OXFORD

Briefings in Bioinformatics, 2024, bbae465
https://doi.org/10.1093/bib/bbae465
Problem Solving Protocol

Multi-Cover Persistence (MCP)-based machine learning for polymer property prediction

Yipeng Zhang[1], Cong Shen[2], Kelin Xia[1,*]

[1]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore
[2]Department of Mathematics, National University of Singapore, Singapore 119076, Singapore

*Corresponding author. E-mail: xiakelin@ntu.edu.sg

Abstract

Accurate and efficient prediction of polymers properties is crucial for polymer design. Recently, data-driven artificial intelligence (AI) models have demonstrated great promise in polymers property analysis. Even with the great progresses, a pivotal challenge in all the AI-driven models remains to be the effective representation of molecules. Here we introduce Multi-Cover Persistence (MCP)-based molecular representation and featurization for the first time. Our MCP-based polymer descriptors are combined with machine learning models, in particular, Gradient Boosting Tree (GBT) models, for polymers property prediction. Different from all previous molecular representation, polymer molecular structure and interactions are represented as MCP, which utilizes Delaunay slices at different dimensions and Rhomboid tiling to characterize the complicated geometric and topological information within the data. Statistic features from the generated persistent barcodes are used as polymer descriptors, and further combined with GBT model. Our model has been extensively validated on polymer benchmark datasets. It has been found that our models can outperform traditional fingerprint-based models and has similar accuracy with geometric deep learning models. In particular, our model tends to be more effective on large-sized monomer structures, demonstrating the great potential of MCP in characterizing more complicated polymer data. This work underscores the potential of MCP in polymer informatics, presenting a novel perspective on molecular representation and its application in polymer science.

Keywords: molecular representation; multi-cover persistence; machine learning; polymer data analysis

Introduction

Polymers have been widely used in various fields from daily life to agriculture and engineering [1–3]. In particular, the development of functional polymers has further expanded the use of synthetic polymers, which are crucial for advanced industrial technology, agricultural production, and medical devices [4–6]. The importance of polymers property prediction and polymer design and discovery can not be overestimated. The emerging field of polymer informatics provides insights into polymer data analysis. Especially, the recent great development of artificial intelligence (AI) models has demonstrated great potential for polymer design and discovery [7, 8].

There are mainly two types of AI models used in polymer data analysis, i.e. fingerprints/descriptors-based machine learning models and end-to-end deep learning models. The fingerprints/descriptors-based models make use of feature vectors calculated based on polymer sequential, structural, physical, or chemical information [9, 10]. These polymer features are then used in quantitative structure activity/property relationships (QSARs/QSPRs) models or combined with machine learning models, such as support vector machine, Gradient Boosting Tree (GBT), etc. In general, these models are suitable for small/medium-sized dataset and has better interpretability. Deep learning approaches are end-to-end models and are free from feature engineering. Among polymer deep learning models, polymer sequence-based Natural Language Processing (NLP) approaches are very popular [11, 12]. In these models, Simplified Molecular-Input Line-Entry system (SMILES) [13] is used to represent polymer structure as a sequence, then NLP tools are employed. Models like TransPolymer [11] and polyBERT [12] are built on this idea, while TransPolymer uses Transformer model and polyBERT uses BERT architecture. Recently, polymer Graph Neural Networks have been developed and used in various polymers property analysis: Zeng et al. [14] use Graph Convolutional Neural Networks to predict the dielectric constant and energy bandgap of polymers. Gurnani et al. [15] combine a graph neural network with multitask learning and other advanced deep learning techniques to created polyGNN for prediction of polymers property.

Deeply rooted in algebraic topology, topological data analysis (TDA) [16, 17] has demonstrated its great power in molecular representation and featurization[18–21], and their combination with learning models have achieved great successes in various steps of drug design, including protein-ligand binding affinity prediction [18,22–24], protein stability change upon mutation prediction [20,25], toxicity prediction [26–28], solvation free energy prediction [29,30], partition coefficient and aqueous solubility [31], binding pocket detection [32], protein mutation analysis [33–35], and drug discovery [36]. These models have demonstrated great advantages

Received: April 29, 2024. **Revised:** August 7, 2024. **Accepted:** September 5, 2024

over traditional molecular representations in D3R Grand challenge [37,38]. TDA-based models have also achieved great successes in material data analysis [39–48]. Mathematically, the key idea of TDA is persistent homology, which tracks the change of the topological invariant, i.e. homology generators, from simplicial complexes over a filtration process. The topological invariants are highly abstract and characterize the intrinsic data properties; thus, have a better transferability for data-driven learning models. Recently, multi-cover persistence (MCP) (or persistent multi-cover) model have been developed and demonstrated great potential for data representation and characterization [49]. Different from traditional TDA models, MCP employs a series of Delaunay slices and Rhomboid tiling; thus, it can extract richer topological and geometric information.

In this work, we develop persistent multi-cover based molecular representation and featurization, and combine them with machine learning models for polymers property prediction. Our MCP-based polymer representation can effective characterize the data information. In particular, the Delaunay slices, which is homotopic equivalent to Multi-covers, reveal more crucial molecular structure information. By modeling molecular structures as Delaunay slices of varying orders and computing their persistent homology barcodes as molecular descriptors, we integrate these descriptors into machine learning models for predicting polymer properties. Specifically, we employ a GBT model combined with Persistent Multi-cover features. Our models can outperform traditional fingerprint-based models and has similar accuracy with geometric deep learning models, on benchmark dataset. In particular, our model tends to be more effective on large-sized monomer structures, demonstrating the great potential of MCP in characterizing more complicated polymer data.

Results

Persistent Multi-cover (MCP)-based molecular representation

The key concept in our MCP model is multi-cover, which is associated with the region encompassed by a collection of spheres in a given space. This collection consists of spheres centered at each point within a specified point set, as illustrated in Fig. 1(A). For a formal definition, readers are directed to the Method section. An inherent characteristic of the Multi-cover is the notion of *order*, defined as the degree of overlap among these spheres. There is a particular interest in the structure of high-order Multi-covers, or k-fold covers, which represent regions overlapped by at least k spheres. This focus enables an analysis of the denser regions within the point cloud. The k-fold cover becomes increasingly representative of these high-density areas as the order k increases. Edelsbrunner [49] introduces various methods to characterize the shape of k-fold covers, thereby facilitating the application of the Multi-cover approach in data analysis. Specifically, he introduces the concept of the Delaunay slice, which can capture the entire shape information of the k-fold cover as illustrated in Fig. 1(C). We utilized Delaunay slices to compute the persistent homology barcode of the k-fold cover. Additionally, Rhomboid tiling can be employed to describe the relationship between multi-cover spaces of different orders k.

Multi-cover approach can be used for the characterization of molecular data. This method represents a significant advancement over traditional topological approaches due to its enhanced capability to extract geometric information closely correlated with molecular data. For example, comparing with the classical Vietoris-Rips complex, the classical Vietoris-Rips complex, which relies on pairwise distances between points, the Multi-cover method, through its implementation of Delaunay slices and Sliced Rhomboid tiling, emphasizes the analysis of the radii of minimal circumcircles of point subsets. This shift in focus allows for a richer extraction of geometric information, as the radius of a minimal circumcircle provides a more informative metric than simple point-to-point distances. Specifically, for a pair of points, the radius of their minimal circumcircle directly corresponds to the distance between them. For a triplet of points, this radius reflects not only the pairwise distances but also the included angle among them. Furthermore, when considering four points, the radius of their minimal circumcircle encompasses information about the dihedral angles in addition to the distances and included angles. These geometric information implicitly reflects molecular structure information, such as bond length, bond angle, and dihedral angle, which are pivotal for the structural, dynamic, and functional analysis of molecular systems across various disciplines, playing a crucial role in understanding and predicting the physical and chemical properties of molecules. Given this foundational significance, the complexes derived through the Multi-cover methodology offer a robust framework for capturing the structural and topological nuances of molecular systems.

MCP model is used for the representation and featurization of molecular structures and interactions at various different scales. Molecules are organized in a hierarchical manner, existing at multiple scales from atoms to molecules and complexes. The range of interactions between these entties varies from strong covalent bonds to weaker forces like hydrogen bonds, Van der Waals forces, and electrostatic interactions. Thus, representing molecules at multiple scales is crucial for understanding their structure and interactions. To achieve this, we apply a process known as filtration, commonly used in persistent modeling techniques such as homology/cohomology [16,17,50], persistent spectral [21,51,52], and persistent functions [53–59], to create layered molecular graphs. The Multi-cover naturally holds a bi-filtration structure, with parameter radius r and parameter order k. To calculate the filtration that fixes k and varies r in practice, we apply Voronoi tessellation to decompose the k-fold cover. Then, we construct an order-k Delaunay slice from the decomposed Multi-cover (see Fig. 1**C**). Increasing (or decreasing) r enlarges (or reduce) the Multi-cover space, adding (or removing) edges, triangles, and higher-dimensional base structures in the Delaunay slice, which is a finite simplicial structure that captures the topology of the k-fold cover. On the other hand, varying k adjusts the focus towards higher-order structures. This aspect of filtration is intricately described using Rhomboid tiling. The formal definitions of Voronoi tessellation, Delaunay slice, and Rhomboid tiling are provided in the Method section. As we change these parameters, certain structures in Multi-cover appear and vanish, which is recorded using the persistent homology barcode [60,61]. This barcode records the emergence and dissolution of structures, providing a comprehensive representation of the molecular framework [19,20,37,38,62–64]. The persistent homology barcode, corresponding to MCP model with parameter k fixed, can be efficiently calculated by several software tools [62,65,66].

MCP-based machine learning model

Our MCP-based learning model has several steps, including polymer data preprocessing, multi-cover persistence representation, MCP feature generation, and machine learning. The flowchart of constructing our model is shown in Fig. 2. The original polymer data consists of a list of SMILES strings representing the monomers of the polymer. We utilize RDKit to convert the SMILES

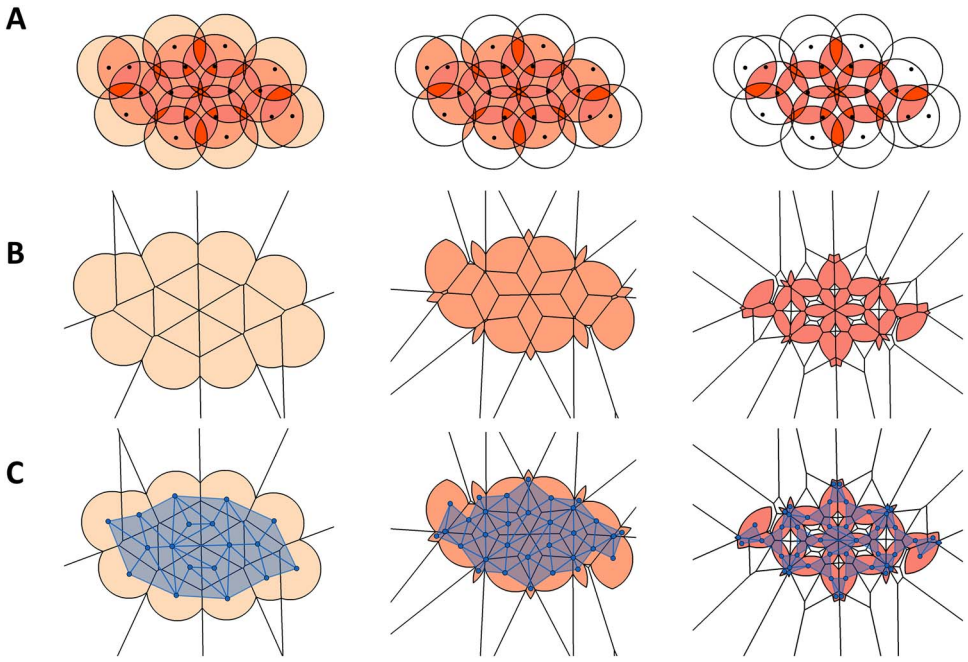

Figure 1. Illustration of the Multi-cover Representation of a Terephthalic Acid Molecule. (A) The 1-fold, 2-fold, and 3-fold covers of spheres associated with each atom, with each sphere having a radius of 1.5 Å. (B) Visualization of the 1-fold, 2-fold, and 3-fold cover spaces shown in (A), along with the corresponding order-1, order-2, and order-3 Voronoi tessellations. (C) Illustration of the order-1, order-2, and order-3 Delaunay slices, constructed from the segments decomposed by the order-1, order-2, and order-3 Voronoi tessellations, respectively.

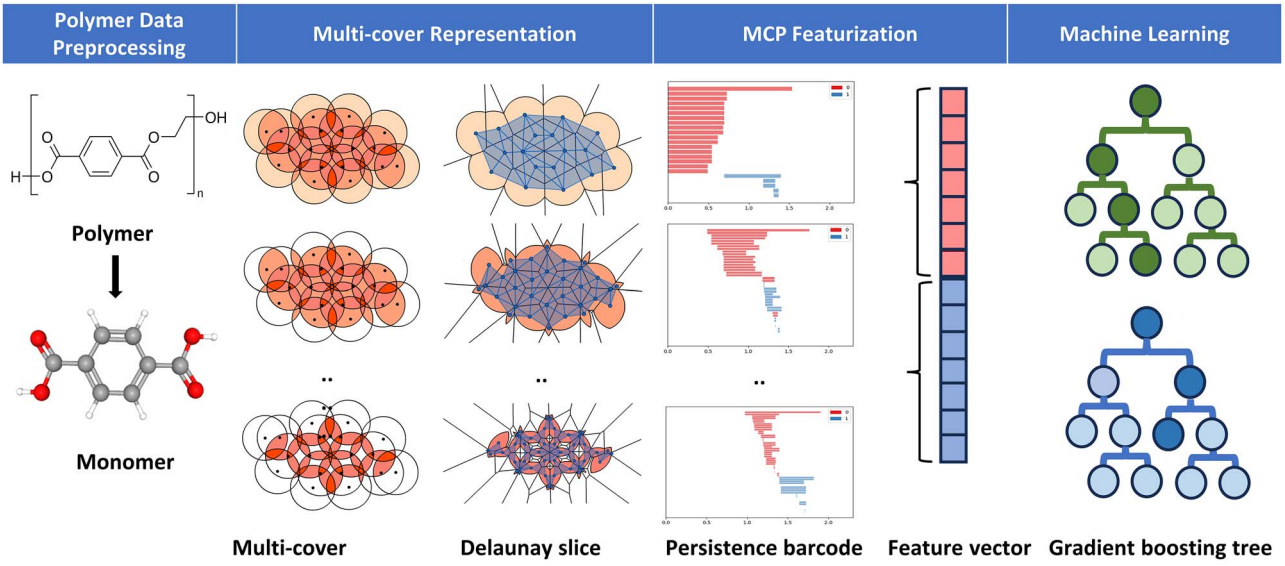

Figure 2. Flowchart illustrating the MCP-based machine learning models used in this study for polymers property predictions. In this work, we atomwisely compute the persistent Multi-cover barcode of a given polymer structure. The second slot of the flowchart illustrates the Multicover structure of monomer of Polyethylene Terephthalate when the radius of the circle is 1.5Å and its Delaunay slices with order 1,2,$\cdots$. The polymer structures are represented as vectors of statistics of persistent barcodes (homology dimension = 0, 1), which serve as features for the machine learning models. In this work, an XGBoost model is applied to the MCP-based features for polymers property prediction.

strings into molecule and add necessary hydrogen atoms, then get coordinates of the atoms in each monomer. Subsequently, the software *rhomboidtiling* developed in the paper [67] is applied to generate Delaunay slices of varying orders from these coordinates. For the computation of the persistent barcode of these Delaunay slices, we employ the GUDHI library. The resulting MCP feature vector comprises statistics of the bars in the barcode. To predict the property from this MCP-based feature, we employ eXtreme Gradient Boosting Tree (XGBoost) model.

Element-specific representation is employed in our MCP-based polymer representation. More specifically, we consider eight atom combinations, including {{all atoms}, {C}, {C, N}, {C, O}, {C, N, O}, {N, O}, {all atoms excluding C and H}, {all atoms excluding H}}. Furthermore, it employs four types of simplicial complex structures to represent polymer configurations, comprising four types of Delaunay slices (orders 1 through 4). For each Delaunay slice, we analyze 18 statistical measures derived from its persistent homology barcode. Specifically, we calculate the average, standard deviation, maximum, minimum, sum, and count of the distributions for the birth, death, and persistence values across all barcode entries. Consequently, the dimensionality of our MCP-based molecular descriptor totals

1152 = 8(atom combinations) × 4 (order of Delaunay slices) × 2 (homologydimensions) × 6(statistics) × 3(birth, death, and persistence of bars).

In our investigation of the MCP-based feature set, we evaluated several machine learning models, including Multilayer Perceptrons, Convolutional Neural Networks, Gradient Boosting Trees (GBT), and Random Forest (RF). Our analysis revealed that GBT models, specifically the eXtreme Gradient Boosting (XGBoost) implementation, outperformed others in terms of computational efficiency and prediction accuracy. Consequently, XGBoost was selected for further analysis. The performance of the predicted Power Conversion Efficiency (PCE) is quantitatively assessed using the average Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2) over a 5-fold cross-validation process. The 5-fold cross-validation was generated using random sampling, and the process was repeated 10 times, each with different random sampling. The Gradient Boosting Tree (GBT) model underwent 10 iterations, with the median RMSE and R^2 values for the 10 runs of 5-fold cross-validation serving as the ultimate results.

Here we consider multiple baseline models: RF models using Extended Connectivity Fingerprints (ECFP-RF) [68], artificial neural networks using ECFP (ECFP-ANN) [69], long short-term memory (LSTM) networks [70], and TransPolymers [11]. The performance metrics for these models are derived from the article introducing TransPolymer [11]. Additionally, we train models such as PolyBERT [12], Geometric-enhanced molecular representation learning (GEM) [71], and Molecular Geometric Deep Learning (Mol-GDL) [72] on the organic photovoltaic (OPV) dataset for comparison. All aforementioned models utilize 5-fold cross-validation through random sampling to obtain their respective results. For a fair comparison with these models, a combined feature consist of our MCP feature and three additional polymer properties: Highest Occupied Molecular Orbital in electronvolts (eV), Bandgap in eV, and molecular weight (Mw) in kilograms per mole (kg/mol) was also evaluated, as these propeties are also employed in models such as TransPolymer. Detailed configurations of the XGBoost model are presented in Supplementary Information. Table 1 summarizes the prediction outcomes using MCP features, the combined feature (MCP+3P), and other baseline models on the OPV dataset, indicating that our MCP-based model achieves state-of-the-art results. Interestingly, we observed a significant enhancement in overall model performance when our feature was fused with the three aforementioned chemical properties, highlighting the remarkable complementarity of our topology feature with chemical properties.

We have compared our model with molecular fingerprint based models on polymer datasets with relatively small-sized monomers. Compared to models including Extended Connectivity Fingerprints (ECFP) and polymer genome (PG) [73], our MCP models demonstrate superior predictive accuracy. Referencing Table 2, it is evident that our model achieves the highest performance across all datasets, with the sole exception of the smallest dataset, Eea, comprising only 368 data points. Notably, within the Xc dataset, models utilizing other fingerprint yield only negative R-squared values for their predictions, whereas our model achieves a notable result of 0.43.

Table 1. Performances of models in predicting polymers' PCE based on the OPV database [69], including Natural language processing-based models (LSTM [70], Transpolymer [11] and PolyBERT [12]), geometry-based graph neural network models (GEM [71] and Mol-GDL [72]) and fingerprint-based models(RandomForest using ECFP(ECFP-RF) [68], and ANN using ECFP(ECFP-ANN) [69]). Evaluation metrics are RMSE and Coefficient of determination (R^2). Evaluation metrics listed here are the average scores based on 5-fold cross-validation. We use bolder notations to denote the best result. Result marked with * was obtained using the model and methodology in its originally articles

Models	RMSE	R^2
TransPolymer [11]	1.92	0.32
PolyBERT [12]	2.02*	0.30*
GEM [71]	2.27*	0.26*
Mol-GDL [72]	1.89*	0.35*
ECFP-RF [68]	1.92	0.27
ECFP-ANN [69]	2.03	0.20
LSTM [70]	2.34	0.00
MCP	1.87	0.36
MCP+3P	**1.77**	**0.43**

Table 2. Performances of fingerprint-based models on multiple polymer databases [11], including ECFP, PG, and our MCP fingerprints. For PG fingerprint, two model including Gaussian process (GP) and ANN are utilized. Evaluation metrics is Coefficient of determination (R^2). Evaluation metrics listed here are the average scores based on 5-fold cross-validation. We use bolder notations to denote the best result. All the results of the baseline models come from the article introducing TransPolymer [11]

Models	Eea	EPS	Nc	Ei	Egb	Egc	Xc	OPV
ECFP-RF	0.70	0.50	0.56	0.57	0.66	0.65	−0.29	0.27
PG-GP	**0.90**	0.68	0.79	0.77	0.91	**0.90**	<0	NA
PG-ANN	0.87	**0.71**	0.78	0.74	0.89	0.89	<0	NA
MCP	0.84	**0.71**	**0.80**	**0.78**	**0.92**	**0.90**	**0.43**	**0.43**

Discussion

Deep learning models like TransPolymer have demonstrated proficiency in predicting properties of polymers composed of small monomers. However, their performance diminishes when applied to datasets containing predominantly medium-sized molecules, such as the OPV dataset. In contrast, our MCP-based model exhibits significant improvements over these models for such datasets. This enhancement is attributed to the model's precise characterization of molecular shapes across varying orders, with a notable efficacy for medium-sized molecules possessing complex geometries. The features generated by our model capture broader shape information compared to other models.

Further evaluation of the MCP model's performance on datasets comprising small monomers revealed that, although it did not surpass TransPolymer, the performance gap was small. Figure 3 compares the performance of our model with other models in different dataset with different average number of atoms. To be specific, The average number of atoms in dataset Eea, EPS, Nc, Ei, Egb, Egc, Xc, OPV is 23.5, 24, 24, 24, 25, 39, 53, 163, respectively. Analysis of the figure reveals a nuanced relationship between the average number of atoms in a dataset and the performance of our model relative to competing models. Specifically, for datasets where the average number of atoms is below 25, our model ranks as the third to fifth best in terms of performance. Notably, the performance gap between our model and the top-performing model in this atom count range is still small. As the average number of atoms increases to between 30 and 60, our model's performance improves to rank as the second best. This trend is most pronounced in datasets with a significantly higher average number of atoms, where our model not only emerges as the best

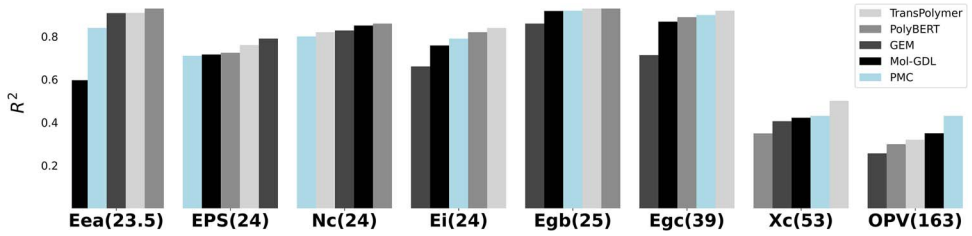

Figure 3. Evaluating MCP Models Versus Top Deep Learning Models: a Barplot Comparison on multiple Polymer Databases. The numbers in brackets indicate the average number of atoms per molecule in the databases. As the average number of atoms increases, the performance ranking of our model also improves. For the OPV dataset, our model's performance significantly surpasses that of other models. Evaluation metrics is Coefficient of determination (R^2).

performing model but also significantly surpasses other models in effectiveness. These results suggests a potential trend: as the size of the molecule per dataset increases, our model tends to be more effective. While these current observation is based on specific datasets, we are optimistic about its general validity. The advantage of our MCP approach lies in using topological invariants across different scales to represent and characterize polymer structures. These invariants capture intrinsic topological properties of the structures and can improve performance for complex data. This is consistent with prior works in TDA for molecular data [18–20]. Currently, we have only validated the superiority of our model over other models on the OPV dataset. We hope to find more polymer datasets, primarily containing medium-sized molecules, in the future to make this finding more robust. In conclusion, our MCP approach shows promise in capturing essential geometric and topological information across a diverse range of polymer sizes, potentially serving as a valuable tool for predicting polymer properties.

Methods

Persistent homology

*homology group*Homology groups are algebraic structures that capture the topological features of a geometric object. For example, 0-dimensional homology group captures the number of connected components in the object, and 1-dimensional homology group can reflect the information about holes in the object.

persistent homology barcode Persistent homology extends the concept of homology groups by examining how these topological features persist across multiple scales. This is done through a process called filtration, which is a nested sequence of simplicial complexes $K(\epsilon)$, each associated with a scale parameter ϵ. As ϵ increases, more simplices (vertices, edges, triangles, etc.) are added, and the topological features of the complexes evolve. Persistent homology tracks the appearance and disappearance of homology generators across the filtration. The result is a barcode that records the 'birth' and 'death' times (values of ϵ at which a generator appears and disappears). This barcode is a widely-used feature in TDA.

Multi-cover model

*k-fold cover*The k-fold cover refers to a specific subspace of $\mathbb{R}^d$ which is covered by at least k balls of radius r centered at the locally finite set X (see Fig. 1(A)):

$$\mathbf{Cover}_{r,k}(X) := \{y \in \mathbb{R}^d \mid \exists k \text{ points} \in X \text{ within a distance } r \text{ from } y\}$$

In the context of TDA, the application of $\mathbf{Cover}_{r,k}(X)$ necessitates its transformation into a finite complex. This conversion is essential for leveraging the topological and geometric properties of $\mathbf{Cover}_{r,k}(X)$ within computational frameworks. René Corbet [74] introduces the concept of the order-k Delaunay slice as an appropriate complex to encapsulate the topological structure of $\mathbf{Cover}_{r,k}(X)$:

$$\mathbf{Del}_{r,k}(X) := \mathrm{Nrv}\left(\{\mathbf{Cover}_{r,k}(X) \cap \mathrm{dom}(Q) \mid Q \subset X, |Q| = k\}\right),$$

where $\mathbf{Nrv}(C)$ denotes the nerve complex of a family of sets $C = (U_i)_{i\in I}$ [75] and $|Q|$ is the number of elements in Q. The domain $\mathrm{dom}(Q)$ is given by

$$\mathrm{dom}(Q) := \{p \in \mathbb{R}^d \mid \|p - x\| \leq \|p - y\|, \forall x \in Q, \forall y \in X \setminus Q\},$$

representing the region closer to every point in Q than to any point in $X \setminus Q$. The set

$$\mathbf{Vor}_k(X) := \left\{\mathrm{dom}(Q) \mid Q \subset X, |Q| = k\right\}$$

is recognized as the order-k Voronoi tessellation, which generalizes the well-known Voronoi diagram [76]. In essence, the order-k Delaunay slice represents the nerve of the convex segments of the k-fold cover dissected by the Voronoi diagram.

Equivalently, the order-k Delaunay slice, $\mathbf{Del}_{r,k}$, can be characterized by the following condition:

Vertices $v_{Q_1}, v_{Q_2}, \ldots, v_{Q_s}$ form an $(s-1)$-simplex in $\mathbf{Del}_{r,k}$ if and only if there exists a sphere S with radius $\leq r$ satisfying:

- The union $(Q_1 \cup Q_2 \cup \cdots \cup Q_s) = \{p \in X \mid p \text{ is inside } S \text{ or on } S\}$,
- The intersection $(Q_1 \cap Q_2 \cap \cdots \cap Q_s)$ contains the subset $\{p \in X \mid p \text{ is inside } S\}$,

where $Q_1, Q_2, \ldots, Q_s$ are distinct subsets of X, each containing exactly k points. Here, v_Q denotes the vertex in $\mathbf{Del}_{r,k}$ corresponding to the convex segment $\mathbf{Cover}_{r,|Q|}(X) \cap \mathrm{dom}(Q)$. This condition serves as a foundational principle for the practical construction of high-order Delaunay slices from a given point set. By leveraging this criterion, we systematically generate complex topological structures, enabling a deeper analysis of the underlying spatial relationships within the dataset.

Multi-Cover Persistence The concept of Multi-cover introduces a natural bi-filtration structure (see Fig. 1 **B**), characterized by the following relationships:

$$\mathbf{Cover}_{r,k}(X) \subseteq \mathbf{Cover}_{s,k}(X), \quad \forall s \geq r$$

$$\mathbf{Cover}_{r,k}(X) \subseteq \mathbf{Cover}_{r,l}(X), \quad \forall l \leq k$$

Within the scope of this paper, our focus primarily rests on the first filtration, wherein the parameter k is held constant and r is

varying. According to the *Persistence Nerve Theorem*, there exists a diagram of objectwise homotopy equivalences between this specific filtration and the filtration of the order-k Delaunay slice [74]. This implies that these two filtrations exhibit isomorphic persistent homology. Consequently, the order-k Delaunay slice can be utilized to compute the persistent homology barcode of k-fold cover in practical scenarios.

Rhomboid tiling Nevertheless, the Delaunay slice lacks a natural filtration when the parameter r remains fixed while k varies. This limitation indicates that Delaunay slices do not adequately represent the comprehensive bi-filtration structure inherent to the Multi-cover concept. A novel complex known as rhomboid tiling, is proposed [49], and it models a homotopy-equivalent bi-filtration structure aligned with that of the Multi-cover.

The motivation behind Rhomboid tiling lies in its correspondence to high-order α-shapes, which employ spheres of radius α to discern the shape of a point set X [77]. More precisely, for a sphere $S \subset \mathbb{R}^d$ with radius α, high-order α-shapes are concerned with the points of X that are inside S or on the boundary of S. Consequently, we define the set of points of X that are inside S as $\mathbf{In}_X S := \{p \in X \mid p \text{ is inside } S\}$, and the set of points on the boundary of S as $\mathbf{On}_X S := \{p \in X \mid p \text{ is on } S\}$, respectively.

To facilitate the connection between Delaunay slices of different orders via Rhomboid tiling, it is imperative to initially embed the vertices in Delaunay slices of all orders into $\mathbb{R}^{d+1}$. This embedding is defined by the transformation:

$$v_Q \mapsto y_Q := \left(\sum_{x \in Q} x, -|Q|\right) \in \mathbb{R}^{d+1}$$

where Q denotes a subset of the given point set X, and v_Q represents the vertex in the Delaunay slice corresponding to the convex segment $\mathbf{Cover}_{r,|Q|}(X) \cap \mathrm{dom}(Q)$.

Rhomboid tiling then integrates all points within X that reside inside or on the surface of a sphere S, utilizing their embedded counterparts in $\mathbb{R}^{d+1}$ to construct its polyhedral cells:

$$\mathbf{Rhomb}(X) := \{\rho_X(S) \mid S \text{ is a sphere in } \mathbb{R}^d\},$$

here $\rho_X(S) := \mathrm{conv}\{y_Q \mid \mathbf{In}_X S \subset Q \subset \mathbf{In}_X S \cup \mathbf{On}_X S\}$ denotes the convex hull of the embedding points. We define $r(\rho_X(S))$ as radius of the minimal sphere S' such that $\mathbf{In}_X S' = \mathbf{In}_X S$ and $\mathbf{On}_X S' = \mathbf{On}_X S$.

The bi-filtration structure of (sliced) Rhomboid tiling is formally defined by the equation:

$$\mathbf{SRhomb}_{r,k}(X) := \{\rho_X(S) \cap I_t \mid r(\rho_X(S)) \le r, t \ge k\}$$

here I_t is defined as the set $\{(x_1, x_2, \cdots, x_{d+1}) \in \mathbb{R}^{d+1} \mid t \le -x_{d+1} \le t+1\}$.

The *Persistence Nerve Theorem* offers a foundational basis for proving that the bifiltration $\mathbf{Cover}_{r,k}$ is homotopic equivalent to $\mathbf{SRhomb}_{r,k}(X)$ [74]. Moreover, Edelsbrunner et al. [67] have developed computational tools for the bifiltration of the sliced Rhomboid tiling associated with a point set. Consequently, this facilitates the practical computation of 2-parameter persistent homology for the Multi-cover.

Dataset details

Table 3 provides a summary of the eight datasets discussed in this study. These datasets can be accessed at the provided GitHub repository (https://github.com/ZhangYipeng01/MCP) or as detailed in the referenced work [11].

Table 3. Summary of all datasets

Dataset	Property	# Data	Data split
Eea [78]	electron affinity	368	5-fold cross-validation
EPS [78]	dielectric constant	382	5-fold cross-validation
Nc [78]	refractive index	382	5-fold cross-validation
Ei [78]	ionization energy	370	5-fold cross-validation
Egb [78]	bandgap (bulk)	561	5-fold cross-validation
Egc [78]	bandgap (chain)	3380	5-fold cross-validation
Xc [78]	crystallization tendency	432	5-fold cross-validation
OPV [69]	power conversion efficiency	1203	5-fold cross-validation

Table 4. Average performance and sample standard deviation (Std) of our model across 10 different random data splits

Datasets	Eea	EPS	Nc	Ei	Egb	Egc	Xc	OPV
Average R^2	0.84	0.71	0.80	0.78	0.92	0.90	0.43	0.43
Std	0.006	0.012	0.012	0.013	0.003	0.002	0.010	0.009

Cross-validation

We apply 5-fold cross-validation to all datasets to evaluate our model's performance. To ensure robustness, we perform random sampling to generate 10 different data splits. The final result is obtained by averaging the model's performance across these 10 splits. Table 4 presents the average performance and the sample standard deviation of our model across these splits. The relatively small sample standard deviation indicates that variations in data splits do not significantly affect our model's performance.

Polymer data processing

We implement data augmentation by considering different monomer units for a polymer as illustrated in Fig. 4. We rotate the SMILES strings repetitively by assuming the two ends are "connected". For each polymer, we generate four augmented monomer SMILES strings in both training and validation datasets. To predict the polymer property, the averaged value of its four augmented monomers is employed. That is our model will give four predicted values for the same polymer but with four augmented monomer structures, its final result is the average value.

In datasets Eea, EPS, Nc, Ei, Egb, Egc, and Xc, the SMILES strings contain an asterisk (*), which serves as a placeholder. A capping process is needed to transfer the SMILES string into a well-defined monomer molecule. In our case, we replace (*) with CH_3 cap. That is to assign a CH_3 to the two ends of the SMILES sequence and generate the corresponding monomer molecule. Note that other different cap approach can also be considered. We then generated the mol file using `rdkit.Chem.MolFromSmiles()` and added hydrogen atoms using `rdkit.Chem.AddHs()`. The Merck Molecular Force Field (MMFF) was employed to obtain the atomic coordinates for further computation.

Machine Learning Details

In our MCP-based learning models, we opt for the XGBoost approach. The hyperparameters utilized for XGBoost are detailed in Table 5:

The hyperparameters used in this study are adapted from our previous works [79]. Specifically, we adjusted the number of estimators to 8000 and the learning rate to 0.001, while keeping the other hyperparameters, such as 'max depth' and 'subsample',

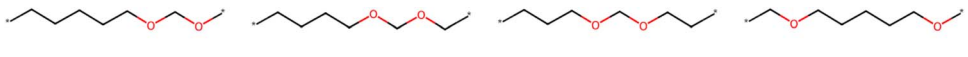

Figure 4. The data augmentation method. All the four monomers represent the same polymer.

Table 5. XGBoost hyperparameters

learning rate	# estimators	max depth	subsample
0.001	8000	6	0.7
colsample bytree	gamma	reg alpha	reg lambda
1	0	0	0

unchanged. All remaining hyperparameters were set to their default values.

Multi-cover based feature

Molecular descriptors have been identified as critical components in the accurate prediction of chemical properties. In this study, we investigate eight atom combinations for feature generation, specifically: {all atoms}, {C}, {C, N}, {C, O}, {C, N, O}, {N, O}, {all atoms excluding C and H}, {all atoms excluding H}. For each specified atom set, we calculate the Delaunay tessellation at four orders of $k = 1, 2, 3, 4$. Subsequently, we analyze the persistent homology barcodes derived from these Delaunay slices at homology dimensions 0 and 1.

To encode the information encapsulated in these persistent homology barcodes into a quantifiable feature vector, we employ a methodology similar to the barcode statistics introduced by Cang et al. [20]. This approach involves defining sets for Birth=$\{b_\alpha\}_{\alpha\in A}$, Death=$\{d_\alpha\}_{\alpha\in A}$, and Persistence=$\{d_\alpha - b_\alpha\}_{\alpha\in A}$ for a given barcode $B = \{[b_\alpha, d_\alpha]\}_{\alpha\in A}$, from which we derive three statistical feature vectors F_b, F_d, and F_p. These vectors encapsulate the statistical attributes—average (avg), standard deviation (std), maximum (max), minimum (min), sum (sum), and count (cnt)—of the distributions of birth, death and persistence value of bars. For instance, the feature vector F_b is composed of six statistics values avg(Birth), std(Birth), max(Birth), min(Birth), sum(Birth), and cnt(Birth).

In summary, for each given point cloud, our methodology yields a feature vector with a dimensionality of $1152 = 8$ (atom combinations) × 4(order of Delaunay slices) × 2(homology dimensions) × 6(statistics) × 3(birth, death, and persistence of bars), providing a robust foundation for subsequent property prediction tasks.

Key Points

Our main contributions in this paper are as follows:

- We propose Multi-Cover persistence (MCP)-based molecular representation and featurization for the first time.
- We develop MCP-based polymer descriptors and combine it with Gradient Boosting Tree (GBT) models for polymers property prediction.
- Our developed MCP machine learning model can outperform traditional fingerprint-based models and has similar accuracy with geometric deep learning models.

Author contributions

K.X. conceived and designed the study. Y.Z. performed the calculation. Y.Z., C.S., and K.X. contributed to the preparation of the manuscript.

Conflict of interest: None declared.

Funding

This work was supported in part by Nanyang Technological University SPMS Collaborative Research Award 2022, Singapore Ministry of Education Academic Research fund (Tier 2 grants MOE-T2EP20220-0010 and MOE-T2EP20221-0003).

Data and code availability

The data and source codes that support the findings of this study are present in the paper. The source codes can be downloaded from https://github.com/ZhangYipeng01/MCP. All the data in this study are publicly accessible. You can find the original datasets in the cited studies. Additionally, both datasets and codes for this study are available on GitHub at https://github.com/ZhangYipeng01/MCP.

References

1. Coates GW, Getzler YDYL. Chemical recycling to monomer for an ideal, circular polymer economy. *Nat Rev Mater* 2020;**5**:501–16. https://doi.org/10.1038/s41578-020-0190-4.
2. Puoci F, Iemma F, Spizzirri UG. *et al*. Polymer in agriculture: a review. *Am J Agric Biol Sci* 2008;**3**:299–314.
3. Spicer CD. Hydrogel scaffolds for tissue engineering: the importance of polymer choice. *Polym Chem* 2020;**11**:184–219. https://doi.org/10.1039/C9PY01021A.
4. Feng L, Zhu C, Yuan H. *et al*. Conjugated polymer nanoparticles: Preparation, properties, functionalization and biological applications. *Chem Soc Rev* 2013;**42**:6620–33. https://doi.org/10.1039/c3cs60036j.
5. Fox ME, Szoka FC, Fréchet JMJ. Soluble polymer carriers for the treatment of cancer: the importance of molecular architecture. *Acc Chem Res* 2009;**42**:1141–51. https://doi.org/10.1021/ar900035f.
6. Lv F, Qiu T, Liu L. *et al*. Recent advances in conjugated polymer materials for disease diagnosis. *Small* 2016;**12**:696–705. https://doi.org/10.1002/smll.201501700.
7. Audus DJ, de Pablo JJ. Polymer informatics: opportunities and challenges. *ACS Macro Lett* 2017;**6**:1078–82. https://doi.org/10.1021/acsmacrolett.7b00228.
8. Chen L, Pilania G, Batra R. *et al*. Polymer informatics: current status and critical next steps. *Mater Sci Eng: R: Rep* 2021a;**144**:100595. https://doi.org/10.1016/j.mser.2020.100595.
9. Pal R, Bourgeois L, Weyland M. *et al*. Chemical fingerprinting of polymers using electron energy-loss spectroscopy. *ACS Omega* 2021;**6**:23934–42. https://doi.org/10.1021/acsomega.1c02939.

10. Wattjes J, Niehues A, Cord-Landwehr S. *et al*. Enzymatic production and enzymatic-mass spectrometric fingerprinting analysis of chitosan polymers with different nonrandom patterns of acetylation. *J Am Chem Soc* 2019;**141**:3137–45. https://doi.org/10.1021/jacs.8b12561.
11. Changwen X, Wang Y, Farimani AB. Transpolymer: a transformer-based language model for polymer property predictions. *NPJ Comput Mater* 2023;**9**:64.
12. Kuenneth C, Ramprasad R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat Commun* 2023;**14**:4099. https://doi.org/10.1038/s41467-023-39868-6.
13. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6. https://doi.org/10.1021/ci00057a005.
14. Zeng M, Kumar JN, Zeng Z. *et al*. Graph convolutional neural networks for polymers property prediction. arXiv preprint arXiv:1811.06231. 2018.
15. Gurnani R, Kuenneth C, Toland A. *et al*. Polymer informatics at scale with multitask graph neural networks. *Chem Mater* 2023;**35**:1560–7. https://doi.org/10.1021/acs.chemmater.2c02991.
16. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geom* 2002;**28**:511–33. https://doi.org/10.1007/s00454-002-2885-2.
17. Zomorodian A, Carlsson G. Computing persistent homology. *Discrete Comput Geom* 2005;**33**:249–74. https://doi.org/10.1007/s00454-004-1146-y.
18. Cang ZX, Wei GW. TopologyNet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput Biol* 2017a;**13**:e1005690. https://doi.org/10.1371/journal.pcbi.1005690.
19. Nguyen DD, Cang ZX, Wei GW. A review of mathematical representations of biomolecular data. *Phys Chem Chem Phys* 2020;**22**:4343–67. https://doi.org/10.1039/C9CP06554G.
20. Cang ZX, Mu L, Wei GW. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput Biol* 2018;**14**:e1005929. https://doi.org/10.1371/journal.pcbi.1005929.
21. Meng Z, Xia K. Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Sci Adv* 2021;**7**:eabc5329. https://doi.org/10.1126/sciadv.abc5329.
22. Nguyen DD, Xiao T, Wang ML. *et al*. Rigidity strengthening: a mechanism for protein–ligand binding. *J Chem Inf Model* 2017;**57**:1715–21. https://doi.org/10.1021/acs.jcim.7b00226.
23. Cang Z, Wei G-W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int J Numer Methods Biomed Eng* 2018;**34**:e2914. https://doi.org/10.1002/cnm.2914.
24. Nguyen DD, Wei GW. AGL-score: algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *J Chem Inf Model* 2019;**59**:3291–304. https://doi.org/10.1021/acs.jcim.9b00334.
25. Cang ZX, Wei GW. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics* 2017c;**33**:3549–57. https://doi.org/10.1093/bioinformatics/btx460.
26. Wu KD, Wei GW. Quantitative toxicity prediction using topology based multi-task deep neural networks. *J Chem Inf Model* 2018;**58**:520–31.
27. Chen D, Gao K, Nguyen DD. *et al*. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat Commun* 2021b;**12**:1–9. https://doi.org/10.1038/s41467-021-23720-w.
28. Jiang J, Wang R, Wei G-W. GGL-tox: geometric graph learning for toxicity prediction. *J Chem Inf Model* 2021a;**61**:1691–700. https://doi.org/10.1021/acs.jcim.0c01294.
29. Wang B, Zhao ZX, Wei GW. Automatic parametrization of nonpolar implicit solvent models for the blind prediction of solvation free energies. *J Chem Phys* 2016;**145**:124110. https://doi.org/10.1063/1.4963193.
30. Wang B, Wang CZ, Wu KD. *et al*. Breaking the polar-nonpolar division in solvation free energy prediction. *J Comput Chem* 2018;**39**:217–33. https://doi.org/10.1002/jcc.25107.
31. Wu KD, Zhao ZX, Wang RX. *et al*. TopP–S: persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J Comput Chem* 2018;**39**:1444–54. https://doi.org/10.1002/jcc.25213.
32. Zhao RD, Cang ZX, Tong YY. *et al*. Protein pocket detection via convex hull surface evolution and associated Reeb graph. *Bioinformatics* 2018;**34**:i830–7. https://doi.org/10.1093/bioinformatics/bty598.
33. Wang M, Cang Z, Wei G-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence* 2020a;**2**:116–23. https://doi.org/10.1038/s42256-020-0149-6.
34. Chen J, Wang R, Wang M. *et al*. Mutations strengthened SARS-CoV-2 infectivity. *J Mol Biol* 2020a;**432**:5212–26. https://doi.org/10.1016/j.jmb.2020.07.009.
35. Wang R, Hozumi Y, Yin C. *et al*. Mutations on COVID-19 diagnostic targets. *Genomics* 2020b;**112**:5204–13. https://doi.org/10.1016/j.ygeno.2020.09.028.
36. Gao K, Nguyen DD, Meihua T. *et al*. Generative network complex for the automated generation of drug-like molecules. *J Chem Inf Model* 2020;**60**:5682–98. https://doi.org/10.1021/acs.jcim.0c00599.
37. Nguyen DD, Cang ZX, Wu KD. *et al*. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R grand challenges. *J Comput Aided Mol Des* 2019a;**33**:71–82. https://doi.org/10.1007/s10822-018-0146-6.
38. Nguyen DD, Gao KF, Wang ML. *et al*. MathDL: mathematical deep learning for D3R grand challenge 4. *J Comput Aided Mol Des* 2019b;**34**:131–47. https://doi.org/10.1007/s10822-019-00237-5.
39. Nakamura T, Hiraoka Y, Hirata A. *et al*. Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology* 2015;**26**:304001. https://doi.org/10.1088/0957-4484/26/30/304001.
40. Hiraoka Y, Nakamura T, Hirata A. *et al*. Hierarchical structures of amorphous solids characterized by persistent homology. *Proc Natl Acad Sci* 2016;**113**:7035–40. https://doi.org/10.1073/pnas.1520877113.
41. Saadatfar M, Takeuchi H, Robins V. *et al*. Pore configuration landscape of granular crystallization. *Nat Commun* 2017;**8**:15082. https://doi.org/10.1038/ncomms15082.
42. Lee Y, Barthel SD, Paweł Dłotko S. *et al*. Quantifying similarity of pore-geometry in nanoporous materials. *Nat Commun* 2017;**8**:1–8. https://doi.org/10.1038/ncomms15396.
43. Krishnapriyan AS, Haranczyk M, Morozov D. Topological descriptors help predict guest adsorption in nanoporous materials. *J Phys Chem C* 2020;**124**:9360–8. https://doi.org/10.1021/acs.jpcc.0c01167.
44. Chen X, Chen D, Weng M. *et al*. Topology-based machine learning strategy for cluster structure prediction. *J Phys Chem Lett* 2020b;**11**:4392–401. https://doi.org/10.1021/acs.jpclett.0c00974.

45. Jiang Y, Chen D, Chen X. *et al*. Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *npj computational materials* 2021b;**7**:28. https://doi.org/10.1038/s41524-021-00493-w.
46. Li S, Liu Y, Chen D. *et al*. Encoding the atomic structure for machine learning in materials science. *Wiley Interdiscip Rev: Computat Mol Sci* 2022;**12**:e1558. https://doi.org/10.1002/wcms.1558.
47. Liu J, Chen D, Pan F. *et al*. Neighborhood path complex for the quantitative analysis of the structure and stability of carboranes. *J Comput Biophys Chem* 2023;**22**:503–11. https://doi.org/10.1142/S2737416523500229.
48. Chen D, Liu J, Jie W. *et al*. Path topology in molecular and materials sciences. *J Phys Chem Lett* 2023;**14**:954–64. https://doi.org/10.1021/acs.jpclett.2c03706.
49. Edelsbrunner H, Osang G. The multi-cover persistence of Euclidean balls. *Discrete Comput Geom* 2021;**65**:1296–313. https://doi.org/10.1007/s00454-021-00281-9.
50. Verri A, Uras C, Frosini P. *et al*. On the use of size functions for shape analysis. *Biol Cybern* 1993;**70**:99–107. https://doi.org/10.1007/BF00200823.
51. Wang R, Nguyen DD, Wei G-W. Persistent spectral graph. *Int J Numer Methods Biomed Eng* 2020c;**36**:e3376. https://doi.org/10.1002/cnm.3376.
52. Liu X, Feng H, Jie W. *et al*. Persistent spectral hypergraph based machine learning (PSH-ML) for protein-ligand binding affinity prediction. *Brief Bioinform* 2021;**22**:bbab127. https://doi.org/10.1093/bib/bbab127.
53. Bauer U, Schönlieb CB, Wardetzky M. Total variation meets topological persistence: a first encounter. *AIP Conf Proc* 2010;**1281**:1022.
54. Xia KL, Zhao ZX, Wei GW. Multiresolution topological simplification. *J Comput Biol* 2015a;**22**:887–91. https://doi.org/10.1089/cmb.2015.0104.
55. Xia KL, Zhao ZX, Wei GW. Multiresolution persistent homology for excessively large biomolecular datasets. *J Chem Phys* 2015b;**143**:10B603_1. https://doi.org/10.1063/1.4931733.
56. Merelli E, Rucco M, Sloot P. *et al*. Topological characterization of complex systems: using persistent entropy. *Entropy* 2015;**17**:6872–92. https://doi.org/10.3390/e17106872.
57. Xia KL, Li ZM, Mu L. Multiscale persistent functions for biomolecular structure characterization. *Bull Math Biol* 2018;**80**:1–31. https://doi.org/10.1007/s11538-017-0362-6.
58. Wee JJ, Xia K. Forman persistent ricci curvature (FPRC) based machine learning models for protein-ligand binding affinity prediction. *Brief Bioinform* 2021;**22**:bbab136. https://doi.org/10.1093/bib/bbab136.
59. Wee JJ, Xia K. Ollivier persistent ricci curvature-based machine learning for the protein–ligand binding affinity prediction. *J Chem Inf Model* 2021b;**61**:1617–26. https://doi.org/10.1021/acs.jcim.0c01415.
60. Carlsson G, Zomorodian A, Collins A. *et al*. Persistence barcodes for shapes. *Int J Shape Model* 2005;**11**:149–87. https://doi.org/10.1142/S0218654305000761.
61. Ghrist R. Barcodes: the persistent topology of data. *Bull Am Math Soc* 2008;**45**:61–75.
62. Pun CS, Lee SX, Xia K. Persistent-homology-based machine learning: a survey and a comparative study. *Artif Intell Rev* 2022;**55**:5169–5213. https://doi.org/10.1007/s10462-021-10080-7.
63. Wei GW. Mathematics at the eve of a historic transition in biology. *Comput Math Biophys* 2017;**5**:138–41. https://doi.org/10.1515/mlbmb-2017-0009.
64. Xia KL, Wei GW. A review of geometric, topological and graph theory apparatuses for the modeling and analysis of biomolecular data. arXiv preprint arXiv:1612.01735. 2016.
65. Fasy BT, Kim J, Lecci F. *et al*. Introduction to the R package TDA. arXiv preprint arXiv:1411.1830. 2014.
66. Wasserman L. Topological data analysis. *Annu Rev Stat Appl* 2018;**5**:501–32. https://doi.org/10.1146/annurev-statistics-031017-100045.
67. Edelsbrunner H, Osang G. A simple algorithm for higher-order delaunay mosaics and alpha shapes. *Algorithmica* 2023;**85**:277–95. https://doi.org/10.1007/s00453-022-01027-6.
68. Cereto-Massagué A, Ojeda MJ, Valls C. *et al*. Molecular fingerprint similarity search in virtual screening. *Methods* 2015;**71**:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.
69. Nagasawa S, Al-Naamani E, Saeki A. Computer-aided screening of conjugated polymers for organic solar cell: classification by Random Forest. *The Journal of Physical Chemistry Letters* 2018;**9**:2639–46. https://doi.org/10.1021/acs.jpclett.8b00635.
70. Simine L, Allen TC, Rossky PJ. Predicting optical spectra for optoelectronic polymers using coarse-grained models and recurrent neural networks. *Proc Natl Acad Sci* 2020;**117**:13945–8.
71. Fang X, Liu L, Lei J. *et al*. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* 2022;**4**:127–34. https://doi.org/10.1038/s42256-021-00438-4.
72. Shen C, Luo J, Xia K. Molecular geometric deep learning. *Cell Rep Methods* 2023;**3**:100621. https://doi.org/10.1016/j.crmeth.2023.100621.
73. Kim C, Chandrasekaran A, Huan TD. *et al*. Polymer genome: a data-powered polymer informatics platform for property predictions. *J Phys Chem C* 2018;**122**:17575–85. https://doi.org/10.1021/acs.jpcc.8b02913.
74. Corbet R, Kerber M, Lesnick M. *et al*. Computing the multicover bifiltration. *Discrete Comput Geom* 2023;**70**:376–405. https://doi.org/10.1007/s00454-022-00476-8.
75. Matoušek J, Björner A, Ziegler GM. *et al*. Using the Borsuk-Ulam theorem: lectures on topological methods in combinatorics and geometry. Springer; 2003.
76. Boots B, Okabe A, Sugihara K. Spatial tessellations. *Geogr Inf Syst* 1999;**1**:503–26.
77. Krasnoshchekov D, Polishchuk V. Order-k α-hulls and α-shapes. *Inf Process Lett* 2014;**114**:76–83. https://doi.org/10.1016/j.ipl.2013.07.023.
78. Kuenneth C, Rajan AC, Tran H. *et al*. Polymer informatics with multi-task learning. *Patterns* 2021;**2**:100238. https://doi.org/10.1016/j.patter.2021.100238.
79. Liu X, Feng H, Lü Z. *et al*. Persistent tor-algebra for protein–protein interaction analysis. *Brief Bioinform* 2023. Oxford University Press;**24**:bbad046. https://doi.org/10.1093/bib/bbad046.