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

https://doi.org/10.1093/bib/bbaa411
Problem Solving Protocol

Hypergraph-based persistent cohomology (HPC) for molecular representations in drug design

Xiang Liu, Xiangjun Wang, Jie Wu and Kelin Xia

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

Abstract

Artificial intelligence (AI) based drug design has demonstrated great potential to fundamentally change the pharmaceutical industries. Currently, a key issue in AI-based drug design is efficient transferable molecular descriptors or fingerprints. Here, we present hypergraph-based molecular topological representation, hypergraph-based (weighted) persistent cohomology (HPC/HWPC) and HPC/HWPC-based molecular fingerprints for machine learning models in drug design. Molecular structures and their atomic interactions are highly complicated and pose great challenges for efficient mathematical representations. We develop the first hypergraph-based topological framework to characterize detailed molecular structures and interactions at atomic level. Inspired by the elegant path complex model, hypergraph-based embedded homology and persistent homology have been proposed recently. Based on them, we construct HPC/HWPC, and use them to generate molecular descriptors for learning models in protein–ligand binding affinity prediction, one of the key step in drug design. Our models are tested on three most commonly-used databases, including PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016, and outperform all existing machine learning models with traditional molecular descriptors. Our HPC/HWPC models have demonstrated great potential in AI-based drug design.

Key words: molecular descriptor; machine learning; hypergraph-based persistent cohomology; drug design.

Introduction

Artificial intelligence (AI) based drug design has great potential to significantly change the landscape of pharmaceutical industries [1–6]. In fact, traditional drug design approaches are not only laborious and time-consuming, but also inefficient and high-cost. Currently, only 10%-14% of drug candidates, which manage to enter clinical trials, can reach the market as medicines [1]. It takes more than 10 years and costs about $2.6 billion to develop a new market-approval prescription medicine [2]. With the ever-increasing accumulation of chemical and biomolecular data, data-driven AI models will usher in an era of faster, cheaper and more-efficient drug design and drug discovery [2]. In fact, medical imaging analysis for radiology, pathology and other medical specialties, have already undergone revolutionary changes with deep learning [7, 8]. Moreover, AI techniques have gradually been applied to the whole drug design process, from target discovery, lead discovery, lead optimization, preclinical development, to the final three phases of clinical trials. Researchers and biopharmaceutical companies are leading the revolution of AI in drug design [1]. For instance, Obama's Cancer Moonshot initiative uses AI for personalization of treatment and early diagnosis. Currently, machine learning and deep learning models have delivered significant better results in molecular docking [9, 10], binding affinity prediction [11, 12], toxicity prediction [13], as well as various quantitative

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.
Prof. Xiangjun Wang is a full professor at Nankai University in China. His research interests include algebraic topology and topological data analysis.
Prof. Jie Wu is a full professor at Hebei normal University in China. His research interests include algebraic topology and topological data analysis. He has been awarded National Science Award at 2007.
Prof. 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: 27 October 2020; **Received (in revised form):** 2 December 2020

structure-activity relationship (QSAR) models [14, 15]. Further progresses from chemical data accumulation, access to more computational power, and development of highly efficient learning algorithms, will pave the way for AI-based drug design to fundamentally change the landscape of drug design and drug discovery [5, 6].

With the excitement and opportunities come challenges. Currently, one of the central challenges for machine learning models in drug design is molecular featurization, which is to identify or design appropriate molecular descriptors or fingerprints [16–19]. In fact, featurization is a long-standing issue for chemical informatics and bioinformatics [14, 15]. Traditional molecular/chemical descriptors are structural and physical properties obtained from structural geometry, chemical conformation, chemical graph, structure topology, as well as molecular formula, hydrophobicity, steric properties and electronic properties [14, 15]. These descriptors are widely used in QSAR and learning models. Recently, a series of mathematical models from algebraic topology, combinatorial topology and differential geometry, have been proposed for molecular representations [20–32]. Unlike traditional molecular descriptors [14, 15], these models use highly abstract fundamental mathematical invariants, thus they can capture deeper and more intrinsic molecular properties [24]. Featurization with higher level of abstraction and generalization has great advantages in machine learning and deep learning models [24]. Significant better results have been achieved using learning models with these advanced mathematical representations, for various aspects of drug design, including protein–ligand binding affinity prediction [20], protein stability changes upon mutation [24] and toxicity prediction [25].

Here, we present the first hypergraph-based molecular representation, hypergraph-based (weighted) persistent cohomology (HPC/HWPC), and HPC/HWPC-based machine learning models for drug design. Our HPC/HWPC models are developed from the recently-proposed hypergraph-based homology and persistent homology models [33–35], which are motivated by the elegant path complex models [36–39]. Mathematically, hypergraph provides a more generalized topological representation, compared with traditional graph and simplicial complex representations, which are widely-used in material, chemical and biological models. Further, molecular descriptors are obtained from HPC/HWPC models, and combined with gradient boosting tree (GBT) model. Our models are tested on three well-established databases, including PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. Our HPC/HWPC-based GBT model can outperform all learning models with traditional molecular descriptors, for protein–ligand binding affinity prediction.

Results

Biomolecular hypergraph representation

A proper characterization of biomolecular interactions between protein–protein, protein–ligand, protein-DNA/RNA and others, is of essential importance for drug design. Motivated by the success of the element-specific persistent homology models [24], we propose the first element-specific biomolecular hypergraph representation for biomolecular interactions at atomic level. The key idea is to describe element-specific atom-pair interactions as different kinds of hyperedges.

Here we consider the protein–ligand interactions. Since protein is usually much larger than ligand, only the binding core region is consider, which is made of all the atoms within a certain cutoff distance of ligand. The binding core region is decomposed into 36 types of atom-pair combinations, made from four protein atom-sets and nine ligand atom-sets (see Materials and methods). The atom-set from protein and ligand is denoted 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 vector, and N_P and N_L the respective total numbers. An element-specific hypergraph $(V_{\mathcal{H}}, \mathcal{H})$ is composed of vertex set $V_{\mathcal{H}} = V_P \cup V_L$ and hyperedge set $\mathcal{H}$. Since a vertex can be viewed as a 0-hyperedge, we use $\mathcal{H}$ to denote hypergraph for simplicity. In our protein–ligand based hypergraph, we define an n-hyperedge in $\mathcal{H}$ as

$$\sigma_n = \begin{cases} \{\mathbf{v}_0,\mathbf{v}_1,...,\mathbf{v}_n\};\mathbf{v}_k\in V_{\mathcal{H}}(0\leqslant k\leqslant n),\exists i,j\in[0,n],\mathbf{v}_i\in V_P,\mathbf{v}_j\in V_L, & n>0 \\ \{\mathbf{v}_0\};\mathbf{v}_0\in V_{\mathcal{H}}, & n=0. \end{cases} \qquad (1)$$

In our model, an n-hyperedge σ_n is composed of $n + 1$ vertices (or atoms) from either protein atom-set or ligand atom-set, with one condition that, when $n > 0$, at least two vertices of the n-hyperedge σ_n are not from the same molecule, i.e., one from protein and the other from ligand. Note that any vertex in $V_{\mathcal{H}}$ is a 0-hyperedge σ_0. All these hyperedges form the hypergraph $\mathcal{H}$. Figure 1 **A** shows a hypergraph-based representation for a protein-ligand complex (PDBID 3P2E). Only a small region of binding core part (ligand and protein region within 5.0 Å of ligand) is considered. A total 36 element-specific hypergraphs are constructed from the 36 atom combinations. The hyperedges are denoted as ellipses, with 1-hyperedges in red and 2-hyperedges in blue. The 36 types of hypergraphs provide a detailed representation of protein-ligand interactivity at atomic level.

Hypergraph-based persistent cohomology

The key component of persistent models, including persistent homology/cohomology [40–42], persistent spectral [43] and persistent function [44], is the filtration process. For any system, a multiscale representation can be generated through a filtration process. In our hypergraph-based persistent cohomology, a filtration value (or "birth time") is assigned to each hyperedge, so that with the increase (or decrease) of filtration value, a series of nested hypergraphs can be generated. Note that "birth time" and "death time" are used in persistent barcodes, as illustrated in Figure 2, to represent the starting and ending value of a barcode.

In our protein–ligand complex based hypergraph model, an interactive distance between two atoms $\mathbf{v}_i$ and $\mathbf{v}_j$ is defined as

$$d(\mathbf{v}_i, \mathbf{v}_j) = \begin{cases} \|\mathbf{v}_i-\mathbf{v}_j\|, & \text{if } \mathbf{v}_i\in \mathbf{V}_P,\mathbf{v}_j\in \mathbf{V}_L \text{ or } \mathbf{v}_i\in \mathbf{V}_L,\mathbf{v}_j\in \mathbf{V}_P \\ g(\mathbf{v}_i,\mathbf{v}_j), & \text{otherwise.} \end{cases} \qquad (2)$$

Here $\|\mathbf{v}_i - \mathbf{v}_j\|$ is the Euclidean distance between the two atoms. Note that $\mathbf{v}_i$ and $\mathbf{v}_j$ are coordinate vectors. Function $g(\mathbf{v}_i, \mathbf{v}_j)$ is the interactive distance between atoms $\mathbf{v}_i$ and $\mathbf{v}_j$ from the same molecule, i.e., both from protein or both from ligand. Its detailed setting can be found in Eq.(5) (see Materials and methods). With the interactive distance, the filtration value for a hyperedge

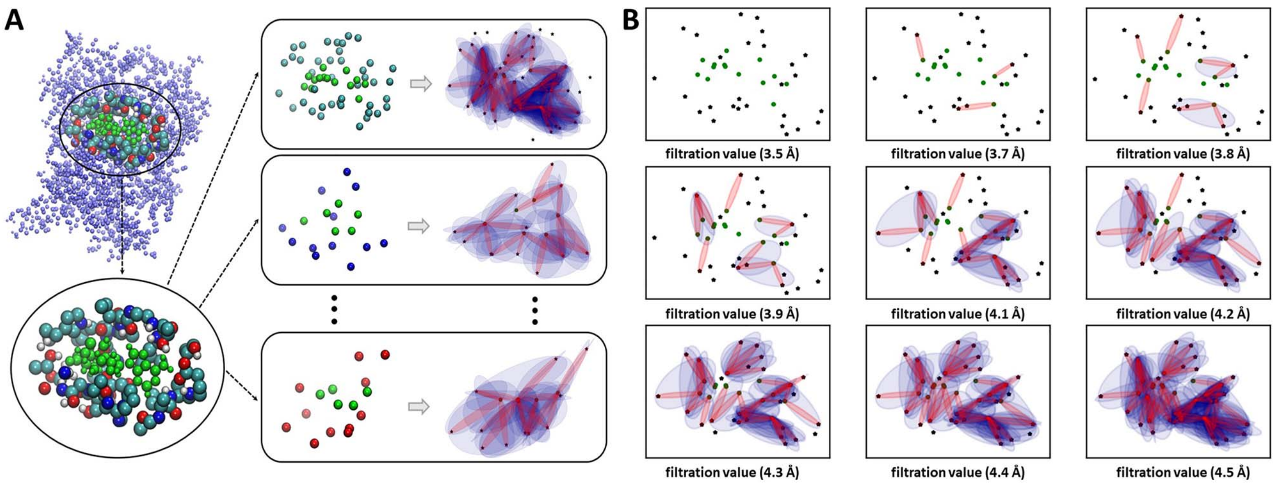

Figure 1. **A.** Illustration of the element-specific hypergraph representation for protein–ligand interactions, using a protein–ligand complex (PDBID 3P2E). The core region is decomposed into four protein atom-sets (C, N, O, S) and nine ligand atom-sets (C, N, O, S, P, F, Cl, Br, I), which are combined to form 36 element-specific combinations, such as C-C, N-S, O-Br, etc. A hypergraph is generated from each element-specific combination and a total 36 hypergraphs are used for the representation of protein–ligand interactions at molecular level. We only show hyperedges with dimensions 0, 1 and 2. A 0-hyperedge is a black vertex (for protein atom) or green vertex (for ligand atom), a 1-hyperedge is represented as a red ellipse and a 2-hyperedge is denoted as a blue ellipse. Note that each ellipse has at least one black vertex and one blue vertex, which means each n-hyperedge ($n > 0$) has at least one vertex from protein and one vertex from ligand. **B.** Illustration of the hypergraph-based filtration process. We consider the hypergraphs from the C-C combination of protein–ligand complex (PDBID 3P2E) as in (**A**). Each hyperedge is associated with a filtration value or 'birth time'. A series of nested hypergraphs are generated during the filtration process.

$\sigma_n = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_n\}$ is defined as,

$$f(\sigma_n) = \begin{cases} \max_{0\le i<j\le n} d(\mathbf{v}_i, \mathbf{v}_j), & n > 0 \\ 0, & n = 0. \end{cases} \tag{3}$$

The hypergraph with the above filtration values naturally generates a well-defined filtration process. Figure 1 **B** demonstrates a hypergraph based filtration process for protein–ligand C-C combination (PDBID 3P2E). It can be seen that as the increase of filtration value, a series of nested hypergraphs can be generated.

Mathematically, the definition of hypergraph homology is nontrivial. Different hypergraph homology definitions have been considered [45–48]. Motivated by the elegant definition of path complex [36–39], embedded homology has been proposed for hypergraphs recently [33]. Different from previous models, embedded homology is found to be consistent for both infimum chain complex and supremum chain complex derived from hypergraph (see Materials and methods). Here we consider embedded homology for our protein–ligand complex based hypergraph models. With the filtration parameter defined in Eq. (3), hypergraph persistent homology can be derived and persistent barcodes [49] can be generated. For simplicity, we omit the word 'embeded', and call hypergraph based embedded homology, persistent embedded homology and persistent embedded cohomology (will be discussed later), as hypergraph homology, persistent homology and persistent cohomology, respectively. Figure 2 **A**, **B** and **C** demonstrate the persistent barcodes of bipartite-graph-based persistent homology (**A**, see Materials and methods) and hypergraph-based persistent homology (**B**, **C**). For bipartite-graph-based persistent homology, the lack of simplices with dimension larger than 1, results in the forever-persisting β_1 barcodes and absence of β_2 barcodes. The introduction of high dimensional hyperedges up to dimension 2 (**B**) and 3 (**C**) has recovered the missing higher dimensional topological information.

Further, we propose hypergraph-based persistent cohomology (HPC) and hypergraph-based weighted persistent cohomology (HWPC) (see Materials and methods). Our HPC model is a generalization of persistent embedded homology to its cohomology counterpart. More importantly, our HWPC can incorporate geometrical, physical, chemical and biological information into HPC model through the use of weights. Mathematically, weights can be defined on hyperedges, hypergraph boundary operators and hypergraph cohomology generators. They can be chosen as atomic types, atomic partial charge, hydrophobic and hydrophilic properties, and other physical, chemical or biological properties. Persistent cohomology enriched barcodes [32, 50] with the incorporated information are generated from HWPC models.

In our protein–ligand complex based HWPC, we consider a new type of weight on hyperedges, that is graph centrality (see Eqs. (6) and (7) in Materials and methods). A scale parameter η controls the "influence range" for the graph centrality. A smaller η value means that each vertex only interacts with vertices within its local region, thus it results in a smaller centrality value. Although a larger η value means global interactions for each nodes and results in a larger centrality value. Further, weights can be defined on cohomology generators and enriched barcodes are obtained (see Eqs. (8) and (9) in Materials and methods). Figure 2 **D** shows two persistent cohomology enriched barcodes for the C-C combination (from complex 3IP5) with two different η values, i.e., 2.5 Å (left-side subgraph) and 10.0 Å (right-side subgraph). We linearly normalize the weight values on the enriched barcode into [0, 1]. The colors represent the values on each enriched barcode. Note that larger graph centrality values result in larger values on enriched barcodes.

HPC-based machine learning

Persistent barcodes from HPC and HWPC can be discretized into feature vectors. Many methods have been proposed [24], including barcode statics, algebraic functions and tropical functions,

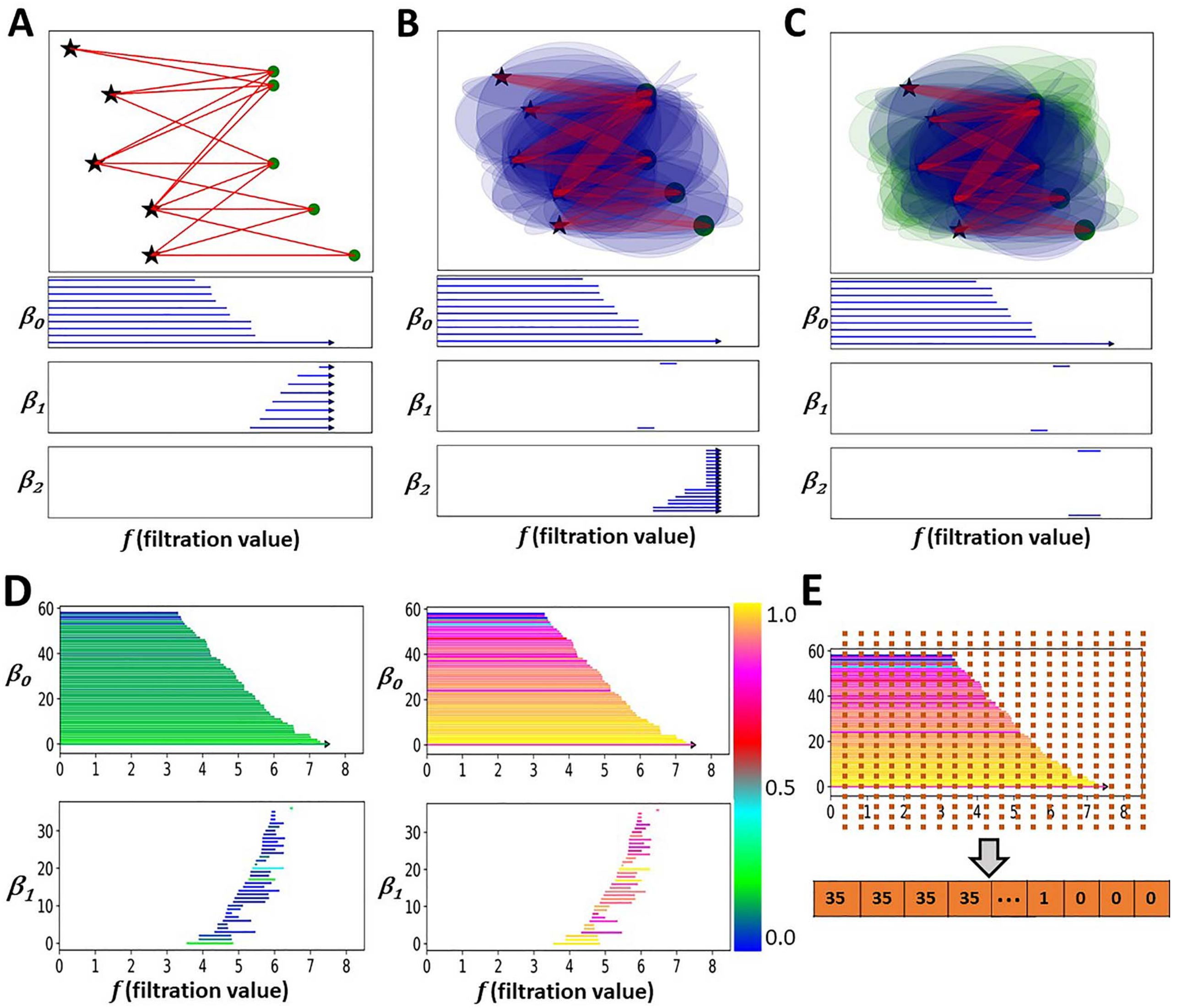

Figure 2. The comparison of persistent barcodes from bipartite graph (**A**), HPC with hyperedges up to dimension 2 (**B**), HPC with hyperedges up to dimension 3 (**C**). The bipartite graph is derived from interactive distance matrix in Eq. (4) with protein atoms represented as black stars and ligand atoms represented as green dots. Its higher dimensional topological information (β_2) is missing and the β_1 barcodes are forever-persistent because of the lack of simplices with dimension higher than 1. In our HPC, by adding 2-hyperedges (represented as blue ellipses, each blue ellipse contains only three atoms), we can 'kill' all the forever-persistent β_1 barcodes, i.e., assign them with 'death' times, and capture β_2 information. However, if we only add hyperedges with dimension lower than three, we will get forever-persistent β_2 barcodes as in (**B**). By adding 3-hyperedges (represented as green ellipses, each green ellipse contains only four atoms), we can 'kill' these forever-persistent β_2 barcodes as in (**C**). **D** Illustration of the HWPC-based enriched barcodes for C-C pair of PDBID 3IP5. **E** Illustration of the molecular descriptors from the discretization of the persistent barcodes.

binning approaches, persistent codebook, persistent paths and siganture features, and 2D/3D representations. Here we consider the binning approach [24]. As illustrated in Figure 2 **E**, the filtration region is discretized into equal-sized bins. The total number of the barcodes (i.e., Betti numbers for HPC) or the sum of the weight values of enriched barcodes (for HWPC) within each bin, is used as molecular descriptors. A large-sized molecular fingerprint is usually obtained in our HPC-based machine learning models.

The use of a systematically-generated large-sized molecular descriptors/fingerprints is found to be more efficient for learning models in chemical and biological data analysis [59]. Essentially, large-sized feature vector can have a better characterization of molecular structures and interactions and facilitate a better transferability for machine learning models. Decision-tree-based models, such as random forest and GBT are usually considered in large-sized fingerprint cases, as they are more robust against overfitting problem.

HPC-ML for protein ligand binding affinity prediction

A drug design process covers various steps from target discovery, lead discovery, lead optimization, preclinical development and three phases of clinical trials. Among these steps, one of the key issues is to identify ligands (drugs) that have higher binding affinity with the target biomolecules. During the past few decades, a variety of empirical, physics-based, knowledge-based and machine-learning-based models are proposed [10, 11, 24]. The databank PDBbind (www.pdbbind.org.cn) is established to systematically evaluate and compare their performance for

Table 1. The PCCs and RMSEs (pK_d/pK_i) for our HPC/HWPC-GBT model in three test cases, i.e., PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016

Dataset	HPC	HPC-HWPC(η_1)	HPC-HWPC(η_2)	HPC-HWPCs(η_1 & η_2)
PDB-2007	0.813(1.423)	0.823(1.418)	0.827(1.395)	0.829(1.403)
PDB-2013	0.770(1.508)	0.780(1.498)	0.779(1.486)	0.784(1.483)
PDB-2016	0.810(1.359)	0.825(1.322)	0.825(1.324)	0.831(1.307)

Table 2. Detailed information of the three PDBbind databases, i.e., PDB-v2007, PDB-v2013 and PDB-v2016.

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

Table 3. The parameters for our GBT model

No. of estimators	Learning rate	Max depth	Subsample
40 000	0.001	9	0.7
Min_samples_split	Loss function	Max features	Repetitions
2	Least square	SQRT	10

protein–ligand binding affinity prediction [55]. Three of the most commonly-used datasets are PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. For each dataset, the core set is regarded as the test set, all entries in refined set except the ones in the core set form the training set. The detailed data information can be found in Table (2).

In HPC/HWPC based molecular descriptor model, we use 10.5 Å as the cut-off distance to extract the binding core region. The filtration range is chosen as [2.0 Å, 7.5Å] and bin size as 0.1 Å. We consider both β_0 and β_1 for HPC and two HWPC models, i.e., one for local interactions with scale parameter $\eta_1 = 2.5$ Å and the other for global interactions with $\eta_2 = 10.0$ Å. In this way, a feature vector of size 3960=36(combinational types)× 55(bin size)× 2(β_0 and β_1) is generated for each protein–ligand complex. In our HPC/HPWC models, we consider four types of featurizations with molecular descriptors from only HPC (vector size 3960), HPC and local HWPC with $\eta_1 = 2.5$ Å (vector size 7920), HPC and global HWPC with $\eta_2 = 10.0$ Å (vector size 7920), HPC and multiscale HWPCs with both η_1 and η_2 (vector size 11 880), respectively. With large-sized molecular fingerprints, we make use of GBT model to alleviate overfitting problem. The detailed setting of GBT parameters are presented at Table 3. The Pearson correlation coefficients (PCCs) and root mean square error (RMSEs), between predicted binding affinities and experimental ones for the three test sets, are calculated and listed in Table 1. Note that the unit for RMSE is pK_d/pK_i, instead of Kcal/mol. To have a better understanding of the performance of our models, we compare our models with traditional molecular descriptor based learning models[12, 51–58]. The PCCs result are illustrated in Figure 3. Note that 10 independent repetitions are conducted and the medians of the 10 PCCs and RMSEs are used as the performance measurements of our HPC/HWPC-GBT model. It can be seen that our model can achieve state-of-the-art results, and has a better performance than traditional molecular descriptor based machine learning models, for protein–ligand binding affinity prediction.

Discussion

The representability of molecular descriptors or fingerprints is of essential importance for machine learning models in material, chemical and biological data analysis. Mathematical invariants from algebraic topology and differential geometry provide a highly effective way of structure representation, as they can characterize the intrinsic information. Moreover, their persistent formulations, including persistent homology/cohomology, persistent spectral and persistent functions, can preserve intrinsic information at various different scales, i.e., a multiscale intrinsic representation. Molecular descriptors from these persistent models can have a much better performance in machine learning models.

The generalization of simplicial complex into hypergraph has provided more flexibility in the topological representation of molecular structures and interactions. Other than embedded homology and persistent homology/cohomology formula for hypergraph, discrete Morse theory [34] and Hodge decomposition model [35] have also been developed on hypergraph. Similarly, molecular descriptors can be derived from these hypergraph-based models and persistent-hypergraph-based models. The incorporation of these intrinsic molecular descriptors into machine learning models will significantly boost the learning performance in the analysis of molecular data from materials, chemistry and biology.

Materials and methods

Biomolecular topological representations

Traditionally, biomolecular structures and interactions are usually modeled as graphs or networks. Graphs are widely used in atom-covalent-bond representation, Gaussian network model, anisotropic network model, protein–protein interaction networks, among others. Recently, simplicial complex based biomolecular representation is proposed and used in the study of biomolecular structure, flexibility, dynamics, function and drug design [20–32]. Mathematically, simplicial complex is a generalization of graph, which contains only 0-simplexes (nodes) and 1-simplexes (edges). Simplicial complex is made of higher dimensional simplexes, such as 2-simplex (triangle), 3-simplex (tetrahedron), etc. In this way, simplicial complex can model more complicated relations between not only two atoms, but also groups of atoms represented as simplexes.

Hypergraph is a further generalization of simplicial complex. Simply speaking, hypergraph is composed of hyperedges and each hyperedge is a set of atoms. Figure 4 illustrates the different topological representations, including graph, simplicial complex and hypergraph for a ligand Uracil. It can be seen that graph only considers interactions between two atoms, i.e., they either interact with each other and an edge is form between them or they do not interact with each other and no edge is form. Simplicial complex can characterize more complicated relationships between simplexes, including upper adjacent, lower adjacent,

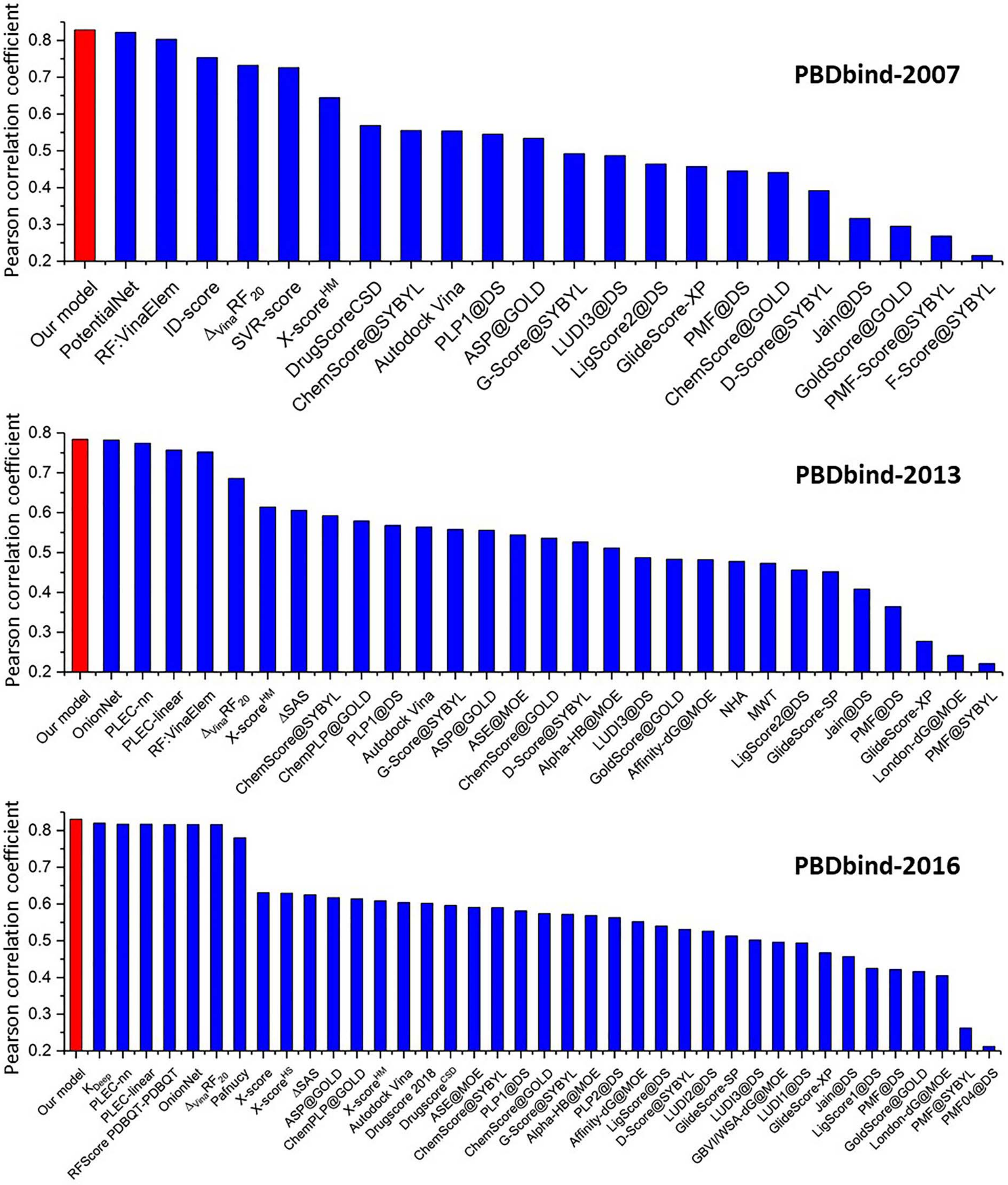

Figure 3. The comparison of PCCs between our combined HPC/HWPC-GBT model and traditional molecular descriptor based models[12, 51–58], for the prediction of protein-ligand binding affinity. The PCCs are calculated based on the core set (test set) of PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016.

parallel neighbour, among others. Since each simplex can be regarded as a hyperedge, any simplicial complex is just a special type of hypergraph and same relationships characterized by simplicial complex can be described using hypergraph representation. Moreover, hypergraph does not require the completeness of simplexes under boundary operator [33], thus it can characterize the most general relations.

Element-specific molecular interaction representation

The characterization and representation of molecular interactions at molecular level is of great importance for molecular structure, flexibility, dynamics and function analysis. Recently, element-specific interaction model has been developed for protein–ligand interaction analysis [10, 20, 24]. In it, a protein is decomposed into four individual atom-sets of C, N, O and S,

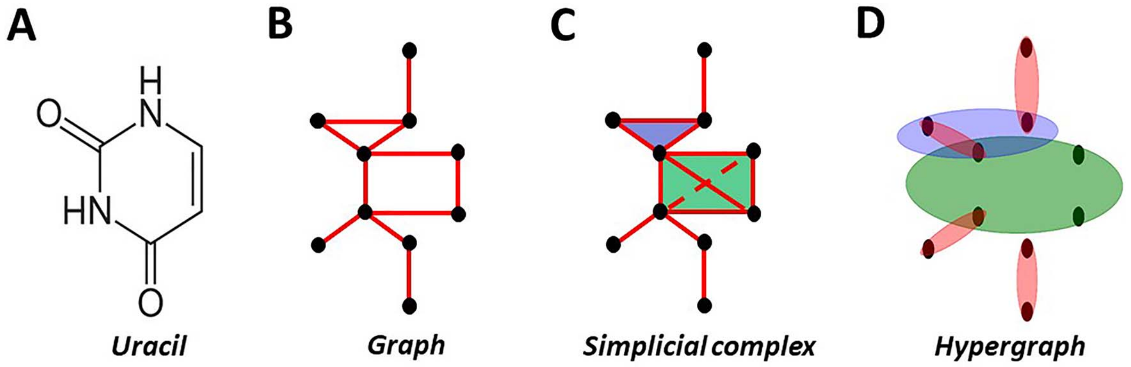

Figure 4. Three topological representations, i.e., graph (B), simplcial complex (C) and hypergraph (D) for ligand Uracil (A). Note that hypergraph is composed of hyperedge and each hyperedge is just a set of vertices. Hypergraph provides the most general topological representation.

respectively. A ligand is decomposed into nine atom-sets of C, N, O, S, P, F, Cl, Br and I, respectively. Molecular interactions are characterized by 36 different atom-pair combinations, between protein atom-sets and ligand atom-sets. For instance, we can take C atom-set from both protein and ligand to form the C-C combination set. Further, the connection topology within each combination set can be described by an interaction matrix [20, 24] as follows,

$$M(m_i, m_j) = \begin{cases} \|\mathbf{v}_i - \mathbf{v}_j\|, & \text{if } \mathbf{v}_i \in \mathbf{V}_P, \mathbf{v}_j \in \mathbf{V}_L \text{ or } \mathbf{v}_i \in \mathbf{V}_L, \mathbf{v}_j \in \mathbf{V}_P \\ \infty, & \text{otherwise.} \end{cases} \tag{4}$$

Here $\mathbf{v}_i$ and $\mathbf{v}_j$ are coordinates for the i- and j-th atoms, and $\|\mathbf{v}_i - \mathbf{v}_j\|$ is their Euclidean distance. Notations m_i and m_j are the indexes of i- and j-th atoms in matrix M, respectively. Two sets V_P and V_L are composed of the respective protein and ligand atom coordinates. Only connections (or interactions) between protein atoms and ligand atoms are considered. Connections between atoms within either protein or ligand are ignored by setting the distance as ∞. Based on the interaction matrix in Eq. (4), Vietoris–Rips complexes can be generated by using Euclidean distance as filtration parameter. These Vietoris–Rips complexes do not contain n-simplexes with ($n > 1$). This is due to the reason that, for any $n+1(n > 1)$ atoms taken from the combination set, at least two atoms of them will come from the same molecule, either protein or ligand. The distance (or filtration value) of the two atoms is ∞, thus they can never 'connect' with each other and no n-complex will be generated.

Mathematically, Vietoris–Rips complexes from the above interaction matrix are just bipartite graphs as illustrated in Figure 2 **A**. Higher dimensional homology information, such as β_2, is not captured in the above model. To recover these information, we propose hypergraph-based molecular representation.

Hypergraph-based molecular interaction representation

Mathematically, a hypergraph is a pair $(V_{\mathcal{H}}, \mathcal{H})$. Here $V_{\mathcal{H}}$ is a set of vertices and $\mathcal{H}$ is a subset of power set $\Delta[V_{\mathcal{H}}]$, which is the collection of all the nonempty subsets of $V_{\mathcal{H}}$. In our model, hyperedges are defined among protein and ligand atoms as Eq. (1). Essentially, a hyperedge has to contain at least one atom from protein and another atom from ligand, except all 0-hyperedges.

The "length" of hyperedge is defined as interactive distance in Eq. (2), and can be used as the filtration parameter. In our protein–ligand complex based hypergraph model, function $g(\mathbf{v}_i, \mathbf{v}_j)$, between two atoms $\mathbf{v}_i$ and $\mathbf{v}_j$ from the same molecule, is defined as follows,

$$g(\mathbf{v}_i,\mathbf{v}_j) = \begin{cases} \max_{\mathbf{v}_k \in V_P, \{\mathbf{v}_i,\mathbf{v}_j,\mathbf{v}_k\} \in \mathcal{H}} \{\|\mathbf{v}_i,\mathbf{v}_k\|, \|\mathbf{v}_j,\mathbf{v}_k\|\} + d_0, & \text{if } \mathbf{v}_i,\mathbf{v}_j \in \mathbf{V}_L \\ \max_{\mathbf{v}_k \in V_L, \{\mathbf{v}_i,\mathbf{v}_j,\mathbf{v}_k\} \in \mathcal{H}} \{\|\mathbf{v}_i,\mathbf{v}_k\|, \|\mathbf{v}_j,\mathbf{v}_k\|\} + d_0, & \text{if } \mathbf{v}_i,\mathbf{v}_j \in \mathbf{V}_P \end{cases} \tag{5}$$

here $d_0 \geq 0$ is a constant value. Note that even though we assign an interactive distance (or 'length') between two atoms from the same molecule, these two atoms can never form a hyperedge in our protein–ligand complex based hypergraph model.

However, filtration values (or "birth time") of n-hyperedges ($n > 1$) in Eq. (3) rely on these interactive distances. More specifically, the filtration value of an n-hyperedge ($n > 1$) is the largest interactive distances between any two atoms (vertices) within the n-hyperedge. If the two atoms are from different molecules, their interactive distance is just their Euclidean distance. If the two atoms are from the same molecule, their interactive distance is the largest "length" of all the possible 1-hyperedges that contains one of the two atoms, plus constant d_0. In this way, interactive function $g(\mathbf{v}_i, \mathbf{v}_j)$ directly determines filtration values of n-hyperedges ($n > 1$) that contain atoms $\mathbf{v}_i$ and $\mathbf{v}_j$ (from the same molecule). For instance, a set of 2-hyperedges $\{\{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_i\} | i \notin \{0, 1\}, \mathbf{v}_0 \in \mathbf{V}_P, \mathbf{v}_1 \in \mathbf{V}_P, \mathbf{v}_i \in V_{\mathcal{H}}\}$ contains two common atoms from protein, all these 2-hyperedges share the same filtration value $g(\mathbf{v}_0, \mathbf{v}_1)$.

Further, the constant d_0 can be used to adjust the filtration value of a hyperedge. To guarantee a well-defined filtration process, i.e., hypergraphs generated at later-stage of filtration have to include the ones produced at early-stage of filtration, constant d_0 has to be nonnegative. As illustrated in Figure 5, different d_0 values result in different persistent barcodes. In general, a larger d_0 value means a larger filtration value for all 2-hyperedges, thus β_1 generators will be 'killed' at a much later stage of filtration, compared with smaller d_0 situations. With the increase of d_0 value, it can be seen that β_1 barcodes are systematically elongated, and some new β_1 barcodes are generated.

Hypergraph-based persistent cohomology

For a hypergraph $\mathcal{H}$, its associated simplicial complex $K_{\mathcal{H}}$ is defined as the smallest simplicial complex such that the hyperedges of $\mathcal{H}$ is a subset of the simplices of $K_{\mathcal{H}}$ [33, 45]. The orientation of a hypergraph can be induced from its associated simplicial complex, i.e., the orientation of a hyperedge is the same as its associated simplex [33, 45]. Let G be an Abelian group

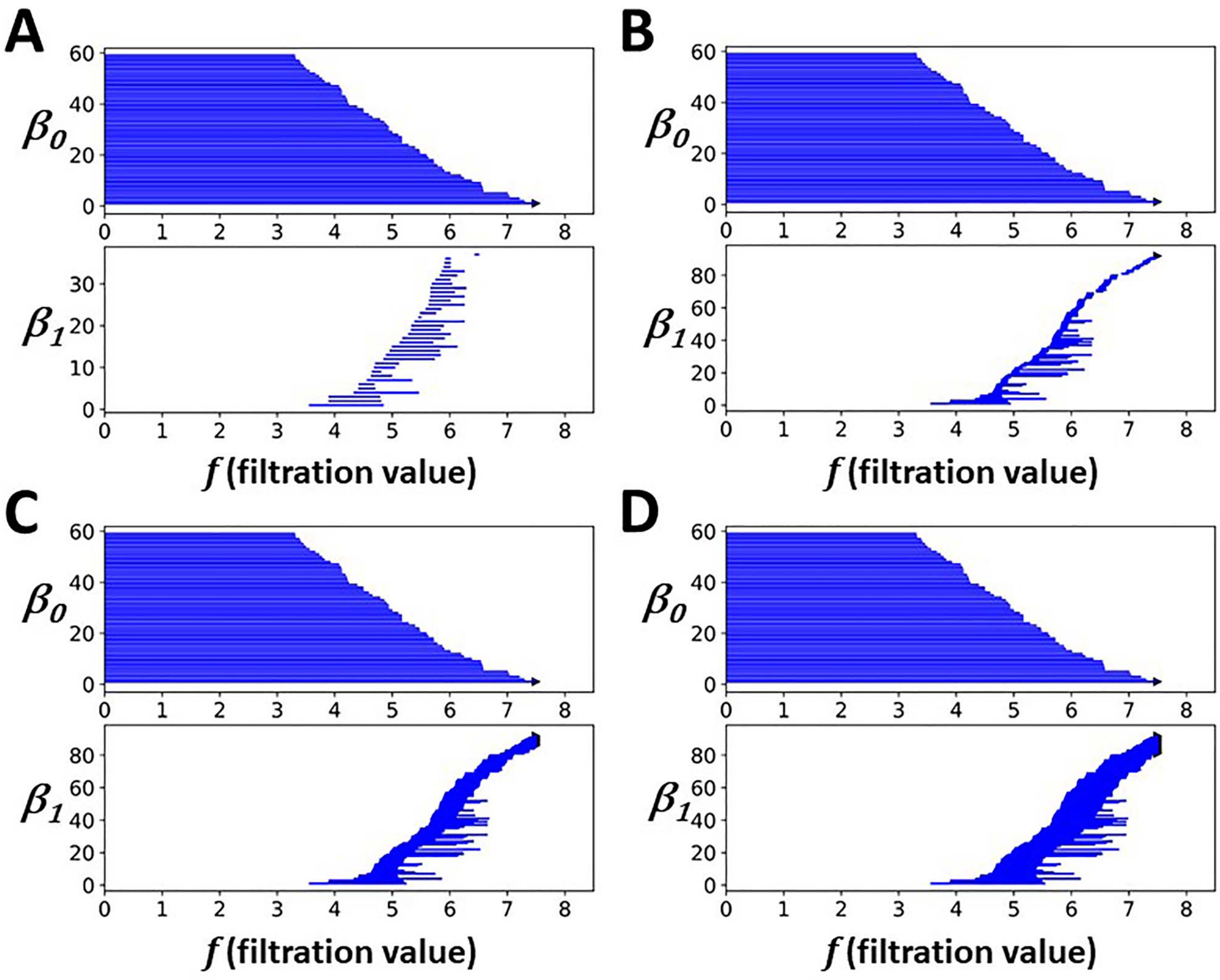

Figure 5. HPC-based persistent barcodes for C-C combination set of PDBID 3IP5 with different d_0 values. From (**A**) to (**D**), d_0 values are 0Å, 0.1Å, 0.4Å and 0.7Å, respectively. It can be seen that with the increase of d_0 value, more β_1 bars are generated and the lengthes of β_1 bars get longer.

and S a nonempty finite set, we use G(S) to denote the collection of linear combinations of the elements in S with coefficients in G. The n-hyperedge group is denoted as $G(\mathcal{H}_n)$ with $\mathcal{H}_n$ the set of all n-hyperedges in $\mathcal{H}$.

Let the vertex set of a hypergraph $\mathcal{H}$ be totally ordered. Following the ideas of simplicial homology, one may want to define the boundary operator ∂_n for n-hypergraph group $G(\mathcal{H}_n)$ as $\partial_n(\sigma_n) = \sum_{i=0}^{n}(-1)^i\sigma_{n-1}^i$ Here $\sigma_n = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_n\}$ is an n-hyperedge, and $(n-1)$-hyperedge $\sigma_{n-1}^i = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, ..., \mathbf{v}_n\}$ is generated by removing the vertex $\mathbf{v}_i$ from hyperedge σ_n. However, there is a serious problem in such a setting that the $(n-1)$-hyperedge σ_{n-1}^i may not always exist in $\mathcal{H}_{n-1}$. With these considerations, various different homology definitions have been proposed for hypergraphs [45–48].

Recently, inspired by the elegant path complex [36–39], embedded homology and persistent embedded homology have been proposed [33]. In these models, an n-th infimum chain group

$$\mathrm{Inf}_n(\mathcal{H}) = \mathrm{Inf}_n(G(\mathcal{H}_\star), G((K_{\mathcal{H}})_\star)) = G(\mathcal{H}_n) \cap \partial_n^{-1}(G(\mathcal{H}_{n-1})),$$

and an n-th supremum chain group

$$\mathrm{Sup}_n(\mathcal{H}) = \mathrm{Sup}_n(G(\mathcal{H}_\star), G((K_{\mathcal{H}})_\star)) = G(\mathcal{H}_n) + \partial_{n+1}(G(\mathcal{H}_{n+1})),$$

are considered. Note that $G(\mathcal{H}_\star) = \{G(\mathcal{H}_0), G(\mathcal{H}_1), G(\mathcal{H}_2)...\}$ is a sequence of hyperedge groups, and $G((K_{\mathcal{H}})_\star) = \{G((K_{\mathcal{H}})_0), G((K_{\mathcal{H}})_1), G((K_{\mathcal{H}})_2), ...\}$ is a sequence of chain groups from the associated simplicial complex. Since $G(\mathcal{H}_\star) \subseteq G((K_{\mathcal{H}})_\star)$, the above boundary operators are well-defined in $G((K_{\mathcal{H}})_\star)$. More importantly, it has been proved that the homology of $\mathrm{Inf}_\star(\mathcal{H})$ and $\mathrm{Sup}_\star(\mathcal{H})$ are isomorphic [33]. In this way, the embedded homology of hypergraph $\mathcal{H}$ can be defined as,

$$H_n(\mathcal{H}) = H_n(\mathrm{Inf}_\star(\mathcal{H})) = H_n(\mathrm{Sup}_\star(\mathcal{H})).$$

Similarly, persistent (embedded) homology can be generated for a hypergraph-based filtration process.

Here we propose (embedded) cohomology for hypergraph $\mathcal{H}$. Similar to the above process, we can define infimum cochain complexes $\mathrm{Inf}^\star(\mathcal{H})$ and supremum cochain complexes $\mathrm{Sup}^\star(\mathcal{H})$, and prove that the cohomology of $\mathrm{Inf}^\star(\mathcal{H})$ and $\mathrm{Sup}^\star(\mathcal{H})$ are isomorphic (see SI). The embedded cohomology of hypergraph $\mathcal{H}$

can be defined as,

$$H^n(\mathcal{H}) = H^n(\mathrm{Inf}^*(\mathcal{H})) = H^n(\mathrm{Sup}^*(\mathcal{H})).$$

More detailed provement process can be found in the SI.

Computationally, the embedded cohomology is calculated based on the supremum cochain complexes $\mathrm{Sup}_\star(\mathcal{H})$. This is because in our models, the supremum cochain complexes are equal to the corresponding cochain complexes of its associated simplicial complex. The hypergraph-based embedded cohomology is isomorphic to the simplicial cohomology of its associated simplicial complex. In this way, we only need to generate the associated simplicial complex and calculate its persistent cohomology information. We consider only β_0 and β_1 information for our protein–ligand complex based hypergraphs. This process is computationally much easier, as associated simplicial complex is just the clique complex of the hypergraph (See SI).

Hypergraph-based weighted persistent cohomology

Weighted persistent cohomology has been proposed to incorporate more structure, physical, chemical and biological information into a unified representation, i.e., persistent cohomology enriched barcode [32, 50]. Different from all previous models, here we consider a new weight scheme, derived from graph centrality and flexibility-rigidity indexes, for persistent cohomology. More specifically, we define the weights for a 0-hyperedge $\sigma_0 = \{\mathbf{v}_i\}$ as,

$$w(\sigma_0) = \begin{cases} \sum_{\mathbf{v}_k \in V_L} e^{-\frac{\|\mathbf{v}_k - \mathbf{v}_i\|^2}{\eta^2}}, & \mathbf{v}_i \in V_P \\ \sum_{\mathbf{v}_k \in V_P} e^{-\frac{\|\mathbf{v}_k - \mathbf{v}_i\|^2}{\eta^2}}, & \mathbf{v}_i \in V_L \end{cases} \quad (6)$$

and an 1-hyperedge $\sigma_1 = \{\mathbf{v}_i, \mathbf{v}_j\}$ as,

$$w(\sigma_1) = e^{-\frac{\|\mathbf{v}_i - \mathbf{v}_j\|^2}{\eta^2}}. \quad (7)$$

The scale parameter η controls the influence range of the nodes. Smaller η values mean local interactions and larger η values mean global interactions.

Further, we define the weight for a 0-cohomology generator δ^0 as,

$$w(\delta^0) = \frac{\sum_{\sigma_0^i \in \mathcal{H}_0} \delta^0(\sigma_0^i) * w(\sigma_0^i)}{\sum_{\sigma_0^i \in \mathcal{H}_0} \delta^0(\sigma_0^i)}, \quad (8)$$

and an 1-cohomology generator δ^1 as,

$$w(\delta^1) = \frac{\sum_{\sigma_1^i \in \mathcal{H}_1} \delta^1(\sigma_1^i) * w(\sigma_1^i)}{\sum_{\sigma_1^i \in \mathcal{H}_1} \delta^1(\sigma_1^i)}. \quad (9)$$

Note that Z_2 is used in computation, and the term $\delta^1(\sigma_1^i)$ is either 0 or 1. For persistent cohomology enriched barcodes, each barcode represents a generator and is colored by the weight values defined above.

Computationally, our persistent cohomology enriched barcodes are calculated based on the associated simplicial complex. For simplicity, the weights defined in Eqs. (6) and (7) are generalized to any 0-simplex and 1-simplex, respectively. Note that distance between two atoms from the same molecule is defined as Eq. (2). Similarly, weights for cohomology generators in Eqs. (8) and (9) are also extended to associated simplicial complex counterpartners.

Key Points

Our main contributions in this paper are as follows:

- To better represent molecular structures and interactions, we propose the first hypergraph model for molecular representation at atomic level.
- To characterize the multiscale information within molecules, we introduce a filtration process and propose hypergraph-based (weighted) persistent cohomology.
- Persistent properties from HPC/HWPC are used as molecular descriptors and combined with machine learning models, in particular, gradient boosting tree (GBT) model.
- Our HPC/HWPC-GBT models are tested on three well-established databases, including PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016. Our models can outperform all machine learning models with traditional molecular descriptors, for protein-ligand binding affinity prediction.

Code and data availability

The PDBbind databases were obtained from http://pdbbind.org.cn. The codes implemented for the hypergraph persistent cohomology and HPC-GBT models can be found in http://github.com/LiuXiangMath/Hypergraph-based-Persistent-Cohomology.

Authors' contributions statement

K.X. designed research; K.X., J.W., X.W. 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 RG31/18 and RG109/19, Tier 2 MOE2018-T2-1-033. The second author was supported by Natural Science Foundation of China (NSFC grant no. 11871284). The third author was supported by Natural Science Foundation of China (NSFC grant no. 11971144) and High-level Scientific Research Foundation of Hebei Province.

Supporting Information (SI)

Hypergraph-based persistent cohomology

Our definition of hypergraph-based persistent (embedded) cohomology is based on the dual relation between supremum/infimum chain complex and supremum/infimum cochain complex. Their homology relations can be naturally induced from universal coefficient theorem.

Let R is a principal ideal domain, G_R is a R-module, C is a chain complex of free R-module with boundary maps R-module homomorphisms. If we denote the n-th homology of the chain complex as $H_n(C,R)$, cohomology $H^n(C,G_R)$ of cochain complex $Hom_R(C,G_R)$ can be determined by the split short exact sequence,

$$0 \to Ext_R(H_{n-1}(C,R),G_R) \to H^n(C,G_R) \to Hom_R(H_n(C,R),G_R) \to 0,$$

which means that,

$$H^n(C,G_R) \cong Ext_R(H_{n-1}(C,R),G_R) \oplus Hom_R(H_n(C,R),G_R).$$

Definition 1.1. **Infimum cochain complex** Given a hypergraph $\mathcal{H}$, the infimum chain complex of $\mathcal{H}$ with coefficient R is $Inf_\star(\mathcal{H},R)$=$Inf_\star(R(\mathcal{H}_\star),R((K_\mathcal{H})_\star))$, we define the infimum cochain complex $Inf^\star(\mathcal{H},G_R)$ with coefficient G_R as

$$\mathrm{Inf}^n(\mathcal{H},G_R) = (\mathrm{Inf}_n(\mathcal{H},R))^* = (\mathrm{Inf}_n(R(\mathcal{H}_\star),R((K_\mathcal{H})_\star)))^*,$$

which the dual of $\mathrm{Inf}_n(R(\mathcal{H}_\star),R((K_\mathcal{H})_\star))$.

Definition 1.2. **Supremum cochain complex** Given a hypergraph $\mathcal{H}$, the supremum chain complex of $\mathcal{H}$ with coefficient R is $Sup_\star(\mathcal{H},R)$=$Sup_\star(R(\mathcal{H}_\star),R((K_\mathcal{H})_\star))$, we define the Supremum cochain complex $Sup^\star(\mathcal{H},G_R)$ with coefficient G_R as

$$\mathrm{Sup}^n(\mathcal{H},G_R) = (\mathrm{Sup}_n(\mathcal{H},R))^* = (\mathrm{Sup}_n(R(\mathcal{H}_\star),R((K_\mathcal{H})_\star)))^*,$$

which the dual of $\mathrm{Sup}_n(R(\mathcal{H}_\star),R((K_\mathcal{H})_\star))$.

For a hypergraph $\mathcal{H}$, its infimum chain complex $Inf_\star(\mathcal{H},R)$ is the largest subchain complex of the chain complex of $K_\mathcal{H}$ that is contained in the graded modules $R(\mathcal{H}_\star)$. Supremum chain complex $Sup_\star(\mathcal{H},R)$ is the smallest subchain complex of the chain complex of $K_\mathcal{H}$ that contains $R(\mathcal{H}_\star)$ as a graded modules.

Theorem 1.3. Let R be a principal ideal domain. Given a hypergraph $\mathcal{H}$, the cohomology of $Sup^\star(\mathcal{H},R)$ and the cohomology of $Inf^\star(\mathcal{H},R)$ are isomorphic.

Proof. From the universal coefficient theorem, we have

$$H^n(Sup^\star(\mathcal{H},G_R)) \cong Ext_R(H_{n-1}(Sup_\star(\mathcal{H},R)),G_R) \oplus Hom_R(H_n(Sup_\star(\mathcal{H},R)),G_R),$$

$$H^n(Inf^\star(\mathcal{H},G_R)) \cong Ext_R(H_{n-1}(Inf_\star(\mathcal{H},R)),G_R) \oplus Hom_R(H_n(Inf_\star(\mathcal{H},R)),G_R).$$

From [33], we have $H_n(Sup_\star(\mathcal{H},R)) \cong H_n(Inf_\star(\mathcal{H},R))$, so that

$$Ext_R(H_{n-1}(Sup_\star(\mathcal{H},R)),G_R) \cong Ext_R(H_{n-1}(Inf_\star(\mathcal{H},R)),G_R),$$

$$Hom_R(H_n(Sup_\star(\mathcal{H},R)),G_R) \cong Hom_R(H_n(Inf_\star(\mathcal{H},R)),G_R).$$

In this way, we have,

$$H^n(Sup^\star(\mathcal{H},G_R)) \cong H^n(Inf^\star(\mathcal{H},G_R))$$

■

Definition 1.4. **Embedded cohomology of hypergraph** Let R be a principal ideal domain. Given a hypergraph $\mathcal{H}$, we define the n-th embedded cohomology with coefficients in an R-module G_R of $\mathcal{H}$ as

$$H^n(\mathcal{H},G_R) = H^n(\mathrm{Sup}^\star(\mathcal{H},G_R)) = H^n(\mathrm{Inf}^\star(\mathcal{H},G_R)).$$

In our computation, the coefficient R and G_R are both Z_2. For simplicity, we denote $Inf_\star(\mathcal{H},R)$, $Sup_\star(\mathcal{H},R)$, $Inf^\star(\mathcal{H},G_R)$, $Sup^\star(\mathcal{H},R)$, and $H^n(\mathcal{H},G_R)$, as $Inf_\star(\mathcal{H})$, $Sup_\star(\mathcal{H})$, $Inf^\star(\mathcal{H})$, $Sup^\star(\mathcal{H})$, and $H^n(\mathcal{H})$, respectively.

Note that in our definition of hypergraph-based embedded cohomology, we make use of universal coefficient theorem and require the coefficient domains to be principal ideal domain R and R-module G_R. These conditions can be extended into more general situations and the definition of hypergraph-based embedded cohomology can be attained without using universal coefficient theorem.

Supremum cochain complexes generated from protein–ligand complex model

Our protein–ligand complex-based hypergraph $\mathcal{H}$ is a very special kind of hypergraph. If we consider Z_2 coefficient, its supremum cochain complex $\mathrm{Sup}_n(\mathcal{H})$ coincides with cochain complex of the associated simplicial complex $K_\mathcal{H}$. Therefore, its embedded cohomology is the simplicial cohomology of $K_\mathcal{H}$. Note that in the following sections, we only consider Z_2 situation.

Theorem 1.5. For an m-dimension hypergraph $\mathcal{H}$, its embedded cohomology is the simplicial cohomology of $K_\mathcal{H}$, if it satisfies the following two conditions.

1. For each vertex of $\mathcal{H}$, it is a 0-hyperedge.
2. For each n-hyperedge $\sigma_n = \{\mathbf{v}_0,\mathbf{v}_1,...,\mathbf{v}_n\}(1 \leqslant n \leqslant m)$ of $\mathcal{H}$, it has an associated ("face") set $\{\sigma^0_{n-1},\sigma^1_{n-1},...,\sigma^n_{n-1}\}$ with $\sigma^i_{n-1}(0 \leqslant i \leqslant n) = \{\mathbf{v}_0,...,\mathbf{v}_{i-1},\mathbf{v}_{i+1},...,\mathbf{v}_n\}$ generated from σ_n by removing the vertex $\mathbf{v}_i$. It only allows at most one ("face") element (among the $n+1$ elements) from the set is not an $(n-1)$-hyperedge of $\mathcal{H}$.

Note that for any hypergraph that satisfies the above two conditions, its embedded cohomology will be the same as the cohomology of the associated simplicial complex.

Since the coefficient is Z_2, we have $\{\mathbf{v}_i,\mathbf{v}_j\} = -\{\mathbf{v}_i,\mathbf{v}_j\} = \{\mathbf{v}_j,\mathbf{v}_i\}$, which means that for any two n-hyperedges σ_n, $\sigma_{n'}$ $(0 \leqslant n \leqslant m)$, if the vertices of σ_n and $\sigma_{n'}$ are same. Then, $\sigma_n = \sigma_{n'}$.

Lemma 1. If a hypergraph $\mathcal{H}$ satisfies the two conditions in the above theorem and $\sigma_t = \{\mathbf{v}_0,\mathbf{v}_1,...,\mathbf{v}_t\}$ is a t-simplex of $K_\mathcal{H}$ but not a t-hyperedge of $\mathcal{H}$, then there exists a $(t+1)$-hyperedge $\sigma_{t+1} = \{\mathbf{v}_0,\mathbf{v}_1,...,\mathbf{v}_t,\mathbf{v}_{t+1}\} \in \mathcal{H}_{t+1}$ and $t+1$ t-hyperedges $\{\sigma^0_t,\sigma^1_t,...,\sigma^t_t\} \subset \mathcal{H}_t$ where $\sigma^i_t(0 \leqslant i \leqslant t) =$

$\{\mathbf{v}_0, ..., \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, ..., \mathbf{v}_{t+1}\}$ which is generated by removing $\mathbf{v}_i$ from σ_{t+1}.

Proof. From the second condition of the theorem, we can see that for each n-hyperedge σ_n of $\mathcal{H}$, either all its $(n-1)$-"faces" exist, or there is only one $(n-1)$-"face" does not exist, assume it is σ^k_{n-1}. So $\mathcal{H}$ will be a simplicial complex(denote as $K^+_{\mathcal{H}}$) if we add these missed σ^k_{n-1} into $\mathcal{H}$. It also can be seen that any simplicial complex that $\mathcal{H}$ can embed is bigger than $K^+_{\mathcal{H}}$. So $K^+_{\mathcal{H}}$ is the smallest simplicial complex that $\mathcal{H}$ can embed, i.e., the associated simplicial complex of $\mathcal{H}$. As a result, the difference between $\mathcal{H}$ and $K_{\mathcal{H}}$ are just these missing "faces" σ^k_{n-1}. Here $\sigma_t \in (K_{\mathcal{H}})_t, \sigma_t \notin \mathcal{H}_t$, so σ_t is just such a σ^k_{n-1}. So from condition 2 we can see that there exists a $(t+1)$-hyperedge $\sigma_{t+1} = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_t, \mathbf{v}_{t+1}\} \in \mathcal{H}_{t+1}$ and $t+1$ t-hyperedges $\{\sigma_t, \sigma^1_t, ..., \sigma^t_t\} \subset \mathcal{H}_t$ where $\sigma^i_t (0 \leqslant i \leqslant t) = \{\mathbf{v}_0, ..., \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, ..., \mathbf{v}_{t+1}\}$. ■

Proof of Theorem 1.3. We have $H^n(\mathcal{H}) = H^n(\mathrm{Sup}^\star(\mathcal{H}))$. From supremum chain group definition, we get that $Sup_n(\mathcal{H}) = Z_2(\mathcal{H}_n) + \partial_{n+1}(Z_2(\mathcal{H}_{n+1}))$, therefore, we have that $Sup^n(\mathcal{H}) = (Z_2(\mathcal{H}_n) + \partial_{n+1}(Z_2(\mathcal{H}_{n+1})))^*$, we claim that

$$(Z_2(\mathcal{H}_n) + \partial_{n+1}(Z_2(\mathcal{H}_{n+1})))^* = (Z_2((K_{\mathcal{H}})_n))^*$$

If the claim is true, it directly follows that the embedded cohomology of $\mathcal{H}$ is just the simplicial cohomology of $K_{\mathcal{H}}$. ■

Proof of the Claim. It suffices to prove that

$$Z_2(\mathcal{H}_n) + \partial_{n+1}(Z_2(\mathcal{H}_{n+1})) = Z_2((K_{\mathcal{H}})_n)$$

Here, we have three cases.

1. $n = 0$, $Z_2(\mathcal{H}_0) + \partial_1(Z_2(\mathcal{H}_1)) = Z_2(\mathcal{H}_0) = Z_2((K_{\mathcal{H}})_0)$ since the 0-hyperedge set of $\mathcal{H}$ is same with the 0-simplex set of $K_{\mathcal{H}}$.
2. $0 < n < m$, firstly we have $Z_2(\mathcal{H}_n) + \partial_{n+1}(Z_2(\mathcal{H}_{n+1})) \subset Z_2((K_{\mathcal{H}})_n)$, for $\{Sup_\star(Z_2(\mathcal{H}_\star), Z_2((K_{\mathcal{H}})_\star)), \partial_\star\}$ is a subchain complex of $\{Z_2((K_{\mathcal{H}})_\star), \partial_\star\}$. Hence we only need to prove that each element of $(K_{\mathcal{H}})_n$ which is not in $\mathcal{H}_n$ can be represented by $Z_2(\mathcal{H}_n) + \partial_{n+1}(Z_2(\mathcal{H}_{n+1}))$. For an element $\sigma_n = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_n\}$ of $(K_{\mathcal{H}})_n$ which is not in $\mathcal{H}_n$, from **lemma 1**, we get that there exists an $(n+1)$-hyperedge $\sigma_{n+1} = \{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_n, \mathbf{v}_{n+1}\} \in \mathcal{H}_{n+1}$ and $n+1$ n-hyperedges $\{\sigma^0_n, \sigma^1_n, ..., \sigma^n_n\} \subset \mathcal{H}_n$ where $\sigma^i_n (0 \leqslant i \leqslant n) = \{\mathbf{v}_0, ..., \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, ..., \mathbf{v}_{n+1}\}$ which is generated by removing $\mathbf{v}_i$ from σ_{n+1}. We can see that $\{\sigma^0_n, \sigma^1_n, ..., \sigma^n_n\}$ and $\partial_{n+1}(\sigma_{n+1})$ are all in $Z_2(\mathcal{H}_n) + \partial_{n+1}(Z_2(\mathcal{H}_{n+1}))$ and it is obvious that we can get σ_n by the linear combination of $\{\sigma^0_n, \sigma^1_n, ..., \sigma^n_n\}$ and $\partial_{n+1}(\sigma_{n+1})$ with proper coefficients.
3. $n = m$, $Z_2(\mathcal{H}_m) + \partial_{m+1}(Z_2(\mathcal{H}_{m+1})) = Z_2(\mathcal{H}_m) = Z_2((K_{\mathcal{H}})_m)$ since the dimension of $\mathcal{H}$ is m and the m-hyperedge set of $\mathcal{H}$ is same with the m-simplex set of $K_{\mathcal{H}}$. ■

Our protein–ligand complex based hypergraph defined in Eq. (1) satisfies the two conditions of Theorem 1.3. Therefore, its supremum chain complex is just the chain complex of $K_{\mathcal{H}}$.

References

1. Smalley E. AI-powered drug discovery captures pharma interest. *Nature* 2017; **35**:604–5.
2. Fleming N. Computer-calculated compounds. *Nature* 2018; **557**(7707): S55–7.
3. Mak K-K, Pichika MR. Artificial intelligence in drug development: present status and future prospects. *Drug Discov Today* 2019; **24**(3): 773–80.
4. Chan HCS, Shan H, Dahoun T, *et al*. Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci* 2019; **40**(8): 592–604.
5. Chen H, Engkvist O, Wang Y, *et al*. The rise of deep learning in drug discovery. *Drug Discov Today* 2018; **23**(6): 1241–50.
6. Ekins S. The next era: deep learning in pharmaceutical research. *Pharm Res* 2016; **33**(11): 2594–603.
7. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017; **19**:221–48.
8. Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; **42**:60–88.
9. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 2010; **26**(9): 1169–75.
10. Khamis MA, Gomaa W, Ahmed WF. Machine learning in computational docking. *Artif Intell Med* 2015; **63**(3): 135–52.
11. Ain QU, Aleksandrova A, Roessler FD, *et al*. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* 2015; **5**(6): 405–24.
12. Jiménez J, Skalic M, Martinez-Rosell G, *et al*. K_DEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model* 2018; **58**(2): 287–96.
13. Mayr A, Klambauer G, Unterthiner T, *et al*. Deeptox: toxicity prediction using deep learning. *Front Environ Sci* 2016; **3**(80).
14. Puzyn T, Leszczynski J, Cronin MT. *Recent advances in QSAR studies: methods and applications*, Vol. **8**. Springer Science & Business Media, 2010.
15. Lo YC, Rensi SE, Torng W, *et al*. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 2018; **23**(8): 1538–46.
16. Schütt KT, Glawe H, Brockherde F, *et al*. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Physical Review B* 2014; **89**(20):205118.
17. Ramprasad R, Batra R, Pilania G, *et al*. Machine learning in materials informatics: recent applications and prospects. *npj Comput Mater* 2017; **3**(1):54.
18. Isayev O, Fourches D, Muratov EN, *et al*. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem Mater* 2015; **27**(3): 735–43.
19. Huan TD, Mannodi-Kanakkithodi A, Ramprasad R. Accelerated materials property predictions and design using motif-based fingerprints. *Physical Review B* 2015; **92**(1): 014106.
20. 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.
21. Cang ZX, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int J Numer Method Biomed Eng* 2017. doi: 10.1002/cnm.2914.

22. 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.
23. 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.
24. 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.
25. Wu KD, Wei GW. Quantitative toxicity prediction using topology based multi-task deep neural networks. *J Chem Inf Model* 2018; **58**(2): 520–31 page 10.1021/acs.jcim.7b00558.
26. 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.
27. 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.
28. Nguyen DD, Wei GW. DG-GL: differential geometry-based geometric learning of molecular datasets. *Int J Numer Method Biomed Eng* 2019; **35**(3): e3179.
29. Nguyen DD, Gao KF, Wang ML, *et al*. MathDL: mathematical deep learning for D3R grand challenge 4. *J Comput Aided Mol Des* 2019;1–17.
30. 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.
31. Grow C, Gao KF, Nguyen DD, *et al*. Generative network complex (GNC) for drug discovery. *arXiv preprint arXiv:191014650* 2019.
32. Cang ZX, Wei GW. Persistent cohomology for data with multicomponent heterogeneous information. *SIAM Journal on Mathematics of Data Science* 2020; **2**(2): 396–418.
33. Bressan S, Li JY, Ren SQ, *et al*. The embedded homology of hypergraphs and applications. *arXiv preprint arXiv:161000890* 2016.
34. Ren SQ, Wang C, Wu CY, *et al*. A discrete Morse theory for hypergraphs. *arXiv preprint arXiv:180407132* 2018.
35. Ren SQ, Wu CY, Wu J. Hodge decompositions for weighted hypergraphs. *arXiv preprint arXiv:180511331* 2018.
36. Grigor'yan A, Muranov Y, Yau ST. Graphs associated with simplicial complexes. *Homol Homotopy Appl* 2014; **16**(1): 295–311.
37. Grigor'yan A, Lin Y, Muranov Y, *et al*. Cohomology of digraphs and (undirected) graphs. *Asian J Math* 2015; **19**(5): 887–932.
38. Grigor'yan A, Jimenez R, Muranov Y, *et al*. On the path homology theory of digraphs and Eilenberg–Steenrod axioms. *Homol Homotopy Appl* 2018; **20**(2): 179–205.
39. Grigor'yan A, Jimenez R, Muranov Y, *et al*. Homology of path complexes and hypergraphs. *Topol Appl* 2019; **267**:106877.
40. Verri A, Uras C, Frosini P, *et al*. On the use of size functions for shape analysis. *Biol Cybern* 1993; **70**(2): 99–107.
41. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geom* 2002; **28**:511–33.
42. Zomorodian A, Carlsson G. Computing persistent homology. *Discrete Comput Geom* 2005; **33**:249–74.
43. Meng ZY, Xia KL. Persistent spectral based machine learning (PerSpect ML) for drug design. *arXiv preprint arXiv:200200582* 2020.
44. Bergomi MG, Ferri M, Vertechi P, *et al*. Beyond topological persistence: starting from networks. *arXiv preprint arXiv:190108051* 2019.
45. Parks AD, Lipscomb SL. *Homology and Hypergraph Acyclicity: A Combinatorial Invariant for Hypergraphs*. Technical report: NAVAL SURFACE WARFARE CENTER DAHLGREN VA, 1991.
46. Chung F, Graham RL. Cohomological aspects of hypergraphs. *Trans Am Math Soc* 1992; **334**(1): 365–88.
47. Emtander E. Betti numbers of hypergraphs. *Commun Algebra* 2009; **37**(5): 1545–71.
48. Johnson J. Hypernetworks of complex systems. In: *International Conference on Complex Sciences*. Springer, 2009, 364–75.
49. Ghrist R. Barcodes: the persistent topology of data. *Bullet Am Math Soc* 2008; **45**(1): 61–75.
50. Silva VD, Morozov D, Vejdemo-Johansson M. Persistent cohomology and circular coordinates. *Discrete Comput Geom* 2011; **45**(4): 737–59.
51. Liu J, Wang RX. Classification of current scoring functions. *J Chem Inf Model* 2015; **55**(3): 475–82.
52. 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. *Mol Inform* 2015; **34**(2–3): 115–26.
53. 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.
54. 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.
55. 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.
56. 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.
57. Feinberg EN, Sur D, Wu ZQ, *et al*. Potentialnet for molecular property prediction. *ACS Central Sci* 2018; **4**(11): 1520–30.
58. 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.
59. Pattanaik L, Coley CW. Molecular representation: going long on fingerprints. *Chem* 2020; **6**(6): 1204–7.