OXFORD

Briefings in Bioinformatics, 2022, **23**(2), 1–15
https://doi.org/10.1093/bib/bbac024
Problem Solving Protocol

Persistent spectral based ensemble learning (PerSpect-EL) for protein–protein binding affinity prediction

JunJie Wee and Kelin Xia

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

Abstract

Protein–protein interactions (PPIs) play a significant role in nearly all cellular and biological activities. Data-driven machine learning models have demonstrated great power in PPIs. However, the design of efficient molecular featurization poses a great challenge for all learning models for PPIs. Here, we propose persistent spectral (PerSpect) based PPI representation and featurization, and PerSpect-based ensemble learning (PerSpect-EL) models for PPI binding affinity prediction, for the first time. In our model, a sequence of Hodge (or combinatorial) Laplacian (HL) matrices at various different scales are generated from a specially designed filtration process. PerSpect attributes, which are statistical and combinatorial properties of spectrum information from these HL matrices, are used as features for PPI characterization. Each PerSpect attribute is input into a 1D convolutional neural network (CNN), and these CNN networks are stacked together in our PerSpect-based ensemble learning models. We systematically test our model on the two most commonly used datasets, i.e. SKEMPI and AB-Bind. It has been found that our model can achieve state-of-the-art results and outperform all existing models to the best of our knowledge.

Keywords: protein–protein interaction, Hodge Laplacian, persistent spectral, molecular featurization, ensemble learning

Introduction

A wide range of biological processes and mechanisms, including cell proliferation, signaling, metabolism, immune system and protein transport, are governed or coordinated by the complex networks of protein–protein interactions (PPIs) [20, 21]. The great size and diversity of PPIs offer a highly selective and tunable way to modulate protein activities and pathways. Protein mutations and genetic variations can affect protein folding and stability, change the binding affinities of protein interactions and consequently lead to disease and drug resistance [51]. The understanding of PPIs, in particular PPI upon mutations, is vital to various biomedical applications, including disease-associated mutation analyses, drug design and therapeutic intervention [20, 21]. Experimentally, various methods have been developed to determine protein assembly structures at different resolutions. Among them, atomic resolution tools include X-ray crystallography, nuclear magnetic resonance and cryo-electron microscopy, and residual resolution tools include cross-linked mass spectrometry, hydrogen/deuterium exchange and mutagenesis. Further, PPI binding affinity and stabilities can be measured by techniques, such as isothermal titration calorimetry, surface plasmon resonance, fluorescence and blue native polyacrylamide gel electrophoresis. However, experimental studies for structures and binding affinity are time-consuming, laborious and expensive. They usually require protein purification. Moreover, the experimental analysis of mutation effects needs both wild-type and mutated proteins, which are very challenging to obtain. Currently, only about 6.5% of the known human interactome has structural information [40].

Fast and efficient computational methods and models have been developed for the analysis of PPIs. In particular, significant efforts have been made for the evaluation of PPI binding affinity upon mutation ($\Delta\Delta G$). These models can be generalized into three categories, including molecular dynamic (MD) based approaches, statistical energy-based models and machine learning models.

Junjie Wee is a Ph.D. student from Nanyang Technological University, Singapore. His research interests are Molecular data analysis, geometry and topological data analysis and machine learning.
Kelin Xia is an assistant professor at Nanyang Technological University, Singapore. His research interests are topological data analysis and Mathematical AI for Molecular Sciences.

Received: November 15, 2021. **Revised:** December 30, 2021. **Accepted:** January 17, 2022

MD-based models, including FoldX [22], Rosetta [29], zone equilibration of mutants (ZEMu) [13], single amino acid mutation based change in binding free energy (SAAMBE) [49] and others [20], usually characterize the binding affinity of PPIs with various physical energy terms, including van der Waals interactions, electrostatic energies, hydrogen bonds, solvation energy, etc. Mutation effects are considered by modeling conformational changes with rotamer and structure ensemble approaches. Different from the MD models, statistical energy-based approaches, including BindProfX [68], BeAtMuSiC [12], contact potentials [38], Profile-score [58] and Dcomplex [32] use various intermolecular potentials extracted from experimental structures to study PPI binding affinity. These intermolecular potentials can be constructed based on the contacts at atomic, residual or other coarse-grained levels. Recently, data-driven machine learning models have achieved state-of-the-art results in PPI analysis [54], due to great advancements in computational power, learning models and data accumulation.

Various PPI databases are established during the past few decades, including Alanine scanning energetics database (ASEdb) [59], PPIs thermodynamic database (PINT) [30], structural kinetic and energetic database of mutant protein interactions (SKEMPI) [37], database of binding affinity change upon mutations (DACUM) [18], antibody-bind database (AB-Bind) [55], protein–protein complex mutation thermodynamics (PROXiMATE) [26] and kinetic and thermodynamic database of mutant protein interactions (dbMPIKT) [31]. Recently, an updated version SKEMPI 2.0 has been constructed [24]. It combines several databases including SKEMPI, AB-Bind, PROXiMATE and dbMPIKT, together with manually curated data from the literature. In total, it has 7085 mutations, including about 3000 single point alanine mutations, about 2000 single point non-alanine mutations and roughly 2000 multiple mutations, on various types of protein complexes, such as protease-inhibitor, antibody-antigen and TCR-pMHC complexes. With the ever-increasing PPI data, a great amount of data-driven learning models have been developed [20, 54], including mCSM [52], ELASPIC [57], BindProf [2], MutaBind [69], iSEE [19], MuPIPR [71], ProAffiMuSeq [25], GeoPPI [34], etc. In general, these data-driven models can be classified into two types: featurization-based machine learning models and end-to-end deep learning models. For the first type, different types of PPI information from sequences, inter-residue interactions, evolutionary conservation, dynamic properties, energy terms, pharmacophore descriptors, structure-based descriptors and others are used as input features for machine learning models, such as support vector machine, random forest, gradient boost tree, etc. Note that these input features are manually generated by using mathematical, physical, chemical and biological models. For end-to-end learning models, proteins are usually represented as surfaces, graphs or networks with embedded vectors or one-hot-vectors [3, 16]. The intrinsic features for PPIs are automatically learned and implicitly represented in deep learning models. The most commonly used deep learning models for PPIs are graph neural networks and geometric learning models. Even with the great advancements, generating a highly efficient molecular featurization, which is key to the performance of learning models, remains a challenging problem [35, 50].

Recently, advanced mathematics, in particular topological data analysis (TDA) [15, 72], is used in molecular representation and featurization [4, 7, 36, 43]. Their combination with learning models have achieved great success in various steps of drug design, including protein-ligand binding affinity prediction [6–8, 47, 48], protein stability change upon mutation prediction [4, 5], toxicity prediction [9, 27, 65], solvation free energy prediction [61, 62], partition coefficient and aqueous solubility [66], binding pocket detection [70] and drug discovery [17]. Outstanding performance has been consistently achieved in D3R Grand challenge [44–46]. In particular, TopNetTree has demonstrated greater power in predicting binding affinity change upon mutation [63]. It has outperformed all existing models and provided great insights for the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) mutations [10, 64]. Motivated by the great success, we have proposed persistent spectral based machine learning models (PerSpect-ML) and use them in drug design [33, 36]. Mathematically, spectral models, including spectral graph theory [11, 56], spectral simplicial complex [1, 14, 23, 41] and spectral hypergraph [33], study the topological properties with algebraic tools, including characteristic polynomial, eigenvalues, eigenvectors and other eigenspectrum properties. The spectral information is used for the characterization of biomolecular structures and interactions [33, 36].

Here, we propose persistent spectral based ensemble learning (PerSpect-EL) models for PPI binding affinity change upon mutation, for the first time. PPIs at the molecular level are represented by a series of simplicial complexes generated from a designed filtration process. Hodge (or combinatorial) Laplacian (HL) matrixes can be systematically constructed on these simplicial complexes. The persistence and variation of the spectral attributes, which are statistical and combinatorial properties of Laplacian eigenvalues, are used as input feature vectors for 1D convolutional neural network (CNN) models. To combine the contributions from these different types of PerSpect attributes, we use an ensemble model and stack the individual 1D CNNs together. Moreover, a series of precalculated physical properties of PPIs are used as auxiliary features and further incorporated into our PerSpect stacking models. Our PerSpect ensemble model is trained and tested on SKEMPI and AB-Bind datasets, which are the two most widely used datasets for PPI binding affinity change upon mutation. It has been found that our model can outperform all existing models, as far as we know. Our PerSpect ensemble learning models have great potential in the analysis of PPIs.

Results

Persistent spectral theory

Molecular structures and interactions can be described by different topological representations, including graphs, simplicial complexes and hypergraphs. Persistent spectral models, including persistent spectral graph, persistent spectral simplicial complex and persistent spectral hypergraph, can be constructed accordingly based on these different representations [33, 36]. Here, we focus on PerSpect simplicial complex models. A simplicial complex, which is a generalization of graphs, is made up of simplices. Geometrically, a simplex can be a node (0-simplex), edge (1-simplex), triangle (2-simplex), tetrahedron (3-simplex) or other n-dimensional counterpart (n-simplex). Topological invariants, in particular, Betti numbers β, can be evaluated from the simplicial complex. In general, β_0 is the number of connected components, β_1 is the number of circles or loops and β_2 is the number of voids/cavities. For each simplicial complex, a set of HL matrices (L_0, L_1, L_2 and higher order L_k) can be constructed. The number (multiplicity) of zero eigenvalues for $k-$**th** dimensional HL matrix L_k equals to β_k. Moreover, non-zero eigenvalues and their eigenvectors, such as Fiedler value (algebraic connectivity) and Fiedler vector, can be used for a more detailed characterization of the 'geometric' properties of the structure.

One of the key concepts for Persistent spectral theory is the filtration process, during which a series of simplicial complexes (or other topological representations) at various scales are systematically generated. Based on these simplicial complexes, a series of HL matrices can be generated and their spectral information can be evaluated. PerSpect models focus on the persistence and variation of eigen spectral information during a filtration process. More specifically, statistical and combinatorial properties of the eigen spectrum can be calculated for each HL matrix, and the change of these attributes during the filtration is defined as PerSpect attributes, which includes persistent (zero-) multiplicity, persistent mean, persistent maximal, etc. Note that persistent (zero-) multiplicity is exactly the Betti curve [36]. Figure 1 illustrates simplexes (**A**), Betti numbers (**B**), a Vietoris-Rips complex (**C**), a filtration process (**D**) and its corresponding Hodge–Laplacian matrices (**E**).

PerSpect for PPI

Characterization of PPI Similar to other biomolecular interactions, such as those between protein–ligand, DNA–ligand and protein–DNA, protein interaction regions or binding domains are usually much smaller than the size of the protein complex. For two proteins $\mathcal{P}_1$ and $\mathcal{P}_2$, and their protein–protein complex $\mathcal{P}_{1,2} = \mathcal{P}_1 \cup \mathcal{P}_2$, we define their interaction domains as follows:

(1) $\mathcal{P}_1^B$: atoms from the binding site of $\mathcal{P}_1$
(2) $\mathcal{P}_2^B$: atoms from the binding site of $\mathcal{P}_2$

In this way, PPIs at the molecular level can be characterized by the various interactions between atoms of these two regions. More specifically, there are two general types of interactions, i.e. atom interactions between and within two regions. Two types of simplicial complexes are employed accordingly to model them.

We denote an atom coordinate as $\mathbf{r}$, and the two binding domains as $\mathcal{P}_1^B = \{\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_{N_{\mathcal{P}_1}}}\}$ and $\mathcal{P}_2^B = \{\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_{N_{\mathcal{P}_2}}}\}$, respectively. To describe the atom interactions between two protein regions, we consider an interactive distance matrix with size $(N_{\mathcal{P}_1}+N_{\mathcal{P}_2}) \times (N_{\mathcal{P}_1}+N_{\mathcal{P}_2})$ as follows:

$$I_{\mathcal{B}}(m_i, m_j) = \begin{cases} \|\mathbf{r}_i - \mathbf{r}_j\|, & \mathbf{r}_i \in \mathcal{P}_1^B, \mathbf{r}_j \in \mathcal{P}_2^B \text{ or } \mathbf{r}_i \in \mathcal{P}_2^B, \mathbf{r}_j \in \mathcal{P}_1^B \\ \infty, & \text{otherwise.} \end{cases}, \quad (1)$$

where $\|\mathbf{r}_i - \mathbf{r}_j\|$ is the Euclidean distance, and m_i and m_j are the indexes of atom $\mathbf{r}_i$ and ligand atom $\mathbf{r}_j$, respectively. Based on the interactive distance matrix, Vietoris–Rips complexes are constructed and atom interactions between two regions are encoded into PerSpect features.

Further, to characterize the atom interactions within binding regions, we consider three atom sets, including $\mathcal{P}_1^B$, $\mathcal{P}_2^B$ and $\mathcal{P}_{1,2}^B = \mathcal{P}_1^B \cup \mathcal{P}_2^B$. For each atom set, its Alpha complexes are systematically constructed based on the filtration of atom radius. From these Alpha complexes, HL matrixes can be generated and their PerSpect features are used as features.

As an important step in effective characterization for PPIs, element-specific combinations [6–8, 47, 48] are employed in our model. Essentially, a protein structure can be decomposed into a series of atom sets, including carbon (C), nitrogen (N) and oxygen (O). The topological features from the combination of these atom sets can be used in the characterization of different types of interactions between and within proteins. In our model, both interactive element-specific combinations and general element-specific combinations are considered. For interactive element-specific models, a total of nine combinations {C–C, C–N, C–O, N–C, N–N, N–O, O–C, O–N, O–O} between $\mathcal{P}_1^B$ and $\mathcal{P}_2^B$ in both wild and mutant types are considered. Vietoris–Rips complexes are generated based on them. Further, a total of seven general element-specific combinations, including {{C}, {N}, {O}, {C,N}, {C,O}, {N,O}, {C,N,O}}, are considered, and alpha complexes are generated based on them. PerSpect features from the corresponding HL matrixes are used as molecular descriptors.

Characterization of PPI on mutation The study of the PPI binding affinity upon mutation is of great importance, especially for the understanding of the mutation effects of SARS-CoV-2. Normally, mutations happens at one or several mutation sites. Here, we focus on the single mutation situation. Let the mutation site of a protein–protein complex $\mathcal{P}_{1,2}$ to be $\mathcal{P}_1^{MS}$, and its neighboring atoms to be $\mathcal{P}_1^{MN}$. More specifically, we have

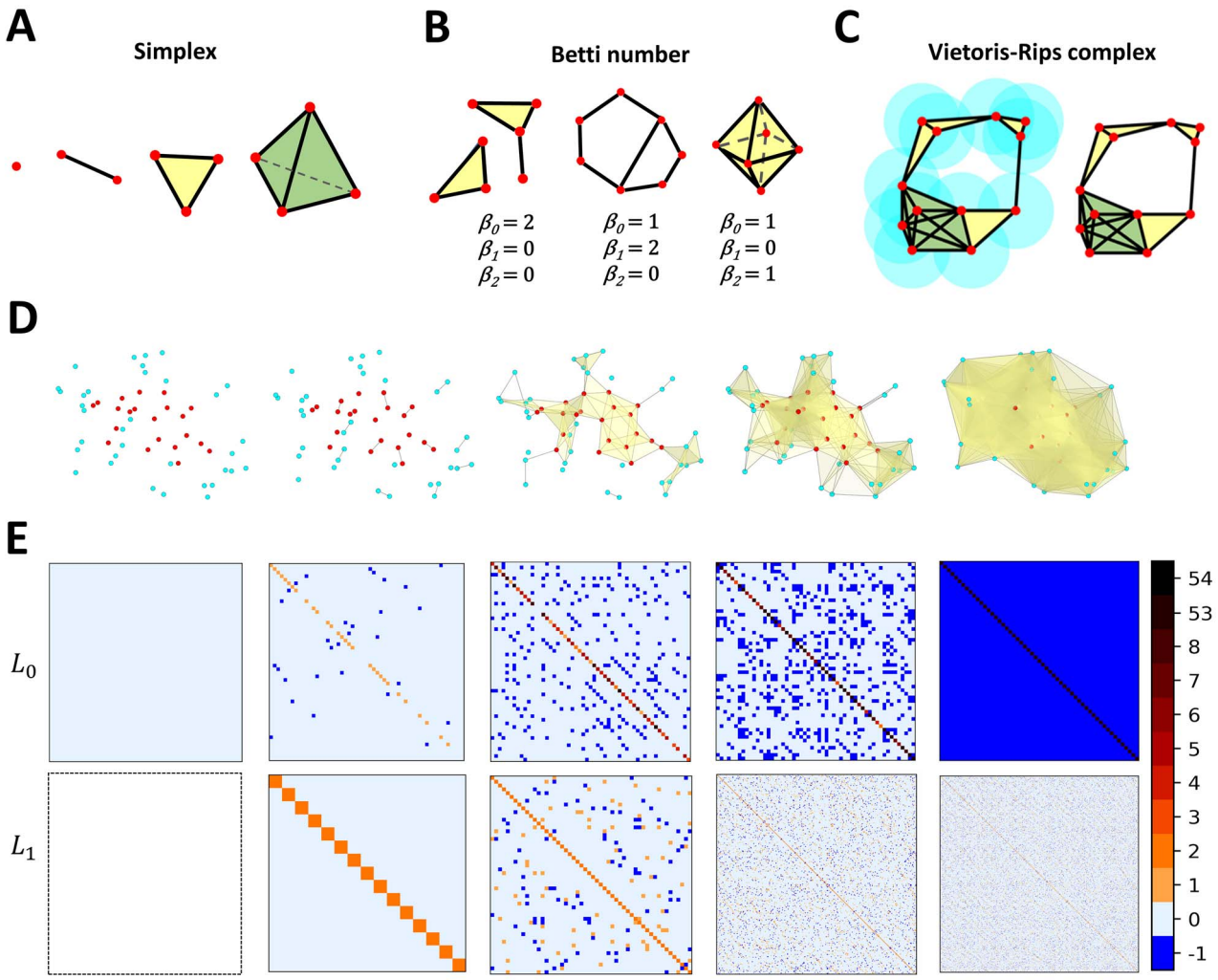

Figure 1. Illustration of fundamental concepts in PerSpect. **A** Examples of k-simplices as a point (0-simplex), an edge (1-simplex), a triangle (2-simplex) and a tetrahedron (3-simplex) geometrically. **B** Geometric meanings of Betti numbers. β_0 is the number of connected components, β_1 is the number of circles or loops, and β_2 is the number of voids or cavities. **C** Illustration of a Vietoris–Rips complex. Let all the vertices associate with pre-defined same-sized spheres. A k-simplex is formed if any two spheres are overlapped with each other. **D** A filtration process for PDBID: 3BN9. We consider the mutation site with mutation ID: HQ100aV. The atoms in red are from wild-type residue tyrosine, while the cyan atoms are within 3Å from tyrosine. **E** The generated combinatorial HLs for **D**. Only the combinatorial HLs L_0 and L_1 are illustrated. L_0 starts off as a zero matrix (isolated point cloud) and gradually transforms into a matrix with all non-diagonal entries -1, representing a complete graph. For L_1, the matrix starts to appear when 1-simplex starts to form in filtration process. The number of non-diagonal non-zero entries in L_1 increases in initial stages of filtration process but slowly converging to zero toward the end, resulting in a diagonal matrix.

(1) $\mathcal{P}_1^{MS}$: atoms of the mutation site
(2) $\mathcal{P}_1^{MN}$: neighboring atoms within 12Å from the mutation site ($\mathcal{P}_1^{MS}$)

To characterize the atom interactions between the mutation site and its neighboring atoms, we can define an interactive matrix $I_{\mathcal{M}}(m_i, m_j)$, similar to Eq. (1), as follows:

$$I_{\mathcal{M}}(m_i, m_j) = \begin{cases} \|\mathbf{r}_i - \mathbf{r}_j\|, & \text{if } \mathbf{r}_i \in \mathcal{P}_1^{MS}, \mathbf{r}_j \in \mathcal{P}_1^{MN} \\ & \text{or } \mathbf{r}_i \in \mathcal{P}_1^{MN}, \mathbf{r}_j \in \mathcal{P}_1^{MS}. \\ \infty, & \text{otherwise.} \end{cases} \quad (2)$$

Here, $\|\mathbf{r}_i - \mathbf{r}_j\|$ is the Euclidean distance, and m_i and m_j are the indexes of atom $\mathbf{r}_i$ and atom $\mathbf{r}_j$, respectively. Similar to the characterization of PPIs, both interactive and general element-specific combinations are considered, and the corresponding Vietoris–Rips and Alpha complexes are generated. Molecular descriptors can be obtained from the corresponding HL matrices. More details can be found in Materials and Methods.

Figure 2 **A** shows the zero-eigenvalue persistent multiplicities of L_0 and L_1 before and after mutation. A clear topological variation is observed in the mutation structures. Note that persistent multiplicities (of zero eigenvalues) correspond to Betti curves. Other persistent attributes from the non-zero eigenvalues would reveal more detailed 'geometric' information of the structures. Figure 2 **B** shows the persistent mean, minimum, maximum and standard deviation (of all non-zero eigenvalues) in L_0 and L_1 between wild and mutant structures. In general, these persistent attributes change with respect

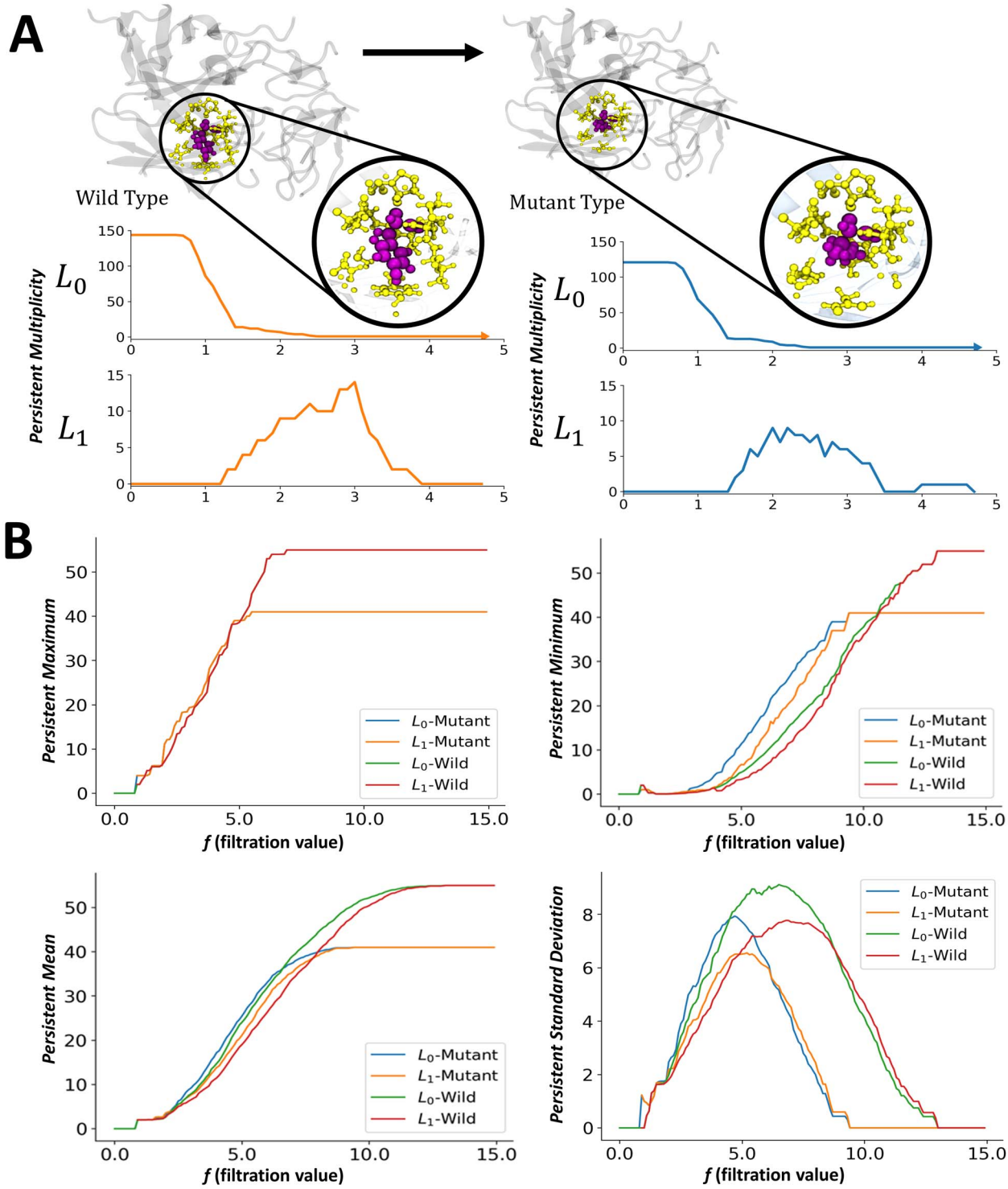

Figure 2. Illustration of persistent attributes in wild type and mutation type of protein 3BN9. The residue tyrosine in wild type is mutated into valine residue (HQ100aV). **A** The zero-eigenvalue persistent multiplicities between two types. They are generated by using all atoms within 5Å within the mutation site. **B** Illustration of persistent maximum, minimum, mean and standard deviation generated from L_0 and L_1.

to filtration values. Different patterns can also be clearly observed between the wild and mutant type.

PerSpect ensemble learning for PPIs binding affinity changes prediction

We apply an ensemble learning model which stacks three different types of base learner models together, as shown in Figure 3. The first type of base learners are 11 persistent attributes from L_0 (Rips complex) with 1D CNN models, which are denoted as **A.1** to **A.11**. The second type of base learner models are PerSpect features from Alpha complexes with 1D CNN models, which are denoted as **B.1** and **B.2**. The third type of base learner are auxiliary features with gradient boosting tree (GBT) model, which

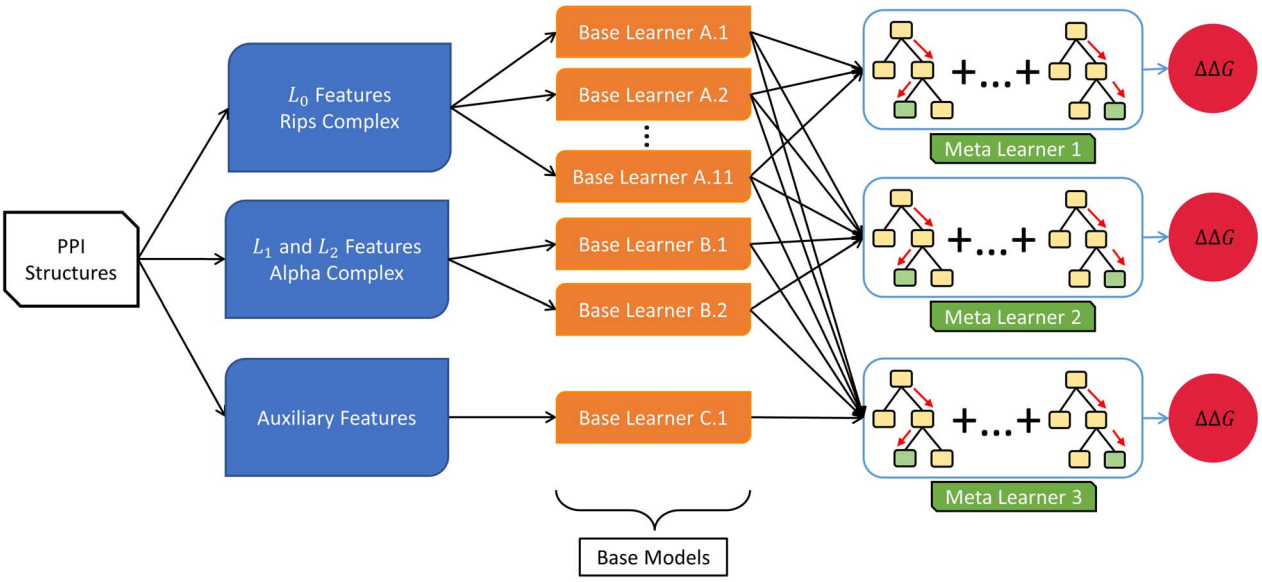

Figure 3. Illustration of the three PerSpect-EL models. The three types of base learners are based on PerSpect features from Rips complex, PerSpect features from Alpha complex and auxiliary features. The first two types of base learners use 1D CNN models, while the third type uses GBT. Their individual output predictions are stacked together as input into Meta Learners for a final prediction. Meta Learner 1 (PerSpect-EL (**M1**)) uses 11 base learners from Rips complexes. Meta Learner 2 (PerSpect-EL (**M2**)) uses 12 base learners from both Rips complexes and Alpha complexes. Meta Learner 3 (PerSpect-EL (**M3**)) uses all 14 base learners.

is denoted as **C.1**. These auxiliary features are the same as those in TopNetTree [63]. The details can be found in Materials and Methods.

In our PerSpect-EL model, we stack all the above three types of base learner models together and denote the model as PerSpect-EL(**M3**). We have also studied another two ensemble learning models PerSpect-EL (**M1**) and PerSpect-EL (**M2**). The PerSpect-EL (**M1**) includes only the first type of base learners, while PerSpect-EL (**M2**) contains in it the first two types of base learners, i.e. excluding auxiliary feature based GBT model. Computationally, all 13 base learner with 1D CNN models have been optimized with the same architecture and hyperparameters. The detailed architecture the 1D CNN model can be found in Materials and Methods.

To validate the performance of our models, we consider the two most commonly used datasets, namely, SKEMPI-1131 and AB-Bind [55, 63].

Performance on SKEMPI-1131 The SKEMPI dataset contains 3047 protein–protein heterodimeric complexes with experimentally determined structures recorded with binding affinity changes due to mutations. The structures are collected from various scientific literature and consist of both single-point and multi-point mutations. From a total of 2317 single point mutations, a subset of 1131 non-redundant interface structures are selected. This 1131 structures are known as the SKEMPI-1131 dataset, which has been commonly used as the benchmark for various prediction models, including TopNetTree, BindProfX, Profile-score, FoldX, SAAMBE, BeAtMuSic and Dcomplex. We test our PerSpect-EL models using a similar 10-fold cross-validation as in previous models. Figure 4 demonstrates the performance of our PerSpect-EL models and Table 1 lists the results for all the state-of-the-art models. It can be seen that our PerSpect-EL (M3) model has a PCC of 0.853, which is better than all existing models, as far as we know. For PerSpect-EL (M1) and PerSpect-EL (M2) models, they have better performance than nearly all the existing model, except TopNetTree, which also uses advanced mathematical invariants as their molecular features. Figure 4 **C** illustrates the detailed predictions of PerSpect-EL (M3) by mutated residue types. Our PerSpect-EL (M3) model can achieve great accuracy in all the predictions.

Further, a detailed comparison, between PerSpect-EL (M3) predictions and experiments, for the average

Table 1. Comparison of PerSpect-EL models with existing state-of-the-art models on SKEMPI-1131 dataset. The PerSpect-EL models include PerSpect-EL (M1) with PCC 0.804±0.004 and RMSE 1.454±0.0138 kcal/mol, PerSpect-EL (M2) with PCC 0.813±0.003 and RMSE 1.430±0.0966 kcal/mol and PerSpect-EL (M3) with PCC 0.853±0.002 and RMSE 1.303±0.00726 kcal/mol

Models	PCC
PerSpect-EL (M3)	**0.853**
TopNetTree	0.850
PerSpect-EL (M2)	**0.813**
PerSpect-EL (M1)	**0.804**
BindProfX	0.738
Profile-score + FoldX	0.738
Profile-score	0.675
SAAMBE	0.624
FoldX	0.457
BeAtMuSic	0.272
Dcomplex	0.056

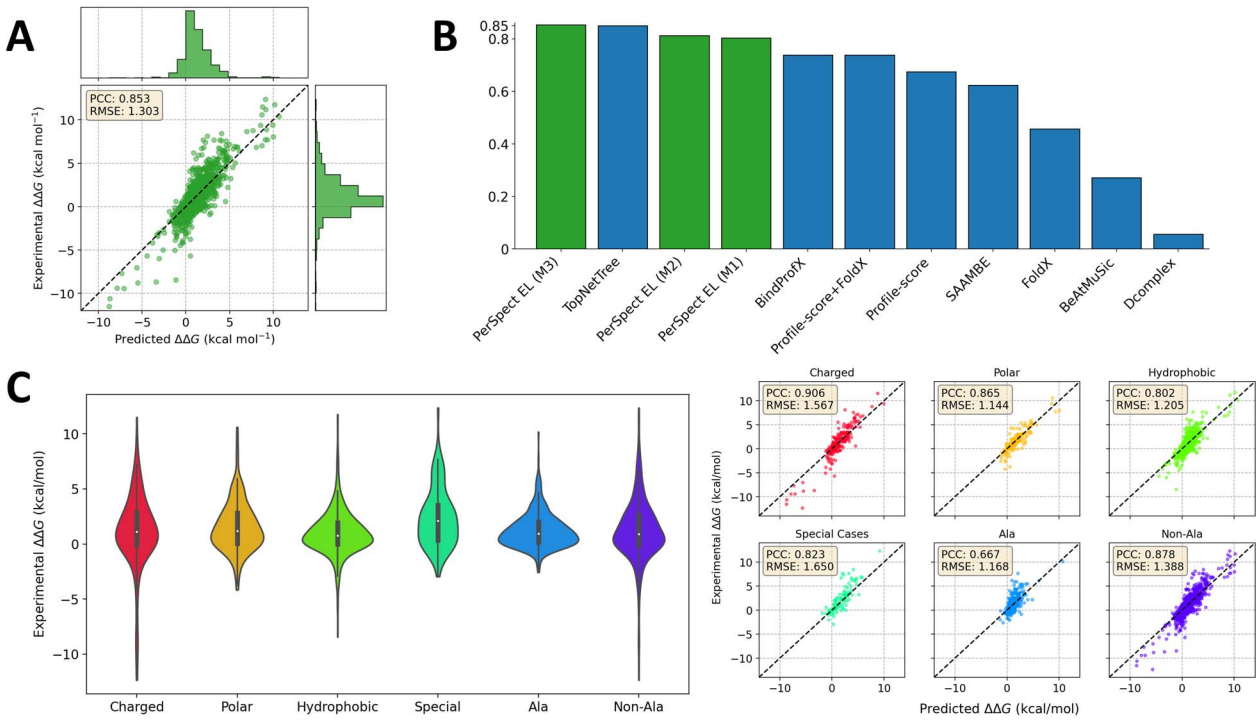

Figure 4. Illustration of PerSpect-EL model performance with SKEMPI-1131 data. **A** Comparison between the experimental binding affinity changes (kcal/mol) with predicted binding affinity changes (kcal/mol) from PerSpect-EL (M3) model; **B** Comparison of PerSpect-EL models with existing state-of-the-art prediction models; **C** Breakdown of predicted binding affinity changes (kcal/mol) by mutation types and by alanine/non-alanine mutations.

and variance of $\Delta\Delta G$ in each type of residue-to-residue mutation is presented in Figure S1. More specifically, Figure S1 **A** shows the comparison of the average of $\Delta\Delta G$ between our predictions and experimental results. Note that a reverse mutation will result in a negative $\Delta\Delta G$. In the residue-to-residue mutation matrix for average $\Delta\Delta G$, the upper triangle region and lower triangle region have exactly the same absolute values but with the opposite sign. A highly consistent pattern between the two residue-to-residue mutation matrices can be observed, indicating that our predictions are extremely accurate. Figure S1 **B** shows the comparison of the variance of $\Delta\Delta G$. Similarly, the values in the two matrices are highly consistent with each other. Further, we can find some interesting results in the mutation matrices. For instance, all the mutations from the other types of residues to alanine (A) result in a positive binding affinity change. This may due to the reason that the alanine is a small-sized residue, thus it has a relatively higher stability than the other large-sized residues. We can also observe that mutations from the other types of residues to tyrosine (Y) or leucine (L) are associated with negative binding affinity change, mainly due to higher instabilities of the two residue types. Moreover, negative binding affinity changes occur in mutations from charged residues to uncharged polar residues (e.g. D to T), mainly due to the influence of the interactions between charged residues.

Performance on AB-Bind S645 The AB-Bind dataset consists of 1101 mutation data entries [55]. Only 645 mutation points across 29 antibody-antigen complexes are single-point mutations. This single-point mutation subset is known as AB-Bind S645, which consists of 20% stabilizing mutations and 80% non-stabilizing ones. A total of 87 of the 645 single-point mutations are homology structures, while 27 single-point mutations result in the non-binding situations. For these 27 non-binders, their binding affinity changes are set to be a constant value 8 kcal/mol, and they are considered as outliers in the dataset. It has been found that these outliers can severely worsen the performance of learning models [63].

It has been found that our PerSpect-EL (M3) model can achieve an average PCC of 0.59 for the 10-fold cross-validation on AB-Bind S645 dataset, including the non-binders. As demonstrated in Table 2, our results are better than all existing models except TopNetTree [63]. However, the same test without non-binders shows that our model can achieve a better performance than all existing models, with an average PCC of 0.70. A similar blind test on 87 homology structures shows that our model can deliver an average PCC of 0.63 with RMSE 1.144 kcal/mol, which is the best results as far as we know. Figure S2 shows the performance of PerSpect-EL (M3) on AB-Bind S645 data. The predictions are grouped by mutation types in Figure S2 **D** and by mutation regions in Figure S2 **E**.

Discussion

Data representations and featurizations are of essential importance to all learning models. Advanced mathematical tools, which characterize molecular intrinsic structural and physical properties, have demonstrated great potential to significantly improve the efficiency of

Table 2. Comparison of PerSpect-EL models with existing state-of-the-art models on AB-Bind S645 dataset. The PerSpect-EL models include PerSpect-EL (M3) with PCC 0.59±0.0242 and RMSE 1.593±0.0341 kcal/mol, PerSpect-EL (M2) with PCC 0.59±0.0246 and RMSE 1.593±0.0353 kcal/mol and PerSpect-EL (M1) with PCC 0.57±0.0246 and RMSE 1.642±0.0376 kcal/mol

Models	Average PCC	
	with non-binders	**w/o non-binders**
TopNetTree	0.65	0.68
PerSpect-EL (M3)	**0.59**	**0.70**
PerSpect-EL (M2)	**0.59**	**0.70**
PerSpect-EL (M1)	**0.57**	**0.66**
TopGBT	0.56	–
mCSM-AB	0.53	0.56
TopCNN	0.53	–
Discovery Studio	0.45	–
mCSM-PPI	0.35	-
FoldX	0.34	–
STATIUM	0.32	–
DFIRE	0.31	–
bASA	0.22	–
dDFIRE	0.19	–
Rosetta	0.16	–

machine learning models for molecular data analysis. In our PerSpect-EL models, Hodge Laplacian based spectral information is used for PPI representation and featurization for the first time. A multiscale representation is achieved in our model through a sequence of HL matrices at various different scales in a specially designed filtration process. Molecular structural and interactional features are generated from PerSpect attributes, which are statistical and combinatorial properties of spectrum information from these HL matrices. Each PerSpect attribute is input into a 1D CNN, and these CNN networks are stacked together in our PerSpect-based ensemble learning models. To the best of our knowledge, this is the first time Hodge theory has been used in ensemble learning models for PPI binding affinity upon mutations.

Methods

Topological Representations

Graph Graph or network models have been applicable to various material, chemical and biological structures and systems. Atoms and bonds are commonly interpreted as vertices and edges in such models. Mathematically, a graph representation can be defined as $G(V, E)$, where $V = \{v_i; i = 1, 2, \cdots, N\}$ is the vertex set with size N. The edges in G forms another set $E = \{e_{ij} = (v_i, v_j); 1 \leq i < j \leq N\}$. Note that graph invariants contain graph properties that does not change under graph isomorphisms (bijective mapping between two graphs). Some common graph invariants are graph order, size, clique number (clique is maximal set of nodes that is complete) and chromatic index.

Simplicial complex A simplicial complex is the extension of graph networks by including its higher dimensional counterparts such as triangles and tetrahedrons. An n-dimensional simplicial complex contains up to n-dimensional simplices. Every simplex has a finite set of vertices and can be viewed geometrically as a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), a tetrahedron (3-simplex) and in general, as a k-dimensional counterpart (k-simplex). More precisely, a k-simplex $\sigma^k = \{v_0, v_1, v_2, \cdots, v_k\}$ is defined as a convex hull formed by its $k+1$ affinely independent points $v_0, v_1, v_2, \cdots, v_k$ as follows:

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

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

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

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

Here, $[v_0, v_1, v_2, \cdots, \hat{v}_i, \cdots, v_k]$ is an oriented $(k-1)$-simplex that is generated from σ^k with v_i removed. The boundary operator maps every simplex to its faces and satisfies the equation $\partial_{k-1}\partial_k = 0$. There are many common types of simplicial complexes such as Vietoris–Rips complex, *Čech* complex, Alpha complex and clique complex. Figure 5 illustrates the graph and simplicial complex (Vietoris–Rips) for Protein PDBID:1PGB.

For simplicity of notations, we denote $\sigma_j^{k-1} \subset \sigma_i^k$ to represent that σ_j^{k-1} is a face of σ_i^k and denote $\sigma_j^{k-1} \sim \sigma_i^k$ if they have the same orientation, i.e. similarly oriented. Furthermore, we say that two k-simplices σ_i^k and σ_j^k are upper adjacent (resp. lower adjacent) neighbors, denoted as $\sigma_i^k \widetilde{\sigma}_j^k$ (resp. $\sigma_i^k \smile \sigma_j^k$), if they are both faces of a common $(k+1)$-simplex (resp. they both share a common $(k-1)$-simplex as their face). In addition, if the orientations of their common lower simplex are the same, it is called similar common lower simplex ($\sigma_i^k \smile \sigma_j^k$ and $\sigma_i^k \sim \sigma_j^k$). On the other hand, if the orientations are different, it is called dissimilar common lower simplex ($\sigma_i^k \smile \sigma_j^k$ and $\sigma_i^k \nsim \sigma_j^k$). The (upper) degree of a k-simplex σ_i^k, denoted as $d(\sigma_i^k)$, is the number of $(k+1)$-simplices, of which σ_i^k is a face.

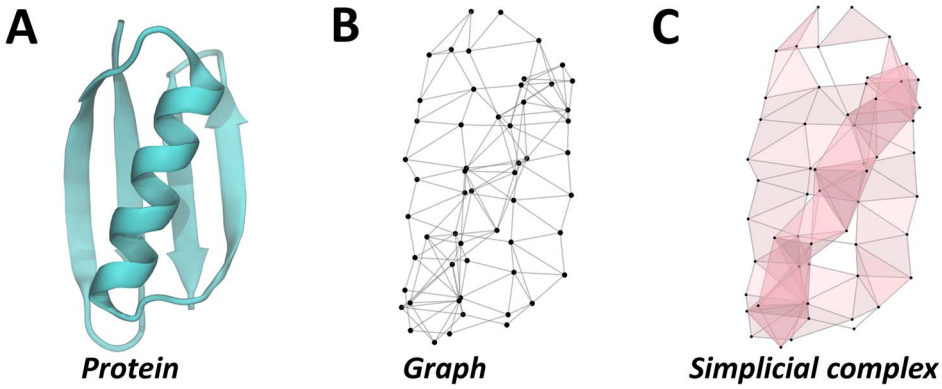

Figure 5. Illustration of graph and simplicial complex based topological representations for protein PDBID:1PGB. Note that only C_α atoms are used in graph and simplicial complex construction.

Spectral theories

The characterization, identification, comparison and analysis of structure data, from material, chemical and biological systems, are usually highly complicated due to the high dimensionality and complexity it contains. Spectral graph theory provides reduction in data dimensionality and complexity by generating compact spectral information using connectivity matrices, which originates from the structural data. These connectivity matrices consists of incidence matrix, adjacency matrix, (normalized) Laplacian matrix and Hessian matrix. Spectral information includes eigenvalues, eigenvectors, eigenfunctions and other related properties, such as Cheeger constant, edge expansion, vertex expansion, graph flow, graph random walk and heat kernel of graph. Spectral graph theory has been consistently generalized for spectral simplicial complexes where higher order connectivity matrices can be considered [1, 14, 23, 53].

Spectral Graph In spectral graph theory, a graph $G(V, E)$ can be easily represented via both an adjacency matrix and Laplacian matrix [11, 39, 56, 60]. The adjacency matrix A contains the connectivity information between any two vertices via the following:

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

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

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

Laplacian matrix, also known as the admittance matrix and Kirchoff matrix, is defined as $L = D - A$. More specifically, it can be written as

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

The Laplacian matrix has many fundamental properties. It is always positive semidefinite, which implies that all its eigenvalues are always non-negative. In fact, the number of zero eigenvalues i.e. its multiplicity, is equivalent to its topological invariant β_0, which counts the number of connected components in the graph. The second smallest eigenvalue, also known as the Fieldler value of the Laplacian matrix of the graph, describes the connectivity information of the graph. Note that Fiedler value is non-zero if and only if the graph is connected. Furthermore, the Fiedler eigenvector can decompose the graph into two well-connected subgraphs. As such, the Fiedler eigenvector and other eigenvectors corresponding to non-zero eigenvalues can also be used as features in spectral clustering. The eigenvalues and eigenvectors also forms the eigenspectrum and the study of spectral graph theory is based on the underlying properties of eigenspectrum.

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

Spectral simplicial complex The spectral simplicial complex theory is then extended to the studying of the spectral properties of HL matrices, which are constructed based on a simplicial complex instead of a graph [1, 14, 23, 42, 53]. For an oriented simplicial complex, its kth boundary matrix B_k can be defined as follows:

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

These boundary matrices satisfy the condition that $B_k B_{k+1} = 0$. The kth HL matrix can then be written as

$$L_k = \begin{cases} B_1 B_1^T, & \text{if } k = 0 \\ B_k^T B_k + B_{k+1} B_{k+1}^T, & \text{if } k \geq 1. \end{cases}$$

Furthermore, if the highest order of the simplicial complex K is n, then the nth HL matrix is $L_n = B_n^T B_n$. The above HL matrices can be explicitly described in terms of the simplex relations. More precisely, L_0 can be described as

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

which is equivalent to the graph Laplacian. Furthermore, when $k > 0$, L_k can be expressed as

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

The eigenvalues of combinatorial Laplacian matrices are independent of the choice of the orientation [23]. Furthermore, the multiplicity of zero eigenvalues, i.e. the total number of zero eigenvalues, of L_k corresponds to the kth Betti number β_k.

One can denote the kth lower HL matrix as $L_k^{\downarrow} = B_k^T B_k$ and the upper HL matrix as $L_k^{\uparrow} = B_{k+1} B_{k+1}^T$. These matrices have also been found to contain several interesting spectral properties [1]. First, the eigenvectors associated with non-zero eigenvalues of $L_k^{\uparrow}$ are orthogonal to the eigenvectors from the non-zero eigenvalues of $L_k^{\downarrow}$. Next, the non-zero eigenvalues of L_k are either the eigenvalues of $L_k^{\downarrow}$ or those of $L_k^{\uparrow}$. Consequently, the eigenvectors associated with non-zero eigenvalues of L_k are either the eigenvectors of $L_k^{\downarrow}$ or those of $L_k^{\uparrow}$.

We consider an oriented simplicial complex K_1 as in Figure 6. Its boundary operators are

$$B_1 = \begin{array}{c} \\ [1] \\ [2] \\ [3] \\ [4] \end{array} \begin{array}{c} \begin{array}{ccccc} [12] & [13] & [23] & [24] & [34] \end{array} \\ \begin{pmatrix} -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \quad B_2 = \begin{array}{c} \\ [12] \\ [13] \\ [23] \\ [24] \\ [34] \end{array} \begin{array}{c} \begin{array}{c} [234] \end{array} \\ \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 1 \end{pmatrix} \end{array}$$

Here, the vertex v_i is denoted as [i], 1-simplex $[v_i, v_j]$ is denoted as $[i,j]$ and 2-simplex $[v_i, v_j, v_k]$ is denoted as $[i,j,k]$. The corresponding HL matrices are as follows:

$$L_0 = \begin{array}{c} \\ [1] \\ [2] \\ [3] \\ [4] \end{array} \begin{array}{c} \begin{array}{cccc} [1] & [2] & [3] & [4] \end{array} \\ \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix} \end{array}$$

and

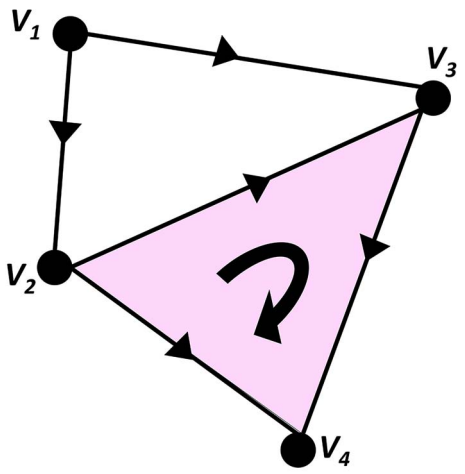

Figure 6. Illustration of the oriented simplex complex K_1.

$$L_1 = \begin{array}{c} \\ [12] \\ [13] \\ [23] \\ [24] \\ [34] \end{array} \begin{array}{c} \begin{array}{ccccc} [12] & [13] & [23] & [24] & [34] \end{array} \\ \begin{pmatrix} 2 & 1 & -1 & -1 & 0 \\ 1 & 2 & 1 & 0 & -1 \\ -1 & 1 & 3 & 0 & 0 \\ -1 & 0 & 0 & 3 & 0 \\ 0 & -1 & 0 & 0 & 3 \end{pmatrix} \end{array}$$

Note that mathematically L_k represents the topological connections in terms of upper and lower adjacency between k-simplexes.

PerSpect Theory

Filtration A multiscale representation is naturally generated via a filtration process [15]. The filtration parameter, denoted as f and key to the filtration process, is usually chosen as sphere diameter for a point cloud data, edge weights from graphs and isovalues in density-based data. Systematic increase (or decrease) of f will naturally induce a sequence of hierarchical topological representations, which can be in the form of simplicial complexes or graphs. The filtration parameter acting on a distance matrix, i.e. a matrix with entries of distance between any two vertices, can be defined with a cutoff value which corresponds to the filtration parameter in filtration process. Essentially, a 1-simplex (or edge) is formed whenever the distance between two vertices are within the filtration parameter. Hence, a gradual consistent increase (or decrease) in filtration parameter generates a series of nested simplicial complexes, with the simplicial complex produced at a smaller filtration parameter being a subset of simplices of the simplicial complex at a larger filtration parameter. With the various definitions and constructions of complexes such as Vietoris–Rips complex, Cech complex, alpha complex, cubical complex, Morse complex and clique complex, a variety of nested simplicial complexes can be constructed.

Persistent Attributes With the generation of eigenvalues from the HL matrices in a filtration process, a collection of 11 persistent attributes is computed to summarize the statistical and combinatorial properties of the eigenvalues in our feature vector. As the number of persistent

Table 3. The detailed information for the Vietoris–Rips complex based L_0 feature generation. Here, $\mathcal{P}_1^B$ and $\mathcal{P}_2^B$ are the binding site atoms of $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, $\mathcal{P}_1^{MS}$ are atoms from the mutation site (note that we assume the mutation site is on $\mathcal{P}_1$) and $\mathcal{P}_1^{MN}$ are the neighboring atoms (within 12Å) of the mutation site. Further, $I_{\mathcal{B}}$ is the interactive distance matrix in Eq. (1), which is constructed based on $\mathcal{P}_1^B$ and $\mathcal{P}_2^B$. $I_{\mathcal{M}}$ is the interactive distance matrix in Eq. (2), which is constructed based on $\mathcal{P}_1^{MS}$ and $\mathcal{P}_1^{MN}$. Moreover, all interactive distance matrices are generated based on the element-specific atom-sets, including carbon(C), nitrogen(N) and oxygen(O). Only 0-dimensional HL matrices are generated for both wild and mutant type

Type	Interaction		Distance	Complex	Dimensions
Wild and mutant	$\mathcal{P}_1^B$	$\mathcal{P}_2^B$	$I_{\mathcal{B}}$	Vietoris–Rips	L_0
Wild and mutant	$\mathcal{P}_1^{MS}$	$\mathcal{P}_1^{MN}$	$I_{\mathcal{M}}$	Vietoris–Rips	L_0

attributes considered for every HL is the same, we then obtain a long feature vector of equal size which can act as our molecular descriptor or fingerprints.

Other than the persistent multiplicity of zero-eigenvalues, we consider a set of persistent attributes for the non-zero eigenvalues. For a set of non-zero eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$, we can define the energy of a simplicial complex as the zeta function

$$\zeta(s) = \sum_{i=1}^{n} \frac{1}{\lambda_i^s} = \sum_{i=1}^{n} e^{-s \log \lambda_i}, s \in \mathbb{C}.$$

The zeta function, which is also introduced in [28], is interesting for particular values of s as it represents specific molecular descriptors. For instance, $\zeta(-m) = \sum_{i=1}^{n} \lambda_i^m, m \in \mathbb{Z}$, refers to the m-th spectral moments of HL matrices. In particular, $\zeta(-1)$ also refers to the Laplacian graph energy. In total, we consider the following persistent attributes/statistics for the featurization of each given set of eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$:

- Multiplicity of zero-eigenvalue
- $\min\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$, also known as the Fiedler value.
- $\max\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$
- $\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i = \frac{1}{n}\zeta(-1)$.
- Standard Deviation
- Laplacian Graph Energy $\zeta(-1)$.
- Generalized Mean Graph Energy $\sum_{i=1}^{n} \frac{|\lambda_i - \bar{\lambda}|}{n}$.
- Spectral second Moment $\zeta(-2)$.
- $\zeta(2) = \sum_{i=1}^{n} \frac{1}{\lambda_i^2}$.
- Quasi-Wiener Index $(n+1)\zeta(1)$.
- Spanning Tree Number $\log[\frac{1}{n+1} \cdot \prod_{i=1}^{n} \lambda_i]$.

PPIs binding affinity changes prediction with PerSpect-EL

PerSpect-based PPI characterization The AB-Bind S645 PDB files can be downloaded from the TopNetTree data. The SKEMPI-1131 PDB files can be downloaded from SKEMPI database (https://life.bsc.es/pid/skempi2/). The 'scap' utility inside the Jackal software [67] is used to generate all the mutated structures needed in AB-Bind S645 and SKEMPI-1131 datasets. For a given backbone in the structure, the scap utility predicts the side-chain conformations and the profix utility fixes any missing atoms and residues in the raw pdb files.

In our PerSpect models, we consider molecular features from two types of simplicial complexes, i.e. Vietoris–Rips complex and Alpha complex. As illustrated in Figure 3, the L_0 features are generated from Vietoris–Rips complex, and L_1 and L_2 features are generated from Alpha complex. The detailed information for Vietoris–Rips based L_0 feature generation is listed in Table 3. We use interactive distance matrices $I_{\mathcal{B}}$ in Eq. (1) and $I_{\mathcal{M}}$ in Eq. (2) to generate a series of Vietoris–Rips complex. Note that all interactive distance matrices are based on three types of element-specific atom-sets, including carbon(C), nitrogen(N) and oxygen(O). Only L_0 features are generated for both wild and mutant type. Computationally, Laplacian matrices L_0 are generated with a step size of 0.25 Å. A summary of 48 HL matrices are generated from each filtration process (with a filtration size 12 Å). The total feature size (for Vietoris–Rips based features) is 19 008 = 11(attributes)×48(stepsize)×9(atom-atom combinations)×2(wild and mutant types)×2(mutation and binding sites).

The detailed information for Alpha-complex based L_1 and L_2 feature generation is listed in Table 4. Different from interactive distance matrix based Vietoris–Rips complex, Alpha complexes are directly generated from six different types of atom sets, including $\mathcal{P}_1^B$, $\mathcal{P}_2^B$, $\mathcal{P}_1^B \cup \mathcal{P}_2^B$, $\mathcal{P}_1^{MS}$, $\mathcal{P}_1^{MN}$ and $\mathcal{P}_1^{MS} \cup \mathcal{P}_1^{MN}$. Note that seven element-specific models are considered, including {{C}, {N}, {O}, {C,N}, {C,O}, {N,O}, {C,N,O}}. Computationally, a filtration process from 1.0 to 9.0Å is considered and a total of 160 HL matrices are generated (for L_1 and L_2). Only the persistent multiplicities of L_1 and L_2 are used and the total size of Alpha-complex based features is 13 440=160(stepsize)×7(element-specific models)×6(atom sets)×2(wild and mutant types).

PerSpect-EL models In our PerSpect-EL models, the base learners use both GBT and 1D CNN models. The detailed hyperparameters for GBT is demonstrated in Table 5. The general architecture for 1D CNN models are demonstrated in Figure 7. The CNN hyperparameters are as follows:

- Parametric ReLU on all four convolutional layers.
- Dropout(0.1)
- First layer weight initialized by he_normal.
- Rest of weights initialized by lecun_uniform
- Batch size = 8, 2000 epochs.
- Adam Optimizer with lr=1e-4.

Table 4. The detailed information for the Alpha-complex based L_1 and L_2 feature generation. Here, $\mathcal{P}_1^B$ and $\mathcal{P}_2^B$ are the binding site atoms of $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively, $\mathcal{P}_1^{MS}$ are atoms from the mutation site and $\mathcal{P}_1^{MN}$ are the neighboring atoms (within 12Å) of the mutation site. Alpha complexes are constructed based on these different types of atom sets. Moreover, element-specific models, including {{C}, {N}, {O}, {C,N}, {C,O}, {N,O}, {C,N,O}}, are considered. Both L_1 and L_2 are generated for wild- and mutant-type complexes

Type	Point Cloud	Distance	Complex	Dimensions
Wild and mutant	$\mathcal{P}_1^B$	Euclidean	Alpha	L_1/L_2
Wild and mutant	$\mathcal{P}_2^B$	Euclidean	Alpha	L_1/L_2
Wild and mutant	$\mathcal{P}_1^B \cup \mathcal{P}_2^B$	Euclidean	Alpha	L_1/L_2
Wild and mutant	$\mathcal{P}_1^{MS}$	Euclidean	Alpha	L_1/L_2
Wild and mutant	$\mathcal{P}_1^{MN}$	Euclidean	Alpha	L_1/L_2
Wild and mutant	$\mathcal{P}_1^{MS} \cup \mathcal{P}_1^{MN}$	Euclidean	Alpha	L_1/L_2

Table 5. Hyperparameter settings for the GBT. These hyperparameters are used in Base Learner C.1 and all Meta Learners

No. of estimators	Max depth	Minimum sample split	Learning rate
4000	7	2	0.01
Loss function	Max features	Subsample size	Repetition
least square	square root	0.7	10 times

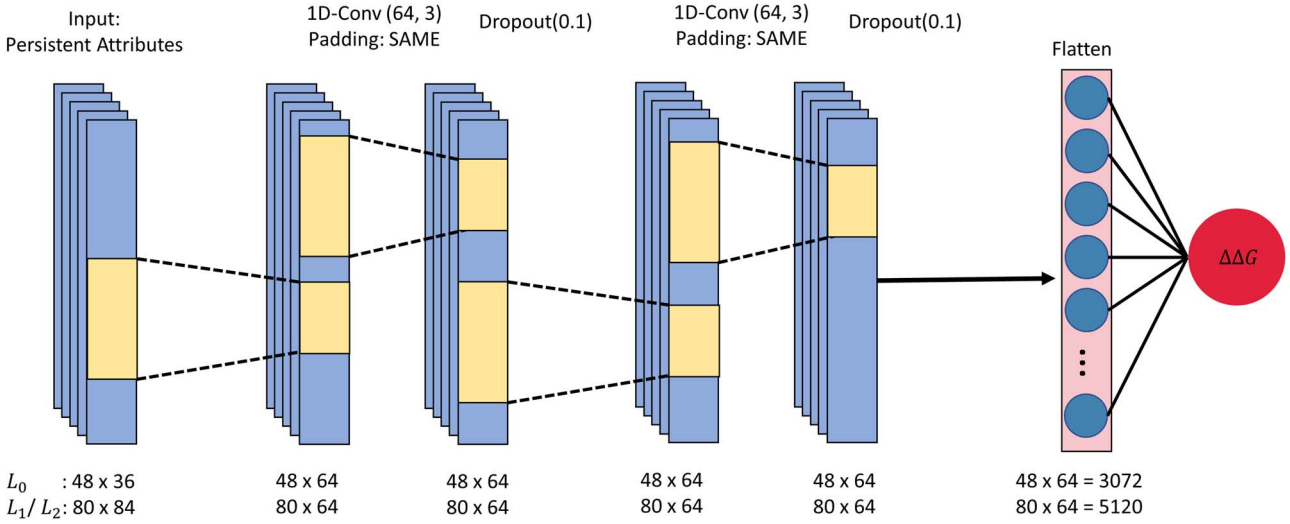

Figure 7. Details of CNN architectures for Base Learner A and Base Learner B in Figure 3. Each persistent attribute generates an individual Base Learner. Two different types of CNN architectures are considered, one for L_0 based features and the other for L_1 and L_2 based features.

Further, auxiliary features, which contain chemical and physical descriptors, are also considered in our PerSpect-EL (M3) models. The auxiliary features act as input features for a GBT base learner model. The prediction outputs from this base learner model are then stacked with the outputs from base learner CNN models, serving as inputs for the meta-learner. The auxiliary features have feature size of 707, which contains mostly atom-level and residue-level features [63]. A more detailed discussion of the auxiliary features can be found in the Supplementary information.

Key Points

Our main contributions in this paper are as follows:

- We develop persistent spectral (PerSpect) based protein–protein interaction (PPI) representation and featurization.
- We propose PerSpect-based ensemble learning (PerSpect-EL) models for PPI binding affinity prediction for the first time.
- We test our model on the two most commonly used datasets, i.e. SKEMPI and AB-Bind. It has been found that our model can achieve state-of-the-art results and outperform all existing models to the best of our knowledge.
- Our model demonstrates great potential in the analysis of mutation effects in PPI binding affinity.

Code and Data Availability

The PerSpect-EL models can be found in https://github.com/ExpectozJJ/PerSpect-Ensemble-Learning. The dataset SKEMPI-1131 can be found in https://life.bsc.es/pid/skempi2 and the dataset AB-Bind S645 can

be found in https://github.com/ExpectozJJ/PerSpect-Ensemble-Learning/tree/main/AB-BindS645.

Author contributions statement

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

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

The computational work for this article was partially done on resources of the National Supercomputing Computer, Singapore (https://www.nscc.sg). This work was supported in part by Nanyang Technological University Startup Grant M4081842.110, Singapore Ministry of Education Academic Research fund Tier 1 RG109/19, Tier 2 MOE-T2EP20120-0013 and MOE-T2EP20220-0010.

References

1. Barbarossa S, Sardellitti S. Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing* 2020; **68**:2992–3007.
2. Brender JR, Zhang Y. Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS Comput Biol* 2015;**11**(10):e1004494.
3. Bronstein MM, Bruna J, Cohen T, *et al.* Geometric deep learning: Grids, groups, graphs, geodesics, and gauges arXiv preprint arXiv:2104.13478. 2021.
4. 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.
5. 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.
6. 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* 10.1002/cnm.29142017.
7. 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.
8. 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.
9. Dong C, Gao K, Nguyen DD, *et al.* Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat Commun* 2021;**12**(1):1–9.
10. Chen J, Wang R, Wang M, *et al.* Mutations strengthened SARS-CoV-2 infectivity. *J Mol Biol* 2020;**432**(19):5212–26.
11. Chung F. *Spectral graph theory*. American Mathematical Society, 1997.
12. Dehouck Y, Kwasigroch JM, Rooman M, *et al.* BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Res* 2013;**41**(W1):W333–9.
13. Dourado DFAR, Flores SC. A multiscale approach to predicting affinity changes in protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics* 2014;**82**(10):2681–90.
14. Eckmann B. Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici* 1944;**17**(1):240–55.
15. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geom* 2002;**28**:511–33.
16. Gainza P, Sverrisson F, Monti F, *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**(2):184–92.
17. Gao K, Nguyen DD, Meihua T, *et al.* Generative network complex for the automated generation of drug-like molecules. *J Chem Inf Model* 2020;**60**(12):5682–98.
18. Geng C, Vangone A, Bonvin AMJJ. Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Engineering, Design and Selection* 2016;**29**(8):291–9.
19. Geng C, Vangone A, Folkers GE, *et al.* iSEE: interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function, and Bioinformatics* 2019;**87**(2):110–9.
20. Geng C, Xue LC, Roel-Touris J, *et al.* Finding the $\delta\delta$g spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2019;**9**(5):e1410.
21. Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. *PLoS Comput Biol* 2012;**8**(12):e1002819.
22. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 2002;**320**(2):369–87.
23. Horak D, Jost J. Spectra of combinatorial Laplace operators on simplicial complexes. *Advances in Mathematics* 2013;**244**:303–36.
24. Jankauskaitė J, Jiménez-García B, Dapkūnas J, *et al.* SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 2019;**35**(3):462–9.
25. Jemimah S, Sekijima M, Michael M, *et al.* ProAffiMuSeq: sequence-based method to predict the binding free energy change of protein–protein complexes upon mutation using functional classification. *Bioinformatics* 2020;**36**(6):1725–30.
26. Sherlyn Jemimah K, Yugandhar, and M Michael Gromiha. PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics* 2017;**33**(17): 2787–8.
27. Jiang J, Wang R, Wei G-W. GGL-Tox: geometric graph learning for toxicity prediction. *J Chem Inf Model* 2021;**61**(4):1691–700.
28. Knill O. The dirac operator of a graph arXiv preprint arXiv:1306.2166. 2013.
29. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proc Natl Acad Sci* 2002;**99**(22):14116–21.
30. Kumar MDS, Gromiha MM. PINT: protein–protein interactions thermodynamic database. *Nucleic Acids Res* 2006;**34**(suppl_1): D195–8.
31. Liu Q, Chen P, Wang B, *et al.* dbMPIKT: a web resource for the kinetic and thermodynamic database of mutant protein interactions arXiv preprint arXiv:1708.01857. 2017.
32. Liu S, Zhang C, Zhou H, *et al.* A physical reference state unifies the structure-derived potential of mean force for protein

folding and binding. *Proteins: Structure, Function, and Bioinformatics* 2004;**56**(1):93–101.
33. Liu X, Wang XJ, Wu J, *et al.* Hypergraph based persistent cohomology (HPC) for molecular representations in drug design. *Brief Bioinform* 2021.
34. Liu X, Luo Y, Li P, *et al.* Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS Comput Biol* 2021;**17**(8):e1009284.
35. Lo YC, Rensi SE, Torng W, *et al.* Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 2018;**23**(8):1538–46.
36. Meng Z, Xia K. Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Sci Adv* 2021;**7**(19):eabc5329.
37. Moal IH, Fernández-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics* 2012;**28**(20):2600–7.
38. Moal IH, Fernandez-Recio J. Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation. *Journal of Chemical Theory and Computation* 2013;**9**(8):3715–27.
39. Mohar B, Alavi Y, Chartrand G, *et al.* The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications* 1991;**2**(871–898):12.
40. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. *Nat Methods* 2013;**10**(1):47–53.
41. 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.
42. Mukherjee S, Steenbergen J. Random walks on simplicial complexes and harmonics. *Random structures & algorithms* 2016;**49**(2):379–405.
43. Nguyen DD, Cang ZX, Wei GW. A review of mathematical representations of biomolecular data. *Phys Chem Chem Phys* 2020.
44. 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.
45. Nguyen DD, Cang ZX, Wu KD, *et al.* Wei. 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.
46. 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.
47. 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.
48. 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.
49. Petukh M, Dai L, Alexov E. Saambe: webserver to predict the charge of binding free energy caused by amino acids mutations. *Int J Mol Sci* 2016;**17**(4):547.
50. Puzyn T, Leszczynski J, Cronin MT. *Recent advances in QSAR studies: methods and applications*, Vol. **8**. Springer Science & Business Media, 2010.
51. Rebsamen M, Kandasamy RK, Superti-Furga G. Protein interaction networks in innate immunity. *Trends Immunol* 2013;**34**(12):610–9.
52. Rodrigues CHM, Myung Y, Pires DEV, *et al.* mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res* 2019;**47**(W1):W338–44.
53. Schaub MT, Benson AR, Horn P, *et al.* Random walks on simplicial complexes and the normalized hodge 1-Laplacian. *SIAM Review* 2020;**62**(2):353–91.
54. Shi Q, Chen W, Huang S, *et al.* Deep learning for mining protein data. *Brief Bioinform* 2021;**22**(1):194–218.
55. Sirin S, Apgar JR, Bennett EM, *et al.* AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Sci* 2016;**25**(2):393–409.
56. Spielman DA. Spectral graph theory and its applications. In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 2007, 29–38.
57. Strokach A, Lu TY, Kim PM. ELASPIC2 (EL2): combining contextualized language models and graph neural networks to predict effects of mutations. *J Mol Biol* 2021;**433**(11):166810.
58. Szilagyi A, Zhang Y. Template-based structure modeling of protein–protein interactions. *Curr Opin Struct Biol* 2014;**24**:10–23.
59. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 2001;**17**(3):284–5.
60. Von Luxburg U. A tutorial on spectral clustering. *Statistics and computing* 2007;**17**(4):395–416.
61. 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.
62. Wang B, Zhao ZX, Wei GW. Automatic parametrization of nonpolar implicit solvent models for the blind prediction of solvation free energies. *J Chem Phys* 2016;**145**(12):124110.
63. Wang M, Cang Z, Wei G-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence* 2020;**2**(2):116–23.
64. Wang R, Hozumi Y, Yin C, *et al.* Mutations on COVID-19 diagnostic targets. *Genomics* 2020;**112**(6):5204–13.
65. Wu KD, Wei GW. Quantitative toxicity prediction using topology based multi-task deep neural networks. *J Chem Inf Model* 10.1021/acs.jcim.7b005582018.
66. 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.
67. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* 2001;**311**(2): 421–430–405.
68. Xiong P, Zhang C, Zheng W, *et al.* Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J Mol Biol* 2017;**429**(3):426–34.
69. Zhang N, Chen Y, Lu H, *et al.* MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience* 2020;**23**(3):100939.
70. 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.
71. Zhou G, Chen M, Ju CJT, *et al.* (eds). Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR genomics and bioinformatics* 2020;**2**(2):lqaa015.
72. Zomorodian A, Carlsson G. Computing persistent homology. *Discrete Comput Geom* 2005;**33**:249–74.
73. Beibei L, Wang B, Zhao R, *et al.* ESES: Software for eulerian solvent excluded surface. *J Comput Chem* 2017;**7**(38):446–66.
74. Dolinsky TJ, Czodrowski P, Li H, *et al.* PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res* 1 July 2007;**35**(suppl_2):W522–5.

75. Chen D, Chen Z, Chen C, *et al.* MIBPB: A software package for electrostatic analysis. *J Comput Chem* 2011;**32**:756–70.
76. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins: Structure, Function, and Bioinformatics* 2008;**73**(3):765–83.
77. Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, and Z. Yaoqi. Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In *Prediction of protein secondary structure* (pp. 55–63). Humana Press, New York, NY.