Persistent Magnitude for the Quantitative Analysis of the Structure and Stability of Carboranes

Wanying Bi*,†, Xin Fu†, Jingyan Li† and Jie Wu†,‡

*School of Mathematical Sciences, Hebei Normal University, No. 20 South Second Ring East Road, Shijiazhuang 050024, P. R. China
†Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, No. 544, Hefangkou Village, Huabei Town, Huairou District, Beijing 101408, P. R. China
‡Corresponding author. E-mail: wujie@bimsa.cn

ABSTRACT: Magnitude, similar to concepts like volume, cardinality or Euler characteristic, has become a key focus in combinatorics and topology. Recent advancements in topological data analysis and persistent homology have emphasized its importance. Persistent magnitude, a newly highlighted invariant introduced by Govc and Hepworth, has emerged as a notable subject of interest. In this work, we apply persistent magnitude to analyze and predict the stability of closo-carborane structures. First, we assess the stability of carboranes by employing cross-validation with different magnitude features. The Pearson correlation coefficients for stability predictions using three distinct magnitude features are 0.900, 0.882 and 0.883, respectively. These results are comparable to the Pearson correlation coefficient of 0.881 obtained when using a single feature based on persistent homology. Second, the utilization of magnitude features to predict the HOMO, LUMO and HOMO–LUMO gaps of carboranes involves conducting eight gradient boosting regressions for each scenario. The lowest correlation coefficients observed are 0.9056, 0.9385 and 0.9427, respectively. These findings highlight the promising performance of persistent magnitude features in the analysis of material structure and stability.

KEYWORDS: Persistent homology; persistent magnitude; structure; stability; carboranes.

1. INTRODUCTION

“Magnitude” is a novel numerical invariant initially proposed by Leinster,1 analogous to the Euler characteristic of a category.2 Similar to how cardinality represents the number of elements in a set, magnitude can express the size of a finite metric space. When examining graphs, we note that they can be regarded as metric spaces, where the distance between vertices is defined as the length of the shortest path connecting them. Given that many models in data analysis are represented as graphs, it becomes pertinent to consider the concept of magnitude within the context of graphs.3 Hypergraphs are generalizations of graphs. The notion of magnitude for hypergraphs has also been proposed and studied.4 Govc and Hepworth introduced the concept of persistent magnitude (PM), a novel numerical invariant for sufficiently well-behaved graded persistence modules.5 Building upon this work, the concept of alpha magnitude has also been introduced.6 The continual development of the theory of PM showcased in these works underscores its significant potential applications in mathematical sciences.

While the original persistent homology method has been successfully applied to the study of single-element clusters7 and some multi-element clusters, these studies either involve systems with only one element or neglect differences between different atoms...
during the analysis, thus overlooking crucial physicochemical information. In this work, we introduce the theory of PM to extract functional features of materials, aiming to address these limitations. We compare this approach with the functional information obtained through persistent homology.

Carborane, known for its stable cluster structures, forms a polyhedral cluster composed of borohydrides. Within this structure, it is possible to replace one or more borohydride B(H) vertices with C(H) units.8 Extensive research, both experimental and theoretical, has been conducted on the structure of these closo-carboranes, specifically those denoted as C\textsubscript{2n}B\textsubscript{n−2}H\textsubscript{n}10,11 Particular emphasis is placed on the stability of C\textsubscript{2n}B\textsubscript{n−2}H\textsubscript{n} (n = 5–20), especially concerning the increasing number of B atoms incorporated into the structure.5,12–18

In this study, we confirm that the proposed PM method is validated by using the cluster structure dataset of C\textsubscript{2n}B\textsubscript{n−2}H\textsubscript{n} (n = 5–20). For carboranes, our analysis is based on the fundamental concept that both bonding and nonbonding interactions within the structure can be utilized to estimate orbital energy. The structure’s relative stability is expressed directly in terms of relative energy. In this research, we use the system’s relative energy to characterize the structure’s stability,12 predicting structural stability by forecasting the relative energy. Simultaneously, we employ PM to extract the topological features of these structures, successfully quantifying the relationship between structural stability and topological features through a linear regression model.

We specifically utilized the PM method to capture the structural information of clusters. Through the application of linear regression and boosted tree regression models, we successfully made accurate predictions of relative energy. Moreover, we established a meaningful relationship between structural stability and PM. During cross-validation, our approach demonstrated precise predictions of relative energy, HOMO, LUMO and HOMO−LUMO gaps for the C\textsubscript{2n}B\textsubscript{n−2}H\textsubscript{n} (n = 5–20) system, highlighting a notable high correlation between the structural features extracted by PM and the properties of the clusters. Additionally, we conducted a comparative analysis with results obtained from persistent homology.

This paper is organized as follows. Section 2 will review relevant theoretical concepts, including simplicial complexes, persistent homology, magnitude and PM. We will then detail the molecular structure analysis method based on PM. Section 3 provides specific applications of PM, including the stability of closo-carborane structures and the prediction of HOMO, LUMO and HOMO−LUMO gaps. Finally, we present a conclusion in Sec. 4.

2. THEORY AND METHODS

In this section, we will review the fundamental concepts of topological data analysis (TDA), including simplicial complexes, homology and persistent homology. Additionally, we will introduce the magnitude and PM. PM is the key tool in our work, and it refines persistent homology, enhancing the barcode representation. Finally, we apply the PM method to the molecular structure of closo-carborane, predicting the stability and structure of closo-carborane molecules. We believe that this theoretical framework and methodology are applicable to other molecules and materials.

2.1. Simplicial complex and homology

A simplex is a geometric shape that generalizes from a point (0-simplex) to a line segment (1-simplex), a triangle (2-simplex) and even a tetrahedron (3-simplex). Essentially, it is like viewing these familiar shapes from a higher-dimensional perspective. In geometry, a simplex \( \sigma_n = \{a_0, ..., a_n\} \), where \( n \) represents its dimension, is a convex region enclosed by \( n + 1 \) vertices, that is

\[
\sigma_n = \left\{ x \mid x = \sum_{i=0}^{n} t_i a_i, \right. \\
\left. \text{where } \sum_{i=0}^{n} t_i = 1, 0 \leq t_i \leq 1, i = 0, ..., n \right\}.
\] (1)

In an intuitive sense, a simplicial complex behaves like a building, with simplices serving as the building blocks stacked together in a particular arrangement. Mathematically, a simplicial complex, denoted as \( K \), is essentially a collection of simplices satisfying two fundamental conditions:

1. Every face of a simplex within \( K \) is also a simplex of \( K \).
2. The intersection of any two simplices in \( K \) results in a face or an empty set.

Essentially, these conditions ensure a well-defined combination of simplices in the simplicial complex, much like the role of adhesive in constructing a building.

Let \( K \) be a simplicial complex. Within this context, we define a \( p \)-chain as an integer linear combination of \( p \)-simplices in \( K \), which can be denoted as \( c = \sum_{i=0}^{n} a_i \sigma_i \),
where \( n \) is the number of \( p \)-simplices in \( \mathcal{K} \) and \( a_i \in \mathbb{Z}_m \). Here, \( \mathbb{Z}_m \) is a cyclic group of integers modulo \( m \), where \( m \) is a prime number. For instance, \( \mathbb{Z}_2 \) comprises only two elements, 0 and 1. In this work, we employ \( \mathbb{Z}_2 \) coefficients, a choice that significantly reduces computational complexity. Subsequently, a \( p \)-chain group, denoted as \( C_p(\mathcal{K}; \mathbb{Z}_p) \), is a free abelian group consisting of all \( p \)-chains on the simplicial complex \( \mathcal{K} \). For the sake of simplicity, \( C_p(\mathcal{K}) \) is commonly used in place of \( C_p(\mathcal{K}; \mathbb{Z}_p) \).

Now, let us introduce another crucial concept in homology theory: the boundary operator. These operators play a pivotal role in describing and quantifying the boundaries of homology chains within topological spaces. The boundary operator is a map \( \partial : C_p(\mathcal{K}) \to C_{p-1}(\mathcal{K}) \). If \( \sigma = [v_0, \ldots, v_p] \) is a simplex with \( p > 0 \), we define

\[
\partial_p \sigma = \partial_p [v_0, \ldots, v_p] = \sum_{i=0}^{p} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_p],
\]

where \( \hat{v}_i \) means that this term \( v_i \) is omitted and \( \mathbb{Z}_2 \) is used. Specially, when \( p = 0 \), \( \partial_0 \sigma = 0 \). An essential property of boundary operators is that they act as a null operation when applied twice, implying that a closed boundary exhibits no further boundaries, represented by

\[
\partial_{p-1} \circ \partial_p = 0.
\]

The boundary operator \( \partial_p : C_p(\mathcal{K}) \to C_{p-1}(\mathcal{K}) \) has a kernel, which is referred to as the group of \( p \)-cycles and is denoted as \( Z_p(\mathcal{K}) \). It is a subgroup of \( C_p(\mathcal{K}) \). The image of \( \partial_{p+1} : C_{p+1}(\mathcal{K}) \to C_p(\mathcal{K}) \) defines the group of \( p \)-boundaries, denoted as \( B_p(\mathcal{K}) \). Due to the property of the boundary operator, which yields zero when applied twice, we conclude that \( B_p(\mathcal{K}) \) is a subgroup of \( Z_p(\mathcal{K}) \). In other words, \( B_p(\mathcal{K}) \subseteq Z_p(\mathcal{K}) \subseteq C_p(\mathcal{K}) \). Given this, we can construct a quotient group

\[
H_p(\mathcal{K}; \mathbb{Z}_p) = Z_p(\mathcal{K})/B_p(\mathcal{K}) = \ker \partial_p/\text{im} \partial_{p+1},
\]

which eliminates boundary factors and keeps nontrivial and valuable topological information. For simplicity, we use \( H_p(\mathcal{K}) \) for \( H_p(\mathcal{K}; \mathbb{Z}_p) \).

Another important topological invariant is called Betti number, which is defined by

\[
\beta_p = \text{rank}(H_p(\mathcal{K})) = \text{rank}(Z_p(\mathcal{K})) - \text{rank}(B_p(\mathcal{K})), \quad p \geq 0.
\]

It is worth noting that the generators of \( H_0(\mathcal{K}) \), \( H_1(\mathcal{K}) \) and \( H_2(\mathcal{K}) \) correspond to the connected components, loops or holes and voids within a topological space, respectively. The Betti number counts the generators of the homology group. For instance, \( \beta_0 \) denotes the number of connected components in a simplicial complex, \( \beta_1 \) signifies the count of loops or holes in a space, and \( \beta_2 \) represents the number of voids in a space, respectively.

### 2.2. Persistent homology and barcodes

Persistent homology is a method in TDA used to analyze the topological features and geometric shapes of a dataset or space.\(^{19}\) It aims to identify and track the persistence of these features across different scales.

Let \( X \) be a collection of discrete points in Euclidean space. For any positive real number \( \epsilon \), the Vietoris–Rips complex \( \mathcal{R}_\epsilon \) on the point set \( X \) is defined as a simplicial complex where each simplex is a subset of the point set \( X \), satisfying that the distance between any two points in that simplex is not greater than \( \epsilon \). For a family of positive real numbers \( \epsilon_1, \ldots, \epsilon_n \), we obtain a filtration of simplicial complexes as follows:

\[
\mathcal{R}_{\epsilon_1} \hookrightarrow \mathcal{R}_{\epsilon_2} \hookrightarrow \ldots \hookrightarrow \mathcal{R}_{\epsilon_n}.
\]

This family of simplicial complexes provides a multiscale representation of the point set. For any \( i \leq j \), we have an inclusion \( \mathcal{R}_{\epsilon_i} \hookrightarrow \mathcal{R}_{\epsilon_j} \). It induces a homomorphism of homology group \( H_i(\mathcal{R}_{\epsilon_i}) \rightarrow H_i(\mathcal{R}_{\epsilon_j}) \). The \((i, j)\)-persistent homology is defined by

\[
H^{ij}(\mathcal{R}) := \text{im}(H_j(\mathcal{R}_{\epsilon_i}) \rightarrow H_j(\mathcal{R}_{\epsilon_j})).
\]

For the Vietoris–Rips complex, our distance parameter \( \epsilon \) typically starts from 0, where \( \mathcal{R}_0 \) corresponds to the discrete point set. Its zero-dimensional Betti number equals the number of discrete points, while higher-dimensional Betti numbers are all 0. As the filtration parameter increases, points in \( X \) gradually connect, leading to the appearance of one-dimensional and higher-dimensional simplices in the complex \( \mathcal{R}_\epsilon \). As a result, the structure of the simplicial complex \( \mathcal{R}_\epsilon \) becomes more complex.

As points in \( X \) connect, the number of connected components gradually decreases, meaning some connected components disappear, resulting in a continual decrease in the zero-dimensional Betti number. Simultaneously, with the appearance of one-dimensional and higher-dimensional simplices, the simplicial complex \( \mathcal{R}_\epsilon \) may exhibit the generation of one-dimensional or higher-dimensional homology generators. For example, if \( \alpha \) is a generator appear at
The graph’s magnitude is like the size of a shape in geometry, similar to volume for a three-dimensional object. It might not be immediately clear what the magnitude of a graph really means in a simple, visual sense. Rather, it seems to be a function that reveals the profound, intrinsic features of a graph.

As shown in Fig. 1, the carbon structure of benzene can be visualized as a regular hexagon. The corresponding magnitude function for the hexagon diagram can be determined as follows:

$$M(t) = \frac{6(e^{-t} - 1)}{(e^{-3t} - 1)(e^{-t} + 1)}.$$  

Subsequently, we can plot the magnitude curve of the hexagon, as illustrated in Fig. 1(c). The magnitude of a graph can be loosely understood as the volume or size of a graph in the topological sense. It serves as a more precise and nuanced graph invariant. The information conveyed by the magnitude is distinctly different from that conveyed by existing graph invariants, such as the Tutte polynomial or the graph spectrum. In comparison, the magnitude provides a unique and refined characterization of a graph’s structural features.

### 2.4. Persistent magnitude

The concept of “persistent magnitude” was introduced within the framework of TDA and persistent homology. It serves as a mathematical tool to quantify significant features within topological data. In persistent homology, the persistence of topological features is elucidated by altering the resolution of the dataset, typically achieved by incrementing the radius parameter. This process results in the generation of a sequence of barcodes, where each barcode corresponds to a
topological feature and conveys information about its duration and intensity. The term “persistent magnitude” pertains to the weighted and symbolic enumeration of these barcode representations. Subsequently, we can derive the PM function, which examines how the magnitude of persistence modules evolves with respect to the parameter $t$. This function furnishes detailed insights into the temporal variations of magnitude within the persistence modules.

Through persistent homology, we derive a series of barcodes, each comprising bars denoted as $[a, b)$, where $a$ and $b$ mark the initiation and culmination of persistence, respectively. For each bar $[a, b)$, its weight is $(e^{-a} - e^{-b})$, which is a positive real number. The PM for a finitely presented persistence module $M$ is defined as the sum of all bar weights within these barcodes, that is

$$|M| = \sum_i \left( e^{-a_i} - e^{-b_i} \right). \quad (11)$$

Additionally, each bar is associated with a dimension, often represented as an integer denoted by $d_i$. In the calculation of PM, the dimension of the bar also affects its symbol, conventionally expressed as $(-1)^d$. In general, for a finitely presented graded persistence module $M_*$, the PM of $M_*$ is computed as follows:

$$|M_*| = \sum_i (-1)^d_i \left( e^{-a_i} - e^{-b_i} \right). \quad (12)$$

Here, $d_i$ denotes the dimension of the $i$-th bar. Now, building upon the previously defined notation, we can introduce the concept of the PM function for a finitely presented persistence module $M$. This function is denoted as $[0, \infty) \to \mathbb{R}$, where $t \mapsto |tM|$. In practical terms, if we have obtained barcodes from data, the PM function can be expressed as

$$|tM| = \sum_i \left( e^{-a_i} - e^{-b_i} \right), \quad (13)$$

where again by convention $e^{-\infty} = 0$.

As shown in Fig. 2, for each closo-carborane structure, we can obtain a corresponding magnitude function. We observe that the magnitude function is a monotonically increasing function, and as the parameter $t$ increases, the magnitude function converges and approximates the number of atoms in the corresponding closo-carborane structure. The values, slopes and patterns of the magnitude function can all serve as features for our data analysis. We will apply these features to analyze the structure of molecules.

2.5. Structure–stability modeling

In this section, we will employ the PM method to analyze molecular structures. The PM function will give us a range of magnitude features, each reflecting different aspects of the data’s topological and geometric

![Fig. 2](image-url) (Color online) (a) The structure representation of $C_2B_3H_5$; (b) the structure representation of $C_2B_8H_{10}$; (c) the structure representation of $C_2B_{15}H_{17}$; (d) the magnitude function of $C_2B_3H_5$; (e) the magnitude function of $C_2B_8H_{10}$; (f) the magnitude function of $C_2B_{15}H_{17}$. 

DOI: 10.1142/S2737416524400027

characteristics. As we may know, molecular structure determines molecular function, such as stability and energy. In this work, we will validate how PM can be utilized to predict the stability of closo-carboranes $C_2B_{n-2}H_n$ ($n = 5–20$).

The relative energy of closo-carboranes is calculated based on the difference in single-point energy, and can predict the structure. Therefore, we use it as the target. Through the additional energy of BH, we can obtain the relative energy of $C_2B_{n-2}H_n$ ($n = 5–20$) as follows:

$$
\Delta H_{add} = E(C_2B_{n-2}H_n) - E(C_2B_{n-2}H_5) - (n - 5)E(BH),
$$

(14)

where $E(BH)$ is the difference in energy between $B_6H_{10}$ and $B_7H_5$.

We will present two applications of the magnitude features in the carbon borane dataset. The first application involves predicting the stability of closo-carboranes. We set the energy of $C_2B_nH_2$ as the baseline value and adopted the most stable structure, which was reoptimized using the DFT method. In the second application, we conducted data augmentation, randomly generating 2193 structural data of $C_2B_{n-2}H_n$ ($n = 5–20$), and calculated the relevant HOMO values, LUMO values and their differences between each structure, namely, the HOMO–LUMO gap value. Later, we will provide more specific details.

The first application is to validate the PM method by utilizing the most stable structure of $C_2B_{n-2}H_n$ ($n = 5–20$). We obtain the PM curve from the barcodes obtained through persistent homology, and here we only consider the case of $H_5$. Further we establish the relationship between features and stability obtained through PM curves. When investigating the computed relative energy and the predicted energy, the presence of multiple features can lead to overfitting, and there may be correlations among these features. Therefore, we will perform separate linear regressions for different features to study the correlation between the magnitude features and the computed relative energy. In this part, we consider the PM function given by the zero-dimensional Betti numbers, written as $M(t) = \sum_i (e^{-a_i t} - e^{-b_i t})$. Here, $(a_i, b_i)$ runs across all the bars of zero-dimensional barcodes. It is worth noting that when $b_i$ is considered as infinity, $e^{-b_i t}$ becomes zero. In this work, we consider the following features:

1. The ratio of the slope of the magnitude curve at the origin to the number of atoms, denoted as $M'(0)/N$.
2. The ratio of the value of the magnitude curve at the origin to the number of atoms, denoted as $M(0)/N$.
3. The ratio of the value of the magnitude curve when approaching stability (curve slope < 0.05) to the number of atoms, denoted as $M_{\text{stability}}/N$.

After obtaining these three features, we performed leave-one-out cross-validation using linear regression. We utilized $M'(0)/N$, $M(0)/N$ and $M_{\text{stability}}/N$, respectively, as inputs for the model. This approach further accentuates the correlation between structural stability and the three features we have extracted.

The second task involves predicting the HOMO, LUMO and HOMO–LUMO gaps of closo-carboranes using the magnitude features. In this application, we utilize the PM function given by zero- to two-dimensional homology information

$$
M(t) = \sum_i (-1)^d \left( e^{-a_i t} - e^{-b_i t} \right). 
$$

(15)

Here, $(a_i, b_i)$ represents the interval from birth to death of the bar, and $d$ denotes the Betti number dimension corresponding to that bar. We set the filtration parameter from 0 Å to 4 Å, with a filtration interval of 0.04. In this manner, for each magnitude curve, we obtain a 100-dimensional vector. These vectors serve as the primary features for predicting the HOMO, LUMO and HOMO–LUMO gaps of closo-carboranes. During the prediction process, we employed a gradient boosting regression model for five-fold cross-validation. In this model, we set the parameters as follows: learning_rate = 0.05, max_depth = 7, min_samples_leaf = 1, min_samples_split = 5, subsample = 0.4 and n_estimators = 100. For HOMO, LUMO and HOMO–LUMO, we conducted eight separate runs. The commonly used parameters in this regression model to reflect prediction accuracy and effectiveness are the Pearson correlation coefficient (PCC), root mean square error (RMSE) and the mean absolute error (MAE). The PCC is defined by the following formula:

$$
PCC = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}.
$$

(16)

Here, $x_i$ and $y_i$ are the prediction result and the quantum mechanic calculation result, respectively. The terms $\bar{x}$ and $\bar{y}$ are the average predicted result and the average value of the quantum mechanic calculation result, respectively. Besides, $n$ denotes the number of computed data.
2.6. Computational methods

We conducted all our calculations using the Gaussian 16 program. First, we optimized all stable structures of closo-carboranes $C_xB_yH_{z+n}$ $(n = 5–20)$ and closo-borane dianions, as reported in the literature, using the B3LYP/6-311G(d, p) level of theory. Subsequently, we determined the relative energies of these different structures through single-point energy calculations. Besides, our code for computing magnitude function and data analysis is obtained with the help of ChatGPT, and one can refer to Ref. 26. Finally, we visually represented all structures using the VESTA software.

3. RESULTS AND DISCUSSION

3.1. PM-based stability analysis for closo-carboranes

In this section, we aim to predict the stability of closo-carboranes using magnitude features. Our objective is to verify the strong correlation between various magnitude features and the stability of closo-carboranes. Prior to this, we review the predictions of closo-carboranes stability based on persistent homology features. Specifically, we utilize the ratio of the average bar length of finite-length bars corresponding to zero-dimensional persistent homology to the number of atoms. Next, applying the PM theory outlined in Sec. 2, we obtain a series of magnitude curves. We employ the values and slopes of these magnitude curves as features to predict the relative energy of closo-carboranes. For each closo-carborane of the form $C_xB_yH_{z+n}$ $(n = 5–20)$, we have identified the most stable structure along with its relative energy. Consequently, we can employ 16 different most stable structures to validate the efficacy of the PM and persistent homology methods. In this study, we employ the simplest linear regression model $y=kx+b$ to investigate the correlation between magnitude features and the relative energy of closo-carboranes. Here, $k$ and $b$ represent the coefficients, and $x$ denotes the magnitude feature under consideration. As discussed in Sec. 2.5, we primarily focus on the following three magnitude features: (1) The ratio of the slope of the magnitude curve at the origin to the number of atoms, denoted as $M'(0)/N$. (2) The ratio of the value of the magnitude curve at the origin to the number of atoms, denoted as $M(0)/N$. (3) The ratio of the value of the magnitude curve when approaching stability (curve slope <0.05) to the number of atoms, denoted as $M_{\text{stability}}/N$. Here, $M(t)$ is the magnitude function and $N$ is the number of atoms of a closo-carborane.

The corresponding PCCs, RMSE and MAE are presented in Table 1. The PCCs for these tree magnitude features are all greater than 0.88, indicating a high correlation between various magnitude features and the relative energy of closo-carboranes. While we cannot guarantee that the magnitude curve provides a higher PCC compared to other methods, our analysis suggests that the magnitude curve itself consistently exhibits a strong correlation with the relative energy. We believe that employing more sophisticated and precise methods for analyzing this curve could lead to improved predictive results.

Figure 3 depicts that the performance of magnitude features is comparable to that of the average bar length derived from persistent homology. It is noteworthy that the prediction for the closo-carborane $C_2B_3H_5$, with its energy set as the baseline value, performs exceptionally well in Fig. 3(d). This indicates that, while the correlation between various features and the relative energy of closo-carboranes is approximately the same, these features exhibit distinct performances in predicting the relative energy for different closo-carboranes.

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

To further validate the effectiveness of the magnitude features, we will employ a PM-based method to predict the HOMO, LUMO and HOMO–LUMO gap of closo-carboranes. We will compute the magnitude function for 2193 closo-carborane structures, incorporating zero- to two-dimensional persistent homology information. Subsequently, we obtain a function value at intervals of 0.04 from the magnitude function within the range of 0–4 Å, resulting in a 100-dimensional vector. We then apply the gradient boosting tree algorithm to these 2193 vectors, leading to the predictive results illustrated in Fig. 4.
In light of the inherent variability in the results produced by the gradient boosting tree algorithm, it is important to note that the predicted outcomes may exhibit differences across multiple runs. To address this, we conducted 8 independent runs for the prediction of HOMO, LUMO and HOMO–LUMO gap, respectively. The ensuing Tables 2–4 provide a comprehensive overview of the PCC, RMSE and MAE for each set of predictions, shedding light on the algorithm's predictive stability.

From the data, it can be observed that our predictive performance is relatively stable, with PCCs consistently
exceeding 0.90. Notably, for both LUMO and HOMO–LUMO gap predictions, the PCCs have reached 0.94. These findings underscore the effectiveness of the PM-based method in predicting the structure of closo-carboranes. This implies that the PM-based method holds significant potential for applications in predicting material stability and functionality.

4. CONCLUSION

Building upon persistent homology, we employed the method of PM to extract structural features. In the methodology and theoretical aspects, we introduced the theory of magnitude and PM. Subsequently, to validate this approach, we conducted an analysis of closo-carborane structures. In the first task, we extracted three features from the zero-dimensional homology, applying linear regression to each feature separately, resulting in PCCs of 0.900, 0.882 and 0.883. Additionally, a comparison was made with a single feature extracted from persistent homology, yielding a PCC of 0.881. We observed substantial potential in the PM method for capturing information about different elements within the structure. In the second task, extracted the features from zero- to two-dimensional homology, and employed the PM method to extract feature information, combined with the GBRT algorithm for predicting the HOMO, LUMO and HOMO–LUMO gaps of closo-carboranes $C_2B_nH_{n+2}$ ($n = 5–20$). Each prediction was repeated eight times, with the lowest correlation coefficients being 0.9056, 0.9385 and 0.9427, respectively. In this work, we demonstrated the potential of PM in analyzing multi-element systems. Our results indicate that the PM-based method exhibits remarkable effectiveness in studying and analyzing molecular structures and functionalities. This method holds broader prospects for applications in molecular structure analysis. In future research, we intend to incorporate advanced machine learning models and apply PM to more intricate multi-element systems, striving for swift and precise predictions.

DATA AND CODE AVAILABILITY

The data and source code obtained in this work are publicly available in the GitHub repository: https://github.com/Wanyingbear/persistent-magnitude.
AUTHOR CONTRIBUTIONS

Wanying Bi carried out the detailed method and algorithms, wrote the first draft, and revised the manuscript. Xin Fu fixed the mathematical theory and revised the manuscript. Jingyan Li supervised the project, Jie Wu proposed the ideas, supervised the project, and acquired funding. The code was written with the assistance of ChatGPT.

FUNDING INFORMATION

The authors are supported by the Natural Science Foundation of China (NSFC Grant No. 11971144) and the start-up research fund from BIMSA.

CONFLICT OF INTEREST

The authors declare no competing interests.

ORCID

Jingyan Li https://orcid.org/0000-0002-5805-6520
Jie Wu https://orcid.org/0000-0003-2445-6750

References


