OXFORD

*Briefings in Bioinformatics*, 22(5), 2021, 1–12

https://doi.org/10.1093/bib/bbab127
Problem Solving Protocol

# Persistent spectral hypergraph based machine learning (PSH-ML) for protein-ligand binding affinity prediction

Xiang Liu, Huitao Feng, Jie Wu and Kelin Xia

Corresponding author: Dr Kelin Xia, Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, 637371 Singapore; E-mail: xiakelin@ntu.edu.sg

## Abstract

Molecular descriptors are essential to not only quantitative structure activity/property relationship (QSAR/QSPR) models, but also machine learning based chemical and biological data analysis. In this paper, we propose persistent spectral hypergraph (PSH) based molecular descriptors or fingerprints for the first time. Our PSH-based molecular descriptors are used in the characterization of molecular structures and interactions, and further combined with machine learning models, in particular gradient boosting tree (GBT), for protein-ligand binding affinity prediction. Different from traditional molecular descriptors, which are usually based on molecular graph models, a hypergraph-based topological representation is proposed for protein–ligand interaction characterization. Moreover, a filtration process is introduced to generate a series of nested hypergraphs in different scales. For each of these hypergraphs, its eigen spectrum information can be obtained from the corresponding (Hodge) Laplacain matrix. PSH studies the persistence and variation of the eigen spectrum of the nested hypergraphs during the filtration process. Molecular descriptors or fingerprints can be generated from persistent attributes, which are statistical or combinatorial functions of PSH, and combined with machine learning models, in particular, GBT. We test our PSH-GBT model on three most commonly used datasets, including PDBbind-2007, PDBbind-2013 and PDBbind-2016. Our results, for all these databases, are better than all existing machine learning models with traditional molecular descriptors, as far as we know.

**Key words:** Persistent spectral hypergraph; Machine learning; Hodge Laplacian; Drug design

## Introduction

The importance of drug design and discovery to our everyday life cannot be overemphasized. However, traditional drug design approaches are laborious, time-consuming, inefficient and expensive. Currently, to develop one new market-approval prescription medicine takes more than 10 years and costs about $2.6 billion [1]. This is because the process of drug design and discovery is highly complicated and covers a variety of steps from target discovery, lead discovery, lead optimization, preclinical development, to the final three phases of clinical trials [1]. Recently, machine learning based models have significantly outperformed traditional models in protein-ligand (PL) binding affinity prediction[2–10], which is one of key steps of drug design. With the ever-increasing accumulation of chemical and biomolecular data, data-driven artificial intelligence (AI) models will usher in an era of faster, cheaper and more-efficient drug design and drug discovery [1].

**Xiang Liu** is a master student from Nankai University in China. He is a visiting student in Nanyang Technological University from Dec 2019 to June 2020.
**Huitao Feng** is a full professor at Nankai University and Mathematical Science Research Center at Chongqing University of Technology in China. His research interests are in differential geometry and topological data analysis.
**Jie Wu** is a full professor at Hebei normal University in China. His research interests are in algebraic topology and topological data analysis. He has been awarded Singapore National Science Award at 2007.
**Kelin Xia** is an assistant professor at Nanyang Technological University, Singapore. His research interests are topological data analysis, molecular based mathematical biology, and machine learning.

**Submitted:** 12 January 2021; **Received (in revised form):** 14 March 2021

There are two general classes of models in AI-based drug design. One is end-to-end deep learning models and the other is feature engineering or featurization based machine learning models. As end-to-end deep learning models, graph neural networks (GNNs), in particular, graph convolution neural networks, have been designed to process non-Euclidean structural data. GNN can work directly on molecular graph representations, which are obtained from molecular structures or derived from molecular Simplified Molecular Input Line Entry Specification(SMILES) sequences. GNN models have achieved great performance in the prediction of various molecular properties in drug discovery [9, 11–17]. Among these models, AquaSol [11] makes use of directed acyclic graph based recursive neural networks (DAG-RNN) to predict molecular solubility. Molecular graph convolutions with only features of atom type, bond type and graph distance have been found to achieve comparable results with deep learning models using more complicated handcrafted features [13]. In DeepVS [12], atom and amino acid embedding together with an effective atom context representation that takes into considerations of protein-ligand complex properties has been incorporated into graph convolutional network (GCN) model. To design an end-to-end representation learning model, a combination framework of compound SMILES-sequence based GCN model and protein sequence-based CNN model has been introduced for the prediction of compound–protein interactions [15]. In DeepChemStable model [16], an attention-based graph convolution network is introduced to dynamically learn graph structural features, and has achieved great success in chemical stability prediction. To retain the molecular spatial connection information, a convolution spatial graph embedding layer (C-SGEL) based GCN model is introduced for molecular property prediction [17].

Feature engineering is widely used in machine learning models in material, chemical and biological data analysis. The essential idea is to represent and characterize molecular structures and functions through molecular descriptors and fingerprints [18, 19], which are usually generated from structural, physical and chemical properties. More than 5000 types of descriptors are proposed and are widely used in quantitative structure acitivity/property relationship (QSAR/QSPR) models. These descriptors can be classified into one-dimensional (1D), two-dimensional (2D), three-dimensional (3D) and four-dimensional (4D) [18, 19]. Molecular properties can be systematically organized into large-sized vectors, known as molecular fingerprints. Various molecular fingerprints are developed, including substructure-key-based fingerprints [20], path-based fingerprints [21, 22], circular fingerprints [23], pharmacophore fingerprints [24, 25] and autoencoded fingerprints from learning models [26–30]. Recently, advanced mathematical tools, in particular topological data analysis (TDA) [31, 32], are used in molecular representations [33–35]. Their combination with learning models have achieved great success in various steps of drug design, including protein-ligand (PL) binding affinity prediction [33, 36–39], protein stability change upon mutation prediction [35, 40], toxicity prediction [41], solvation free energy prediction [42, 43], partition coefficient and aqueous solubility [44], binding pocket detection [45] and drug discovery [46]. Outstanding performance has been consistently achieved in D3R Grand challenge [47–49].

Motivated by the great success of TDA-based mathematical representations in drug design, we have proposed persistent spectral based machine learning (PerSpect ML) [50]. Mathematically, spectral models study the topological properties with algebraic tools, including characteristic polynomial, eigenvalues, eigenvectors, etc. Based on spectral graph theory [51, 52] and spectral simplicial complex [53–56], we propose persistent spectral graph and persistent spectral simplicial complex for the characterization of biomolecular structures and interactions [50]. Topologically, graphs and simplicial complexes can be further generalized into hypergraphs. Hypergraph is composed of hyperedges, which are defined as non-zero sets of vertices, and are the generalization of edges and simplexes. Recently, we have developed hypergraph-based embedded persistent cohomology for molecular representation and further combined with machine learning models [57].

Here we propose persistent spectral hypergraph (PSH) and PSH-based machine learning (PSH-ML) for the first time. Our PSH-ML models are used in protein-ligand binding affinity prediction. Different from previous molecular models, molecular structures and interactions at atomic level are represented as hypergraphs. Based on the supremum chain group from embedded persistent homology, we develop our spectral hypergraph model. Further, a filtration process is introduced to generate a series of nested hypergraphs. The persistence and variation of hypergraph spectral information is called PSH. Persistent attributes are generated from PSH and used as molecular descriptors in machine learning models, in particular, gradient boosting tree model (GBT). GBT is chosen to reduce overfitting as our model has large-sized molecular descriptors. Our PSH-GBT models are tested on three well-established PL binding affinity datasets, i.e. PDB-v2007, PDB-v2013 and PDB-v2016, from PBDbind databank. The state-of-the-art results are obtained for all these datasets.

## Results

### PSH-based biomolecular characterization

#### *Biomolecular hypergraph representation*

A hypergraph $(V_{\mathcal{H}}, \mathcal{H})$ is composed of vertex set $V_{\mathcal{H}}$ and hyperedge set $\mathcal{H}$, which is a collection of non-empty subsets of $V_{\mathcal{H}}$. Hyperedges can be viewed as a generalization of edges (in graph) and simplexes (in simplicial complex). Compared with graph model, both simplicial complex and hypergraph can characterize higher dimensional topological connections, such as three edges share the same triangle or four triangles share the same tetrahedron. However, the major difference between hypergraph and simplicial complex is the completeness of boundary operators for their chain groups. If one defines a boundary operator $\partial_k$, as in chain groups of simplicial complexes, for $k$-hypergraph group $G(\mathcal{H}_k)$ as $\partial_k(\sigma_k) = \sum_{i=0}^{k}(-1)^i\sigma_{k-1}^i$, with $\sigma_k = \{v_0, v_1, ..., v_k\}$ a $k$-hyperedge and $(k-1)$-hyperedge $\sigma_{k-1}^i = \{v_0, v_1, ..., v_{i-1}, v_{i+1}, ..., v_k\}$ generated by removing the vertex $v_i$ from hyperedge $\sigma_k$, a serious problem will be encountered as one or more $(k-1)$-hyperedge $\sigma_{k-1}^i$ may not always exist in the $k-1$ hyperedge set $\mathcal{H}_{k-1}$. Stated differently, boundary operators from the simplicial complex chain groups will not be well defined for hyperedge groups. Roughly speaking, boundary elements of simplexes are always elements of the simplicial complex, while boundary elements of hyperedges may not within the hypergraph.

Figure 1 demonstrates the comparison between graph, simplicial complex and hypergraph representations for protein 1ALF. Only the backbone $C_\alpha$ atom and nitrogen (N) atoms are considered. There are totally 38 atoms made from 19 $C_\alpha$ atoms and 19 N atoms. The graph is generated using a cutoff distance 3.5 Å. That is an edge is formed between any two atoms if their distance is smaller or equal to 3.5 Å. The simplicial complex is generated

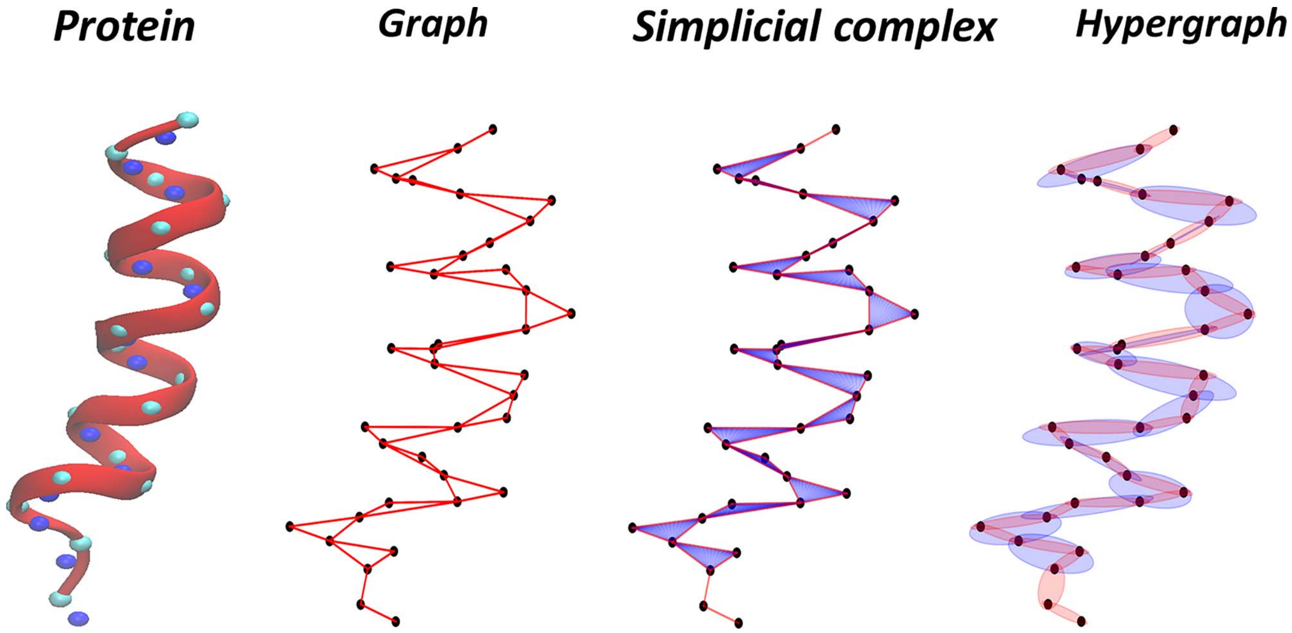

**Figure 1.** The comparison of three different topological representations, i.e. graph, simplicial complex and hypergraph, for protein 1ALF. Only the backbone $C_\alpha$ atoms and N atoms are considered. Graph is composed of only 0-simplexes (vertices) and 1-simplex (edges). In contrast, simplicial complex can have higher dimensional simplexes, such as 2-simplexes (triangles in blue color). Hypergraph is made of hyperedges, which are non-empty subsets of vertices and can be viewed as a generalization of edges and simplexes. 1-hyperedges and 2-hyperedges are represented as red and blue ellipses, respectively.

using the Viertoris–Rips complex with the same cutoff distance. For hypergraph, a $k$-hyperedge ($k > 0$) is formed among $k+1$ atoms if they satisfy two conditions. First, there must be one $C_\alpha$ atom and one N atom. Second, the distance between any two atoms are smaller or equal to 3.5 Å. It can be seen that in the protein hypergraph, each 2-hyperedge (blue ellipse) contains in it only two 1-hyperedges (red ellipses), thus hyperedge groups are not complete under traditional chain boundary operators.

There are various different ways to define molecular hypergraphs. For instance, we can define any covalent (or noncovalent) bond as 1-hyperedge, and any two adjacent covalent (or noncovalent) bonds as 2-hyperedge. Moreover, hypergraph can be used in molecular interaction characterization. Hyperedge can be naturally generated from molecules that are interacted with each other. A detailed discussion of the hypergraph-based molecular interaction models is in Method.

## PSH for biomolecular representation

PSH is the combination of spectral hypergraph models and filtration-based persistent models. Mathematically, there are four spectral hypergraph models [56, 58–60, 60–64], which are derived from different types of hypergraph Laplacian matrixes. The first model is based on hypergraph incidence matrix [58–62]. Together the derived vertex diagonal matrix, hypergraph adjacent matrix and Laplacian matrix can be generated. The second one is from the associated simplicial complex of hypergraph [55, 56, 65]. Through a clique expansion, a hypergraph induces a unique clique complex, from which (combinatorial) Hodge Laplacian (HL) matrixes can be generated. The third one is based on the tensor representation of hypergraph [60, 63, 66]. For an $m$-uniform hypergraph, an adjacency tensor can be generated accordingly and spectral information can be derived from either H-eigenvalue tensor or Z-eigenvalue tensor. The last spectral hypergraph is proposed in this paper based on the supremum and infimum chain complexes of the hypergraph [64, 67–69]. From these chain groups, boundary matrixes and associated HL matrixes can be defined. A detailed description of these models can be found in Method.

The filtration process is key to persistent models, including persistent homology/cohomology [31, 32, 70], persistent spectral [50] and persistent function [71]. Mathematically, a nested sequence of topological representations, including graphs, simplicial complexes, hypergraphs, etc, can be generated following the increase or decrease of a certain filtration value. In PSH, a series of Laplacian matrixes are generated from hypergraph sequences. From these Laplacian matrixes, spectral information can be derived. The persistence and variance of the spectral information is called persistent spectral.

Figure 2 shows a hypergraph-based filtration process for protein 1ALF (**a**), and the HL matrixes of dimension 0 (**b**) and dimension 1 (**c**). We consider the 0-dimension (Dim (0)) and 1-dimension (Dim (1)) supremum HL matrixes as in Eq.(3). Five different filtration values, i.e. 0.0 Å, 2.5 Å, 3.5 Å, 4.5 Åand 5.0 Å, are considered. Since the total number of atoms is 38, the size of Dim (0) HL matrixes are always $38 * 38$. With the increase of filtration value, off-diagonal entries of Dim (0) HL decrease from 0 directly to -1 and eventually all off-diagonal entries become -1. At the same time, diagonal entries gradually increase until all of them reach 37. Note that 1-hyperedges are generated only between $C_\alpha$ atoms and N atoms, but extra '1-hyperedge elements' are generated from the boundary group of $G(\mathcal{H}_2)$ in the supremum chain group $\mathrm{Sup}_1(\mathcal{H})$ as indicated in Eq. (2).

Further, the size of Dim (1) HL matrix consistently increases with the filtration until it reaches the maximum size of $703 * 703$ ($703 = C_{38}^2$). The diagonal entries will consistently increase and their largest value is 38, which is summation of upper degree 36 and lower degree 2. Note that the upper degree of an 1-hyperedge is the number of 2-hyperedges that contain the 1-hyperedge.

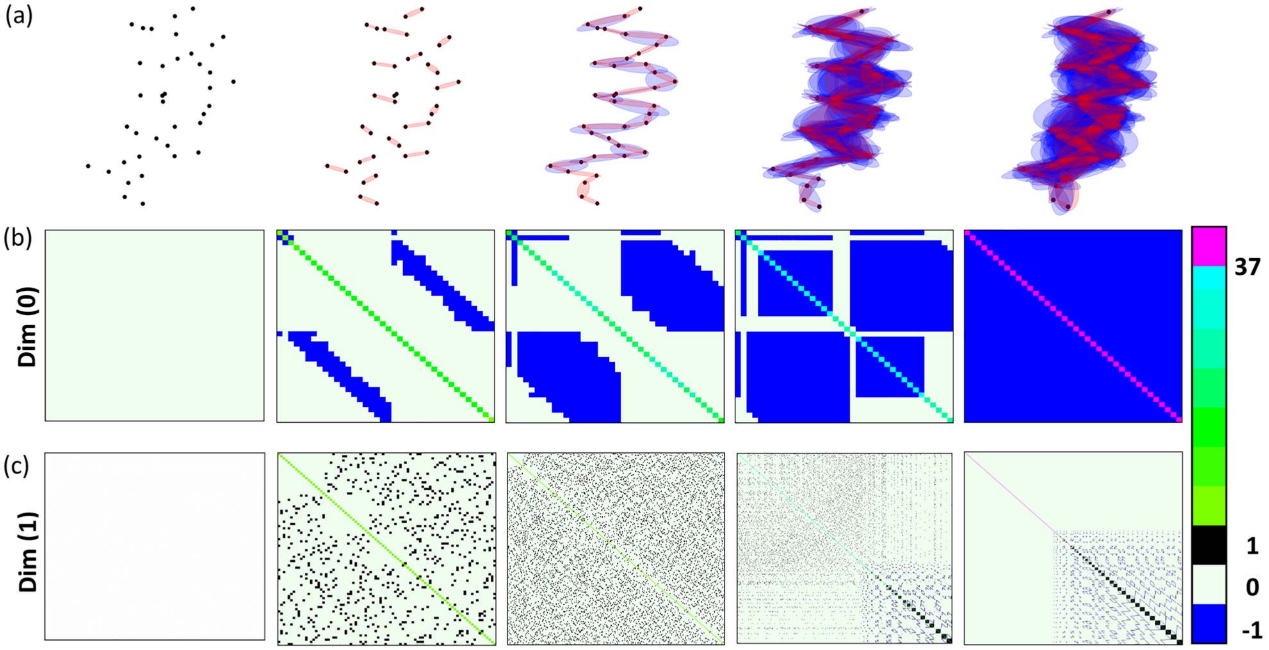

**Figure 2.** The illustration of hypergraph-based filtration process for protein 1ALF (**a**) and the corresponding HL matrixes of Dim (0) (**b**) and Dim (1) (**c**). Five different filtration values, i.e. 0.0 Å, 2.5 Å, 3.5 Å, 4.5 Å and 5.0 Å, are considered. With the increase of filtration value, off-diagonal entries of Dim (0) HL matrix decrease from 0 to -1 until all of them become -1. Diagonal entries gradually increase to 37. For Dim (1) HL, its matrix sizes consistently increase until it reaches 703*703. There are two types of diagonal entries. One corresponds to 1-hyperedges (totally 361) and has the largest value as 38. The other is for '1-hyperedge elements' (totally 342) from boundary groups and has the largest value as 21.

Since each 1-hyperedge contains one $C_\alpha$ atom and one N atom, any of the rest 18 $C_\alpha$ atoms or 18 N atoms can be added to form a 2-hyperedge. In this way, there are totally 36 2-hyperedges. Among them, 18 2-hyperedges are composed of two $C_\alpha$ atoms and one N atom and the other 18 2-hyperedges are composed of two N atoms and one $C_\alpha$ atom. Further, only 361 ($19*19$) diagonal entries, corresponding to the 1-hyperedges formed between 19 $C_\alpha$ atoms and 19 N atoms, can finally reach 38. The rest 342 diagonal entries can only reach the maximum value 21, which is summation of upper degree 19 and lower degree 2. These entries correspond to '1-hyperedge elements' from the boundary group of $G(\mathcal{H}_2)$. These '1-hyperedge elements' are formed within $C_\alpha$ atom-set or within N atom-set. For each atom-set, there are 171 ($C_{19}^2$) '1-hyperedge elements'. The '1-hyperedge element' forms a 2-hyperedge only with a different type of atom, thus its largest upper degree is 19.

### *PSH-based machine learning*

As stated above, HL matrixes from the hypergraph filtration process are usually of different sizes. The corresponding eigenvalues or eigenvectors cannot be directly used as input for machine learning models. To solve the problem, persistent attributes, which are statistical and combinatorial properties of eigenvalues, are proposed as PSH-based molecular descriptors or fingerprints. More specifically, we consider statistical properties, such as multiplicity of zero-eigenvalue, number of nonzero-eigenvalue, maximum, minimum, average, standard deviation and sum of eigenvalues from the HL matrix. The persistence of these eigenvalue attributes are defined as persistent multiplicity of zero-eigenvalue, persistent number of nonzero-eigenvalue, persistent maximum, persistent minimum, persistent average, persistent standard deviation, persistent sum, respectively. Essentially, persistent attributes can be viewed as a function over filtration value. Mathematically, the multiplicity of zero-eigenvalue from HL matrixes corresponds to the Betti number of embedded homology. Other eigenvalue attributes [18], such as spectral moments, quasi-Wiener index, spanning tree number, etc, can also be used as descriptors.

Figure 3 illustrates five persistent attributes including persistent multiplicity (**b**), persistent maximum (**c**), persistent minimum (**d**), persistent mean (**e**), persistent standard derivation (**f**), for the hypergraph-based filtration process of protein 1ALF in Figure 2. The persistent barcode for the embedded homology is illustrated in Figure 3 (**a**). It can be seen that persistent multiplicity equals to persistent Betti number (or Better curve), which is the summation of bacodes at each filtration value. Further, Dim (0) persistent maximum, persistent minimum and persistent mean all consistently increase with the filtration value until they reach the same largest value 38. Dim (0) persistent standard deviation increases at early stage of filtration until it reaches peak value at filtration value around 25 Å, then it begins to decrease until it stabilizes at value 0. Moreover, Dim (1) persistent attributes have more complicated properties. Dim (1) persistent maximum has the same pattern as Dim (0) persistent maximum. Dim (1) persistent minimum and persistent mean increase firstly at early stage of filtration, then plummet suddenly at filtration value of 16.6 Å, after that they consistently increase again until reaching their largest values of 19 for persistent minimum and 29.73 for persistent mean. Note that 19 is the smallest diagonal value and 29.73 is the average of diagonal values, of Dim (1) HL matrix. Dim (1) persistent standard deviation also increases at early stage of filtration. After a sudden increase at the same filtration value 16.6 Å, it begins to decrease and reaches its local minimum at filtration value around 25 Å, then it consistently increase until reaching the largest value. Note that

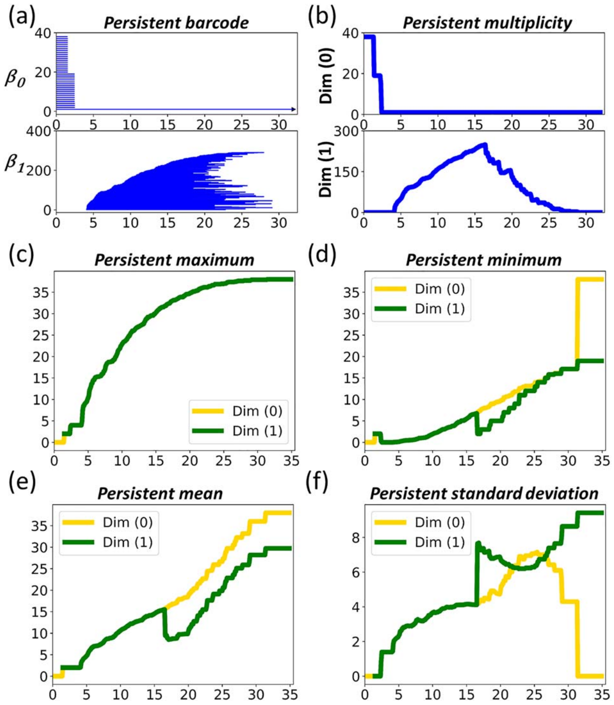

**Figure 3.** Five persistent attributes obtained from the hypergraph-based filtration process of protein 1ALF in Figure 2. They are persistent multiplicity of zero-eigenvalues (**b**), persistent maximum (**c**), persistent minimum (**d**), persistent mean (**e**), persistent standard derivation (**f**). Note that persistent multiplicity of zero-eigenvalues is equal to persistent Betti number (or Betti curve), which is summation of persistent bacodes (**a**) at each filtration value.

16.6 Å is the filtration value when 2-hyperedges begin to appear. Interestingly, Dim (1) persistent multiplicity reaches its largest value just before 16.6 Å and begin to decrease at exactly 16.6 Å.

### PSH-ML for protein-ligand binding affinity prediction

Our PSH-ML is applied to protein-ligand binding affinity prediction, which is one of the key steps in drug design and discovery. Three most commonly used protein-ligand databases, including PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016, are considered [72]. The detailed data can be downloaded from PDBbind website, and the size of training and test sets can be found in Table 1. Element-specific representation [33, 35–37, 40] is used to characterize the detailed PL interactions at molecular level. The essential idea is to decompose protein molecule into four atom-sets and ligand molecule into nine atom-sets. In this way, 36 atom–atom combinations can be constructed with one atom-set from protein and another atom-set from ligand, and a total 36 types of element-specific hypergraphs can be constructed from these atom-atom combinations. An example of element-specific hypergraph representation for protein–ligand interactions can be found in Figure 4. For each atom-atom combination, a hypergraph model is constructed. Essentially, each 1-hyperedge should have an atom (as vertex) from protein and the other from ligand. For a 2-hyperedge, it should have at least one atom from protein and one atom from ligand. Speaking differently, 2-hypedge cannot be formed between three protein atoms or three ligand atoms. Details can be found in Method.

Computationally, the protein binding core region is defined as the protein domain that is within 10Å cutoff distance of ligand. The filtration region is from 2.0Å to 7.0Å with step 0.1Å. We only consider seven persistent attributes, from supremum HL matrixes of Dim (0,1), as listed below,

- persistent multiplicity of zero-eigenvalue
- persistent number of nonzero-eigenvalue
- persistent maximum
- persistent minimum
- persistent average
- persistent standard deviation
- persistent sum

Only Dim (0) spectral information is used for ligand. Two PSH-GBT models are considered. The first model employs features only from PL complexes. There are totally 25 200=36(atom combinations)*50(filtration values)*2(Dim (0,1) dimensional HLs for PL complex)*7(persistent attributes) features in this model. The second model combines features from PL complexes and features only from ligands. This model contains 37 800=36(atom combinations)*50(filtration values)*3(Dim (0,1) HLs for PL complex and Dim (0) HL for ligand)*7(persistent attributes) descriptors. Note that GBT model is employed to reduce the overfitting problem from large-sized feature vector, and the detail parameters of the GBT model are listed in Table 2.

Table 3 shows our PSH-GBT results of Pearson correlation coefficient (PCC) and root mean square error (RMSE), with unit $pK_d/pK_i$, for the three test sets of PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. Note that 10 independent regressions are performed and the median values of PCCs and RMSEs are taken as the final performance of our model. Figure 5 illustrates the comparison of the predicted binding affinity from our PSH-GBT model with experimental ones, for the three test sets of PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. It can be seen that the best results are obtained from PDBbind-v2016. This is partially due to the reason that more data are available at 2016 training set. Further, Table 4 lists the 10 cases with the largest prediction errors in each test case, using our model.

To have a better understanding of our model performance, we compare the performance of our PSH-GBT(PL+L) model with more than 40 deep learning models and machine learning models with traditional descriptors [2–10]. The results of the corresponding PCC values are illustrated in Figure 6. It can be seen that our model outperforms all the other models for all three datasets, which demonstrates the great potential of our model in drug design. Note that all these models use network or graph representation for molecular structures or interactions. In comparison, we use hypergraph to represent molecular structures and interactions. Further, we use PSH-based persistent attributes as molecular descriptors. Since our PSH-based molecular features are highly abstract and characterize the intrinsic molecular information, they have a better transferability and the related machine learning models can have a better accuracy compared with traditional descriptors.

## Discussion

Machine learning models have made tremendous progresses in text, video, audio and image data analysis. In particular, convolutional neural network (CNN) models have achieved revolutionary advancements in the analysis of image data. However, molecular data from material, chemical and biological systems are fundamentally different from text and image data, as their properties are usually directly determined by their topological structures.

**Table 1.** Detailed information of the three PDBbind databases, i.e. PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016

| Dataset | Refined set | Training set | Test set (Core set) |
|---|---|---|---|
| PDBbind-v2007 | 1300 | 1105 | 195 |
| PDBbind-v2013 | 2959 | 2764 | 195 |
| PDBbind-v2016 | 4057 | 3772 | 285 |

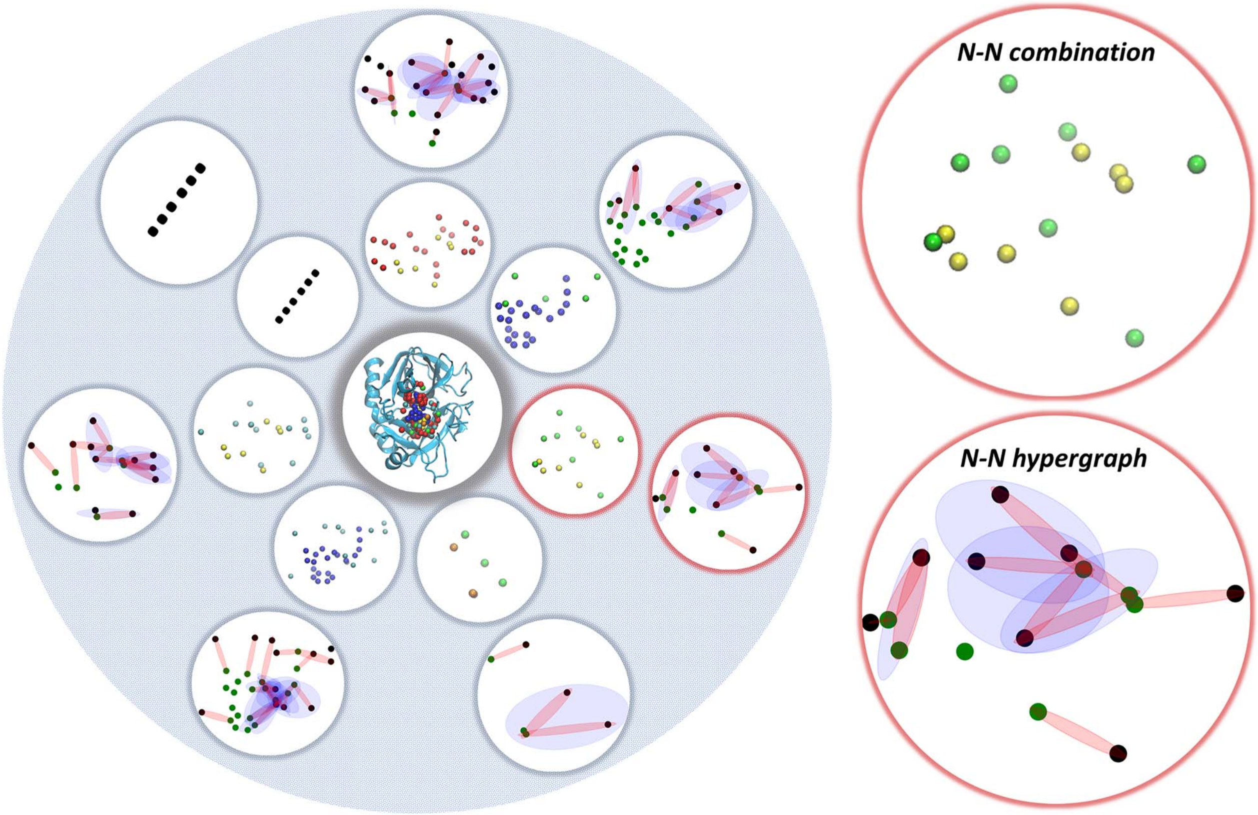

**Figure 4.** The illustration of element-specific hypergraphs for protein–ligand interactions. The protein-ligand complex (PDBid: 1W7G) is decomposed into 36 types of atom–atom combinations. A total of 36 element-specific hypergraphs are constructed based on these combinations.

**Table 2.** The parameter setting for our GBT model.

| | | | |
|---|---|---|---|
| No. of Estimators | Learning rate | Max depth | Subsample |
| 40 000 | 0.001 | 8 | 0.7 |
| Min_samples_split | Loss function | Max features | Repetitions |
| 2 | Least square | SQRT | 10 |

Persistent models, including persistent homology/cohomology, persistent functions and persistent spectral, provide a series of highly effective molecular descriptors that not only preserve intrinsic structural information, but also maintain molecular multiscale properties. However, persistent models are based on graphs and simplicial complexes. Only persistent spectral graph and persistent spectral simplicial complex have been used in molecular representations.

Topologically, hypergraph is a generalization of simplicial complex and has more flexibility in characterizing complicated structures. Recently, we have developed hypergraph-based embedded cohomology and persistent embedded cohomology for molecular structure and interaction description. Here we propose PSH, which can be viewed as a generalization of persistent spectral graph and persistent spectral simplicial complex. Our spectral hypergraph model is based on supremum chain group from the embedded homology model. Mathematically, various different spectral hypergraph models can also be employed and PSH models can be generated correspondingly. Other than protein-ligand binding affinity prediction, our PSH-ML models can also be used in other drug design and drug discovery procedures. Note that our models and GNN models are based on molecular structures and interactions, they cannot be directly used in gene sequence analysis.

## Methods

Our PSH contains two essential components, i.e. spectral hypergraph model and filtration-based persistent representation. Our

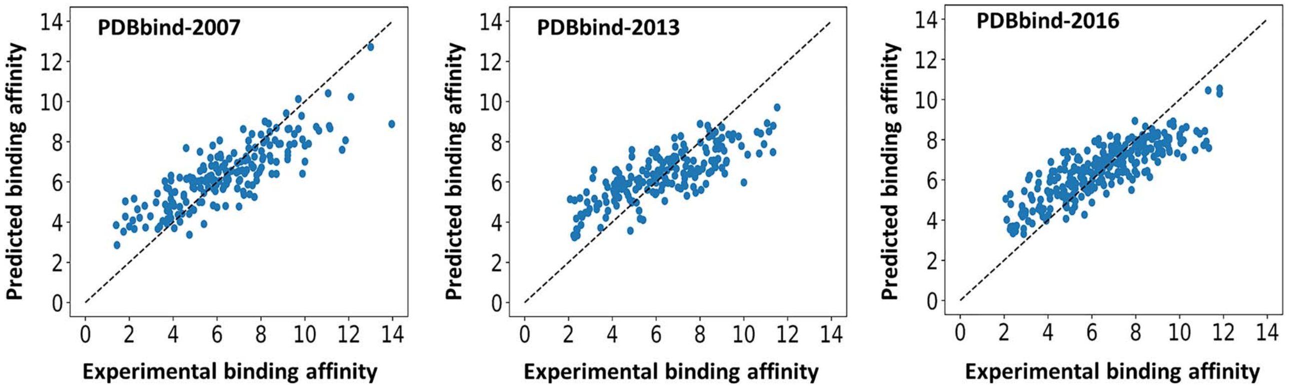

**Figure 5.** The comparison of the predicted binding affinity from our PSH-GBT model with experimental ones, for the test sets of PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. Note that the unit of the binding affinity is $pK_d/pK_i$.

**Table 3.** The PCCs and RMSEs ($pK_d/pK_i$) for our PSH-GBT models in three test cases, i.e. PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. The PSH-GBT(PL) model uses features only from protein-ligand complexes. The PSH-GBT(PL+L) model uses combined features from both PL complexes and Ligands.

| | PSH-GBT(PL) | PSH-GBT(PL+L) |
|---|---|---|
| PDBbind-v2007 | 0.824(1.406) | 0.827(1.400) |
| PDBbind-v2013 | 0.780(1.491) | 0.783(1.482) |
| PDBbind-v2016 | 0.830(1.293) | 0.835(1.280) |

**Table 4.** The list of 10 cases with the largest prediction errors in our PSH-GBT model, for the three test datasets of PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016.

| | PDB IDs |
|---|---|
| PDBbind-v2007 | 7CPA, 1Y6Q, 1SL3, 2DRC, 1Y1M, 1BMA, 1FLR, 1M0N, 1BRA, 2HDQ |
| PDBbind-v2013 | 1IGJ, 2X00, 2YMD, 3MYG, 2XDL, 3AO4, 3UTU, 1PS3, 4TMN, 2VO5 |
| PDBbind-v2016 | 2X00, 5C2H, 5DWR, 3MYG, 2YMD, 1PS3, 4TMN, 3AO4, 2XDL, 3OZT |

PSH can characterize not only topological information, such as Betti number, but also more detailed 'geometric' information embedded in nonzero eigenvalues, thus it can work as an effective model for molecular representation.

## Persistent spectral hypergraph

In general, there are four spectral hypergraph models and all of them are based on hypergraph Laplacian (or Hodge-Laplacian) matrixes [56, 58–60, 60–64]. For simplicity, we denote a hypergraph as $(V_{\mathcal{H}}, \mathcal{H})$, and its vertex set as $V_{\mathcal{H}} = \{v_i; i = 1, 2, ..., N\}$ with $N$ the total number of hypergraph vertices, and its hyperedge set as $\mathcal{H} = \{\sigma_i^k; k = 0, 1, ...; i = 1, 2, ...\}$ with $k$ the hyperedge dimension.

### *Incidence-matrix-based spectral hypergraph*

The simplest way to generate hypergraph Laplacian is to use the incidence matrix [58–62]. If we arrange hyperedges with dimensions larger than 0 into a sequence labeled as $\sigma(1), \sigma(2), ..., \sigma(N_e)$ with $N_e$ the total number of hyperedges, its incidence matrix $\mathbf{H}$ of size $N * N_e$ is defined as follows:

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

From the incidence matrix $\mathbf{H}$, a diagonal matrix $\mathbf{D}_v$ can be defined as

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

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

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

Further, we can have the symmetric normalized hypergraph Laplacian $\mathbf{L}_{\text{sym}} = 2\mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{H}^T\mathbf{D}_v^{-1/2}$ with $\mathbf{I}$ the identity matrix, and the random walk hypergraph Laplacian $\mathbf{L}_{\text{rw}} = 2\mathbf{I} - \mathbf{D}_v^{-1}\mathbf{H}\mathbf{H}^T$.

### *Associated-simplicial-complex-based spectral hypergraph*

Another way to define hypergraph Laplacian is to employ a clique expansion. For each hyperedge, an edge is formed between any pair of vertices within this hyperedge in its clique graph. A clique complex $K_{\mathcal{H}}$ can be generated from the clique graph. This clique complex is the smallest simplicial complex that $\mathcal{H}$ can be embedded into, and it is also called associated simplicial complex. Based on $K_{\mathcal{H}}$, Hodge (combinatorial) Laplacian matrixes can be constructed accordingly [53–56, 73–76].

Computationally, we can assign a certain orientation for the clique complex $K_{\mathcal{H}} = \{\delta_i^k; k = 0, 1, ...; i = 1, 2, ...\}$. Its $k$-th boundary matrix $\mathbf{B}_k$ ($k > 0$) can be defined as follows:

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

Here $\delta_i^{k-1} \subset \delta_j^k$ means that $\delta_i^{k-1}$ is a face of $\delta_j^k$ and $\delta_i^{k-1} \not\subset \delta_j^k$ means the opposite. The notation $\delta_i^{k-1} \sim \delta_j^k$ means the two simplexes have the same orientation, i.e. oriented similarly, and $\delta_i^{k-1} \nsim \delta_j^k$ means the opposite.

The $k$-th HL matrix is defined as follows [55, 75]:

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

Note that $\mathbf{L}_0 = \mathbf{B}_1\mathbf{B}_1^T$ is just the graph Laplacian matrix. More specifically, the $k$-th combinatorial Laplacian matrix ($k > 0$) can be expressed explicitly as [55, 75]

$$L_k(i,j) = \begin{cases} d(\delta_i^k) + k + 1, & \text{if } i{=}j \\ 1, & \text{if } i{\neq}j, \delta_i^k \not\frown \delta_j^k, \delta_i^k \smile \delta_j^k \text{ and } \delta_i^k \sim \delta_j^k \\ -1, & \text{if } i{\neq}j, \delta_i^k \not\frown \delta_j^k, \delta_i^k \smile \delta_j^k \text{ and } \delta_i^k \not\sim \delta_j^k \\ 0, & \text{if } i{\neq}j, \delta_i^k \frown \delta_j^k \text{ or } \delta_i^k \not\smile \delta_j^k. \end{cases}$$

Here $d(\delta_i^k)$ is (upper) degree of $k$-simplex $\delta_i^k$. It is the number of $(k+1)$-simplexes, of which $\delta_i^k$ is a face. Notation $\delta_i^k \frown \delta_j^k$ means the two simplexes are upper adjacent, i.e. they are faces of a common $(k+1)$-simplex, and $\delta_i^k \not\frown \delta_j^k$ means the opposite. Notation $\delta_i^k \smile \delta_j^k$ means the two simplexes are lower adjacent, i.e. they share a common $(k-1)$-simplex as their face, and $\delta_i^k \not\smile \delta_j^k$ means the opposite. Notation $\delta_i^k \sim \delta_j^k$ means the two simplexes have the same orientation, i.e. oriented similarly, and $\delta_i^k \not\sim \delta_j^k$ means the opposite.

### *Tensor-eigenvalue-based spectral hypergraph*

Hypergraph can be represented as a tensor and spectral hypergraph can be generated by using tensor eigenvalue models [60, 63, 66]. One can denote a real $m$-order $n$-dimensional tensor $A$ that consists of $n^m$ real entries as $A_{i_1,\ldots,i_m} \in R$ where $i_j = 1, 2, \ldots, n$ for $j = 1, 2, \ldots, m$. For a vector $x \in R^n$, one can use $x_i$ to denote its components, and $x^{[m]}$ to denotes a vector in $R^n$ such that $x_i^{[m]} = x_i^m$ for all $i$. The tensor product $Ax^{m-1}$ is a vector in $R^n$, whose $i$-th component is

$$\sum_{i_2,\ldots,i_m=1}^{n} A_{i,i_2,\ldots,i_m} x_{i_2} \cdots x_{i_m}.$$

A real number $\lambda$ is an H-eigenvalue of tensor $A$, if there exists a nonzero real vector $x$ that satisfies

$$Ax^{m-1} = \lambda x^{[m-1]}.$$

Here $x$ is an H-eigenvector of A associated with H-eigenvalue $\lambda$.

A real number $\lambda$ is a Z-eigenvalue of tensor $A$, if there exists a nonzero real vector $x$ that satisfies

$$\begin{cases} Ax^{m-1} = \lambda x \\ x^T x = 1. \end{cases}$$

Here $x$ is a Z-eigenvector of A associated with Z-eigenvalue $\lambda$.

A hypergraph $(V_{\mathcal{H}}, \mathcal{H})$ is said to be $m$-uniform ($m \geq 2$), if the cardinal number, i.e. number of vertices, of hyperedges all equals to $m$. The adjacency tensor $\mathcal{A}_{\mathcal{H}}$ for an $m$-uniform hypergraph $\mathcal{H}$ is the symmetric tensor $\mathcal{A}_{\mathcal{H}} = (a_{i_1,\ldots,i_m}) \in R^{[m,n]}$, where $n$ is the number of vertices and $E$ is set of hyperedges,

$$a_{i_1,\ldots,i_m} = \frac{1}{(m-1)!}\begin{cases} 1, & \textit{if } \{i_1, \ldots, i_m\} \in E \\ 0, & \textit{otherwise}. \end{cases}$$

The spectral information can be derived from the adjacency tensor of the hypergraph, using either H-eigenvalue or Z-eigenvalue.

### *Embedded-homology-based spectral hypergraph*

Recently, motivated by the elegant path complex [77–80], embedded homology and persistent embedded homology have been proposed [64, 67–69]. The essential idea of embedded homology is the construction of infimum and supremum chain groups. Here we propose hypergraph HL matrixes based on these chain groups and develop a new spectral hypergraph model.

For a hypergraph $(V_{\mathcal{H}}, \mathcal{H})$, its $k$-hyperedge group is denoted as $G(\mathcal{H}_k)$ with $\mathcal{H}_k$ the set of all $k$-hyperedges in $\mathcal{H}$. Note that $G$ is an Abelian group and $G(\mathcal{H}_k)$ is the collection of linear combinations of $k$-hyperedges in $\mathcal{H}_k$ with coefficients in $G$. We can define a $k$-th infimum chain group as follows:

$$\mathrm{Inf}_k(\mathcal{H}) = \mathrm{Inf}_k(G(\mathcal{H}_\star), G((K_{\mathcal{H}})_\star)) = G(\mathcal{H}_k) \cap \partial_k^{-1}(G(\mathcal{H}_{k-1})),$$

and a $k$-th supremum chain group as follows:

$$\mathrm{Sup}_k(\mathcal{H}) = \mathrm{Sup}_k(G(\mathcal{H}_\star), G((K_{\mathcal{H}})_\star)) = G(\mathcal{H}_k) + \partial_{k+1}(G(\mathcal{H}_{k+1})). \tag{2}$$

Note that $G(\mathcal{H}_\star) = \{G(\mathcal{H}_0), G(\mathcal{H}_1), G(\mathcal{H}_2) \ldots\}$ is a sequence of hyperedge groups, and $G((K_{\mathcal{H}})_\star) = \{G((K_{\mathcal{H}})_0), G((K_{\mathcal{H}})_1), G((K_{\mathcal{H}})_2), \ldots\}$ is a sequence of chain groups from the associated simplicial complex $K_{\mathcal{H}}$, which is from the clique expansion of the hypergraph as stated above.

From infimum chain groups, $k$-th infimum boundary operators can be defined as follows:

$$\partial_k^{Inf} : \mathrm{Inf}_k(\mathcal{H}) \to \mathrm{Inf}_{k-1}(\mathcal{H}).$$

From supremum chain groups, $k$-th supremum boundary operators can be defined as follows:

$$\partial_k^{Sup} : \mathrm{Sup}_k(\mathcal{H}) \to \mathrm{Sup}_{k-1}(\mathcal{H}).$$

Computationally, boundary matrixes can be obtained from the above boundary operators. More specifically, we can denote the $k$-th infimum boundary matrixes as $\mathbf{B}_k^{Inf}$ and the $k$-th supremum boundary matrixes as $\mathbf{B}_k^{Sup}$. Similar to the HL matrixes as in Eq.(1), we can define $k$-th infimum HL matrixes as

$$\mathbf{L}_k^{Inf} = (\mathbf{B}_k^{Inf})^T\mathbf{B}_k^{Inf} + \mathbf{B}_{k+1}^{Inf}(\mathbf{B}_{k+1}^{Inf})^T.$$

and $k$-th supremum HL matrixes as

$$\mathbf{L}_k^{Sup} = (\mathbf{B}_k^{Sup})^T\mathbf{B}_k^{Sup} + \mathbf{B}_{k+1}^{Sup}(\mathbf{B}_{k+1}^{Sup})^T. \tag{3}$$

Note that the multiplicity of zero-eigenvalue of $\mathbf{L}_k^{Inf}$ and $\mathbf{L}_k^{Sup}$ are same, which is just the Betti number corresponding to the embedded homology of $\mathcal{H}$.

## PSH-ML for protein-ligand binding affinity prediction

### *Hypergraph representation for biomolecular interactions*

To characterize the protein–ligand interactions at molecular level, we consider the element-specific molecular representations [33, 35–37, 40]. Essentially, we can decompose protein

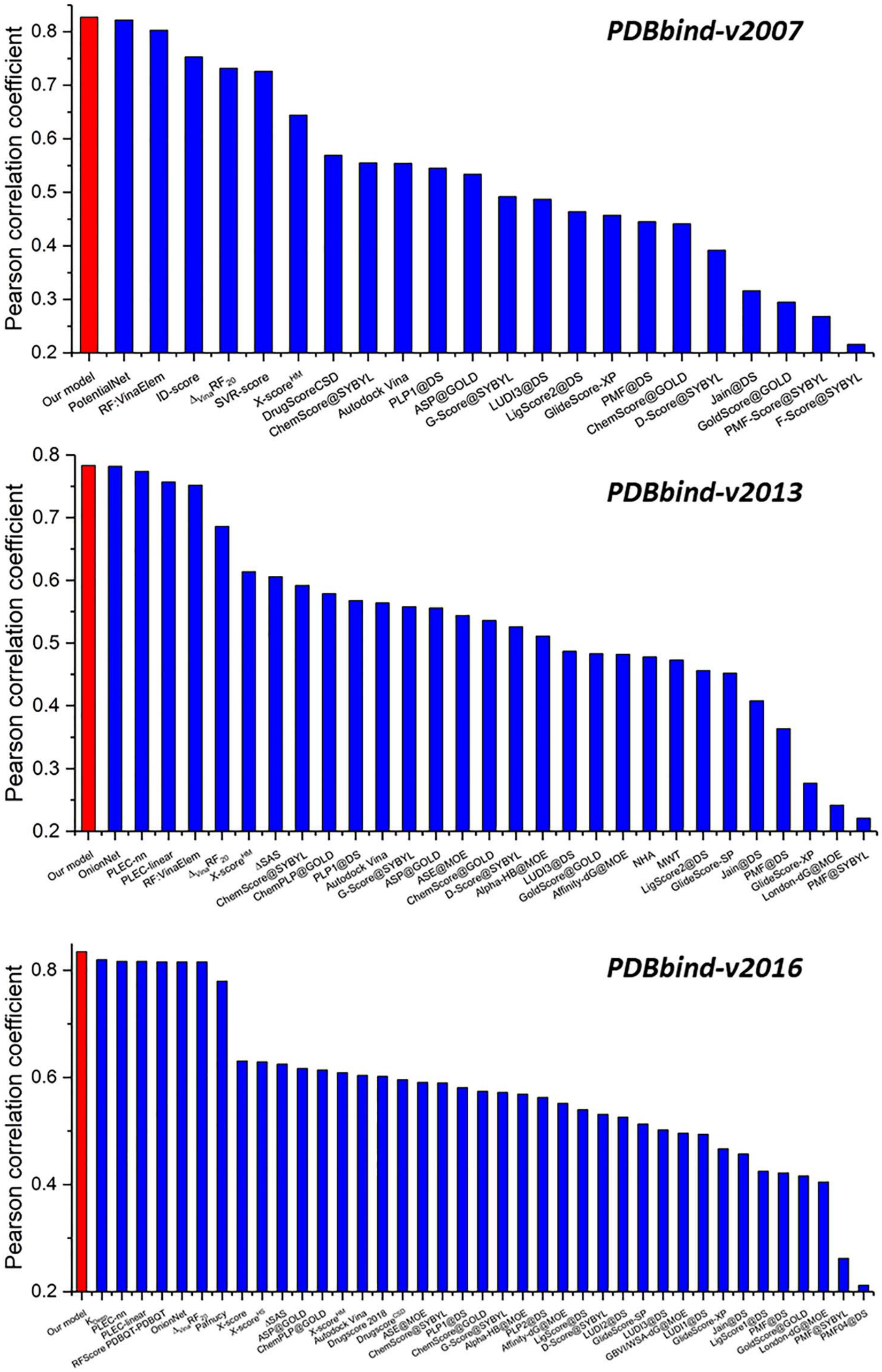

**Figure 6.** The comparison of our PSH-GBT(PL+L) model and machine learning models with traditional molecular descriptors, for the prediction of protein-ligand binding affinity [2–10]. The PCCs are calculated based on the core set (test set) of PDBbind-2007, PDBbind-2013 and PDBbind-2016.

structures into four atom-sets composed of C, N, O and S atoms, respectively, and ligands into nine atom-sets composed of C, N, O, S, P, F, Cl, Br and I atoms, respectively. In this way, a total 36 atom combinations can be constructed with one atom-set from protein and another atom-set from ligand.

Hypergraphs can be constructed from each atom combinations [57]. We denote the atom-set from protein and ligand as $V_P = \{\mathbf{v}_i; i = 1, 2, ..., N_P\}$ and $V_L = \{\mathbf{v}_j; j = 1, 2, ..., N_L\}$, respectively, with $\mathbf{v}_i$ and $\mathbf{v}_j$ the respective $i$-th and $j$-th atom coordinate, and $N_P$ and $N_L$ the respective total numbers. The hypergraph vertex set is $V_{\mathcal{H}} = V_P \cup V_L$. A $k$-hyperedge $\sigma_k$ of hyperset $\mathcal{H}$ can be defined as

$$\sigma_k = \begin{cases} \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_k\}; \mathbf{v}_n \in V_{\mathcal{H}} (0 \leqslant n \leqslant k), \exists i, j \in [0, k], \mathbf{v}_i \in V_P, \mathbf{v}_j \in V_L, & k > 0 \\ \{\mathbf{v}_0\}; \mathbf{v}_0 \in V_{\mathcal{H}}, & k = 0. \end{cases}$$

Note that each vertex is a 0-hyperedge, each $k$-hyperedge ($k > 0$) must contain one vertex (atom) from protein and another vertex (atom) from ligand.

### *PSH representation for biomolecular interactions*

An essential component for PSH is the filtration-based persistent analysis. In order to build a hypergraph-based filtration process, a suitable filtration value (or ‘birth time’) is required for each hyperedge. Here we define a filtration value for a $k$-hyperedge $\sigma_k = \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_k\}$ as follows [57]:

$$f(\sigma_k) = \begin{cases} \max_{0 \le i < j \le k} d(\mathbf{v}_i, \mathbf{v}_j), & k > 0 \\ 0, & k = 0, \end{cases}$$

here interactive distance $d(\mathbf{v}_i, \mathbf{v}_j)$ between atom vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ is defined as [57]

$$d(\mathbf{v}_i, \mathbf{v}_j) = \begin{cases} \max_{\mathbf{v}_k \in V_P, \{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k\} \in \mathcal{H}_2} \{\|\mathbf{v}_i, \mathbf{v}_k\|, \|\mathbf{v}_j, \mathbf{v}_k\|\}, & \text{if } \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}_L \\ \max_{\mathbf{v}_k \in V_L, \{\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k\} \in \mathcal{H}_2} \{\|\mathbf{v}_i, \mathbf{v}_k\|, \|\mathbf{v}_j, \mathbf{v}_k\|\}, & \text{if } \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}_P \\ \|\mathbf{v}_i - \mathbf{v}_j\|, & \text{otherwise} \end{cases}$$

where $||\mathbf{v}_i - \mathbf{v}_j||$ is the Euclidean distance. In this way, a filtration process can be built on hypergraph $\mathcal{H}$, by consistently increasing the filtration values and adding in hyperedges with ‘birth time’ smaller or equal to the filtration value. A hypergraph-based filtration process can be found in Figure 2 (**a**).

From the filtration process, a sequence of nested hypergraphes can be generated as $\mathcal{H}^0 \subset \mathcal{H}^1 \subset \ldots \subset \mathcal{H}^N$ where hypergraph $\mathcal{H}^t$ is generated at the filtration value $t$. Special spectral hypergraph models can be built from these hypergraphs. Here we consider supremum HL matrix from the embedded-homology-based spectral hypergraph model. For hypergraph $\mathcal{H}^t$, a corresponding supremum HL matrix $(\mathbf{L}_k^{Sup})^t$ can be generated. In this way, a series of supremum HL matrixes $\{(\mathbf{L}_k^{Sup})^t | t = 0, 1, \ldots, N\}$ can be generated from these hypergraphes.

The computational cost is mainly from eigen decomposition of Hodge-Laplacian matrixes. We consider eigenvalue solver ‘numpy.linalg.eigvalsh’ from Python library Numpy. The complexity of the eigen decomposition algorithm is about $O(n^3)$ for an $n * n$ matrix. Moreover, the eigen-decomposition process has to be repeated for each filtration value. In this way, our computational cost is higher than persistent homology models [33–35]. However, since we only consider the binding core region, which is much smaller than the entire protein-ligand complex, our computational cost is still affordable. Efficient algorithms from eigen decomposition and matrix operation can be considered to reduce the computational cost, so that our models can be used for more complicated data.

### Key Points

Our main contributions in this paper are as follows:

- We propose the first embedded-homology-based spectral hypergraph and persistent spectral hypergraph (PSH) model.
- The PSH has been applied into the characterization of molecular structure and interaction at atomic level, for the first time.
- We develop PSH-ML models by using the persistent attributes from PSH as molecular descriptors/fingerprints and combining them with machine learning models, in particular, gradient boosting tree (GBT) model.
- Based on the three well-established datasets, including PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016, we test the performance of our PSH-GBT models for protein-ligand binding affinity prediction. Our PSH-GBT model can outperform all machine learning models with traditional molecular descriptors, as far as we know.

## Data and code availability

Data and code can be found from this link https://github.com/LiuXiangMath/Persistent-Spectral-Hypergraph.

## Author contributions statement

K.X. designed research; K.X., H.F., W.J. and X.L. performed research; K.X. and X.L. analyzed data; and K.X. and X.L. wrote the paper.

## Acknowledgments

This work was supported in part by Nanyang Technological University Startup Grant M4081842 and Singapore Ministry of Education Academic Research fund Tier 1 RG109/19, Tier 2 MOE2018-T2-1-033. The third author was supported by Natural Science Foundation of China (NSFC grant no. 11971144) and High-level Scientific Research Foundation of Hebei Province.

## References

1. Fleming N. Computer-calculated compounds. *Nature* 2018;**557**(7707):S55–7.
2. Liu J, Wang RX. Classification of current scoring functions. *J Chem Inf Model* 2015;**55**(3):475–82.
3. Li HJ, Leung KS, Wong MH, *et al*. Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular informatics* 2015;**34**(2–3):115–26.
4. Wójcikowski M, Kukiełka M, Stepniewska-Dziubinska MM, *et al*. Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* 2019;**35**(8):1334–41.
5. Jiménez J, Skalic M, Martinez-Rosell G, *et al*. $\mathrm{K_{DEEP}}$: Protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model* 2018;**58**(2):287–96.

6. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* 2018;**34**(21):3666–74.
7. Su MY, Yang QF, Du Y, *et al*. Comparative assessment of scoring functions: The CASF-2016 update. *J Chem Inf Model* 2018;**59**(2):895–913.
8. Afifi K, Al-Sadek AF. Improving classical scoring functions using random forest: The non-additivity of free energy terms' contributions in binding. *Chem Biol Drug Des* 2018;**92**(2):1429–34.
9. Feinberg EN, Sur D, Wu ZQ, *et al*. Potentialnet for molecular property prediction. *ACS central science* 2018;**4**(11): 1520–30.
10. Boyles F, Deane CM, Morris GM. Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics* 2020;**36**(3):758–64.
11. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J Chem Inf Model* 2013;**53**(7):1563–75.
12. Pereira JC, Caffarena ER, dos Santos CN. Boosting docking-based virtual screening with deep learning. *J Chem Inf Model* 2016;**56**(12):2495–506.
13. Kearnes S, McCloskey K, Berndl M, *et al*. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* 2016;**30**(8):595–608.
14. Gomes J, Ramsundar B, Feinberg EN, *et al*. Atomic convolutional networks for predicting protein-ligand binding affinity arXiv preprint arXiv:1703.10603. 2017.
15. Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 2019;**35**(2): 309–18.
16. Li X, Yan X, Gu Q, *et al*. Deepchemstable: chemical stability prediction with an attention-based graph convolution network. *J Chem Inf Model* 2019;**59**(3):1044–9.
17. Wang X, Li Z, Jiang M, *et al*. Molecule property prediction based on spatial graph embedding. *J Chem Inf Model* 2019;**59**(9):3817–28.
18. Puzyn T, Leszczynski J, Cronin MT. *Recent advances in QSAR studies: methods and applications*, Volume 8. In: *Springer Science & Business Media*, 2010.
19. Lo YC, Rensi SE, Torng W, *et al*. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 2018;**23**(8):1538–46.
20. Durant JL, Leland BA, Henry DR, *et al*. Reoptimization of mdl keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**(6):1273–80.
21. O'Boyle NM, Banck M, James CA, *et al*. Open Babel: An open chemical toolbox. *J Chem* 2011;**3**(1):1–14.
22. Hall LH, Kier LB. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci* 1995;**35**(6):1039–45.
23. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**(5):742–54.
24. Landrum G. *RDKit: Open-source cheminformatics*, 2006.
25. Stiefl N, Watson IA, Baumann K, *et al*. ErG: 2D pharmacophore descriptions for scaffold hopping. *J Chem Inf Model* 2006;**46**(1):208–20.
26. Merkwirth C, Lengauer T. Automatic generation of complementary descriptors with molecular graph networks. *J Chem Inf Model* 2005;**45**(5):1159–68.
27. Duvenaud DK, Maclaurin D, Iparraguirre J, *et al*. Convolutional networks on graphs for learning molecular fingerprints. In: *Advances in neural information processing systems*, 2015, 2224–32.
28. Coley CW, Barzilay R, Green WH, *et al*. Convolutional embedding of attributed molecular graphs for physical property prediction. *J Chem Inf Model* 2017;**57**(8):1757–72.
29. Xu Y, Pei J, Lai L. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *J Chem Inf Model* 2017;**57**(11):2672–85.
30. Winter R, Montanari F, Noé F, *et al*. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* 2019;**10**(6):1692–701.
31. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geom* 2002;**28**:511–33.
32. Zomorodian A, Carlsson G. Computing persistent homology. *Discrete Comput Geom* 2005;**33**:249–74.
33. Cang ZX, Wei GW. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput Biol* 2017;**13**(7): e1005690.
34. Nguyen DD, Cang ZX, Wei GW. A review of mathematical representations of biomolecular data. *Phys Chem Chem Phys* 2020;**22**:4343–67.
35. 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**(1):e1005929.
36. Cang ZX, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering* 2017. doi: 10.1002/cnm.2914.
37. Nguyen DD, Xiao T, Wang ML, *et al*. Rigidity strengthening: A mechanism for protein–ligand binding. *J Chem Inf Model* 2017;**57**(7):1715–21.
38. Cang ZX, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering* 2018;**34**(2):e2914.
39. 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**(7): 3291–304.
40. Cang ZX, Wei GW. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics* 2017;**33**(22):3549–57.
41. Wu KD, Wei GW. Quantitative toxicity prediction using topology based multi-task deep neural networks. *J Chem Inf Model* 2018. doi: 10.1021/acs.jcim.7b00558.
42. Wang B, Zhao ZX, Wei GW. Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies. *J Chem Phys* 2016;**145**(12):124110.
43. Wang B, Wang CZ, Wu KD, *et al*. Breaking the polar-nonpolar division in solvation free energy prediction. *J Comput Chem* 2018;**39**(4):217–33.
44. 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**(20):1444–54.
45. Zhao RD, Cang ZX, Tong YY, *et al*. Protein pocket detection via convex hull surface evolution and associated Reeb graph. *Bioinformatics* 2018;**34**(17):i830–7.

46. Grow C, Gao KF, Nguyen DD, *et al*. Generative network complex (GNC) for drug discovery. *Communications in Information and Systems* 2019;**19**(3):241–77.
47. 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* 2019;**33**(1): 71–82.
48. Nguyen DD, Gao KF, Wang ML, *et al*. MathDL: Mathematical deep learning for D3R Grand Challenge 4. *J Comput Aided Mol Des* 2020;**34**:131–47.
49. 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* 2019;**33**(1): 71–82.
50. Meng ZY, Xia KL. Persistent spectral based machine learning (PerSpect ML) for drug design arXiv preprint arXiv:2002.00582. 2020.
51. Chung F. *Spectral graph theory*. American Mathematical Society, 1997.
52. Spielman DA. Spectral graph theory and its applications. In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS07)*. IEEE, 2007, 29–38.
53. Eckmann B. Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici* 1944;**17**(1):240–55.
54. Muhammad A, Egerstedt M. Control using higher order Laplacians in network topologies. In: *Proc. of 17th International Symposium on Mathematical Theory of Networks and Systems*. Citeseer, 2006, 1024–38.
55. Horak D, Jost J. Spectra of combinatorial Laplace operators on simplicial complexes. *Advances in Mathematics* 2013;**244**:303–36.
56. Barbarossa S, Sardellitti S. Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing* 2020;**68**:2992–3007.
57. Liu X, J WX, Wu J, *et al*. Hypergraph based persistent cohomology (HPC) for molecular representations in drug design. *Briefings in Bioinformatics, accepted* 2020.
58. Feng KQ, Li WCW. Spectra of hypergraphs and applications. *Journal of number theory* 1996;**60**(1):1–22.
59. Sun L, Ji SW, Ye JP. Hypergraph spectral learning for multi-label classification. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, 668–76.
60. Cooper J, Dutle A. Spectra of uniform hypergraphs. *Linear Algebra and its applications* 2012;**436**(9):3268–92.
61. Lu LY, Peng X. High-ordered random walks and generalized Laplacians on hypergraphs. In: *International Workshop on Algorithms and Models for the Web-Graph*. Springer, 2011, 14–25.
62. Barbarossa S, Tsitsvero M. An introduction to hypergraph signal processing. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, 6425–9.
63. Banerjee A, Char A, Mondal B. Spectra of general hypergraphs. *Linear Algebra and its Applications* 2017;**518**:14–30.
64. Bressan S, Li JY, Ren SQ, *et al*. The embedded homology of hypergraphs and applications. *Asian Journal of Mathematics* 2019;**23**(3):479–500.
65. Parks AD, Lipscomb SL. Homology and hypergraph acyclicity: a combinatorial invariant for hypergraphs, tech. rep. In: *NAVAL SURFACE WARFARE CENTER DAHLGREN VA*, 1991.
66. Qi L, Luo Z. *Tensor analysis: spectral theory and special tensors*. SIAM, 2017.
67. Ren SQ, Wu CY, Wu J. Hodge decompositions for weighted hypergraphs. *arXiv preprint arXiv:180511331* 2018.
68. Ren SQ, Wu CY, Wu J. Evolutions of hypergraphs and their embedded homology arXiv preprint arXiv:1804.07132. 2018.
69. Ren SQ, Wu J. Stability of persistent homology for hypergraphs arXiv preprint arXiv:2002.02237. 2020.
70. Verri A, Uras C, Frosini P, *et al*. On the use of size functions for shape analysis. *Biol Cybern* 1993;**70**(2):99–107.
71. Bergomi MG, Ferri M, Vertechi P, *et al*. Beyond topological persistence: Starting from networks arXiv preprint arXiv:1901.08051. 2019.
72. Liu ZH, Y L, L H, *et al*. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 2015;**31**(3):405–12.
73. Mukherjee S, Steenbergen J. Random walks on simplicial complexes and harmonics. *Random structures & algorithms* 2016;**49**(2):379–405.
74. Parzanchevski O, Rosenthal R. Simplicial complexes: spectrum, homology and random walks. *Random Structures & Algorithms* 2017;**50**(2):225–61.
75. Shukla S, Yogeshwaran D. Spectral gap bounds for the simplicial Laplacian and an application to random complexes. *Journal of Combinatorial Theory, Series A* 2020;**169**: 105134.
76. Torres JJ, Bianconi G. Simplicial complexes: higher-order spectral dimension and dynamics arXiv preprint arXiv:2001.05934. 2020.
77. Grigor'yan A, Muranov Y, Yau ST. Graphs associated with simplicial complexes. *Homology, Homotopy and Applications* 2014;**16**(1):295–311.
78. Grigor A. yan, Y. Lin, Y. Muranov, and S. T. Yau, Cohomology of digraphs and (undirected) graphs. *Asian Journal of Mathematics* 2015;**19**(5):887–932.
79. Grigor'yan A, Jimenez R, Muranov Y, *et al*. On the path homology theory of digraphs and Eilenberg–Steenrod axioms. *Homology, Homotopy and Applications* 2018;**20**(2): 179–205.
80. Grigor'yan A, Jimenez R, Muranov Y, *et al*. Homology of path complexes and hypergraphs. *Topology and its Applications* 2019;**267**:106877.