MOLECULAR BIOLOGY

Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction

Zhenyu Meng and Kelin Xia*

Molecular descriptors are essential to not only quantitative structure-activity relationship (QSAR) models but also machine learning–based material, chemical, and biological data analysis. Here, we propose persistent spectral–based machine learning (PerSpect ML) models for drug design. Different from all previous spectral models, a filtration process is introduced to generate a sequence of spectral models at various different scales. PerSpect attributes are defined as the function of spectral variables over the filtration value. Molecular descriptors obtained from PerSpect attributes are combined with machine learning models for protein-ligand binding affinity prediction. Our results, for the three most commonly used databases including PDBbind-2007, PDBbind-2013, and PDBbind-2016, are better than all existing models, as far as we know. The proposed PerSpect theory provides a powerful feature engineering framework. PerSpect ML models demonstrate great potential to significantly improve the performance of learning models in molecular data analysis.

INTRODUCTION

Data-driven learning models are among the most important and rapidly evolving areas in chemoinformatics and bioinformatics (*1*). Greatly benefiting from the accumulation of experimental data, machine learning models have contributed significantly to various aspects of virtual screening in drug design. In particular, machine learning–based models have achieved a better accuracy than traditional physics-, knowledge-, and empirical-based models in protein-ligand binding affinity prediction (*2*–*4*). Featurization, or feature engineering, is key to the performance of machine learning models in material, chemical, and biological systems. More than 5000 molecular descriptors and chemical descriptors have been proposed to characterize the structural, physical, chemical, and biological properties (*5*). These descriptors capture information from molecular formula, fragments, motifs, topological features, geometric features, conformation properties, hydrophobicity, electronic properties, steric properties, etc. They are widely used in quantitative structure-activity relationships (QSARs) and quantitative structure-property relationships (QSPRs). These descriptors can be combined to form a fixed-length vector, known as molecular fingerprint. Molecular fingerprints, which can be generated from various software packages, such as RDKit (*6*), Open Babel (*7*), and ChemoPy (*8*), are widely used in machine learning models.

Recently, advanced mathematical models from algebraic topology, differential geometry, and algebraic graph theory have been used for the representation of biomolecular systems (*4*). They have been found to significantly enhance the performance of statistic learning models in various aspects of drug design (*3*). Different from traditional molecular descriptors, three unique kinds of invariants, i.e., topological invariant (Betti numbers), geometric invariant (curvatures), and algebraic graph invariant (eigenvalues), are considered (*3*, *4*). The combination of these invariants with learning models has achieved unprecedented success in various aspects of drug design (*3*), including protein-ligand binding affinity prediction, protein stability change upon mutation prediction, toxicity prediction, partition coefficient and aqueous solubility prediction, and binding pocket detection. These advanced mathematics–based machine learning models have constantly achieved some of the best results in D3R Grand Challenge (*9*, *10*), an annual worldwide competition for drug discovery.

Here, we present a new molecular representation framework, known as persistent spectral (PerSpect), and PerSpect-based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. We combine a filtration process, which will induce a series of nested topological representations (graph, simplicial complex, and hypergraph), with spectral models (spectral graph, spectral simplicial complex, and spectral hypergraph). Molecular descriptors are obtained from PerSpect attributes, which are functions of eigenvalues over the filtration value. Our PerSpect ML can achieve state-of-the-art results in protein-ligand binding affinity prediction.

RESULTS

Biomolecular topological modeling

The structure-function relationship is of essential importance to the analysis of biomolecular flexibility, dynamics, interactions, and functions. As a mathematical tool, topology studies the network and connection information within the data and provides an effective way of structure characterization. There are three basic topological representations, including graph, simplicial complex, and hypergraph. An example of these representations for an Aspirin molecule is given in Fig. 1. Mathematically, a simplicial complex, which is composed of simplexes, can be viewed as a generalization of the graph. A k-simplex is the convex hull made from $k + 1$ vertices and can be viewed geometrically as a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), a tetrahedron (3-simplex), and their k-dimensional counterpart (k-simplex). Note that a graph is composed of only 0-simplexes and 1-simplexes, while a simplicial complex is made from simplexes at different dimensions under certain combinational rules. With hyperedges as its building blocks, a hypergraph is a further generalization of the simplicial complex (see Materials and Methods).

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore.
*Corresponding author. Email: xiakelin@ntu.edu.sg

Recently, topological data analysis (TDA), in particular persistent homology (*11*–*13*), has been used in molecular representations. TDA-based machine learning models have achieved outstanding performance in various aspects of drug design (*3*, *4*). One of key reasons for their successes is the use of the topological invariant, i.e., Betti numbers, as molecular descriptors. As illustrated in Fig. 2C, β_0 is the number of connected components, β_1 is the number of circles or loops, and β_2 is the number of voids or cavities. Note that for the octahedron surface, which is composed of eight triangles (in yellow color), its β_2 value is 1. In Fig. 2D, four simplicial complexes are generated at filtration values of 0.7, 0.9, 1.1, and 1.6. Their β_0 values are 7, 4, 3, and 1, respectively; their β_1 values are 0, 1, 0, and 1, respectively; and their β_2 values are all 0.

The other key reason for the great successes of TDA-based machine learning models is the use of a filtration process. As illustrated in Fig. 2D, the filtration value (denoted as *f*) is defined as the diameter of the spheres assigning to each point of the data. With the increase of filtration value, simplexes are consistently generated and a sequence of nested Vietoris-Rips complexes is obtained. Their Betti numbers can be calculated and visualized by using a persistent barcode, as demonstrated in Fig. 2E. At each filtration value, if we add the bars together (along the green lines), the total number is exactly equal to the Betti numbers. Note that simplicial complexes that are generated at smaller filtration values characterize short-range interactions, while the ones from larger filtration values characterize long-range interactions. Betti numbers from different filtration values represent interaction information from various scales; thus, persistent barcode provides a unified multiscale topological representation of the interactions within a structure.

PerSpect theory

Essentially, TDA studies the topological invariants at multiple scales, while our PerSpect theory studies spectral information from various different scales. Our PerSpect theory covers three basic models, i.e., PerSpect graph (*14*), PerSpect simplicial complex, and PerSpect hypergraph. Mathematically, spectral graph theory (*15*), spectral simplicial complex (*16*–*19*), and spectral hypergraph (*20*–*22*) have been developed on the basis of graph, simplicial complex, and hypergraph (see Materials and Methods). These models use different types of connection matrices, in particular Laplacian matrices, to represent structure connections. Generally speaking, Laplacian matrices from graphs characterize relations between vertices, Hodge Laplacian (or combinatorial Laplacian) matrices from simplicial complexes describe connections between simplexes, and hypergraph Laplacian matrices represent hyperedge connections. On the basis of these matrices, spectral information, including eigenvalues,

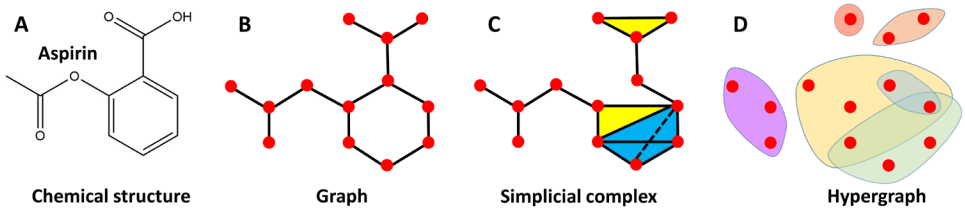

Fig. 1. Three topological representations of the aspirin molecule. (**A**) Chemical structure of aspirin. (**B**) to (**D**) Topological representations of aspirin structure: (B) graph; (C) simplicial complex; and (D) hypergraph. Mathematically, a graph is a simplicial complex with only vertices (0-simplexes) and edges (1-simplexes). The simplicial complex includes higher-dimensional simplexes, such as 2-simplexes (triangles in yellow) and 3-simplexes (tetrahedrons in blue). Hypergraph is a further generalization of the simplicial complex by replacing simplexes with hyperedges.

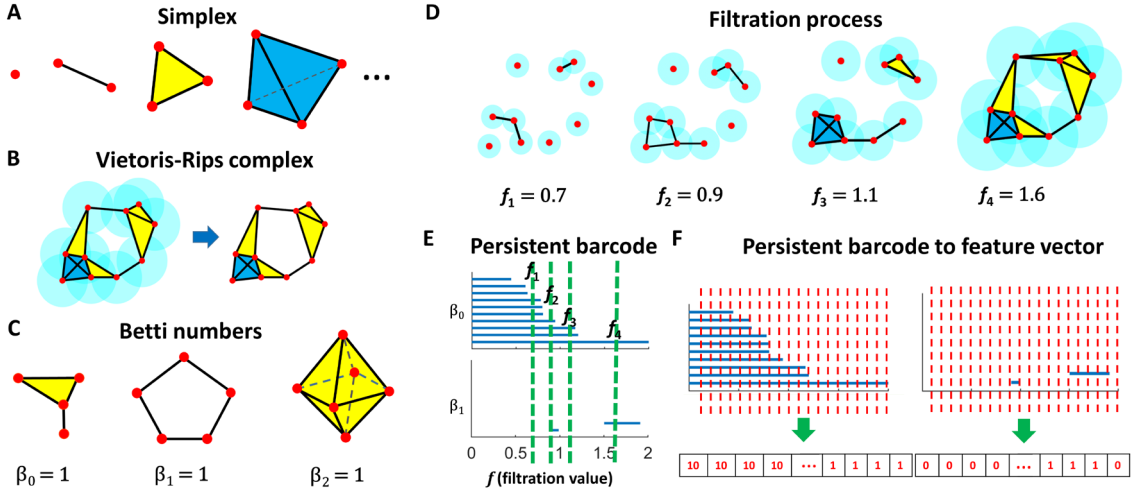

Fig. 2. The illustration of the fundamental concepts in TDA. (**A**) *k*-simplex is a convex hull made from $k+1$ vertices. Geometrically, they can be viewed as a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), and a tetrahedron (3-simplex). (**B**) In Vietoris-Rips complex, spheres are assigned to each data point, and a *k*-simplex is formed among a set of $k+1$ vertices if any two spheres among $k+1$ spheres overlap with each other. (**C**) Geometrically, β_0 is the number of connected components, β_1 is the number of circles or loops, and β_2 is the number of voids or cavities. (**D**) Illustration of a filtration process. Simplicial complexes at four different filtration values represent interactions at four different scales. (**E**) Persistent barcode generated from the filtration process in (D). (**F**) Persistent barcode–based featurization using a binning approach.

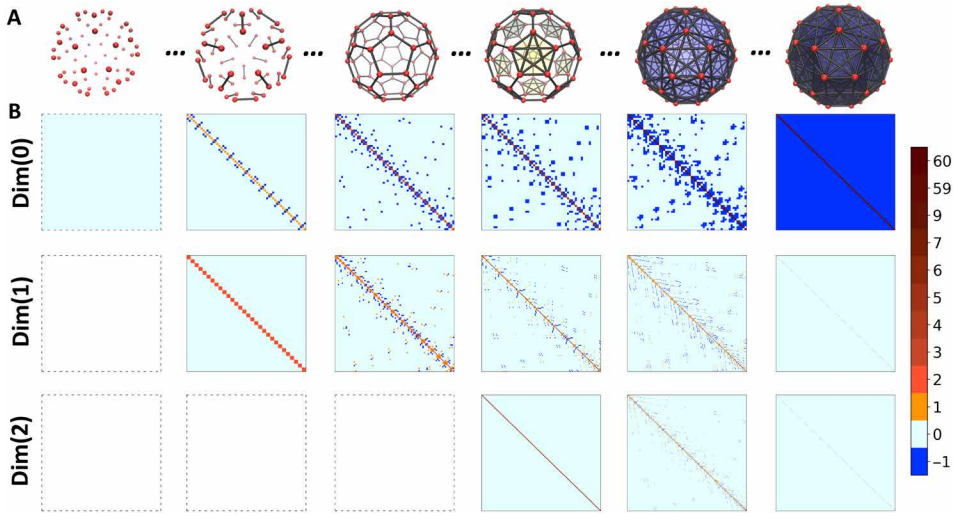

Fig. 3. Illustration of the filtration process and the associated Hodge Laplacian matrices for fullerene C_{60}. (**A**) A series of nested simplicial complexes are generated during the filtration. (**B**) Hodge Laplacian matrices are generated from these simplicial complexes. Hodge Laplacian matrices at Dim(0) to Dim(2) are illustrated. During the filtration, Dim(0) Laplacian matrix changes from all-zero-entry matrix, meaning no connections at all, to a matrix with all off-diagonal entries as −1, representing a complete graph. For Dim(1) and Dim(2) Hodge Laplacian matrices, the total number of their off-diagonal non–zero entries increases at the early stage of filtration, then decreases, and finally goes to zero, resulting in two diagonal matrices.

eigenvectors, characteristic polynomials, and eigenfunctions, can be calculated and used in structure description.

Different from all previous spectral models, our PerSpect theory characterizes the persistence and variation of spectral information at various different scales. A filtration process as in TDA is considered to generate a nested sequence of topological structures, which can be graphs, simplicial complexes, or hypergraphs. From these topological representations, a sequence of connection matrices can be constructed and their spectral information can be calculated. Our PerSpect theory studies spectral information in this series of connection matrices. PerSpect attributes are defined as functions of spectral variables over the filtration value.

PerSpect attributes can be obtained from the statistical and combinatorial properties of spectral information over a filtration process. They can be used to describe both geometric and topological information of structures. For instance, the multiplicity (or number) of Dim(k) (kth dimension) zero eigenvalue is equal to Betti number β_k; thus, persistent multiplicity, which is defined as the multiplicity of Dim(k) zero eigenvalue over a filtration process, is exactly the persistent Betti number or Betti curve (*12*,*13*), which is just the summation of bars at each filtration value as stated above. Basic statistic properties, such as mean, SD, maximum, and minimum, can be used to define four PerSpect attributes, i.e., persistent mean, persistent SD, persistent maximum, and persistent minimum. Other eigenspectral properties can also be incorporated into our PerSpect attributes (see Materials and Methods).

Figure 3 demonstrates a sequence of nested simplicial complexes and Hodge Laplacian matrices for the filtration process of fullerene C_{60}. Vietoris-Rips complex is used, and filtration parameter is chosen as the sphere diameter. It can be seen that, during the filtration process, complexes have been generated and the simplicial complex "grows" from a set of isolated vertices to a fully connected complete graph. The corresponding Laplacian matrices characterize this "expansion" process. We denote $\mathbf{L}_k$ as kth dimensional Hodge Laplacian matrix. For Dim(0), at the very start of the filtration, there are only 60 vertices (0-simplex), and a 60*60 all-zero $\mathbf{L}_0$ matrix is generated according to Eq. 2 (see Materials and Methods). As the filtration value increases, the size of $\mathbf{L}_0$ matrices remains unchanged, while more and more entries with −1 value appear at its off-diagonal part. When the filtration value is large enough, a complete graph is obtained, and a full $\mathbf{L}_0$ matrix, i.e., all diagonal entries are 59 and all off-diagonal entries are −1, is generated according to Eq. 2. For Dim(1), at the early stage of filtration, no edges (1-simplexes) and, thus, no $\mathbf{L}_1$ matrix exist. With edges emerging as the filtration value increases, $\mathbf{L}_1$ matrices are generated. Different from Dim(0) case, the size of $\mathbf{L}_1$ matrices increases with the number of edges. Off-diagonal entries can be 1 and −1 depending on the edge orientation, as in Eq. 3. When the filtration value is large enough, all edges will be either upper adjacent or not lower adjacent; thus, $\mathbf{L}_1$ matrix becomes a diagonal matrix with all its diagonal entries as 60. For Dim(2), no $\mathbf{L}_2$ matrices exist at the beginning stage of filtration, as no 2-simplexes are generated. The size of $\mathbf{L}_2$ matrices increases with the filtration, and the matrix eventually grows into a diagonal matrix with its diagonal entry value 60 according to Eq. 3. Mathematically, higher-dimensional Hodge Laplacian matrices can also be generated.

Furthermore, we can study PerSpect attributes for fullerene C_{60}. Figure 4 shows a comparison between persistent barcode and persistent multiplicity. It can be seen that the persistent multiplicity is equivalent to persistent Betti number or Betti curve. In this way, the persistent homology information is naturally embedded into persistent multiplicity. Figure 5 shows the persistent mean, persistent SD, persistent maximum, and persistent minimum for C_{60}. It can be seen that these PerSpect attributes change with the filtration value. Each variation of PerSpect attributes indicates a certain change of the simplicial complexes. At filtration size 7.10 Å, a complete three-dimensional simplicial complex is achieved, i.e., any four vertices can form a 3-simplex. The corresponding $\mathbf{L}_0$ has eigenvalues 0 and 60. The size for the corresponding $\mathbf{L}_1$ is 1770*1770, and its eigenvalues are all 60. The size for complete corresponding $\mathbf{L}_2$ is

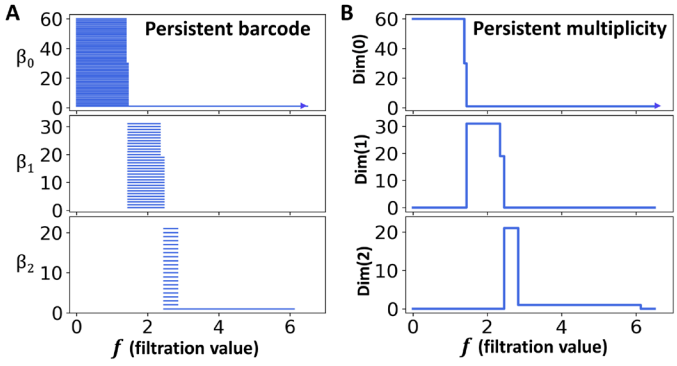

Fig. 4. Comparison of persistent barcodes and persistent multiplicities of fullerene C_{60}. (**A**) The persistent barcodes for fullerene C_{60}. (**B**) The persistent multiplicities for fullerene C_{60}. The Dim(*k*) persistent multiplicity is multiplicity of zero eigenvalues for Dim(*k*) combinatorial Laplacian matrices during a filtration process. Multiplicities of zero eigenvalues are equivalent to Betti numbers. Persistent multiplicity is equivalent to persistent Betti numbers or Betti curves.

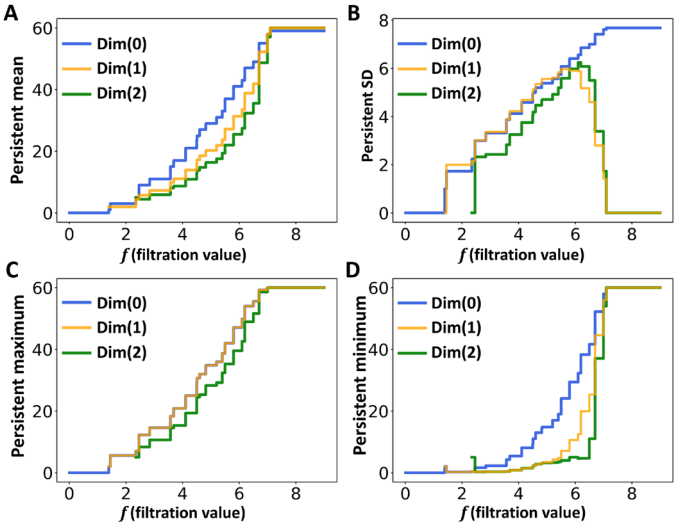

Fig. 5. Illustration of four PerSpect attributes for fullerene C_{60}. (**A**) Persistent mean. (**B**) Persistent SD. (**C**) Persistent maximum. (**D**) Persistent minimum.

34220*34220, and its eigenvalues are also 60. Note that $1770 = C_{60}^2$ and $34220 = C_{60}^3$.

PerSpect ML models

Our PerSpect theory provides a mathematical representation of molecules. PerSpect attributes can naturally work as featurization or feature engineering of molecular structures and interactions. More specifically, molecular descriptors/fingerprints can be obtained from the discretization of PerSpect attributes. Similar to the binning approach as in Fig. 2F, we can decompose the filtration region into equal-sized bins and use PerSpect attribute value at each grid point as an individual feature. These values are combined into a feature vector and further used in various machine learning models, such as support vector machine, random forest, gradient boosting tree (GBT), and convolutional neural network (CNN). Because PerSpect attributes are generated from highly abstract spectral models at multiple scales, PerSpect attribute–based feature vectors can balance between complexity reduction, data simplification, and preservation of intrinsic structure information. A better featurization with a higher transferability is obtained in our PerSpect models; thus, they can boost the performance of learning models in molecular data analysis.

PerSpect ML for protein-ligand binding affinity prediction

The prediction of protein-ligand binding affinity is a key step in drug design and discovery (*2*). An accurate prediction requires a better representation of the interactions between proteins and ligands at the molecular level. Here, the element-specific (ES) topological model is considered to characterize protein-ligand interactions (*3*). Essentially, a molecule can be decomposed into different atom sets, each with only one type of atom. For instance, a protein structure can usually be decomposed into five different atom sets, each containing one type of atom, including hydrogen (H), carbon (C), nitrogen (N), oxygen (O), and sulfur (S). Ligands are usually composed of around 10 types of atom sets. Among them, five types are the same as in protein, and the other five types include phosphorus (P), fluoride (F), chloride (Cl), bromide (Br), and iodine (I). In the ES topological model, protein-ligand interactions are characterized by topological connections between two atom sets, one from protein and the other from ligand. For instance, a C-N graph can be constructed (for protein-ligand interactions) using a C atom set from protein and a N atom set from ligand.

Interactions in the ES topological models can be characterized by distance relationships. In this way, ES-based interactive distance matrix (ES-IDM) (*3*) can be defined as follows

$$M(i,j) = \begin{cases} \|\mathbf{r}_i - \mathbf{r}_j\|, & \text{if } \mathbf{r}_i \in \mathbf{R}_P, \mathbf{r}_j \in \mathbf{R}_L \text{ or } \mathbf{r}_i \in \mathbf{R}_L, \mathbf{r}_j \in \mathbf{R}_P \\ \infty, & \text{otherwise} \end{cases}$$

Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ are coordinates for the *i*th and *j*th atoms, and $\|\mathbf{r}_i - \mathbf{r}_j\|$ is their Euclidean distance. Two sets $\mathbf{R}_P$ and $\mathbf{R}_L$ are atom coordinate sets for protein and ligand, respectively. Note that only connections (or interactions) between protein atoms and ligand atoms are considered in the ES-IDM models. Connections between atoms within either protein or ligand are ignored by setting their distance as ∞, i.e., an infinitely large value. Interactions in the ES topological models can also be modeled using electrostatic properties. ES-based interactive electrostatic matrix (ES-IEM) (*3*) can be defined as follows

$$M_E(i,j) = \begin{cases} \dfrac{1}{1+\exp\left(-\dfrac{c\,q_i q_j}{\|\mathbf{r}_i - \mathbf{r}_j\|}\right)}, & \text{if } \mathbf{r}_i \in \mathbf{R}_P, \mathbf{r}_j \in \mathbf{R}_L \text{ or } \mathbf{r}_i \in \mathbf{R}_L, \mathbf{r}_j \in \mathbf{R}_P \\ \infty, & \text{otherwise} \end{cases}$$

Here, q_i and q_j are partial charges for the *i*th and *j*th atoms, and parameter *c* is a constant value. In this matrix, electrostatic interactions between atoms within either protein or ligand are dismissed by setting their value as ∞ in our ES-IEM models. A filtration process can be generated from both ES-IDM and ES-IEM. The filtration parameter can be chosen as either the distance value or electrostatic value. Simplicial complexes can be generated by using Vietoris-Rips complex. We consider PerSpect simplicial complex model and select 11 PerSpect attributes to generate feature vectors (see Materials and Methods).

To validate our models, we consider the three most commonly used protein-ligand datasets, namely, PDBbind-2007, PDBbind-2013, and PDBbind-2016 (*23*). Three PerSpect-GBT models with features from ES-IDM model, ES-IEM model, and combined ES-IDM and ES-IEM models are considered. An average Pearson correlation coefficient (PCC) of around 0.81 is obtained for all three models in all three datasets. The results are for the test sets and are listed in Table 1. Figure 6 shows the comparison between the predicted binding affinity values with the experimental ones. Furthermore, to have a better understanding of the performance of our models, we compare our PCC results with the state-of-the-art results in literature (*2*, *24*–*31*), as far as we know. The results are illustrated in Fig. 7. It can be seen that our PerSpect-GBT models have achieved the highest PCCs for all three datasets.

Note that our PerSpect-GBT can be applied to various other steps of virtual screening in drug design, including the prediction of solubility, partition coefficient, toxicity, and other properties for drug absorption, distribution, metabolism, excretion, and toxicity (*32*).

DISCUSSION

Advanced mathematical representations that characterize molecular intrinsic structural, physical, and chemical properties provide a solid foundation for molecular function and property analysis. Molecular descriptors obtained from the advanced mathematical representations provide an effective featurization for learning models in material, chemical, and biological data analysis. Compared with traditional featurization, our PerSpect theory has several advantages. First, a multiscale representation is attained through a filtration process. Note that PerSpect attributes capture the eigen information from various different scales through an expansion process, instead of only a special fixed scale as in traditional models. Second, PerSpect models characterize the intrinsic structure properties. Essentially, Betti number, a topological invariant, is incorporated into PerSpect attributes. Third, a balance between the geometric complexities and topological simplification is achieved. PerSpect attributes from non–zero eigenvalues characterize the quantitative geometric information of the structure. Last, it is the first time the Hodge theory has been used in featurization and machine learning models.

Table 1. The PCCs and root mean square errors (in kcal/mol) of our PerSpect simplicial complex–based GBT models on the three test sets of PDBbind-2007, PDBbind-2013, and PDBbind-2016. Three PerSpect-GBT models are considered. Their features are generated from the ES-IDM model, the ES-IEM model, and combined ES-IDM and ES-IEM models (ES-IDM + ES-IEM). The detailed information of the training sets and test sets can be found in Table 2. The detailed setting of GBT parameters can be found in Table 3.

	ES-IDM	ES-IEM	ES-IDM + ES-IEM
PDBbind-2007	0.829 (1.868)	0.816 (1.941)	0.836 (1.847)
PDBbind-2013	0.781 (2.005)	0.786 (1.979)	0.793 (1.956)
PDBbind-2016	0.830 (1.764)	0.832 (1.757)	0.840 (1.724)
Average	0.813 (1.879)	0.811 (1.892)	0.823 (1.842)

MATERIALS AND METHODS

Topological representations

Graph

Graph or network models have been applied to various material, chemical, and biological systems. In these models, atoms and bonds are usually simplified as vertices and edges. Mathematically, a graph representation can be denoted as $G(V, E)$, where $V = \{v_i; i = 1,2, \ldots, N\}$ are vertex set with $N = |V|$ the total number. Here, $E = \{e_i = (v_{i_1}, v_{i_2}); 1 \le i_1 < i_2 \le N\}$ denotes the edge set. Note that graph invariants are graph properties that remain unchanged under graph isomorphism (bijective mapping between two graphs). Typical graph invariants include graph order, size, clique number (clique is a maximal set of nodes that is complete), and chromatic index.

Simplicial complex

A simplicial complex is the generalization of a graph into its higher-dimensional counterpart. The simplicial complex is composed of simplexes. Each simplex is a finite set of vertices and can be viewed geometrically as a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), a tetrahedron (3-simplex), and their k-dimensional counterpart (k-simplex). More specifically, a k-simplex $\sigma^k = \{v_0, v_1, v_2, \cdots, v_k\}$ is the convex hull formed by $k + 1$ affinely independent points $v_0, v_1, v_2, \cdots, v_k$ as follows

$$\sigma^k = \left\{ \lambda_0 v_0 + \lambda_1 v_1 + \cdots + \lambda_k v_k \middle| \sum_{i=0}^{k} \lambda_i = 1; \forall i, 0 \le \lambda_i \le 1 \right\}$$

The ith dimensional face of σ^k ($i < k$) is the convex hull formed by $i + 1$ vertices from the set of $k + 1$ points $v_0, v_1, v_2, \cdots, v_k$. The simplexes are the basic components for a simplicial complex.

A simplicial complex K is a finite set of simplexes that satisfy two conditions. First, any face of a simplex from K is also in K. Second, the intersection of any two simplexes in K is either empty or a shared face. A kth chain group C_k is an Abelian group of oriented k-simplexes σ^k, which are simplexes together with an orientation, i.e., ordering of their vertex set. The boundary operator $\partial_k (C_k \to C_{k-1})$ for an oriented k-simplex σ^k can be denoted as

$$\partial_k \sigma^k = \sum_{i=0}^{k} (-1)^i [v_0, v_1, v_2, \cdots, \hat{v}_i, \cdots, v_k]$$

Here, $[v_0, v_1, v_2, \cdots, \hat{v}_i, \cdots, v_k]$ is an oriented $(k - 1)$–simplex, which is generated by the original set of vertices except v_i. The boundary operator maps a simplex to its faces, and it guarantees that $\partial_{k-1}\partial_k = 0$. There are various kinds of simplical complexes, including Vietoris-Rips complex, Čech complex, alpha complex, and clique complex. Among them, Vietoris-Rips complex is used here, and an example can be found in Fig. 2. Clique complex (also known as flag complex) can be generated directly from a graph or a hypergraph by using a clique expansion.

To facilitate a better description, we use notation $\sigma_j^{k-1} \subset \sigma_i^k$ to indicate that σ_j^{k-1} is a face of σ_i^k and notation $\sigma_j^{k-1} \sim \sigma_i^k$ if they have the same orientation, i.e., oriented similarly. For two oriented k-simplexes, σ_i^k and σ_j^k, of a simplicial complex K, they are upper adjacent, denoted as $\sigma_i^k \cap \sigma_j^k$, if they are faces of a common $(k + 1)$–simplex; they are lower adjacent, denoted as $\sigma_i^k \cup \sigma_j^k$, if they share a common $(k - 1)$–simplex as their face. Moreover, if the orientations of their common lower simplex are the same, it is called similar common lower simplex ($\sigma_i^k \cup \sigma_j^k$ and $\sigma_i^k \sim \sigma_j^k$); if their orientations

Table 2. Details of the three PDBbind databases. The refined sets are composed of training set and test set (core set).

Version	Refined set	Training set	Test set (core set)
v2007	1300	1105	195
v2013	2959	2764	195
v2016	4057	3772	285

Table 3. The setting of parameters for our GBT model.

No. of estimators	Maximum depth	Minimum sample split	Learning rate
40,000	6	2	0.001
Loss function	Maximum features	Subsample size	Repetition
Least square	Square root	0.7	10 times

are different, it is called dissimilar common lower simplex ($\sigma_i^k \cup \sigma_j^k$ and $\sigma_i^k \nsim \sigma_j^k$). The (upper) degree of a k-simplex σ_i^k, denoted as $d(\sigma^k)$, is the number of $(k+1)$–simplexes, of which σ_i^k is a face.

Hypergraph

A hypergraph is a generalization of a graph in which an edge is made of a set of vertices. Mathematically, a hypergraph $(V_\mathcal{H}, \mathcal{H})$ consists of a set of vertices (denoted as $V_\mathcal{H}$) and a set of hyperedges (denoted as $\mathcal{H}$). Each hyperedge contains an arbitrary number of vertices and can be regarded as a subset of $V_\mathcal{H}$. A hyperedge e_i^h is said to be incident with a vertex v_j when the vertex is in the hyperedge, i.e., $v_j \in e_i^h$. Note that a hypergraph can also be viewed as a generalization of the simplicial complex. Moreover, a unique clique complex can be generated from a hypergraph by defining each hyperedge as a clique.

Spectral theories

The characterization, identification, comparison, and analysis of structure data, from material, chemical, and biological systems, are usually complicated because of their high dimensionality and complexity. Spectral graph theory is proposed to reduce the data dimensionality and complexity by studying the spectral information of connectivity matrices, constructed from the structure data. These connectivity matrices include incidence matrix, adjacency matrix, (normalized) Laplacian matrix, and Hessian matrix. Spectral information includes eigenvalues, eigenvectors, eigenfunctions, and other related properties, such as Cheeger constant, edge expansion, vertex expansion, graph flow, graph random walk, and heat kernel of graph. Spectral graph theory has been generalized into spectral simplicial complex (*16*–*19*, *33*) and spectral hypergraph (*20*–*22*).

Spectral graph

In spectral graph theory, a graph $G(V, E)$ is represented by its adjacency matrix and Laplacian matrix (*15*, *34*–*36*). The adjacency matrix $\mathbf{A}$ describes the connectivity information and can be expressed as

$$A(i,j) = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & (v_i, v_j) \notin E \end{cases}$$

The degree of a vertex v_i is the total number of edges that are connected to vertex v_i, i.e., $d(v_i) = \sum_{i \neq j}^{N} A(i,j)$. The vertex diagonal matrix $\mathbf{D}$ can be defined as

$$D(i,j) = \begin{cases} \sum_{i \neq j}^{N} A(i,j), & i = j \\ 0, & i \neq j \end{cases}$$

Laplacian matrix, also known as admittance matrix and Kirchhoff matrix, is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. More specifically, it can be expressed as

$$L(i,j) = \begin{cases} d(v_i), & i = j (5) \\ -1, & i \neq j \text{ and} (v_i, v_j) \in E \\ 0, & i \neq j \text{ and} (v_i, v_j) \notin E \end{cases} \quad (1)$$

The Laplacian matrix has many important properties. It is always positive-semidefinite; thus, all its eigenvalues are non-negative. In particular, the number (multiplicity) of zero eigenvalues is equal to the topological invariant β_0, which counts the number of connected components in the graph. The second smallest eigenvalue, i.e., the first non–zero eigenvalue, is called Fiedler value or algebraic connectivity, which describes the general connectivity of the graph. The corresponding eigenvector can be used to subdivide the graph into two well-connected subgraphs. All eigenvalues and eigenvectors form an eigenspectrum, and spectral graph theory studies the properties of the graph eigenspectrum.

There are two types of normalized Laplacian matrices, including the symmetric normalized Laplacian matrix, which is defined as $\mathbf{L}_{\text{sym}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$, and random walk normalized Laplacian, which is defined as $\mathbf{L}_{\text{rw}} = \mathbf{D}^{-1}\mathbf{L}$.

Spectral simplicial complex

The spectral simplicial complex theory studies the spectral properties of combinatorial Laplacian (or Hodge Laplacian) matrices, which are constructed on the basis of a simplicial complex (*16*–*19*, *33*). For an oriented simplicial complex, its kth boundary (or incidence) matrix $\mathbf{B}_k$ can be defined as follows

$$B_k(i,j) = \begin{cases} 1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \sim \sigma_j^k \\ -1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \text{ and } \sigma_i^{k-1} \nsim \sigma_j^k \\ 0, & \text{if } \sigma_i^{k-1} \not\subset \sigma_j^k \end{cases}$$

These boundary matrices satisfy the condition that $\mathbf{B}_k\mathbf{B}_{k+1} = \mathbf{0}$. The kth combinatorial Laplacian matrix can be expressed as follows

$$\mathbf{L}_k = \mathbf{B}_k^T\mathbf{B}_k + \mathbf{B}_{k+1}\mathbf{B}_{k+1}^T$$

Note that 0th combinatorial Laplacian is

$$\mathbf{L}_0 = \mathbf{B}_1\mathbf{B}_1^T$$

Furthermore, if the highest order of the simplicial complex K is n, then the nth combinatorial Laplacian matrix is $\mathbf{L}_n = \mathbf{B}_n^T\mathbf{B}_n$.

The above combinatorial Laplacian matrices can be explicitly described in terms of the simplex relations. More specifically, $\mathbf{L}_0$ can be expressed as

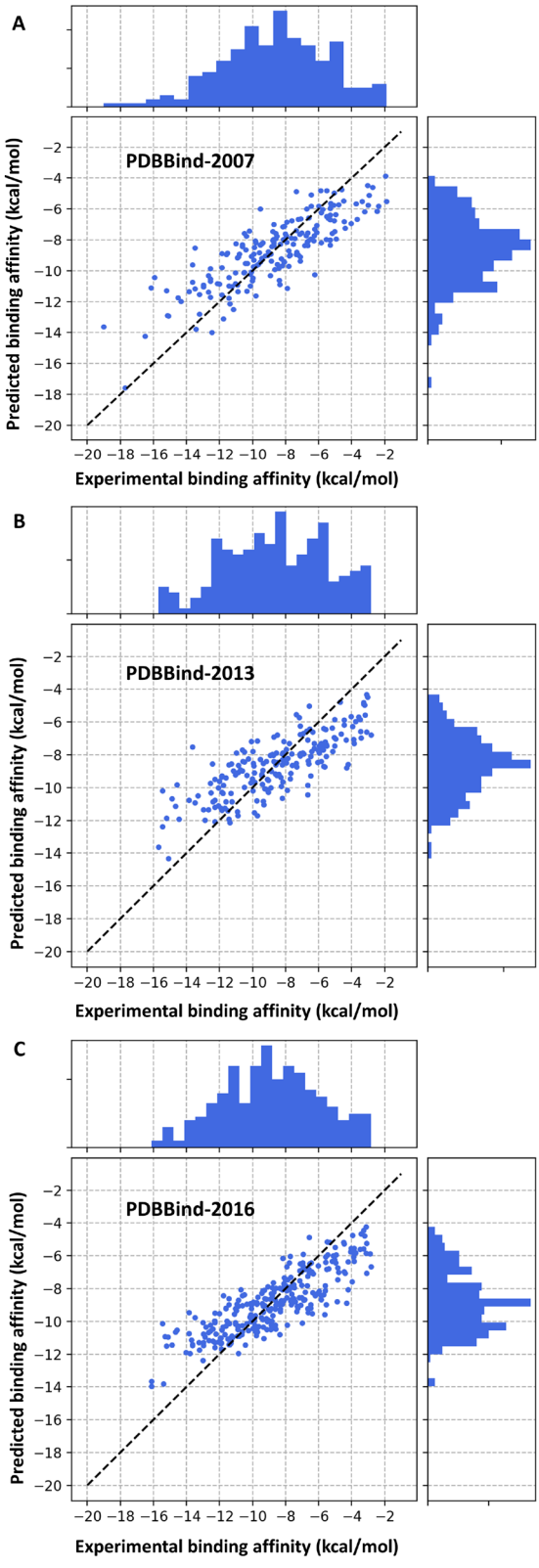

Fig. 6. Comparison of predicted protein-ligand binding affinities and experimental results for the three test sets. (**A**) PDBbind -2007. (**B**) PDBbind-2013. (**C**) PDBbind-2016. The PCCs are 0.836, 0.793, and 0.840, respectively. The root mean square errors are 1.847, 1.956, and 1.724 kcal/mol, respectively.

$$L_0(i,j)=\begin{cases} d(\sigma_i^0), & \text{if } i = j (9)\\ -1, & \text{if } i \neq j \text{ and } \sigma_i^0 \cap \sigma_j^0\\ 0, & \text{if } i \neq j \text{ and } \sigma_i^0 \not\frown \sigma_j^0 \end{cases} \quad (2)$$

It can be seen that this expression is exactly the graph Laplacian as in Eq. 1. Furthermore, when $k > 0$, $\mathbf{L}_k$ can be expressed as

$$L_k(i,j)=\begin{cases} d(\sigma_i^k)+k+1, \text{ if } i = j\\ 1, \text{ if } i \neq j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \cup \sigma_j^k \text{ and } \sigma_i^k \sim \sigma_j^k\\ -1, \text{ if } i \neq j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \cup \sigma_j^k \text{ and } \sigma_i^k \not\sim \sigma_j^k\\ 0, \text{ if } i \neq j, \sigma_i^k \cap \sigma_j^k \text{ or } \sigma_i^k \not\backsim \sigma_j^k \end{cases} \quad (3)$$

The eigenvalues of combinatorial Laplacian matrices are independent of the choice of the orientation (*17*). Furthermore, the multiplicity of zero eigenvalues, i.e., the total number of zero eigenvalues, of $\mathbf{L}_k$ is equal to the kth Betti number β_k.

We can define the kth combinatorial down Laplacian matrix as $\mathbf{L}_k^{\text{down}} = \mathbf{B}_k^T\mathbf{B}_k$ and combinatorial up Laplacian matrix as $\mathbf{L}_k^{\text{up}} = \mathbf{B}_{k+1}\mathbf{B}_{k+1}^T$. These matrices have very interesting spectral properties (*18*). First, eigenvectors associated with non–zero eigenvalues of $\mathbf{L}_k^{\text{down}}$ are orthogonal to eigenvectors from non–zero eigenvalues of $\mathbf{L}_k^{\text{up}}$. Second, non–zero eigenvalues of $\mathbf{L}_k$ are either the eigenvalues of $\mathbf{L}_k^{\text{down}}$ or those of $\mathbf{L}_k^{\text{up}}$. Third, eigenvectors associated with non–zero eigenvalues of $\mathbf{L}_k$ are either eigenvectors of $\mathbf{L}_k^{\text{down}}$ or those of $\mathbf{L}_k^{\text{up}}$.

Spectral hypergraph

Laplacian matrices can also be defined on hypergraph (*20*–*22*). One way to do that is to use a clique expansion, in which a graph is constructed from a hypergraph $(V_{\mathcal{H}}, \mathcal{H})$ by replacing each hyperedge with an edge for each pair of vertices in this hyperedge. A graph Laplacian matrix can then be defined on this hypergraph-induced graph. Note that the clique expansion also generates a clique complex, and Hodge Laplacian matrices can also be constructed based on it.

The other way is to directly use the incidence matrix. In a hypergraph, an incidence matrix $\mathbf{H}$ can be defined as follows

$$H(i,j)=\begin{cases} 1, & \text{if } v_i \in e_j^h\\ 0, & \text{if } v_i \notin e_j^h \end{cases}$$

The vertex diagonal matrix $\mathbf{D}_v$ is

$$D_v(i,j)=\begin{cases} \sum_j H(i,j), & i = j\\ 0, & i \neq j \end{cases}$$

The hypergraph adjacent matrix is then defined as $\mathbf{A} = \mathbf{H}\mathbf{H}^T - \mathbf{D}_v$, and the unnormalized hypergraph Laplacian matrix is defined as

$$\mathbf{L} = 2\mathbf{D}_v - \mathbf{H}\mathbf{H}^T$$

Similar to the graph models, the symmetric normalized hypergraph Laplacian is defined as $\mathbf{L}_{\text{sym}} = 2\mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{H}^T\mathbf{D}_v^{-1/2}$, with $\mathbf{I}$ as the identity matrix. The random walk hypergraph Laplacian is defined as $\mathbf{L}_{\text{rw}} = 2\mathbf{I} - \mathbf{D}_v^{-1}\mathbf{H}\mathbf{H}^T$. Recently, embedded homology,

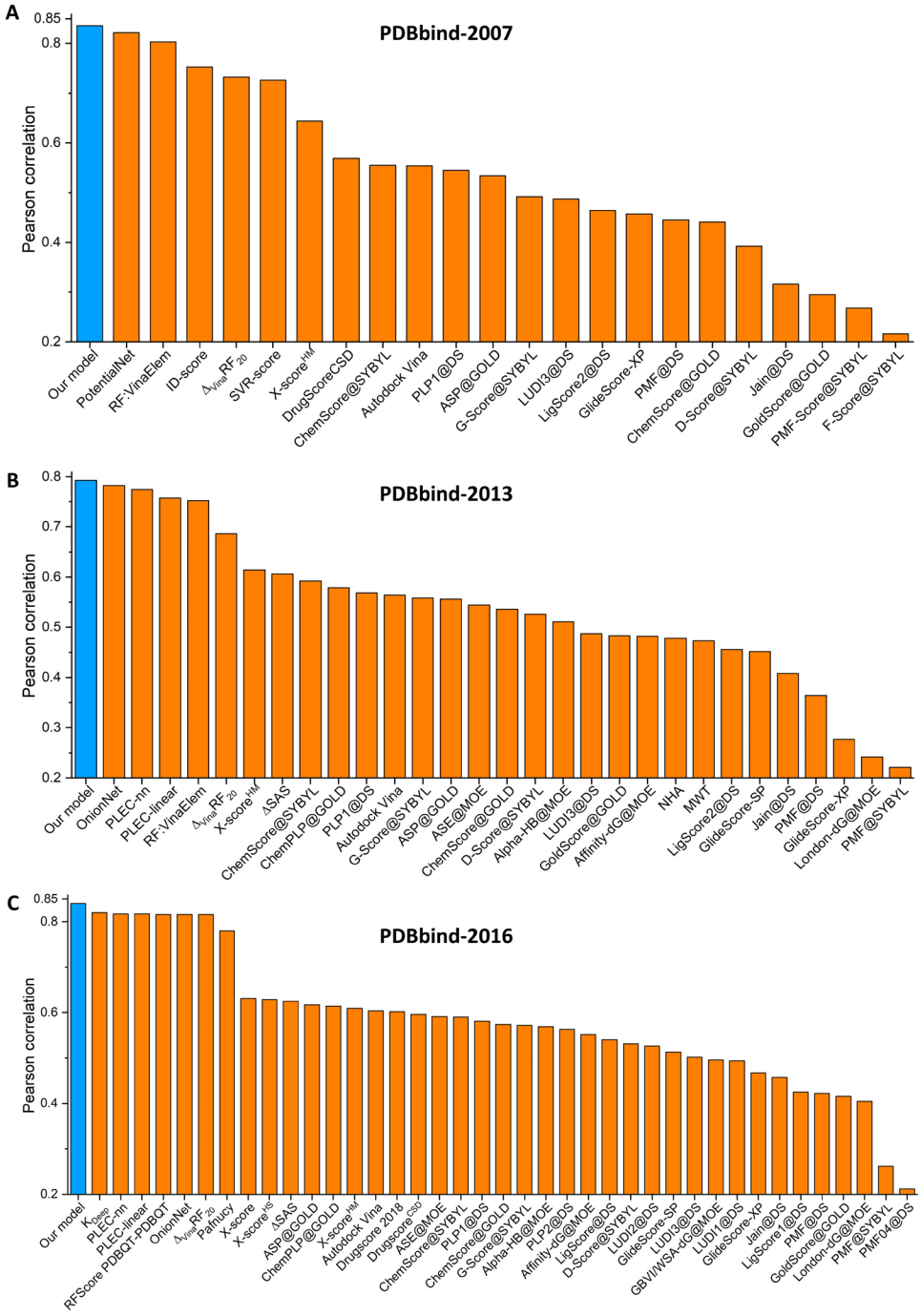

Fig. 7. Performance comparison of our PerSpect simplicial complex–based GBT with the-state-of-art models (*2*, *24*–*31*). We consider three datasets, including (**A**) PDBbind-2007, (**B**) PDBbind-2013, and (**C**) PDBbind-2016 .

persistent homology, and weighted (Hodge) Laplacians have been developed for hypergraphs (*37*, *38*).

PerSpect theory

Filtration

A filtration process naturally generates a multiscale representation (*12*). The filtration parameter, denoted as f and key to the filtration process, is usually chosen as sphere radius (or diameter) for point cloud data, edge weight for graphs, and isovalue (or level set value) for density data. A systematical increase (or decrease) of the value for the filtration parameter will induce a sequence of hierarchical topological representations, which can be not only simplicial complexes but also graphs and hypergraphs. For instance, a filtration operation on a distance matrix, i.e., a matrix with distances between any two vertices as its entries, can be defined by using a cutoff value as the filtration parameter. More specifically, if the distance between two vertices is smaller than the cutoff value, an edge is formed between them. In this way, a systematical increase (or decrease) of

the cutoff value will deliver a series of nested graphs, with the graph produced at a lower cutoff value as a part (or a subset) of the graph produced at a larger cutoff value. Similarly, nested simplicial complexes can be constructed by using various definitions of complexes, such as Vietoris-Rips complex, Čech complex, alpha complex, cubical complex, Morse complex, and clique complex. Nested hypergraphs can also be generated by using a suitable definition of hyperedge.

PerSpect models

The essential idea of our PerSpect theory is to provide a new mathematical representation that characterizes the intrinsic topological and geometric information of the data. Different from all previous spectral models, our PerSpect theory does not consider the eigenspectrum information of the graph, simplicial complex, or hypergraph, constructed from data at a particular scale; instead, it focuses on the variation of the eigenspectrum of these topological representations during a filtration process. Stated differently, our PerSpect theory studies the change of eigenspectrum when the structure representation, i.e., graph, simplicial complex, or hypergraph, grows from a set of isolated vertices to a fully connected topology, according to their inner structure connectivity and a predefined filtration parameter.

Mathematically, a filtration operation will deliver a nested sequence of graphs as follows

$$G^0 \subseteq G^1 \subseteq \cdots \subseteq G^m$$

Here, ith graph G^i is generated at a certain filtration value f_i. Computationally, we can equally divide the filtration region (of the filtration parameter) into m intervals and consider a topological representation at each interval. A series of Laplacian matrices $\{\boldsymbol{L}^i \mid_{i=1,2,\ldots,m}\}$ can be generated from these graphs. Furthermore, a nested sequence of simplicial complexes can also be generated from a filtration process

$$K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m$$

Similarly, the ith simplicial complex K^i is generated at filtration value f_i. Combinatorial Laplacian matrix series $\{\boldsymbol{L}_k^i \mid_{i=1,2,\ldots,m;k=0,1,2,\ldots}\}$ can be constructed from these simplicial complexes. Note that the size of these Laplacian matrices may be different. Moreover, with a suitable filtration process, a nested sequence of hypergraph can be generated as follows

$$H^0 \subseteq H^1 \subseteq \cdots \subseteq H^m$$

Hypergraph Laplacian matrix series $\{\boldsymbol{L}^i \mid_{i=1,2,\ldots,m}\}$ can be constructed accordingly.

Persistent attributes

Other than the multiplicity of zero eigenvalues and non–zero eigenvalue statistic properties, PerSpect attributes can also be generated from various spectral indexes (*5*). For a Laplacian matrix with eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, commonly used spectral indexes include sum of eigenvalues (Laplacian graph energy), sum of absolute deviation of eigenvalues (generalized average graph energy $\sum_{i=1}^{n} |\lambda_i - \bar{\lambda}|$, with $\bar{\lambda}$ as the average eigenvalue), spectral moments ($\sum_{i=1}^{n} \lambda_i^k$, with k as the order of moment), quasi-Wiener index ($\sum_{j=1}^{A} \frac{A+1}{\lambda_j}$, with $\lambda_j > 0$ and A as the number of all non–zero eigenvalues), and spanning tree number $\left(\log\left[\frac{1}{A+1} \cdot \prod_{j=1}^{A} \lambda_j\right]\right)$. Furthermore, other spectral information, including algebraic connectivity, modularity, Cheeger constant, vertex/edge expansion, and other flow, random walk, and heat kernel–related properties, can be generalized into their corresponding PerSpect attributes. Note that other persistent functions have also been considered (*39*). Moreover, physical models, such as cluster expansion and symmetry function (*40*), can be used as generalized persistent functions. Last, note that various normalized (Hodge) Laplacians have been proposed (*17*–*19*). New PerSpect attributes can be generated from these normalized (Hodge) Laplacian matrices.

Protein-ligand binding affinity prediction with PerSpect ML

The three datasets (refined sets) are downloaded from PDBbind (www.pdbbind.org.cn). The core set is used as the test dataset, and the training dataset is the refined set excluding the core set. The detailed data information can be found in Table 2.

In our ES-IDM–based PerSpect simplicial complex models, the distance value is considered as the filtration parameter. The filtration value goes from 0.00 to 25.00 Å. For discretization, Laplacian matrices are generated with a step of 0.10 Å. That is to say, a total of 250 Laplacian matrices are generated from each filtration process. There are, in total, 4*9 = 36 types of ES-IDMs between 4 types of atoms from protein, including C, N, O, and S, and 9 types of atoms from ligand, including C, N, O, S, P, F, Cl, Br, and I. In our ES-IEM–based PerSpect simplicial complex models, the interaction strength is used as the filtration parameter and its value goes from 0.00 to 1.00. In our calculation, the constant c is set to be 100. The Laplacian matrix is generated with a step of 0.01, meaning a total 100 Laplacian matrices for each filtration process. There are 5*10 = 50 types of ES-IEMs, between 5 types of atoms from protein, including H, C, N, O, and S, and 10 types of atoms from ligand, including H, C, N, O, S, P, F, Cl, Br, and I.

Furthermore, we consider 11 PerSpect features as follows: Dim(0) persistent multiplicity (of zero eigenvalue), Dim(1) persistent multiplicity (of zero eigenvalue), persistent maximum, persistent minimum, persistent mean, persistent SD, persistent Laplacian graph energy, persistent generalized mean graph energy, PerSpect moment (second order), persistent quasi-Wiener index, and persistent spanning tree number.

Note that other than the persistent multiplicity, all PerSpect attributes are calculated from Dim(0) Laplacians. To sum up, in our ES-IDMs, there are 36 types of atom combinations as stated above, and the total number of features is 36[types]*250[persistence]*11[eigen feature]. Similarly, there are 50 types of ES-IEMs, and the number of features is 50[types]*100[persistence]*11[eigen feature]. Because we have a large feature vector, decision tree–based models are considered to avoid overfitting. In particular, GBT models have delivered better results in protein-ligand binding affinity prediction. The parameters of GBT are listed in Table 3. Note that 10 independent regressions are conducted, and the medians of 10 PCCs and root mean square errors are computed as the performance measurement of our PerSpect ML model.

REFERENCES AND NOTES

1. Y. C. Lo, S. E. Rensi, W. Torng, R. B. Altman, Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).
2. H. J. Li, K.-S. Leung, M.-H. Wong, P. J. Ballester, Improving AutoDock Vina using random forest: The growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Mol. Inform.* **34**, 115–126 (2015).

3. Z. X. Cang, L. Mu, G.-W. Wei, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLOS Comput. Biol.* **14**, e1005929 (2018).
4. D. D. Nguyen, Z. Cang, G.-W. Wei, A review of mathematical representations of biomolecular data. *Phys. Chem. Chem. Phys.* **22**, 4343–4367 (2020).
5. T. Puzyn, J. Leszczynski, M. T. Cronin, *Recent Advances in QSAR Studies: Methods and Applications* (Springer Science & Business Media, 2010), vol. 8.
6. G. Landrum, *RDKit: Open-Source Cheminformatics* (2006); www.rdkit.org/.
7. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: An open chemical toolbox. *J. Chem.* **3**, 33 (2011).
8. D. S. Cao, Q. S. Xu, Q. N. Hu, Y. Z. Liang, ChemoPy: Freely available python package for computational biology and chemoinformatics. *Bioinformatics* **29**, 1092–1094 (2013).
9. D. D. Nguyen, K. F. Gao, M. L. Wang, G. W. Wei, MathDL: Mathematical deep learning for D3R Grand Challenge 4. *J. Comput. Aided Mol. Des.* **34**, 131–147 (2020).
10. D. D. Nguyen, Z. Cang, K. Wu, M. Wang, Y. Cao, G. W. Wei, Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *J. Comput. Aided Mol. Des.* **33**, 71–82 (2019).
11. A. Verri, C. Uras, P. Frosini, M. Ferri, On the use of size functions for shape analysis. *Biol. Cybern.* **70**, 99–107 (1993).
12. H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification. *Discrete Comput. Geom.* **28**, 511–533 (2002).
13. A. Zomorodian, G. Carlsson, Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274 (2005).
14. R. Wang, D. D. Nguyen, G.-W. Wei, Persistent spectral graph. *Int. J. Numerical Methods Biomed. Eng.* **36**, e3376 (2020).
15. F. Chung, *Spectral Graph Theory* (American Mathematical Society, 1997).
16. B. Eckmann, Harmonische funktionen und randwertaufgaben in einem komplex. *Comment. Math. Helv.* **17**, 240–255 (1944).
17. D. Horak, J. Jost, Spectra of combinatorial Laplace operators on simplicial complexes. *Adv. Math.* **244**, 303–336 (2013).
18. S. Barbarossa, S. Sardellitti, Topological signal processing over simplicial complexes. *IEEE Trans. Signal Process.* **68**, 2992–3007 (2020).
19. M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, A. Jadbabaie, Random walks on simplicial complexes and the normalized hodge 1-Laplacian. *SIAM Rev.* **62**, 353–391 (2020).
20. K. Feng, W.-C. W. Li, Spectra of hypergraphs and applications. *J. Number Theory* **60**, 1–22 (1996).
21. L. Sun, S. W. Ji, J. P. Ye, Hypergraph spectral learning for multi-label classification, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2008), pp. 668–676.
22. L. Y. Lu, X. Peng, High-ordered random walks and generalized Laplacians on hypergraphs, in *International Workshop on Algorithms and Models for the Web-Graph* (Springer, 2011), pp. 14–25.
23. Z. H. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, R. Wang, PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
24. J. Liu, R. Wang, Classification of current scoring functions. *J. Chem. Inf. Model.* **55**, 475–482 (2015).
25. M. Wójcikowski, M. Kukiełka, M. M. Stepniewska-Dziubinska, P. Siedlecki, Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **35**, 1334–1341 (2019).
26. J. Jiménez, M. Skalic, G. Martinez-Rosell, G. De Fabritiis, K_{DEEP}: Protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.* **58**, 287–296 (2018).
27. M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, P. Siedlecki, Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666–3674 (2018).
28. M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, R. X. Wang, Comparative assessment of scoring functions: The CASF-2016 update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).
29. K. Afifi, A. F. Al-Sadek, Improving classical scoring functions using random forest: The non-additivity of free energy terms contributions in binding. *Chem. Biol. Drug Des.* **92**, 1429–1434 (2018).
30. E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar, V. S. Pande, Potentialnet for molecular property prediction. *ACS Cent. Sci.* **4**, 1520–1530 (2018).
31. F. Boyles, C. M. Deane, G. M. Morris, Learning from the ligand: Using ligand-based features to improve binding affinity prediction. *Bioinformatics* **36**, 758–764 (2020).
32. J. Hodgson, ADMET—Turning chemicals into drugs. *Nat. Biotechnol.* **19**, 722–726 (2001).
33. S. Mukherjee, J. Steenbergen, Random walks on simplicial complexes and harmonics. *Random Struct. Algorithms* **49**, 379–405 (2016).
34. D. A. Spielman, Spectral graph theory and its applications, in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* (IEEE, 2007), pp. 29–38.
35. B. Mohar, Y. Alavi, G. Chartrand, O. R. Oellermann, The Laplacian spectrum of graphs. *Graph Theory Comb. Appl.* **2**, 12 (1991).
36. U. Von Luxburg, A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
37. S. Bressan, J. Y. Li, S. Q. Ren, J. Wu, The embedded homology of hypergraphs and applications. arXiv:1610.00890 [math.AT] (2016).
38. S. Q. Ren, C. Y. Wu, J. Wu, Hodge decompositions for weighted hypergraphs. arXiv:1805.11331 [math.AT] (2018).
39. M. G. Bergomi, M. Ferri, P. Vertechi, L. Zuffi, Beyond topological persistence: Starting from networks. arXiv:1901.08051 [math.CO] (2019).
40. J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).

Acknowledgments: We thank the National Supercomputing Centre, Singapore (NSCC) for providing the computing resource. **Funding:** This work was supported, in part, by Nanyang Technological University Startup Grant M4081842.110 and Singapore Ministry of Education Academic Research Fund Tier 1 RG109/19 and Tier 2 MOE2018-T2-1-033. **Author contributions:** K.X. conceived and designed the study. Z.M. and K.X. performed the calculation. K.X. contributed to the preparation of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper. The code can be downloaded from www.github.com/fdmm1989/PersistentHodgeLaplacian. The PDBbind datasets can be downloaded from www.pdbbind.org.cn.

Submitted 29 April 2020
Accepted 18 March 2021
Published 7 May 2021
10.1126/sciadv.abc5329

Citation: Z. Meng, K. Xia, Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci. Adv.* **7**, eabc5329 (2021).