Persistent Tor-algebra for protein-protein interaction analysis

Xiang Liu[a,b], **Huitao Feng**[b,c], **Zhi Lü**[d], **and Kelin Xia**[a,*]

[a]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371; [b]Chern Institute of Mathematics and LPMC, Nankai University, Tianjin, China, 300071; [c]Mathematical Science Research Center, Chongqing University of Technology, Chongqing, China, 400054; [d]Institute of Mathematics, School of Mathematical Sciences, Fudan University, Shanghai, China, 200433

This manuscript was compiled on May 4, 2022

Protein-protein interactions (PPIs) play crucial roles in almost all biological processes from cell-signaling and membrane transport to metabolism and immune systems. Efficient characterization of PPIs at the molecular level is key to the fundamental understanding of PPI mechanisms. Even with the gigantic amount of PPI models from graphs, networks, geometry and topology, it remains as a great challenge to design functional models that efficiently characterize the complicated multiphysical information within PPIs. Here we propose persistent Tor-algebra (PTA) model for a unified algebraic representation of the multiphysical interactions. Mathematically, our PTA is inherently algebraic data analysis (ADA). In it, protein structures and interactions are described as a series of face rings and Tor modules, from which PTA model is developed. The multiphysical information within/between biomolecules are implicitly characterized by PTA and further represented as PTA barcodes. To test our PTA models, we consider PTA-based ensemble learning (PTA-EL) for PPI binding affinity prediction. The two most-commonly used datasets, i.e., SKEMPI and AB-Bind, are employed. It has been found that our model outperforms all the existing models as far as we know. Mathematically, our PTA model provides a highly efficient way for the characterization of molecular structures and interactions.

Persistent Tor-algebra | Protein-protein interactions | Multiphysical interactions | Ensemble learning

Protein-protein interactions (PPIs) are crucial to a wide range of biological processes and mechanisms, including cell metabolism, signaling, protein transport and immune system (1, 2). Protein mutations and genetic variations can influence protein folding, protein stability, and protein interactions, leading to diseases and drug resistance(3). The understanding of PPIs, especially PPIs upon mutations, is of great importance to various biomedical applications, such as analysis of mutation-induced diseases, drug design, and therapeutic intervention (1, 2). Experimentally, various methods and tools have been developed to determine the protein structures, such as X-ray crystallography, nuclear magnetic resonance, cryo-electron microscopy and cross-linked mass spectrometry, and to evaluate the PPI binding affinity and stability, such as isothermal titration calorimetry, surface plasmon resonance, fluorescence, and blue native polyacrylamide gel electrophoresis. However, experimental studies of protein structures and binding affinities are time-consuming, labor-intensive, and expensive. Currently, only about 6.5% of the known human interactome has structural information (4).

Efficient computational methods and models have been proposed for the PPI studies. One of the focuses is the evaluation of PPI binding affinity changes upon mutations ($\Delta\Delta G$). The models can be categorized into three groups, including molecular dynamic (MD) based approaches, statistical energy methods, and data-driven learning ones. MD-based models include FoldX (5), Rosetta (6), zone equilibration of mutants (ZEMu) (7), single amino acid mutation based change in binding free energy (SAAMBE) (8), and others (1). They usually characterize the binding affinity of PPIs with various physical energy terms, including van der Waals interactions, electrostatic energies, hydrogen bonds, solvation energy, etc. Further, mutation effects are considered by modeling conformational changes with rotamer and structure ensemble approaches in these MD-based models. Different from MD methods, statistical energy based PPI models extract various intermolecular potentials from experimental structures based on the contacts at atomic, residual, or other coarse-grained levels. These models include BindProfX (9), BeAtMuSiC (10), contact potentials (11), Profile-score (12), and Dcomplex (13). Recently, data-driven learning models have achieved state-of-the-art results in PPI analysis (14).

The great advancements of machine learning models for PPI analysis is largely due to the accumulation of PPI data in various databanks, including Alanine scanning energetics database (ASEdb) (15), protein-protein interactions thermodynamic database (PINT) (16), structural kinetic and energetic database of mutant protein interactions (SKEMPI) (17), database of binding affinity change upon mutations (DACUM)

Significance Statement

Highly-efficient molecular representations and characterizations are key to all data-driven AI models in molecular sciences. Geometric and topological models have demonstrated their great advantages over traditional models, and have given birth to newly-developed geometric data analysis and topological data analysis (TDA), respectively. Here we develop the first persistent Tor-algebra (PTA) model for a unified algebraic representation and characterization of biomolecular data. Our PTA makes use of face rings and Tor modules to describe the multiphysical information of biomolecular structures and interactions, thus it is inherently “algebraic data analysis" (ADA). The algebraic descriptors can be obtained from our PTA barcodes and further used as features for learning models. These algebraic invariant based molecular features are highly-abstract and characterize the most-intrinsic information of biomolecular data. They have a superior transferability and can be better “understood" by learning models.

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

There is no conflict of interest.

[2]To whom correspondence should be addressed. E-mail: xiakelin@ntu.edu.sg

(18), antibody-bind database (AB-Bind) (19), protein-protein complex mutation thermodynamics (PROXiMATE) (20), and kinetic and thermodynamic database of mutant protein interactions (dbMPIKT) (21). An updated version SKEMPI 2.0 has been constructed recently (22). It combines several databases including SKEMPI, AB-Bind, PROXiMATE, and dbMPIKT, together with manually curated data from the literature. This dataset consists of 7085 mutations on various types of protein complexes, such as protease-inhibitor, antibody-antigen, and TRC-pMHC complexes. More specifically, there are about 3000 single point alanine mutations, about 2000 single point non-alanine mutations, and roughly 2000 multi points mutations. The ever-increasing PPI data has given rise to various data-driven learning models (1, 14), including mCSM (23), ELASPIC (24), BindProf (25), MutaBind (26), iSEE (27), MuPIPR (28), ProAffiMuSeq (29), GeoPPI (30), and so on. In general, these data-driven models can be classified into two types, i.e., featurization-based machine learning models and end-to-end deep learning models. For the machine learning models, 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 deep learning models, proteins are usually represented as surfaces, graphs, or networks with embedded vectors or one-hot-vectors (31, 32). 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, the design of efficient molecular representation and featurization is still a big challenge for all the learning models. (33, 34).

Advanced mathematical tools, in particular topological data analysis (TDA) (35, 36), have been used in molecular representation and featurization (37–40), and their combination with learning models have achieved great successes in various steps of drug design, including protein-ligand binding affinity prediction (37, 41–44), protein stability change upon mutation prediction (39, 45), toxicity prediction (46–48), solvation free energy prediction (49, 50), partition coefficient and aqueous solubility (51), binding pocket detection (52), and drug discovery (53). Outstanding performance has been consistently achieved in D3R Grand challenge (54–56). In particular, TopNetTree has demonstrated great power in predicting binding affinity changes upon mutations (57). It outperformed all existing models and provided great insights for the SARS-CoV-2 mutations (58, 59). One of most prominent properties of these advanced mathematical tools is that they make use of either geometric or topological invariants, which are usually highly abstract and characterize the intrinsic data properties, thus have a better transferability for data-driven learning models.

Here we propose persistent Tor-algebra (PTA) for the first time. Our PTA is inherently "algebraic data analysis" (ADA), which is to represent and characterize data with algebraic concepts and methods. We have also developed the first PTA-based molecular representation and featurization, and PTA-based ensemble learning (PTA-EL) for the prediction of PPI binding affinity change upon mutation. Mathematically, Tor-algebra is a homotopy invariant from homological algebra (60, 61). It provides an algebraic representation and characterization of topological structures in terms of rings, modules, and their homological invariants. In our PTA model, the protein-protein complex structures are modeled as a series of Vietoris-Rips complexes. Face rings and Tor modules are defined on these complexes and persistent Tor-algebra are developed. Algebraic descriptors can be obtained from the PTA barcode and further used as input features for our ensemble learning model. More specifically, 72 PTA features are fed into 72 1D convolutional neural network (CNN) models, separately. Other than these CNN meta learners, a total of 72 gradient boosting tree (GBT) meta learners can be constructed through the combination of PTA features with auxiliary features from molecular physical properties. Our PTA-EL model is systematically validated on the two most-commonly used datasets, i.e., SKEMPI and AB-Bind datasets. It has been found that our model can outperform all existing models for the prediction of PPI binding affinity change upon mutation, as far as we know. Our model also has a great potential for the analysis and design of efficient antibody for SARS-CoV-2.

Results

Persistent Tor-algebra model. Mathematically, Tor-algebra is a powerful homotopy invariant. When two simplicial complexes share the same Tor-algebra, they are homotopy equivalent, i.e., they can continuously deform to each other. Tor-algebra is from the Tor-functor, which one of the central concepts of homological algebra. It is directly related to the moment-angle complex in toric topology. More specifically, Tor-algebra of a simplicial complex is isomorphic to the integral cohomology of the corresponding moment-angle complex. Essentially, Tor-algebra and moment-angle complex characterize the geometric and algebraic aspects of the same simplicial complex, respectively. Computationally, Tor-algebra of a simplicial complex can be seen as a combination of the reduced simplicial cohomology of its subcomplexes. Persistent Tor-algebra is a multiscale model that combines Tor-algebra with a filtration process.

Tor-algebra for simplicial complex. Given a simplicial complex $\mathcal{K}$ with vertex set $\{v_1, v_2, ..., v_m\} = [m]$, its face ring (or the Stanley-Reisner ring) $\mathbb{F}[\mathcal{K}]$ over the coefficient field $\mathbb{F}$ can be naturally constructed (61). The face ring $\mathbb{F}[\mathcal{K}]$ has the $\mathbb{F}$-vector space basis consisting of monomials $v_{j_1}^{\alpha_1}...v_{j_k}^{\alpha_k}$ where α_i is a positive integer and $\{v_{j_1}, ..., v_{j_k}\}$ is a simplex of $\mathcal{K}$. The face ring can uniquely determine its underlying simplicial complex.

For the face ring $\mathbb{F}[\mathcal{K}]$ of simplicial complex $\mathcal{K}$, a Tor module $\mathrm{Tor}_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F})$ can be derived where $\mathbb{F}[m]$ is the polynomial algebra $\mathbb{F}[v_1, v_2, ..., v_m]$ with degree 2 for each v_i. This is called the Tor-algebra of the simplicial complex $\mathcal{K}$. It is a homotopy invariant, and two simplicial complexes share the same Tor-algebra if they can deform to each other. For simplicity, we denote $\mathrm{Tor}_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F})$ as $\mathrm{Tor}(\mathcal{K})$. Tor-algebra $\mathrm{Tor}(\mathcal{K})$ is directly related to the simplicial cohomology of the subcomplexes of $\mathcal{K}$. Mathematically, there is a decomposition $\mathrm{Tor}(\mathcal{K}) = \bigoplus_{i,j>=0} \mathrm{Tor}^{-i,2j}(\mathcal{K})$ where $\mathrm{Tor}^{-i,2j}(\mathcal{K})$ is the $(-i, 2j)$-grade component of $\mathrm{Tor}(\mathcal{K})$. The reduced simplicial cohomology of all the full subcomplexes of $\mathcal{K}$ with j vertices uniquely determines the $(-i, 2j)$-grade component $\mathrm{Tor}^{-i,2j}(\mathcal{K})$.

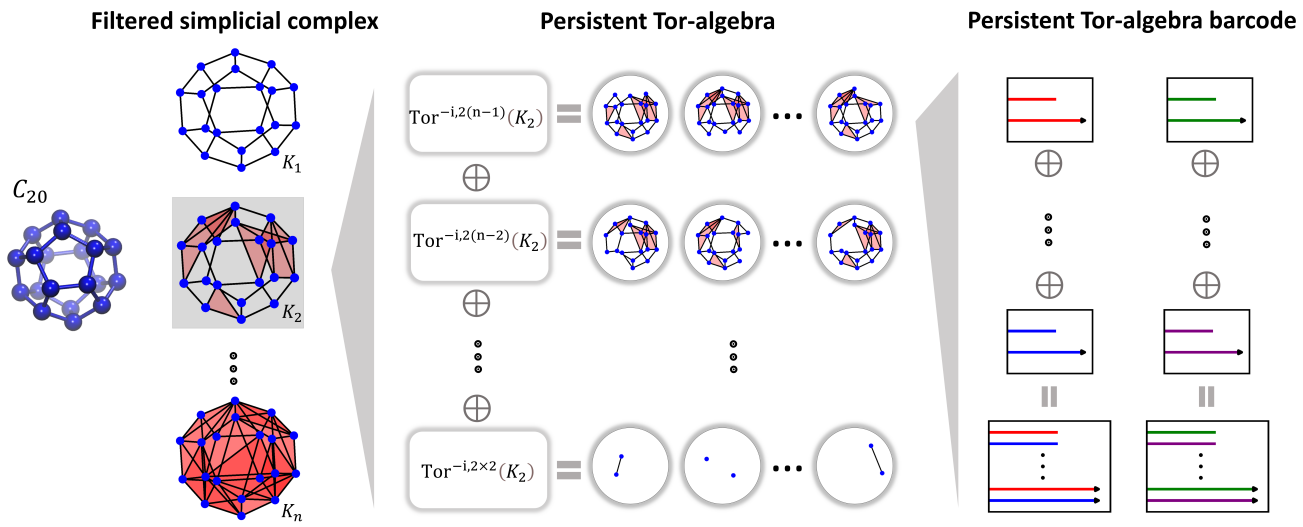

Fig. 1. Persistent Tor-algebra model for C_{20} fullerene molecule. C_{20} is represented as a filtration of simplicial complexes $\mathcal{K}_1 \subset \mathcal{K}_2 \subset ... \subset \mathcal{K}_n$. For a specific simplicial complex $\mathcal{K}_i$, its Tor-algebra $\mathrm{Tor}(\mathcal{K}_i)$ is the direct sum of all the $(-i, 2j)$-grade components $\mathrm{Tor}^{-i,2j}(\mathcal{K}_i)$ with $i, j \geq 0$ as the vertical direction of "Persistent Tor-algebra" part. And a specific component $\mathrm{Tor}^{-i,2j}(\mathcal{K}_i)$ is the direct sum of the reduced simplicial cohomology of the subcomplexes of $\mathcal{K}_i$ where each subcomplex has exactly j vertices as the horizontal direction. Such as the last row of component $\mathrm{Tor}^{-i,2\times 2}(\mathcal{K}_2)$, j is 2 so that each subcomplex has exactly two vertices. The filtration of simplicial complexes $\mathcal{K}_1 \subset \mathcal{K}_2 \subset ... \subset \mathcal{K}_n$ naturally induce filtration process of their subcomplexes. Hence when we get persistent Tor-algebra from the filtration of simplicial complexes, we also get persistent cohomology from the induced filtrations of subcomplexes. From the decomposition, the persistent Tor-algebra barcode is exactly the direct sum of the persistent cohomology barcode of the subcomplexes.

More specifically,

$$\mathrm{Tor}^{-i,2j}(\mathcal{K}) = \bigoplus_{J\subset[m],|J|=j} \hat{H}^{j-i-1}(\mathcal{K}_J, \mathbb{F}) \qquad [1]$$

where $[m]$ is the vertex set $\{v_1, ...v_m\}$, J is a vertex subset, $\mathcal{K}_J$ is a subcomplex of $\mathcal{K}$ obtained by restricting to $J \subset [m]$ and $\hat{H}^{j-i-1}(\mathcal{K}_J, \mathbb{F})$ is the reduced simplicial cohomology of $\mathcal{K}_J$ in dimension $j-i-1$. The decomposition in Equation (1) provides a bridge between Tor-algebra of the entire simplicial complex and the reduced simplicial cohomology from its local subcomplexes. Essentially, the Tor-algebra of a simplicial complex can be regarded as a combination of the reduced simplicial cohomology of its subcomplexes.

The "Persistent Tor-algebra" part of Figure 1 gives an illustration of the decompisition. C_{20} is represented as a series of simplicial complexes. For the simplicial complex $\mathcal{K}_2$, its Tor-algebra $\mathrm{Tor}(\mathcal{K}_2)$ is the direct sum of all possible $(-i, 2j)$-grade component $\mathrm{Tor}^{-i,2j}(\mathcal{K}_2)$ with $i, j \geq 0$ as the vertical direction. And a specific component $\mathrm{Tor}^{-i,2j}(\mathcal{K}_2)$ is the direct sum of the reduced simplicial cohomology of all the subcomplexes of $\mathcal{K}_2$ where each subcomplex has exactly j vertices and the dimension of cohomology is $j-i-1$ as the horizontal direction. Such as the last row of component $\mathrm{Tor}^{-i,2\times 2}(\mathcal{K}_2)$, j is 2 so that all subcomplexes have exactly two vertices.

Persistent Tor-algebra. Motivated by the great success of topological data analysis, in particular, persistent homology model, we develop the persistent Tor-algebra model, for the first time. The key point of all the persistent models, including persistent homology, persistent spectral, persistent curvatures etc., is the filtration process that gives a multiscale characterization of the data (40, 62, 63). Here, we propose persistent Tor-algebra, for the first time, through the combination of Tor-algebra and the filtration process. More specifically, assume we have a filtration of simplicial complex. That is a sequence of nested simplicial complexes connected by inclusions

$$\emptyset = \mathcal{K}_0 \to \mathcal{K}_1 \to ... \to \mathcal{K}_n = \mathcal{K}$$

where $\mathcal{K}_i$ is a subcomplex of $\mathcal{K}_{i+1}$. For each simplicial complex $\mathcal{K}_i$, its face ring $\mathbb{F}(\mathcal{K}_i)$ over coefficient $\mathbb{F}$ can be constructed, which can uniquely determine its underlying simplicial complex. Hence a series of face rings can be derived

$$\mathbb{F}(\mathcal{K}_0) \to \mathbb{F}(\mathcal{K}_1) \to ... \to \mathbb{F}(\mathcal{K}_n)$$

where two adjacent face rings are connected by the homomorphism induced from the inclusion map. Also, for each face ring $\mathbb{F}(\mathcal{K}_i)$, its Tor-algebra $\mathrm{Tor}(\mathcal{K}_i)$ can be constructed and the homomorphism naturally induce homomorphism between the Tor-algebra of two adjacent face rings. Consequently, we get a sequence of Tor-algebra connected by homomorphisms

$$\mathrm{Tor}(\mathcal{K}_0) \to \mathrm{Tor}(\mathcal{K}_1) \to ... \to \mathrm{Tor}(\mathcal{K}_n)$$

We call this sequence of Tor-algebra together with the homomorphisms as the persistent Tor-algebra. Further, by considering a specific graded component of Tor-algebra modules, we can get the persistent Tor-algebra for a given index pair $(-i, 2j)$. We have

$$\mathrm{Tor}^{-i,2j}(\mathcal{K}_1) \to \mathrm{Tor}^{-i,2j}(\mathcal{K}_2) \to ... \to \mathrm{Tor}^{-i,2j}(\mathcal{K}_n)$$

This is called the persistent Tor-algebra of $\mathcal{K}$ in the $(-i, 2j)$-th graded component. These Tor-algebras form a persistent module that can be represented as a persistent barcode or persistent diagram. The bars in persistent barcode and the points in persistent diagram reflect the birth, death and evolution process of the Tor-algebra through the filtration process.

Figure 1 shows a persistent Tor-algebra barcode representation for a filtration of C_{20}. C_{20} is represented as a filtration of

simplicial complexes $\mathcal{K}_1 \subset \mathcal{K}_2 \subset ... \subset \mathcal{K}_n$. From the decomposition, the Tor-algebra of a specific simplicial complex $\mathcal{K}_i$ is the direct sum of the cohomology of the subcomplexes of $\mathcal{K}_i$. The filtration naturally induce filtration processes of their subcomplexes. Consequently, when we get persistent Tor-algebra from the filtration of the simplicial complexes $\mathcal{K}_1 \subset \mathcal{K}_2 \subset ... \subset \mathcal{K}_n$, we also get persistent cohomology from the induced filtration of their subcomplexes. From the decomposition, we know that the persistent Tor-algebra barcode is exactly the direct sum of the persistent cohomology barcode of the subcomplexes.

Persistent Tor-algebra for protein-protein interaction analysis.

Persistent Tor-algebra based molecular representation and featurization. Molecular representation and featurization are of great importance for the analysis of molecular data from material, chemistry and biology. Recently, simplicial complex has been used in molecular representation, especially in drug design, the derived persistent homology theory has shown great power and is being hotly studied. A simplicial complex is a set of vertices, edges, triangles and higher dimensional counterparts which are glued together along their faces. Physically, a vertice can represent a molecular atom, residue, or even the whole molecule. An edge can represent the interactions of various kinds between two vertices including covalent bonds, electrostatic, and other non-covalent forces. The triangles, tetrahedrons and other higher dimensional counterparts can represent "many-body" interactions among several vertices, which characterize the higher dimensional structures of the molecules.

Protein-protein complexes are usually of great sizes, our aim is to predict the binding affinity change ($\Delta\Delta G$) following mutations. And the mutation sites are very small, usually no more than 10 residues. So only protein atoms near mutation sites are considered to reduce computational cost and avoid the irrelevant information.

More specifically, for each protein-protein complex, protein atoms within 10Å of the mutation sites are considered. We use the element-specific representations, six atom combinations are extracted, including $\{C\}$, $\{N\}$, $\{O\}$, $\{C, N\}$, $\{C, O\}$, and $\{N, O\}$. Both the wild type and mutated type of the protein structure are considered. So there are totally 12 atom combinations for each protein-protein complex. For each atom combination, a Vietoris-Rips complex $\mathcal{K}$ and its associated face ring are constructed according to the atom coordinates. Then six persistent Tor-algebra components are computed, the indexes are $(1, 2)$, $(2, 3)$, $(n-1, n)$, $(n-2, n)$, $(n-2, n-1)$ and $(n-3, n-1)$ as (i, j) for $\mathrm{Tor}^{-i,2j}(\mathcal{K})$ where n is the vertex number of $\mathcal{K}$. So totally $72 = 12 \times 6$ persistent Tor-algebra components are generated for each protein-protein complex. These persistent Tor-algebra can be represented as persistent barcode. There are various kinds of methods discretizing the persistent barcode into feature vectors, including binning approaches, barcode statistic, algebraic functions, etc. Here we consider the barcode statistic. More specifically, the filtration region [0Å,10Å] is divided into 40 equal-sized bins with grid size 0.25Å. For the (i, j) values of $(1, 2)$, $(2, 3)$, $(n-1, n)$ and $(n-2, n-1)$, the values of right endpoints of the bars in each bin are considered. And for (i, j) values of $(n-2, n)$ and $(n-3, n-1)$, the values of both the left and right endpoints of the bars in each bin are considered. So 40 sets of real values are generated for each persistent Tor-algebra component with a specific (i, j) index. Then six statistic values, including maximal value, minimal value, average value, sum, standard deviation and the number of elements are considered. Hence a feature of $240 = 40 \times 6$ columns are generated for each persistent Tor-algebra with a specific (i, j) index. In all, for a protein-protein complex, totally 72 topological features of size 240 are generated. Besides these topological features, based on (57), 707 auxiliary features are also considered in our model.

Benchmark datasets. Two datasets, including AB-Bind dataset and SKEMPI dataset, are considered in our benchmark tests. There are 1101 mutational data points with experimental determined binding affinities in AB-Bind dataset (19). Among them, 645 single-point mutations across 29 antibody-antigen complexes are collected and called AB-Bind S645 set, this subset consists of 20% stabilizing mutations and 80% destabilizing ones (57, 64). Note that there are 27 non-binders in this datasets (without experimental binding affinities), and their binding affinity changes are set to be 8 kcal/mol (57, 64). The SKEMPI dataset contains 3047 binding free energy change upon mutations (17), it consists of single-point mutations and multi-points mutations. The 2317 single-point mutation entries are referred to as the SKEMPI S2317 set. A set of 1131 protein-protein complexes from SKEMPI S2317, which have non-redundant interface single-point mutations, are selected and and called SKEMPI S1131 dataset(9). Both AB-Bind S645 dataset and SKEMPI S1131 dataset are widely used in the benchmark of machine learning models for PPIs (57, 64).

Persistent Tor-algebra based ensemble learning models. The framework of our model is illustrated in Figure 2. Firstly, for the PPI data, some simplicial complexes and associated face rings are generated from the atom coordinates. These face rings naturally give an algebraic representation of the structure and interactions of PPI data. Then persistent Tor-algebra are calculated to give quantitative descriptions of the data which can also be represented as persistent barcodes. Finally, these persistent Tor-algebra barcodes are used as molecular descriptors for machine learning models. More specifically, Vietoris-Rips complex is used to give a simplicial complex representation, six types of persistent Tor-algebra barcodes are calculated in our featurization and a stacking ensemble learning model is used to do the prediction.

Stacking is an ensemble machine learning algorithm that uses a meta learner to learn how to best combine the predictions from several base learners. Our persistent Tor-algebra based stacking models consist of two types of base learners, 1D convolutional neural network and gradient boosting tree, and two meta learners. The first type base learners consist of 72 1D CNN models with 72 types of Tor-algebra features as inputs separately. Besides the topological features, some precalculated auxiliary features from physical and chemical properties are combined with these topological features to form another 72 features, and these features are inputs for 72 gradient boosting tree models, the second type base learners.

We stack all the 144 base learners with a gradient boosting tree model and denote the model as PTA-EL(M2). We also studied another stacking model PTA-EL(M1) which only considers the first type base learners, that is only using Tor-algebra based 1D CNN to make the base learner predictions. Computationally, all the 72 1D CNN models have the same

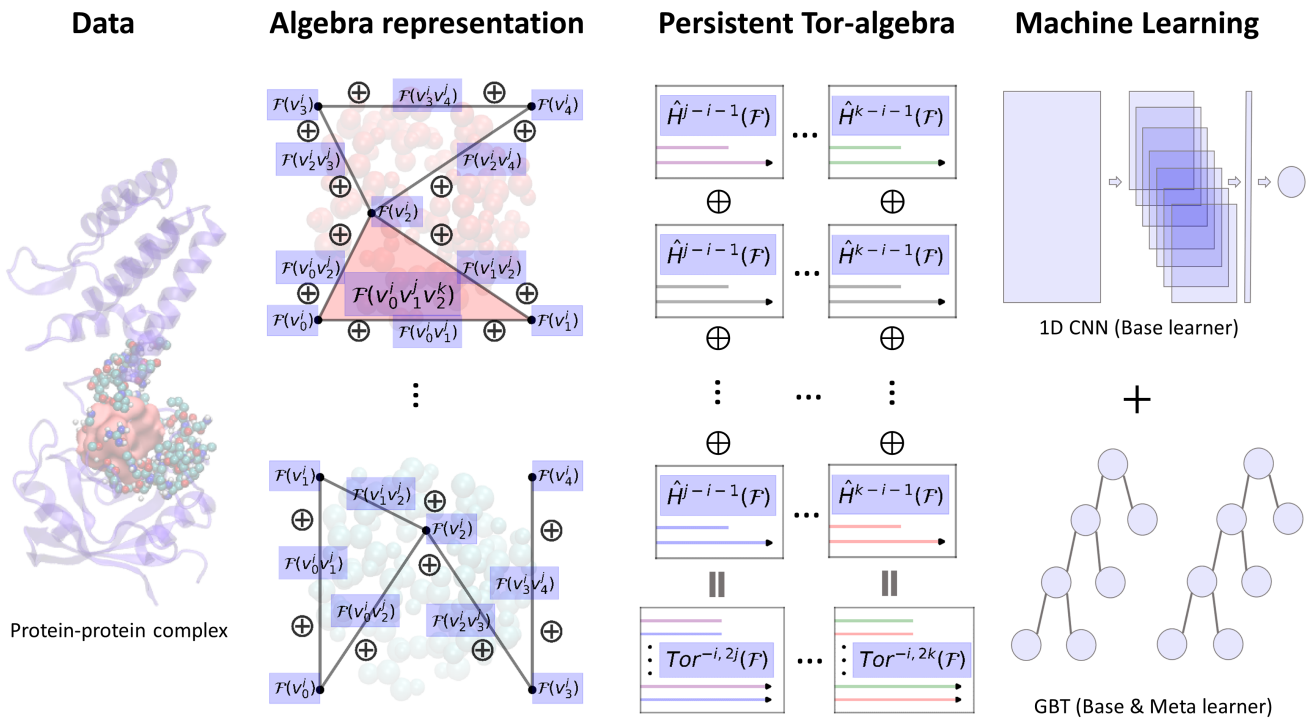

Fig. 2. Illustration of our persistent Tor-algebra based machine learning models. Firstly, simplicial complexes and associated face rings are generated to characterize the structure and interactions of the PPI data. Then, persistent Tor-algebra are calculated to give quantitative featurization of the data, whcih can be represented as persistent barcode. Finally, the persistent Tor-algebra barcode are used as molecular descriptors for machine learning models. More specifically, element specific Vietoris-Rips complex is used to give simplicial complex representation and six types of persistent Tor-algebra barcode are calculated. Stacking ensemble algorithm is used in machine learning, which consists of two types of base learners, 1D CNN and gradient boosting tree, and a meta learner, gradient boosting tree.

architecture and hyperparameters, and all the 72 GBT models also share the same parameters. In the performance assessment of our model, Pearson correlation coefficient (PCC) and root-mean-square error (RMSE) are used to assess the quality of prediction. Tenfold cross-validation is used to do the regression. Ten independent regressions are performed and the average value of PCC and RMSE are used as the measurement of the performance of our model.

Performance on SKEMPI S1131 dataset Table 1 illustrates the PCCs for all the machine learning models, as far as we know, on SKEMPI S1131 dataset using tenfold cross-validation. It

Table 1. Comparison of the performance between our model and other models on SKEMPI S1131 dataset.

Method	PCC
PTA-EL(M2)	**0.866**
TopNetTree	0.850
PTA-EL(M1)	**0.772**
BindProfX	0.738
Profile-score+FoldX	0.738
Profile-score	0.675
SAAMBE	0.624
FoldX	0.457
BeAtMuSic	0.272
Dcomplex	0.056

can be seen that our model has achieved the best results with PCC of 0.866 and RMSE of 1.216 kcal/mol. Detailed information of our prediction results can be found in Figure 3. In Figure 3, comparison between the experimental binding affinity changes and the predicted binding affinity changes of our PTA-EL(M2) model is illustrated in **A**. The mutations are grouped into charged, polar,hydrophobic and special cases according to the mutation types. We also studied the alanine mutations and non-alanine mutations. The distribution of experimental binding affinity changes for different groups is shown in **C**. Prediction results for these different mutation types are illustrated in **D**. Average PCC (RMSE) values of 0.917(1.354), 0.884(1.074), 0.790(1.227), 0.908(1.114), 0.670(1.162) and 0.894(1.254) were achieved for the charged, polar, hydrophobic, special cases, alanine and non-alanine respectively. It can be seen that consistent results are achieved except the hydrophobic and alanine mutations. We believe that the relative inferior performance for hydrophobic mutations and alanine mutations is due to their small data sizes.

Figure 4 shows the comparison of average and variance between experimental binding affinity changes and the predicted binding affinity changes of our PTA-EL(M2) model on SKEMPI S1131. A matrix with x-axis representing the wild type residues and y-axis representing the mutated type residues is used. In the matrix representation, the $\Delta\Delta G$ for a reverse mutation, i.e., from mutated types to wild types, is set to be the opposite values. Hence the residue to residue matrix is an antisymmetric matrix. It can be seen that the experimental-based matrix and prediction-based matrix are highly consistent for both the average and variance binding

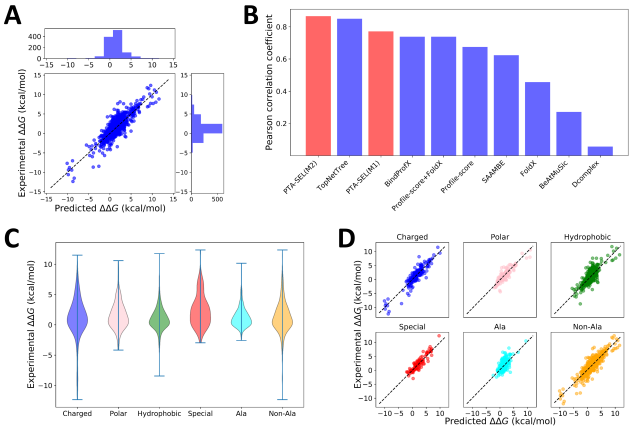

Fig. 3. The performance of our model on SKEMPI S1131 dataset. **A** The comparison between the experimental binding affinity changes (kcal/mol) and predicted binding affinity changes (kcal/mol). **B** The comparison of the performance between our model and other existing models. **C** Distributions of experimental binding affinity changes grouped according to residue mutation types. **D** Prediction results for different groups, with PCC (RMSE) 0.917(1.354), 0.884(1.074), 0.790(1.227), 0.908(1.114), 0.670(1.162) and 0.894(1.254) for the charged, polar, hydrophobic, special cases, alanine and non-alanine respectively.

Table 2. Comparison of the performance between our model and other models on AB-Bind S645.

Method	PCC	
	with nonbinders	without nonbinders
TopNetTree	0.65	0.68
PTA-EL(M2)	**0.61**	**0.72**
PTA-EL(M1)	**0.57**	**0.68**
TopGBT	0.56	-
mCSM-AB	0.53	0.56
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	-

affinity changes, which indicates that our predictions are highly accurate.

Performance on AB-Bind S645 dataset Table 2 lists the PCCs of all machine learning models on AB-Bind S645 dataset, as far as we know. It can be seen that our model ranked second among all the existing models. Note that there are 27 nonbinders that do not follow the general distribution of the other data in the dataset. It has been reported that these nonbinders have a strong negative impact on the prediction model accuracy (57). Our model can rank as first if we exclude these 27 nonbinders from the dataset. More specifically, the PCC increases from 0.61 to 0.72 by excluding these 27 nonbinders. Detailed information of our results can be found in Figure 5. **A** and **C** show the comparisons between experimental $\Delta\Delta G$ and predicted $\Delta\Delta G$ of our model with nonbinders and without nonbinders respectively. The mutations are classified into five groups according to the mutation regions, including core, rim, support, interior and surface. The distribution of

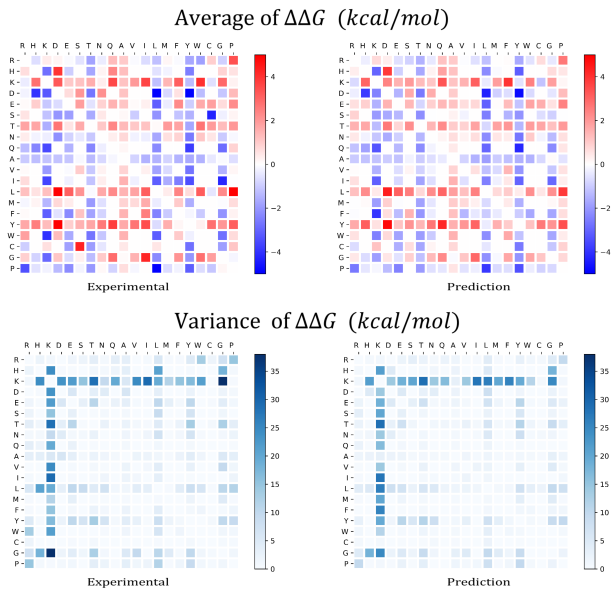

Fig. 4. The comparison of the average and variance between the experimental binding affinity changes (kcal/mol) and predicted binding affinity changes (kcal/mol) for dataset SKEMPI S1131. The residue to residue mutations are illustrated in a matrix. The x-axis represents wild residue types while y-axis is for the mutated residue types. The $\Delta\Delta G$ for a reverse mutation is set to be its opposite value. **(a)** Average binding affinity changes upon mutation (kcal/mol). **(b)** Variance of the binding affinity changes upon mutations (kcal/mol).

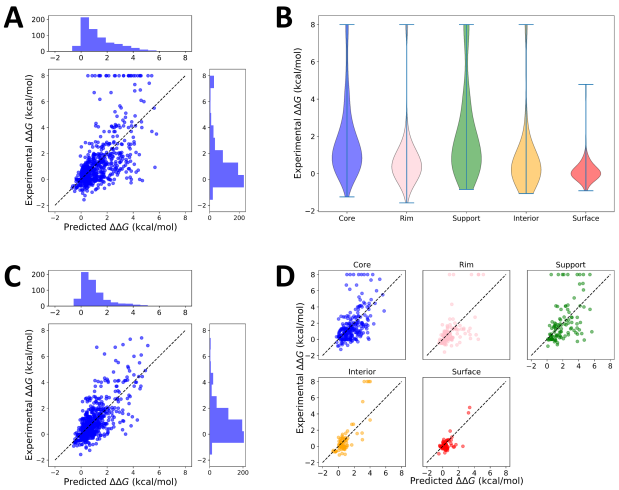

Fig. 5. The performance of our model on AB-Bind S645 dataset. **A** The comparison between the experimental binding affinity changes (kcal/mol) and predicted binding affinity changes (kcal/mol) with nonbinders. **B** Distributions of experimental binding affinity changes grouped according to residue region types. **C** The comparison between the experimental binding affinity changes and predicted binding affinity changes without nonbinders. **D** Prediction results for different groups, with PCC (RMSE) 0.530(1.659), 0.537(1.454), 0.456(2.132), 0.821(1.235) and 0.697(0.607) for core, rim, support, interior, and surface respectively.

their experimental binding affinity changes is shown in **B**. Prediction results on different mutation regions are shown in **D**, average PCC (RMSE) values of 0.530(1.659), 0.537(1.454), 0.456(2.132), 0.821(1.235) and 0.697(0.607) were achieved for the these groups respectively. This result shows that the performance is consistent among different mutation regions except for the support region. We believe the reason is due to the data of support region has too many nonbinders.

Discussion

Efficient transferable molecular representation and featurization are of great importance for machine learning models in material, chemical and biological data analysis. Recently, many mathematical invariants from algebraic topology and differential geometry have been proposed, including persistent homology, persistent spectral, persistent curvatures and other persistent functions. These persistent functions provide a series of highly effective molecular descriptors that not only perserve the intrinsic structure information but also maintain molecular multiscale properties. Molecular descriptors from these mathematical invariants can have a much better performance in machine learning models.

In our persistent Tor-algebra based stacking model, the homotopy invariant, Tor-algebra, is used to give quantitative characterization of the molecular structure. The Tor-algebra of a simplicial complex can be seen as a combination of the local cohomology information of its subcomplexes. Further, by considering a filtration of the simplicial complex, we propose persistent Tor-algebra, which gives a multiscale characterization of the molecular structure. The persistent Tor-algebra can be discretized into molecular descriptors which are further fed into our stacking model. Note that we have two indexes (i, j) in Tor-algebra $\mathrm{Tor}^{-i,2j}$ where index j determines the size of local subcomplex we are considering and index i determines the dimension of the cohomology, so a series of molecular descriptors can be derived by changing indexes i and j. To the best of our knowledge, we propose the persistent Tor-algebra for the first time, and this is the first time that persistent Tor-algebra is used for the molecular representation and featurization, and combined with machine learning for the protein-protein binding affinity changes following mutations.

Our persistent Tor-algebra based machine learning models give a try for "Algebraic data analysis". In topological data analysis, simplicial complex representation is the most commonly used method to characterize the shape of the data. Topological invariant, persistent homology, can be calculated from the simplicial complexes to give quantitative featurization of the data. In our model, face ring is used to give an algebraic representation of the data, and its persistent Tor-algebra is calculated to do featurization of the data, for the first time. Our attempt provides a framework of data analysis from algebraic aspect.

Materials and Methods

Persistent Tor-algebra. In this section, we give construction of persistent Tor-algebra for a simplicial complex. More specifically, we firstly give algebraic construction of Tor for any two modules. Then we turn to the simplicial complex, the face ring is given and its associated Tor is defined as the To-algebra of the simplicial complex. By considering a filtration process of the simplicial complex, we propose the persistent Tor-algebra for simplicial complex. For sections "Tor modules", "Face ring of a simplicial complex" and "Tor-algebra of the simplicial complex", a detailed description can be found in (61).

Tor modules. Assume A is a commutative finitely generated $\mathbb{F}$-algebra with unit, graded by nonnegative even numbers (i.e. $A = \bigoplus_{i>=0} A_i$) and connected (i.e., $A_0 = \mathbb{F}$). The basic

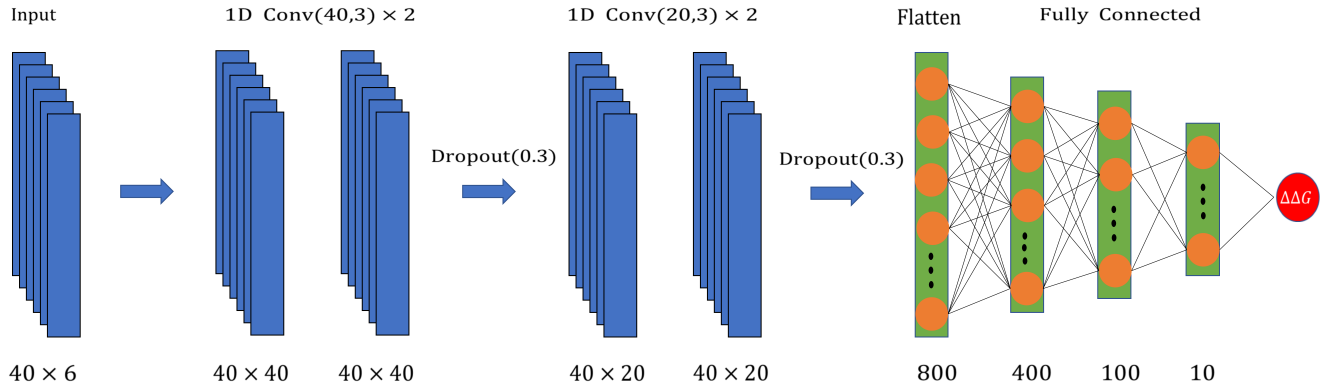

Fig. 6. Details of the 1D CNN models in the first-layer base learners. Each of the 72 persistent Tor-algebra features is fed into a base learner CNN model. Each of the CNN model will predict a binding affinity change that will be combined with auxiliary features to form the input for the second-layer base learners in our stacking model.

example is the polynomial algebra $\mathbb{F}[v_1, v_2, ..., v_m]$ with degree 2 for each v_i. We also assume that all A-modules M are nonnegatively graded and finitely generated, and all module maps are degree-preserving.

Definition 1 *Given an A-module M, a free resolutoin of M is an exact sequence of free A-modules*

$$... \xrightarrow{d_{i+1}} R^{-i} \xrightarrow{d_i} ... \xrightarrow{d_2} R^{-1} \xrightarrow{d_1} R^0 \xrightarrow{d_0} M \to 0$$

Here exact means $Ker(d_i) = Im(d_{i+1})(0 \leqslant i)$. The resolution can be converted into a bigraded $\mathbb{F}$-vector space $R = \bigoplus_i R^{-i} = \bigoplus_{i,j} R^{-i,j}$ where $R^{-i,j}$ is the j-th graded component of the module R^{-i} and the (i, j)-th component of d acts as $d_{i,j} : R^{-i,j} \to R^{-i+1,j}$. We refer to the first grading of R as external; it comes from the indexing of the terms in the resolution and is therefore nonpositive. The second grading of R, which is internal, comes from the grading in the modules R^{-i} and is therefore even and nonnegative.

Now we give the construction of Tor. Assume we have a free resolution of an A-module M, and N is another A-module. Applying the functor $\bigotimes_A N$ to the free resolution, we obtain a cochain complex

$$... \to R^{-i} \bigotimes_A N \to ... \to R^{-1} \bigotimes_A N \to R^0 \bigotimes_A N \to 0$$

The i-th cohomology module of the cochain complex is denoted as $\mathrm{Tor}_A^i(M, N)$. We can also write

$$\mathrm{Tor}_A(M, N) = \bigoplus_{i,j>=0} \mathrm{Tor}_A^{-i,2j}(M, N)$$

where $\mathrm{Tor}_A^{-i,2j}(M, N)$ is the $2j$-th graded component of $\mathrm{Tor}_A^{-i}(M, N)$. Actually, for any A-module M, there is a canonical way to construct a free resolution for M. So the $\mathrm{Tor}_A(M, N)$ can be defined for any A-module M.

Face ring of a simplicial complex.

Definition 2 *Given a simplicial complex $\mathcal{K}$ on the vertex set $\{v_1, ..., v_m\}$, the face ring of $\mathcal{K}$ is the quotient graded ring*

$$\mathbb{F}[\mathcal{K}] = \mathbb{F}[v_1, v_2, ..., v_m]/I_{\mathcal{K}}$$

where $I_{\mathcal{K}} = (V_I | I \notin \mathcal{K})$ is the ideal generated by those monomials V_I for which I is not a simplex of $\mathcal{K}$. (For any vertex subset $I = \{v_{j_1}, ..., v_{j_k}\}$, V_I is the monomial $v_{j_1}...v_{j_k}$). The ideal $I_{\mathcal{K}}$ is known as the Stanley − Reisner ideal of $\mathcal{K}$.

Next we give a more clear description of the face ring.

Theorem 1 *Given a simplicial complex $\mathcal{K}$, its face ring $\mathbb{F}[\mathcal{K}]$ has the $\mathbb{F}$-vector space basis consisting of monomials $v_{j_1}^{\alpha_1} v_{j_2}^{\alpha_2} ... v_{j_k}^{\alpha_k}$ where $\alpha_i > 0$ and $\{j_i, j_2, ..., j_k\} \in \mathcal{K}$.*

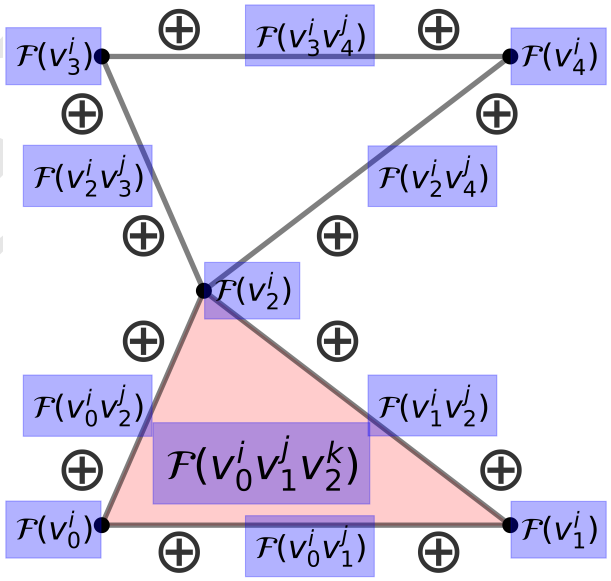

Fig. 7. An example of face ring. The simplicial complex $\mathcal{K}$ has five vertices, six edges and one triangle. Each simplex corresponds to a collection of basis elements of the face ring in the sense of vector space. And the direct sum of all subspaces generated by these basis elements is the whole face ring of $\mathcal{K}$.

Figure 7 gives an example of face ring. The simplicial complex $\mathcal{K}$ has five vertices $\{v_0, v_1, v_2, v_3, v_4\}$, six edges $\{v_0, v_1\}$, $\{v_0, v_2\}$,$\{v_1, v_2\}$, $\{v_2, v_3\}$, $\{v_2, v_4\}$, $\{v_3, v_4\}$ and a triangle $\{v_0, v_1, v_2\}$. In the sense of vector space, each n-simplex $\{v_0, v_1, ..., v_n\}$ corresponds to a collection of basis elements $\{v_0^{i_1}, v_1^{i_2}, ..., v_n^{i_n}\}(i_1, ..., i_n > 0)$. And the direct sum of all these subspaces generated from the basis elements form the whole face ring of $\mathcal{K}$. In this example, the basis elements of the face ring has the form among $\{v_0^i, v_1^i, v_2^i, v_3^i, v_4^i, v_0^i v_1^j, v_0^i v_2^j, v_1^i v_2^j, v_2^i v_3^j, v_2^i v_4^j, v_3^i v_4^j, v_0^i v_1^j v_2^k\}(i, j, k > 0)$

Actually, the face ring determines its underlying simplicial complex.

Theorem 2 (Bruns-Gubeladze) *Let $\mathbb{F}$ be a field, and $\mathcal{K}_1$, $\mathcal{K}_2$ be two simplicial complexes on the vertex sets $[m_1]$, $[m_2]$ respectively. Suppose $\mathbb{F}[\mathcal{K}_1]$ and $\mathbb{F}[\mathcal{K}_2]$ are isomorphic as $\mathbb{F}$-agebra. Then there exists a bijective map $[m_1] \to [m_2]$ which induces an isomorphism between $\mathcal{K}_1$ and $\mathcal{K}_2$.*

Tor-algebra of the simplicial complex. Given a simplicial complex $\mathcal{K}$, its face ring $\mathbb{F}[\mathcal{K}]$ has a $\mathbb{F}[v_1, v_2, ..., v_m]$-module structure via the quotient projection $\mathbb{F}[v_1, v_2, ..., v_m] \to \mathbb{F}[\mathcal{K}]$. Then its Tor-modules can be considered, we have

$$\mathrm{Tor}_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F}) = \bigoplus_{i,j>=0} \mathrm{Tor}^{-i,2j}_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F})$$

where $\mathbb{F}[m] = \mathbb{F}[v_1, v_2, ..., v_m]$ and $\mathrm{Tor}^{-i,2j}_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F})$ is $2j$-th graded component of $\mathrm{Tor}^{-i}_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F})$.

Definition 3 *Given a simplicial complex $\mathcal{K}$, its face ring is denoted as $\mathbb{F}[\mathcal{K}]$. The Tor-algebra of $\mathcal{K}$ is defined as $\mathrm{Tor}_{\mathbb{F}[v_1,v_2,...,v_m]}(\mathbb{F}[\mathcal{K}], \mathcal{F})$. The bigraded Betti number of $\mathbb{F}[\mathcal{K}]$ is defined as*

$$\beta_{-i,2j} = dim(\mathrm{Tor}^{-i,2j}_{\mathbb{F}[v_1,v_2,...,v_m]}(\mathbb{F}[\mathcal{K}], \mathbb{F}))$$

For simplicity, we denote the Tor-algebra of $\mathcal{K}$ as $\mathrm{Tor}(\mathcal{K}) = \bigoplus_{i,j>=0} \mathrm{Tor}^{-i,2j}(\mathcal{K})$. The following fundamental result of Hochster reduces the calculation of the Betti numbers $\beta_{-i,2j}$ to the calculation of reduced simplicial cohomology of full subcomplexes in $\mathcal{K}$.

Theorem 3 (Hochster) *Given a simplicial complex $\mathcal{K}$, its face ring is denoted as $\mathbb{F}[\mathcal{K}]$. We have*

$$\mathrm{Tor}^{-i,2j}_{\mathbb{F}[v_1,v_2,...,v_m]}(\mathbb{F}[\mathcal{K}], \mathcal{K}) = \bigoplus_{J\subset\mathcal{K}, |J|=j} \hat{H}^{j-i-1}(\mathcal{K}_J, \mathbb{F})$$

where $\mathcal{K}_J$ is any full subcomplex of $\mathcal{K}$ obtained by restricting to $J \subset [m]$ and $\hat{H}(\mathcal{K}_J, \mathbb{F})$ is the reduced simplicial cohomology of $\mathcal{K}_J$. We assume $\hat{H}^{-1} = F$.

Persistent Tor-algebra of the simplicial complex. We develop the persistent Tor-algebra for the simplicial complex. Assume we have a filtration of a simplicial complex $\mathcal{K}$, i.e., a sequence of nested simplicial complexes:

$$\emptyset = \mathcal{K}_0 \subset \mathcal{K}_1 \subset \mathcal{K}_2 \subset ... \subset \mathcal{K}_n = \mathcal{K}$$

where each $\mathcal{K}_i$ is a subcomplex of $\mathcal{K}_{i+1}$. We consider the inclusion map from $\mathcal{K}_i$ to $\mathcal{K}_{i+1}$, this simplicial map naturally induces a homomorphism from the face ring of $\mathcal{K}_i$ to the face ring of $\mathcal{K}_{i+1}$. Hence we get a sequence of face rings connected by homomorphisms

$$\mathbb{F}(\mathcal{K}_0) \to \mathbb{F}(\mathcal{K}_1) \to ... \to \mathbb{F}(\mathcal{K}_n)$$

For each face ring $\mathbb{F}(\mathcal{K}_i)$, its Tor-algebra can be generated, so we get a sequence of Tor-algebra

$$\mathrm{Tor}(\mathcal{K}_0) \to \mathrm{Tor}(\mathcal{K}_1) \to ... \to \mathrm{Tor}(\mathcal{K}_n)$$

where two adjacent Tor-algebras are connected by homomorphisms induced from the homomorphisms of face rings. We call this sequence of Tor-algebra modules the persistent Tor-algebra. Further, by considering a specific graded component of Tor-algebra modules, we can get the persistent graded Tor-algebra for a given $(-i, 2j)$. We have

$$\mathrm{Tor}^{-i,2j}(\mathcal{K}_1) \to \mathrm{Tor}^{-i,2j}(\mathcal{K}_2) \to ... \to \mathrm{Tor}^{-i,2j}(\mathcal{K}_n)$$

This is called the persistent Tor-algebra of $\mathcal{K}$ in the $(-i, 2j)$-th graded component. Similar to persistent homology, the persistent Tor-algebra also forms a persistent module and can be represented as a persistent barcode or persistent diagram. The bars in persistent barcode and the points in persistent diagram reflect the "birth", "death", and "persistence" of the Tor-algebra module during the filtration process.

Persistent Tor-algebra based stacking ensemble learning model. Our persistent Tor-algebra based stacking model consists of several GBT models and 1D CNN models. Computationally, all the 72 1D CNN models have the same architecture and hyperparameters, and all the GBT models also share the same parameters. The detailed parameter setting of GBT is listed in Table 3. The general architecture for 1D CNN is demonstrated in Figure 6. The CNN hyperparameters are listed in Table 4:

Table 3. The parameters for gradient boosting tree (GBT) models.

No. of Estimators	Learning rate	Max depth	Subsample
4000	0.01	6	0.7
Min_samples_split	Loss function	Max features	Repetitions
2	Least square	SQRT	10

Table 4. The parameters for convolutional neural network.

Activation function	Dropout	Initialized weights	rest weights
Relu	0.3	he_normal	lecun_uniform
Batchsize	Optimizer	learning rate	Repetitions
16	Adam	1e-5	10

Code availability

Code can be found from this link https://github.com/LiuXiangMath/PTA-SEL.

ACKNOWLEDGMENTS. We would like to thanks Prof. Jie Wu for his great suggestions and Dr. Hao Li for his nice talk on the combinatoric information of polytope.

This work was supported in part by Nanyang Technological University Startup Grant M4081842 and Singapore Ministry of Education Academic Research fund Tier 1 RG109/19, MOE-T2EP20120-0013 and MOE-T2EP20220-0010. The first author was supported by Nankai Zhide Foundation. The second author was supported by Natural Science Foundation of China (NSFC grant no. 11931007, 11221091, 11271062, 11571184).

1. Geng C, Xue LC, Roel-Touris J, Bonvin AM (2019) Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdisciplinary Reviews: Computational Molecular Science* 9(5):e1410.
2. Gonzalez MW, Kann MG (2012) Chapter 4: Protein interactions and disease. *PLoS computational biology* 8(12):e1002819.
3. Rebsamen M, Kandasamy RK, Superti-Furga G (2013) Protein interaction networks in innate immunity. *Trends in immunology* 34(12):610–619.
4. Mosca R, Céol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. *Nature methods* 10(1):47–53.
5. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology* 320(2):369–387.

6. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences* 99(22):14116–14121.
7. Dourado DF, Flores SC (2014) A multiscale approach to predicting affinity changes in protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics* 82(10):2681–2690.
8. Petukh M, Dai L, Alexov E (2016) Saambe: webserver to predict the charge of binding free energy caused by amino acids mutations. *International journal of molecular sciences* 17(4):547.
9. Xiong P, Zhang C, Zheng W, Zhang Y (2017) Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology* 429(3):426–434.
10. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D (2013) BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic acids research* 41(W1):W333–W339.
11. Moal IH, Fernandez-Recio J (2013) Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation. *Journal of Chemical Theory and Computation* 9(8):3715–3727.
12. Szilagyi A, Zhang Y (2014) Template-based structure modeling of protein–protein interactions. *Current opinion in structural biology* 24:10–23.
13. Liu S, Zhang C, Zhou H, Zhou Y (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics* 56(1):93–101.
14. Shi Q, Chen W, Huang S, Wang Y, Xue Z (2021) Deep learning for mining protein data. *Briefings in bioinformatics* 22(1):194–218.
15. Thorn KS, Bogan AA (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17(3):284–285.
16. Kumar MS, Gromiha MM (2006) PINT: protein–protein interactions thermodynamic database. *Nucleic acids research* 34(suppl_1):D195–D198.
17. Moal IH, Fernández-Recio J (2012) SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics* 28(20):2600–2607.
18. Geng C, Vangone A, Bonvin AM (2016) 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* 29(8):291–299.
19. Sirin S, Apgar JR, Bennett EM, Keating AE (2016) AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Science* 25(2):393–409.
20. Jemimah S, Yugandhar K, Michael Gromiha M (2017) PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics* 33(17):2787–2788.
21. Liu Q, Chen P, Wang B, Li J (2017) dbMPIKT: a web resource for the kinetic and thermodynamic database of mutant protein interactions. *arXiv preprint arXiv:1708.01857*.
22. Jankauskaitė J, Jiménez-García B, Dapkūnas J, Fernández-Recio J, Moal IH (2019) SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 35(3):462–469.
23. Rodrigues CH, Myung Y, Pires DE, Ascher DB (2019) mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research* 47(W1):W338–W344.
24. Strokach A, Lu TY, Kim PM (2021) ELASPIC2 (EL2): combining contextualized language models and graph neural networks to predict effects of mutations. *Journal of molecular biology* 433(11):166810.
25. Brender JR, Zhang Y (2015) Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS computational biology* 11(10):e1004494.
26. Zhang N, et al. (2020) MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience* 23(3):100939.
27. Geng C, Vangone A, Folkers GE, Xue LC, Bonvin AM (2019) iSEE: interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function, and Bioinformatics* 87(2):110–119.
28. Zhou G, et al. (2020) Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR genomics and bioinformatics* 2(2):lqaa015.
29. Jemimah S, Sekijima M, Gromiha MM (2020) ProAffiMuSeq: sequence-based method to predict the binding free energy change of protein–protein complexes upon mutation using functional classification. *Bioinformatics* 36(6):1725–1730.
30. Liu X, Luo Y, Li P, Song S, Peng J (2021) Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS computational biology* 17(8):e1009284.
31. Bronstein MM, Bruna J, Cohen T, Veličković P (2021) Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.
32. Gainza P, et al. (2020) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* 17(2):184–192.
33. Puzyn T, Leszczynski J, Cronin MT (2010) *Recent advances in QSAR studies: methods and applications*. (Springer Science & Business Media) Vol. 8.
34. Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. *Drug discovery today* 23(8):1538–1546.
35. Edelsbrunner H, Letscher D, Zomorodian A (2002) Topological persistence and simplification. *Discrete Comput. Geom.* 28(4):511–533.
36. Zomorodian A, Carlsson G (2005) Computing persistent homology. *Discrete Comput. Geom.* 33(2):249–274.
37. Cang ZX, Wei GW (2017) TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLOS Computational Biology* 13(7):e1005690.
38. Nguyen DD, Cang ZX, Wei GW (2020) A review of mathematical representations of biomolecular data. *Physical Chemistry Chemical Physics*.
39. Cang ZX, Mu L, Wei GW (2018) Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology* 14(1):e1005929.
40. Meng Z, Xia K (2021) Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science Advances* 7(19):eabc5329.
41. Cang ZX, Wei GW (2017) Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering* p. 10.1002/cnm.2914.
42. Nguyen DD, Xiao T, Wang ML, Wei GW (2017) Rigidity strengthening: A mechanism for protein–ligand binding. *Journal of chemical information and modeling* 57(7):1715–1721.
43. Cang ZX, Wei GW (2018) Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering* 34(2):e2914.
44. Nguyen DD, Wei GW (2019) AGL-Score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *Journal of chemical information and modeling* 59(7):3291–3304.
45. Cang ZX, Wei GW (2017) Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics* 33(22):3549–3557.
46. Wu KD, Wei GW (2018) Quantitative toxicity prediction using topology based multi-task deep neural networks. *Journal of chemical information and modeling* p. 10.1021/acs.jcim.7b00558.
47. Chen D, et al. (2021) Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nature Communications* 12(1):1–9.
48. Jiang J, Wang R, Wei GW (2021) GGL-Tox: geometric graph learning for toxicity prediction. *Journal of chemical information and modeling* 61(4):1691–1700.
49. Wang B, Zhao ZX, Wei GW (2016) Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies. *The Journal of chemical physics* 145(12):124110.
50. Wang B, Wang CZ, Wu KD, Wei GW (2018) Breaking the polar-nonpolar division in solvation free energy prediction. *Journal of computational chemistry* 39(4):217–233.
51. Wu KD, Zhao ZX, Wang RX, Wei GW (2018) TopP–S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *Journal of computational chemistry* 39(20):1444–1454.
52. Zhao RD, Cang ZX, Tong YY, Wei GW (2018) Protein pocket detection via convex hull surface evolution and associated Reeb graph. *Bioinformatics* 34(17):i830–i837.
53. Gao K, Nguyen DD, Tu M, Wei GW (2020) Generative network complex for the automated generation of drug-like molecules. *Journal of chemical information and modeling* 60(12):5682–5698.
54. Nguyen DD, et al. (2019) Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *Journal of computer-aided molecular design* 33(1):71–82.
55. Nguyen DD, Gao KF, Wang ML, Wei GW (2019) MathDL: Mathematical deep learning for D3R Grand Challenge 4. *Journal of computer-aided molecular design* pp. 1–17.
56. Nguyen DD, et al. (2019) Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *Journal of computer-aided molecular design* 33(1):71–82.
57. Wang M, Cang Z, Wei GW (2020) A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence* 2(2):116–123.
58. Chen J, Wang R, Wang M, Wei GW (2020) Mutations strengthened SARS-CoV-2 infectivity. *Journal of molecular biology* 432(19):5212–5226.
59. Wang R, Hozumi Y, Yin C, Wei GW (2020) Mutations on COVID-19 diagnostic targets. *Genomics* 112(6):5204–5213.
60. Avramov L, Halperin S (1986) Through the looking glass: a dictionary between rational homotopy theory and local algebra in *Algebra, algebraic topology and their interactions*. (Springer), pp. 1–27.
61. Buchstaber VM, Pano TE (2015) *Toric topology*. (American Mathematical Soc.) Vol. 204.
62. Liu X, J. WX, Wu J, Xia KL (2021) Hypergraph based persistent cohomology (HPC) for molecular representations in drug design. *Briefings in Bioinformatics*.
63. Wee J, Xia K (2021) Forman persistent ricci curvature (FPRC) based machine learning models for protein-ligand binding affinity prediction. *Briefings in Bioinformatics, in press*.
64. Pires DE, Ascher DB (2016) mcsm-ab: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic acids research* 44(W1):W469–W473.