistent homology, which treats atoms as equivalent vertices, cannot cope with the asymmetry (polarizability) or unequal interactions between different types of atoms in real systems. Persistent path homology (PPH) designed for directed graphs was proposed by Grigor'Yan *et al.*[34–36] as a promising approach for analyzing protein-ligand interactions,[37] complex diseases[38] and molecules.[39] In our work, to apply PPH to the quantitative analysis of the stability of multi-element carborane clusters, each structure is represented by a directed graph in which the atoms are weighted vertices. In this process, symmetric bidirectional connections are established when describing interactions between atoms of the same type, while weight-based asymmetric unidirectional connections are established when describing interactions between atoms of different types.

In this paper, a quantitative analysis of the structure of $C_2B_nH_{n+2}(n=2\sim22)$ is performed by using PPH. First, the distance-based persistent path homology (DPPH) and the angle-based persistent path homology (APPH) complement each other to give the path homology fingerprints of each multi-element structure. Next, taking o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$ as examples, we distinguish these three multi-element structures by PPH and relate their path homology fingerprints to their properties. Furthermore, we found a linear relationship between the PPH fingerprints of the structure and the enthalpy of formation, PPH and the HUMO-LUMO gap and successfully predicted the stability of close-carboranes with guaranteed accuracy. In addition, we introduce nonlinear models, specifically, the gradient-boosted regression tree (GBRT) model, into the nonlinear prediction of formation enthalpies and HOMO-LUMO gaps with convincing accuracy. This is the first time that PPH has been used for the quantitative analysis of multi-element systems. And the results demonstrate the potential of PPH for quantitative analysis of multi-elemental structures.

2. METHODS

PPH is an extension of the path homology theory[36] that provides a way to measure the persistence of the homology of path complexes over a range of scales. This allows us to study the topological structure of a space in a way that captures the evolution of its path-connected components. It can be used to extract meaningful topological signatures from data and gain insights into the underlying structure of complex systems, particularly those that can be modeled as directed graphs or networks. In this section, we introduce the key concepts related to persistent path homology theory, including paths, path complexes, path homology, filtrations and PPH.

2.1. Path, path complex and path homology

Let V be a nonzero finite set. For a given integer $p\geq0$, the elementary ***p*-path** on V is a sequence $i_0i_1i_2,\ldots,i_p$ of elements in V. Let $\mathbb{K}$ be a field and let $\Lambda_p=\Lambda_p(V)$ be a $\mathbb{K}$-linear space generated by all the element p-paths. We denote $e_{i_0i_1,\ldots,i_p}$ the generator corresponding to the elementary p-paths $i_0i_1,\ldots,i_p$ and the family $\{e_{i_0i_1,\ldots,i_p}, i_0i_1,\ldots,i_p\in V\}$ is a basis of Λ_p over $\mathbb{K}$. Specifically, we make the convention that $\Lambda_{-1}=0$. An element in Λ_p can be uniquely written as

$$\mu=\sum a^{i_0i_1,\ldots,i_p}e_{i_0i_1,\ldots,i_p},\quad a^{i_0i_1,\ldots,i_p}\in\mathbb{K}. \tag{1}$$

For any integer $p\geq0$, a $\mathbb{K}$-linear map $\partial:\Lambda_p\to\Lambda_{p-1}$ is defined as

$$\begin{aligned}\partial e_{i_0}&=0,\quad p=0,\\ \partial e_{i_0i_1,\ldots,i_p}&=\sum_{k=0}^{p}(-1)^k e_{i_0,\ldots,\hat{i}_k,\ldots i_p},\quad p>0,\end{aligned} \tag{2}$$

DOI: 10.1142/S2737416524500339

where $\widehat{i_k}$ indicates the omission of the index i_k, and thus ∂ can be deduced as a boundary operator on $(\Lambda_p)_p$, and $\partial^2=0$.

A **path complex** on V is a nonempty collection $\mathcal{P}$ of elementary paths on V and it satisfies that $i_1i_2,\ldots,i_p\in\mathcal{P}$ and $i_0i_1i_2,\ldots,i_{p-1}\in\mathcal{P}$ for any $i_0i_1,\ldots,i_p\in\mathcal{P}$. Let G be a digraph. The collection of paths on G is a path complex, denoted by $\mathcal{P}(G)$. The paths in $\mathcal{P}$ are called allowed paths, the linear space generated by all allowed paths is denoted as

$$\mathcal{A}_p=\mathcal{A}_p(\mathcal{P})=\left\{\sum_{i_0i_1\ldots i_p\in V}a^{i_0i_1,\ldots,i_p}e_{i_0i_1,\ldots,i_p}\,|\,i_0i_1,\ldots,i_p\in\mathcal{P},a^{i_0i_1,\ldots,i_p}\in\mathbb{K}\right\}. \quad (3)$$

Here, as convention, let $\mathcal{A}_{-1}=0$ be the null space. The space of ∂-invariant p-paths can be deduced by

$$\Omega_{-1}=0,\quad \Omega_p=\Omega_p(\mathcal{P})=\{x\in\mathcal{A}_p|\partial x\in\mathcal{A}_{p-1}\},\quad p\geq 0. \quad (4)$$

Then $(\Omega_p)_p$ is a subchain complex of $(\Lambda_p(V))_p$. The **path homology** is defined by

$$H_p(\mathcal{P};\mathbb{K}):=H_p\left(\Omega_p(\mathcal{P})\right)=\frac{\ker\partial\big|_{\Omega_p}}{\mathrm{im}\,\partial\big|_{\Omega_{p+1}}},\quad p\geq 0. \quad (5)$$

The path homology of a digraph G is that of path complex $\mathcal{P}(G)$. The p-th Betti number of the digraph G is the rank of the homology $H_p(G;\mathbb{K})$, denoted as $\beta_p(G)$.

2.2. Persistent path homology and filtrations

Let $(S,\leq)$ be an order set and $(S,\leq)$ can be regarded as a category with elements in S as objects and all the binary orders as morphisms. A filtration of path complexes means a covariant functor $\mathcal{F}$: $(S,\leq)\to$ **Path** from the category $(S,\leq)$ to the category of path complexes. For each element $a\in S$, $\mathcal{F}_a$ is a path complex such that we have $f_{b,c}\circ f_{a,b}=f_{a,c}$ for $a\leq b\leq c$, where $f_{a,b}:\mathcal{F}_a\to\mathcal{F}_b$ is the morphism induced by $a\to b$. The morphism $f_{a,b}$ induces a morphism of path homology $\tilde{f}_{a,b}:H_p(\mathcal{F}_a;\mathbb{K})\to H_p(\mathcal{F}_b;\mathbb{K})$. The $(a,\ b)$-persistent path topology of $\mathcal{F}$ is defined by

$$H_p^{a,b}(\mathcal{F};\mathbb{K})=\mathrm{im}\left(H_p(\mathcal{F}_a;\mathbb{K})\to H_p(\mathcal{F}_b;\mathbb{K})\right),\quad p\geq 0. \quad (6)$$

The $(a,\ b)$-persistent Betti number is defined as the rank of $H_p^{ab}(\mathcal{F};\mathbb{K})$.

In practice, the path complex is usually defined on digraphs. Let **Digraph** denote as the category of digraphs. A filtration of digraphs is a covariant functor $\Gamma:(S,\leq)\to$ **Digraph** from the category $(S,\leq)$ to the category of digraphs. A filtration of digraphs can induce a filtration of path complexes, which results in the PPH of digraphs. And different filtration can result in different persistence.

2.3. Distance-based filtration

Let $G=(V,E)$ be a digraph, where V represents the set of data points in a metric space $(X,\|\cdot\|)$. Then, there is a weight function $d:E\to\mathbb{R}$ on the edge set E deduced by

$$d(x_1,x_2)=\|x_1-x_2\|,\quad (x_1,x_2)\in E\subset X\times X. \quad (7)$$

Here, we assume the metric space $(X,\|\cdot\|)$ as the Euclidean space with L_2-norm. Then, let $E_t=\{(x,y)\in E|d(x,y)\leq t\}$ and $\mathcal{G}_t=(V,E_t)$. It can be deduced that $\mathcal{G}:(\mathbb{R},\leq)\to$ **Digraph**, $t\mapsto\mathcal{G}_t$ is a filtration of digraphs, which leads to a persistent diagram $\mathcal{D}(\mathcal{G})$ of G.

2.4. Angle-based filtration

Let $G=(V,E)$ be a digraph with V in Euclidean space $\mathbb{R}^3$. The process of the angle-based filtration of digraphs is as follows. Above all, we fix a coordinate system which is unique for a predefined rule. The filtration process is based on a sequence of angles in a prescribed order $(S^2,\leq)$ such that $S^2=\{(\alpha,\beta)|\alpha\in[0,2\pi],\beta\in[0,\pi]\}$ is an ordered set with order deduced by

$$(\alpha,\beta)\leq(\alpha'\beta'),\quad \text{if } (\alpha\leq\alpha') \text{ or } (\alpha=\alpha',\beta\leq\beta'). \quad (8)$$

For discrete case, let m, k be positive integers, we can choose the ordered set S^2 by

$$\left(\frac{2\pi t}{k},\frac{\pi s}{m}\right),\ t=0,\ldots,k-1,s=0,\ldots,m-1. \quad (9)$$

In this work, we set $m=36$, $k=72$ and achieve a filtration based on the angle shown as spiral progress in a polar coordinate system. Finally, we get a filtration of digraphs based on angles $\mathcal{P}:(S^2,\leq)\to$ **Digraph**, $\theta\mapsto\mathcal{P}_\theta$, where $(S^2,\leq)$ is an ordered set on a sphere and $\mathcal{P}_\theta=(V_\theta,E_\theta)$, where $V_\theta=\{x\in V|x\leq\theta\in S^2\}$ and $E_\theta=E\cap(V_\theta\times V_\theta)=\{(x,y)\in E|x,y\in V_\theta\}$.

2.5. Extracting the intrinsic topological features

Each closo-carborane is transformed into a point cloud according to atomic coordinates, as shown in Fig. 1(a),

DOI: 10.1142/S2737416524500339
J. Comput. Biophys. Chem. 2025, 24 (1), 1–11

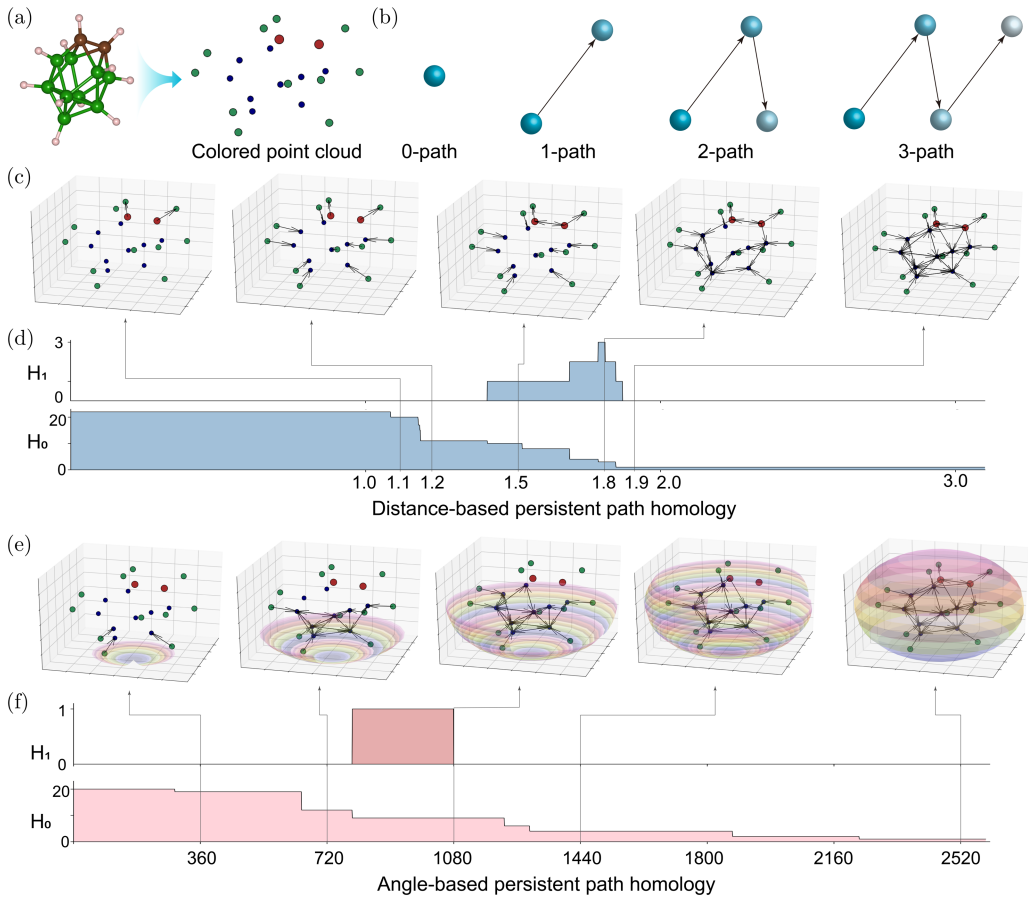

Fig. 1. (Color online) Take $C_2B_9H_{11}$ as an example to show PPH. (a) The structural diagrams of $C_2B_9H_{11}$. The brown balls represent the C atom, the green balls represent the B atom, and the white balls represent the H atom. The point cloud constructed from the atomic coordinates of $C_2B_9H_{11}$. The red points represent the C atom, the blue points represent the B atom and the green points represent the H atom. The larger the electronegativity value of an atom, the larger the volume. (b) Illustration of the basic component that makes up the path complex, p-path, where $p=0$, 1, 2 and 3. (c) Over the filtration, the directional connections between atom pairs gradually increase. Directed connections are connections of atoms with higher electronegativity values to those with lower electronegativity values. (d) The DPPH fingerprints. The vertical axis is the rank of homology groups H_0 and H_1 and the horizontal axis is Euclidean distance. (e) The sphere is divided into 2592 zones according to angle. During the filtration process, which starts at the south pole, each region is filtered clockwise and the directed connections between pairs of atoms gradually increase. Directed connections are those between atoms with a higher electronegativity value and those with a lower electronegativity value. (f) The APPH fingerprints. The vertical axis is the rank of homology groups H_0 and H_1 and the horizontal axis is the region number.

where the element types are labeled with colors. We assign a weight, represented by the volume, to each element type based on its electronegativity value. PPH is applied to extract the Betti-n (β_n) number of the point cloud as a topological invariant representing the geometric information of cluster structure. The β_n is the rank of the homology group H_n. The construction of the homology groups is carried out based on paths, which is the basic building block of the path complex. Figure 1(b) shows the 0-path, 1-path, 2-path and 3-path. Then, a mathematical operation called boundary operator is introduced to obtain the chain complex. Finally, the homology groups $(H_p,H_{p-1},\ldots,H_0)$ are generated from the chain complex. In this paper, Betti-0 (β_0) and Betti-1 (β_1), which generally represent the number of independent components and the number of directed cycles separately, are used for further analysis.

Roughly speaking, the above method extracts only the persistent path homology of the multi-element structure over a fixed geometric relation. Specifically, through the filtration process, we continuously capture the topological invariants of the multi-element structure in a changing scale in the path sense, thus capturing the structural topological information while also preserving the geometric information. Considering the

DOI: 10.1142/S2737416524500339

lack of interatomic distance information, the distance relative to each node in the point cloud is chosen as the filtration parameter. As shown in the left chart of subgraphs in Fig. 1(c), the filtration parameter is 1.1, meaning only directed connections for atomic pairs with Euclidean distances less than 1.1 can be constructed. Under this condition, only two directed connections exist for carbon and hydrogen atoms. In the process of increasing the filtration parameter, when the Euclidean distance between two atoms is not greater than the parameter, a directed edge is formed connecting these two atoms, the direction of which is determined by the atomic weights. Figure S1 records the connectivity relationships for several different filtration parameters, where each dark blue entry represents a directed connection between atoms. Over the filtration process, topological invariants for homology H_0 and H_1 are depicted in Fig. 1(d). The maximum value of β_0 is 22 at the beginning, which corresponds to 22 atoms of the molecule. After that, as the filtration parameters increase, the number of connection relationships also increases, so the independent component decreases, which leads to a monotonic decrease in β_0. The value of β_1 is 1 when the filtration parameter is equal to 1.5 and reaches a maximum value of 3 when the filtration parameter increases to 1.8. This shows that DPPH can embed the positional relations in the structure into PPH.

In order to analyze the structure from an external perspective, we introduce APPH. First, we assume the existence of a sphere whose center is the center of the point cloud. The south pole direction of the sphere is obtained by the vector sum of all paths, and then the horizontal plane of the sphere is obtained by the right-handed spiral rule. Similar to meridians and latitudes, we can divide the sphere into several, in this work we are dividing it into 2592 regions. The angle-based filtration process starts from the south pole and traverses each region clockwise from south to north, as shown in Fig. 1(e). During filtration, each pair of atoms within each traversed region establishes a directed connection whose direction is determined by the weights. The topological invariance of homology groups and point clouds is shown in Fig. 1(f). β_0 reflects the number of independent components in the system, so as the filtration parameter increases, β_0 decreases. β_1 indicates the number of directed cycles and disappears at a filtration parameter of 1080. Based on the above topological approach, the following work will establish the relationship between multi-element structure topological features and attributes.

3. RESULTS

3.1. Persistent path homology analysis for o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$

Different types of atoms will lead to asymmetric interactions between atoms, which is an important factor affecting the macroscopic properties of multi-element structures. o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$ are three different isomers of $C_2B_{10}H_{12}$, respectively, with two carbon atoms adjacent to each other, two carbon atoms separated by a boron atom and two carbon atoms opposite each other. Their structures are shown schematically in the left subfigure of Figs. 2(a)–2(c). They differ in terms of electron distributions, electron affinities and acidic.[40] In terms of thermodynamic stability, p-$C_2B_{10}H_{12}$ is the best, m-$C_2B_{10}H_{12}$ is the second and o-$C_2B_{10}H_{12}$ is the worst. Recently, it has been shown that the thermodynamic stability of the three clusters is related to the relative position of the carbon atoms and that the repulsive effect is significantly weaker with the insertion of a boron atom between two carbon atoms.[41] Furthermore, it has been found that unlike m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$, o-$C_2B_{10}H_{12}$ shows intriguing luminescent properties due to the presence of carbon-carbon bonds.[42] However, persistent homology is unable to effectively distinguish these three isomers. As shown in Fig. S2, there is no significant difference between the barcode of the m-$C_2B_{10}H_{12}$ and the o-$C_2B_{10}H_{12}$. PPH captures the structural differences between these three clusters. The connections highlighted in red are influenced by the position of the carbon atoms, as shown in the right subgraphs of Figs. 2(a)–2(c). Figure S3 documents the linkage relationships, where each dark region represents a directed connection between atoms.

Next, the DPPH fingerprints of o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$ are established (Fig. 2(d)). As shown in Fig. 2(d), the difference in β_0 between the three multi-element structures only appears when the filtration parameter is around 1.7. Therefore, we compare the digraphs of the three clusters when the filtration parameter is 1.7. As shown in Fig. 2(e), it can be seen that the carbon atoms in o-$C_2B_{10}H_{12}$ have more directed connections than the boron atoms in m-$C_2B_{10}H_{12}$, and therefore, the β_0 for o-$C_2B_{10}H_{12}$ is lower than that for m-$C_2B_{10}H_{12}$. In contrast, the carbon atoms in p-$C_2B_{10}H_{12}$ are not attached to boron, so the β_0 of p-$C_2B_{10}H_{12}$ is higher than that of m-$C_2B_{10}H_{12}$. Furthermore, as shown by the green highlighted connection in Fig. 2(e), a directed cycle formed by 1-paths is created in o-$C_2B_{10}H_{12}$ and m-$C_2B_{10}H_{12}$, respectively,

DOI: 10.1142/S2737416524500339
J. Comput. Biophys. Chem. 2025, 24 (1), 1–11

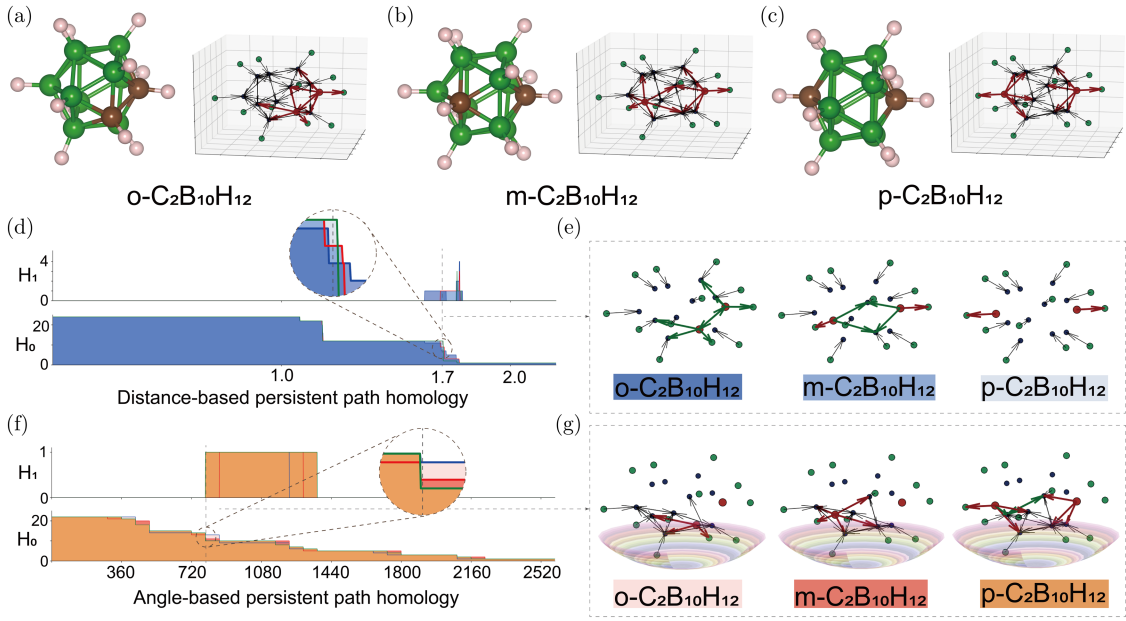

Fig. 2. (Color online) Application of DPPH and APPH to three isomers. (a–c) The structural diagrams and the digraphs of o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$. The brown balls represent the C atom, the green balls represent the B atom and the white balls represent the H atom. In the digraph, the directed connections depending on the relative position of carbon atoms are marked in red. (d) The DPPH fingerprints of o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$. The horizontal coordinate is the filtration parameter, and the vertical coordinate is H_0 and H_1. (e) Digraphs when the filtration parameter is 1.7. Depending on the relative position of carbon atoms, the directed connections are marked in red and the directed cycles are marked in green. (f) The APPH fingerprints of o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$. The horizontal coordinate is the filtration parameter, and the vertical coordinate is the H_0 and H_1. (g) Digraphs when the filtration parameter is 810. Depending on the relative position of carbon atoms, the directed connections are marked in red and the directed cycles are marked in green.

and thus the value of β_1 is 1. In contrast, p-$C_2B_{10}H_{12}$ does not create directed cycles because it lacks a directed connection between carbon and boron atoms, and therefore, β_1 is 0. Thus, this filtration parameter reflects the effect of the position of the carbon atom on the topological features of these three isomers. From the above analysis, it can be found that the differences in the fingerprints correspond to a certain topological feature of the structure, indicating that the path topological features can reflect the intrinsic information of multi-element structures, which may be useful for predicting the properties of the multi-element clusters. For example, the luminescence properties of closo-carborane can be predicted by applying fingerprints to identify the directed connections between carbon atoms.

Figure 2(f) shows the APPH of o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$. As shown in Fig. 2(f), the difference in the relative position of carbon atoms is manifested as a general variation of H_0 and H_1 between the three isomers over the filtration process. The APPH gives a better indication of macroscopic topological features. Unlike DPPH, APPH can distinguish between these three isomers at more different filtration parameters. We choose the digraphs of the three clusters with a filtration parameter of 810 as an example for comparison. As shown in Fig. 2(g), o-$C_2B_{10}H_{12}$ has the most connections, m-$C_2B_{10}H_{12}$ has the next most connections, and p-$C_2B_{10}H_{12}$ has the least connections. The connection highlighted in green is a directed cycle that exists only in p-$C_2B_{10}H_{12}$, a cluster with β_1 of 1.

3.2. Linear prediction for the stability of closo-carboranes

The above results indicate that PPH is well represented in capturing structural features that can qualitatively explain the stability and properties of closo-carboranes. To exemplify the potential of PPH, we constructed 21 structures, namely $C_2B_nH_{n+2}$ ($n=2\sim22$), to quantitatively predict the stability of closed carboranes.

First, we calculate the enthalpy of formation of each structure to quantitatively assess the thermodynamic stability of the structure. We calculate the integration of H_0 with the filtration parameter, i.e., the area enclosed by the β_0 curve, to detect the relationship between the

DOI: 10.1142/S2737416524500339

fingerprints and the enthalpy of formation. The results show a strong linear relationship between the integrals of the DPPH and the APPH fingerprints and the enthalpy of formation (Figs. 3(a) and 3(b)). The Pearson correlation coefficients (PCC) between the integral and enthalpy of formation are 0.999 and 0.993, respectively, indicating that topological features are highly correlated with thermodynamic stability. β_0 represents the number of independent components, and its integral represents a summation of the number of independent components over the filtration process. The enthalpy of formation is a thermodynamic property that represents the change in energy when one mole of a substance forms from its elements in their standard states. When chemical bonds form, energy is released because atoms are more stable in a bonded state than when they are isolated. The number of chemical bonds in a molecule is directly related to its energy. The overall energy of a molecule is the sum of the energies associated with all the bonds within it. Therefore, as the number of atoms bonded in the molecule increases during filtration, we obtain information about the energy in the molecule, i.e., the enthalpy of formation, through the sum of the Betti number.

In the same way, we calculate the integration of H_1 with the filtration parameter, i.e., the area enclosed by β_1 curve, to detect the relationship between the fingerprints and the enthalpy of formation. As a result, we find a strong linear correlation between the APPH fingerprints and the enthalpy of formation, with a PCC

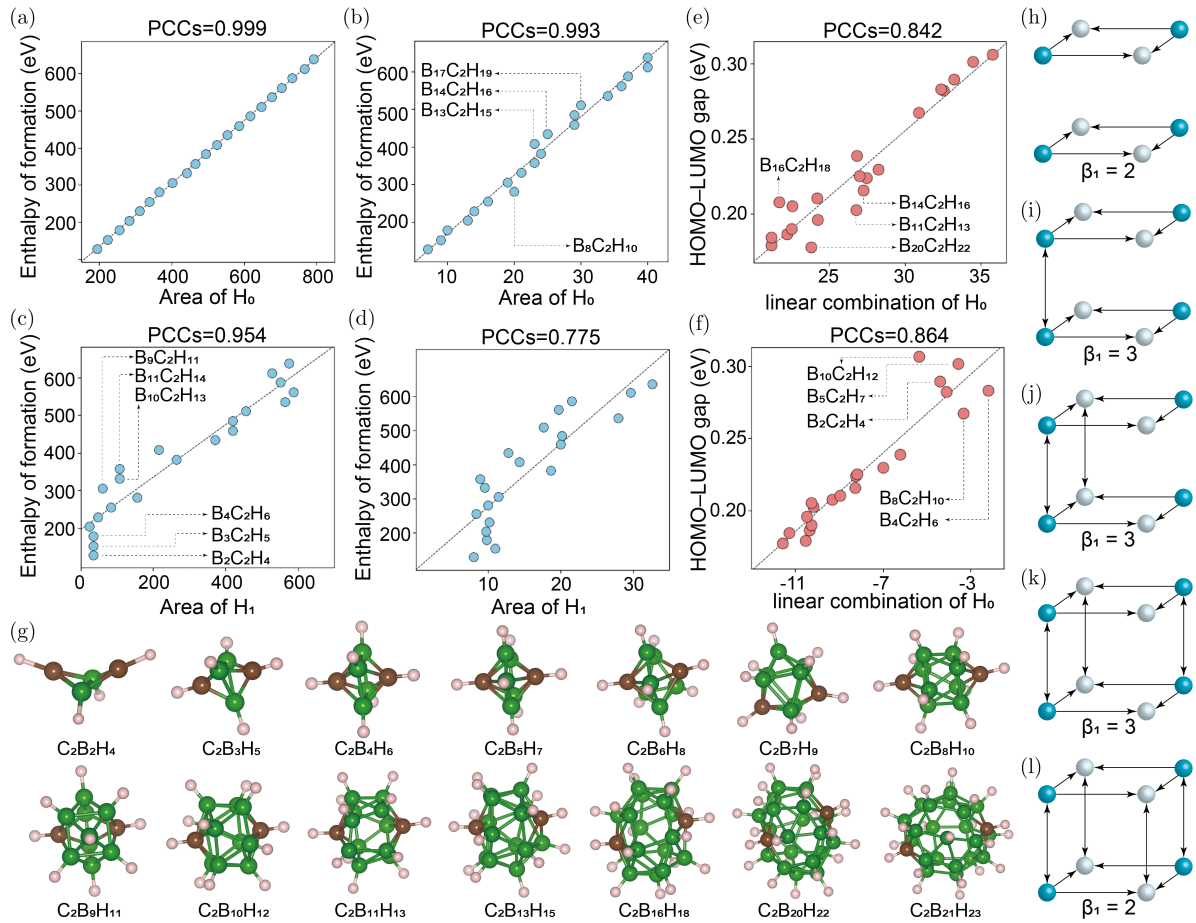

Fig. 3. (Color online) Results of linearly predict the stability of closo-carboranes. (a, b) Linear relations between the integral of H_0 of the DPPH and the APPH fingerprints and the enthalpy of formation. (c) Linear relation between the integral of H_1 of the DPPH and the APPH fingerprints and the enthalpy of formation. (d) Linear relation between the integral of H_0 of traditional persistent homology fingerprint and the enthalpy of formation. (e, f) Linear relations between the linear sum of the values at the inflection point of β_0 curve of the DPPH and the APPH fingerprints and the HOMO-LUMO gap. The points in the above fitting figures are the data points of $C_2B_nH_{n+2}$($n=2\sim22$), and the dotted lines are the fitting line. (g) Closo-carboranes structure diagram of different atomic numbers. The brown balls represent the C atom, the green balls represent the B atom and the white balls represent the H atom. (h–l) Digraphs of two directed cycles show different connection relations and the value of β_1.

DOI: 10.1142/S2737416524500339

of 0.954 (Fig. 3(c)). The value of β_1 represents there are two directed cycles in Fig. 3(h). As shown in Figs. 3(i)–3(k), when a connection between the two directed cycles occurs, a new directed cycle is created ($\beta_1=3$). In Fig. 3(l), when the two directed cycles are fully connected ($\beta_1=2$). Thus, β_1 can be considered as the number of complex transmission paths in the structure, reflecting the distance and interaction between the topological configurations. It can be observed that there are three outliers in Fig. 3(c), $C_2B_2H_4$, $C_2B_3H_5$ and $C_2B_4H_6$. This is because they contain too few atoms to form a directed cycle (Fig. 3(g)). As a comparison, we introduce the traditional persistent homology for $C_2B_nH_{n+2}(n=2\sim22)$, as shown in Fig. S4. Figure 3(d) shows the linear relationship between the integral of H_1 obtained by traditional persistent homology and the enthalpy of formation. The PCC is 0.775, which is lower than that of the PPH. The integral of H_1 quantifies the change in the number of directed cycles within the structure. These directed cycles reflect the action of forces on multiple atoms in the molecule, representing the energy involved in forming a portion of the molecule when a higher-dimensional configuration is achieved. Therefore, the number of directed cycles is expected to have a strong linear relationship with the enthalpy of formation.

As a description of the excitation energy, the HOMO-LUMO gap can be used to assess chemical stability. We calculate the linear sum of the values at the inflection point of the β_0 curve to detect the relationship between the fingerprints and the HOMO-LUMO gap. The results show a linear relationship between the linear sum of both the DPPH and the APPH fingerprint and the HOMO-LUMO gap (Figs. 3(e) and 3(f)). The PCCs between the linear sum and the HOMO-LUMO gap are 0.842 and 0.864, respectively. When using DPPH, a strong linear relationship between the fingerprints of structures containing a large number of atoms and the HOMO-LUMO gap can be seen, while outliers are structures containing a small number of atoms, such as $C_2B_{14}H_{16}$, $C_2B_{16}H_{18}$ and $C_2B_{20}H_{22}$. When using APPH, a strong linear relationship is observed between the fingerprints of structures containing a small number of atoms and the HOMO-LUMO gap, while outliers are structures containing a large number of atoms, such as $C_2B_2H_4$, $C_2B_4H_6$ and $C_2B_5H_7$. Therefore, when predicting the HOMO-LUMO gap, the DPPH and the APPH fingerprints are complementary to each other. The HOMO-LUMO gap indicates the structure's susceptibility to chemical reactions, which can be interpreted as the likelihood of chemical bond reorganization. Information on the changes and bonding of the independent components of the atoms is recorded in the β_0 curve, thus reflecting the stability of the molecular structure.

3.3. Nonlinear prediction for the stability of closo-carboranes

GBRT becomes a powerful nonlinear prediction model with advantages such as high efficiency and less susceptibility to overfit.[43] In contrast to the enthalpy of formation, the HOMO-LUMO gap cannot reach a highly linear correlation with the PPH fingerprints. In this part, we introduce the GBRT to predict the HOMO-LUMO gap for closo-carboranes.

In the above work, we establish oriented connections between pairs of atoms containing covalent bonds in 21 optimized closo-carboranes structures for linear prediction. To apply GBRT to predict the HOMO-LUMO gap, in addition to the 21 optimized structures, we build 2172 structures in the optimization process. Then, we obtain the PPH fingerprints considering the bonding restriction and the PPH fingerprints without considering the bonding restriction, a total of 2193 structures for prediction and the details of GBRT are shown in Table S1. Figures 4(a) and 4(b) show prediction results of the HOMO-LUMO gap when considering bonding restriction. The PCCs between the DPPH and the APPH fingerprints and the HOMO-LUMO gap reach 0.878 and 0.974, respectively, demonstrating that GBRT has higher predictive power than linear prediction. Figures 4(c) and 4(d) show the prediction results of the HOMO-LUMO gap without bonding restriction. The PCCs reach 0.926 and 0.974, respectively. We observe that the PCCs between the DPPH fingerprint when without bonding restriction and the HOMO-LUMO gap are higher than those when bonding restriction is considered. It indicates that GBRT successfully extracts and integrates a large number of complex and cluttered topological features. Similarly, introducing GBRT into the nonlinear prediction of the enthalpy of formation, the PCCs are all around 0.99 (Fig. S5) and the combined results are shown in Table S2, demonstrating the strong generality of PPH of predicting stability.

Finally, we provide a comprehensive comparison of three approaches: the graph-based method, the simple complexes method and the path complex approach. As illustrated in Fig. S6, the graph-based method, when compared to the simple composite shape method, fails to capture topological information beyond the one-dimensional simple complexes. Additionally, it lacks the

DOI: 10.1142/S2737416524500339

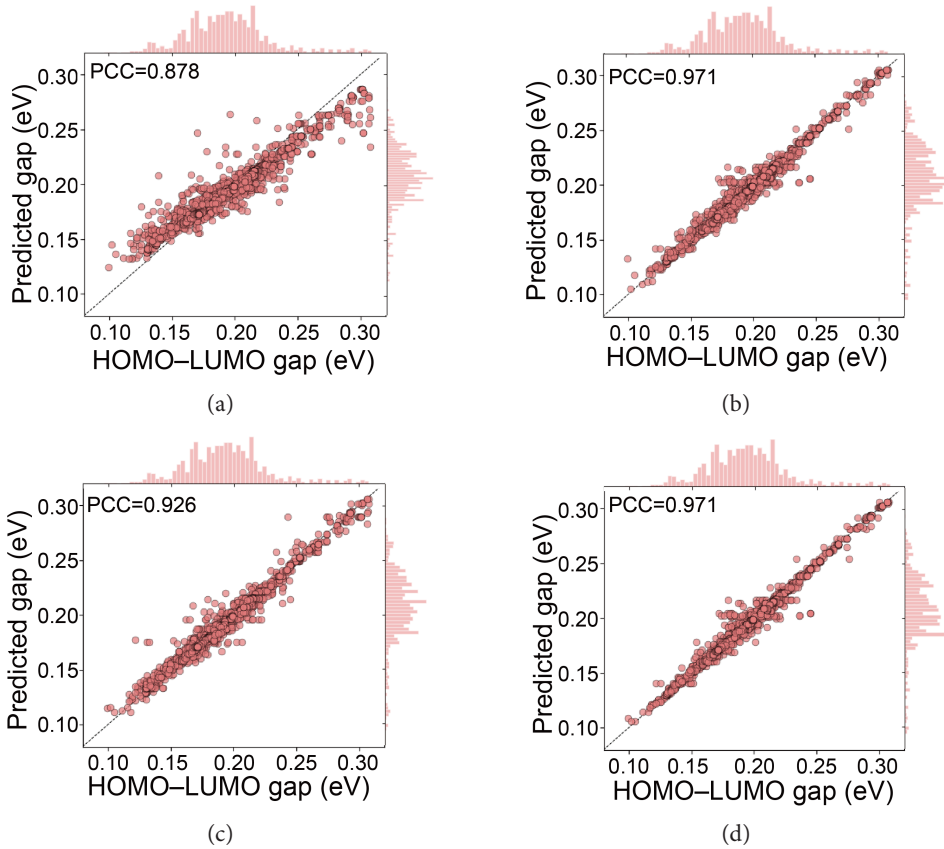

Fig. 4. (Color online) Nonlinear prediction of closo-carboranes stability by GBRT. (a, b) Prediction of the HOMO-LUMO gap by the DPPH and the APPH fingerprints when considering bonding restriction. (c, d) Prediction of the HOMO-LUMO gap by the DPPH and the APPH fingerprints without considering bonding restriction. The points in the above diagram are the data points of $C_2B_nH_{n+2}(n=2\sim22)$, and the dotted lines represent the standard lines when the prediction is completely accurate.

capability to extract crucial details of intramolecular voids and cavities. In contrast, the simple complexes method, when compared to the path complex approach, falls short in recognizing the directionality of bonds and is not suited for extracting data on asymmetric interactions between different atoms. A significant advantage of the path topology method is its innovative feature extraction filtering process. This process leverages the ability of the simple complexes method to refine the topological features inherent to both the simple complexes and path complexes methods. Importantly, this refinement does not compromise the topological significance of the features.

Next, we proceed with a comparative analysis of these three methods in predicting two properties: enthalpy of formation and HOMO-LUMO gap. Figure 3 illustrates the correlation assessments between the simple complex approach and the path complex approach for linear predictions. Notably, the path complex approach exhibits a significantly higher correlation, emphasizing its superior performance in this regard. Meanwhile, we employ a Graph Convolutional Network (GCN) feature extraction method[44] for nonlinear prediction, in contrast to the path complex approach outlined in our manuscript. Figure S7 presents the predicted enthalpy of formation and HOMO-LUMO gap obtained through the GCN feature extraction method under the GBRT model. While the predictions are also favorable, the PCC is still lower than the 0.999 and 0.971 reported in our paper.

4. CONCLUSION

Topology, as a novel method for data analysis, can greatly simplify the complexity of data while retaining critical information and has recently become one of the most promising methods for studying molecular structures. However, the applicability of traditional topological methods is not satisfactory when studying the cluster structure of multiple elements. In this work, we introduce PPH, a more appropriate and effective algebraic topological method, as the new tool for the analysis of carboranes. To validate the proposed method, we first performed a quantitative analysis of the

structure and stability of $C_2B_nH_{n+2}$(n=2~22). Above all, the DPPH and the APPH fingerprints are obtained as the intrinsic features of closo-carborane. In the first task, we distinguish o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$ and qualitatively analyze the properties of $C_2B_{10}H_{12}$ based on PPH. We also use the features extracted from H_0 and H_1 as input to the linear relationship to achieve a prediction accuracy of PCCs above 0.95 for the thermodynamic stability of closo-carborane and around 0.85 for the chemical stability. Then, we introduce GBRT into the nonlinear prediction of the enthalpy of formation and the HOMO-LUMO gap, showing higher PCCs and demonstrating the strong generality of the PPH fingerprint in predicting stability. We believe that when combined with cutting-edge machine learning models, PPH will certainly become a more powerful method for advanced characterization in terms of multi-element structures and pave the way for structural design and new material discovery.

STATEMENT OF USAGE OF ARTIFICIAL INTELLIGENCE

ChatGPT helped in touching up the language in some paragraphs of the paper.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Wang designed the project, modified the method, performed computational studies, wrote the first draft and revised the manuscript. Zhang provided the data. Pan supervised the project, acquired funding and revised the manuscript.

CONFLICT OF INTEREST

The authors declare no competing financial interest.

FUNDING INFORMATION

This work was financially supported by the Guangdong Basic and Applied Basic Research Foundation (2020A1515110843), Young S&T Talent Training Program of Guangdong Provincial Association for S&T (SKXRC202211), Soft Science Research Project of Guangdong Province (No. 2017B030301013), National Natural Science Foundation of China (22109003) and the Major Science and Technology Infrastructure Project of Material Genome Big-science Facilities Platform supported by Municipal Development and Reform Commission of Shenzhen, Soft Science Research Project of Guangdong Province (No. 2017B030301013).

SUPPLEMENTARY INFORMATION

The Supplementary Information are available at: https://www.worldscientific.com/doi/suppl/10.1142/S2737416524500339.

ORCID

Bingxu Wang https://orcid.org/0009-0003-7224-6579
Mingzheng Zhang https://orcid.org/0009-0004-1463-0369
Feng Pan https://orcid.org/0000-0002-1362-4336

References

1. Quan, Y. J.; Xie, Z. W. Controlled Functionalization of O-Carborane via Transition Metal Catalyzed B-H Activation. *Chem. Soc. Rev.* **2019**, *48* (13), 3660–3673.
2. Farras, P. *et al.* Metallacarboranes and Their Interactions: Theoretical Insights and Their Applicability. *Chem. Soc. Rev.* **2012**, *41* (9), 3445–3463.
3. Furue, R. *et al.* Aggregation-Induced Delayed Fluorescence Based on Donor/Acceptor-Tethered Janus Carborane Triads: Unique Photophysical Properties of Nondoped OLEDs. *Angewa. Chem. Int. Ed.* **2016**, *55* (25), 7171–7175.
4. Zhang, J., Wang, X.; Jin, G. X. Polymerized Metallocene Catalysts and Late Transition Metal Catalysts for Ethylene Polymerization. *Coord. Chem. Rev.* **2006**, *250* (1–2), 95–109.
5. Fisher, S. P. *et al.* Nonclassical Applications of closo-Carborane Anions: From Main Group Chemistry and Catalysis to Energy Storage. *Chem. Rev.* **2019**, *119* (14), 8262–8290.
6. Hawthorne, M. F.; Maderna, A. Applications of Radiolabeled Boron Clusters to the Diagnosis and Treatment of Cancer. *Chem. Rev.* **1999**, *99* (12), 3421–3434.
7. Soloway, A. H. *et al.* The Chemistry of Neutron Capture Therapy. *Chem. Rev.* **1998**, *98* (4), 1515–1562.
8. Nunez, R. *et al.* Electrochemistry and Photoluminescence of Icosahedral Carboranes, Boranes, Metallacarboranes, and Their Derivatives. *Chem. Rev.* **2016**, *116* (23), 14307–14378.
9. Housecroft, C. E. Carboranes as Guests, Counterions and Linkers in Coordination Polymers and Networks. *J. Organomet. Chem.* **2015**, *798*, 218–228.
10. Schleyer, P. V. R.; Najafian, K. Stability and Three-Dimensional Aromaticity of closo-Monocarbaborane

Anions, CB n-1Hn-, and closo-Dicarboranes, C2B n-2H n. *Inorg. Chem.* **1998**, *37* (14), 3454–3470.
11. Williams, R. E. Early Carboranes and Their Structural Legacy. *Adv. Organomet. Chem.* **1994**, *36*, 1–55.
12. Zhang, J. J.; Lin, Z. Y.; Xie, Z. W. DFT Studies on Structures, Stabilities, and Electron Affinities of Closo-Supercarboranes C2Bn-2Hn (n = 13–20). *Organometallics* **2015**, *34* (23), 5576–5588.
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. Z.; Ren, S. K.; Xie, Z. W. Transition Metal-Carboryne Complexes: Synthesis, Bonding, and Reactivity. *Acc. Chem. Res.* **2011**, *44* (4), 299–309.
15. 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.
16. Moore, J. E. The Birth of Topological Insulators. *Nature* **2010**, *464* (7286), 194–198.
17. Liu, X.; Xia, K. L. Neighborhood Complex Based Machine Learning (NCML) Models for Drug Design. in *4th International Workshop on Interpretabil of Machine Intelligence in Medical Image Computing* (*iMIMIC*)*/1st International Workshop on Topol Data Analysis and Its Applicat for Medical Data* (*TDA4MedicalData*) *at 24th Int. Conf. Medical Image Computing and Compter Assisted Intervent* (*MICCAI*), Springer, 2021, pp. 87–97.
18. Wu, C. Y. *et al.* Discrete Morse Theory for Weighted Simplicial Complexes. *Topol. Its Appl.* **2020**, *270*, 107038.
19. Liu, X. *et al.* Hypergraph-Based Persistent Cohomology (HPC) for Molecular Representations in Drug Design. *Briefings Bioinf.* **2021**, *22* (5), bbaa411.
20. Edelsbrunner, H.; Letscher, D.; Zomorodian, A. Topological Persistence and Simplification. *Discrete Comput. Geom.* **2002**, *28*, 511–533.
21. Zomorodian, A.; Carlsson, G. Computing Persistent Homology. *Discrete Comput. Geom.* **2005**, *33* (2), 249–274.
22. Carlsson, G. E. Topology and Data. *Bull. Am. Math. Soc.* **2009**, *46*, 255–308.
23. Chen, J. H.; Wei, G. W. Omicron BA.2 (B.1.1.529.2): High Potential for Becoming the Next Dominant Variant. *J. Phys. Chem. Lett.* **2022**, *13* (17), 3840–3849.
24. Cang, Z. X.; Wei, G. W. Persistent Cohomology for Data with Multicomponent Heterogeneous Information. *Siam J. Math. Data Sci.* **2020**, *2* (2), 396–418.
25. Carlsson, G. *et al.* On the Local Behavior of Spaces of Natural Images. *Int. J. Comput. Vis.* **2008**, *76* (1), 1–12.
26. Pachauri, D. *et al.* Topology-Based Kernels with Application to Inference Problems in Alzheimer's Disease. *IEEE Trans. Med. Imaging* **2011**, *30* (10), 1760–1770.
27. Singh, G. *et al.* Topological Analysis of Population Activity in Visual Cortex. *J. Vis.* **2008**, *8* (8), 11.1–8.
28. Meng, Z. Y.; Xia, K. L. Persistent Spectral-Based Machine Learning (PerSpect ML) for Protein-Ligand Binding Affinity Prediction. *Sci. Adv.* **2021**, *7* (19), eabc5329.
29. Hiraoka, Y. *et al.* Hierarchical Structures of Amorphous Solids Characterized by Persistent Homology. *Proc. Natl. Acad. Sci.* **2016**, *113* (26), 7035–7040.
30. Anand, D. V. *et al.* Topological Feature Engineering for Machine Learning Based Halide Perovskite Materials Design. *npj Comput. Mater.* **2022**, *8* (1), 203.
31. Zheng, S. *et al.* Application of Topology-Based Structure Features for Machine Learning in Materials Science. *Chin. J. Struct. Chem.* **2023**, *42* (7), 100120.
32. Xia, K. L. *et al.* Persistent Homology for the Quantitative Prediction of Fullerene Stability. *J. Comput. Chem.* **2015**, *36* (6), 408–422.
33. Chen, X. *et al.* Topology-Based Machine Learning Strategy for Cluster Structure Prediction. *J. Phys. Chem. Lett.* **2020**, *11* (11), 4392–4401.
34. Grigor'yan, A. *et al.* Homologies of Path Complexes and Digraphs. arXiv preprint arXiv:1207.2834, 2012.
35. Grigor'yan, A. *et al.* On the Path Homology Theory of Digraphs and Eilenberg-Steenrod Axioms. *Homol. Homotopy Appl.* **2018**, *20* (2), 179–205.
36. Grigor'yan, A. *et al.* Path Complexes and Their Homologies. *J. Math. Sci.* **2020**, *248* (5), 564–599.
37. Liu, R.; Liu, X.; Wu, J. Persistent Path-Spectral (PPS) Based Machine Learning for Protein–Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* **2023**, *63* (3), 1066–1075.
38. Wu, S. *et al.* The Metabolomic Physics of Complex Diseases. *Proc. Natl. Acad. Sci.* **2023**, *120* (42), e2308496120.
39. Chen, D. *et al.* Path Topology in Molecular and Materials Sciences. *J. Phys. Chem. Lett.* **2023**, *14* (4), 954–964.
40. Hermansson, K.; Wojcik, M.; Sjoberg, S. o-, m-, and p-Carboranes and Their Anions: Ab Initio Calculations of Structures, Electron Affinities, and Acidities. *Inorg. Chem.* **1999**, *38* (26), 6039–6048.
41. Poater, J. *et al.* Too Persistent to Give Up: Aromaticity in Boron Clusters Survives Radical Structural Changes. *J. Am. Chem. Soc.* **2020**, *142* (20), 9396–9407.
42. Ochi, J.; Tanaka, K.; Chujo, Y. Recent Progress in the Development of Solid-State Luminescent o-Carboranes with Stimuli Responsivity. *Angew. Chem.Int. Ed.* **2020**, *59* (25), 9841–9855.
43. Natekin, A.; Knoll, A. Gradient Boosting Machines, a Tutorial. *Front. Neurorob.* **2013**, *7*, 21.
44. Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.