PERSISTENT TOR-ALGEBRA BASED STACKING ENSEMBLE LEARNING (PTA-SEL) FOR PROTEIN-PROTEIN BINDING AFFINITY PREDICTION

Xiang Liu
Chern Institute of Mathematics
Nankai University
Tianjin, China
liuxiangmath@163.com

Kelin Xia
Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
Singapore, 637371
xiakelin@ntu.edu.sg

ABSTRACT

Protein-protein interactions (PPIs) play crucial roles in almost all biological processes. Recently, Data-driven machine learning models have shown great power in the analysis of PPIs. However, efficient molecular representation and featurization are still key issues that hinder the performance of learning models. Here, we propose persistent Tor-algebra (PTA), PTA-based molecular characterization and featurization, and PTA-based stacking ensemble learning (PTA-SEL) for PPI binding affinity prediction, for the first time. More specifically, the Vietoris-Rips complex is used to characterize the PPI structure and its persistent Tor-algebra is computed to form the molecular descriptors. These descriptors then are fed into our stacking model to make the prediction. 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 outperforms all the existing models as far as we know, which demonstrates the great power of our model.

1 INTRODUCTION

Protein-protein interactions (PPIs) play an essential role in a wide range of biological processes and mechanisms, including cell metabolism, signaling, protein transport and immune system (Geng et al., 2019b; Gonzalez & Kann, 2012). The understanding of PPIs, in particular PPIs upon mutations, is significant to various biomedical applications, including disease-associated mutation analysis, drug design, and therapeutic intervention (Geng et al., 2019b; Gonzalez & Kann, 2012). Experimentally, various methods have been developed to determine the protein structures and compute the binding affinities and stabilities of PPI. However, experimental studies are time-consuming and labor-intensive. Various efficient computational methods and models have been proposed for the PPI studies. Particularly, lots of models have been developed for the evaluation of PPI binding affinity changes upon mutations ($\Delta\Delta G$). With the ever-increasing PPI data, a great amount of data-driven learning models have been developed (Geng et al., 2019b; Shi et al., 2021), including mCSM (Rodrigues et al., 2019), ELASPIC (Strokach et al., 2021), BindProf (Brender & Zhang, 2015), MutaBind (Zhang et al., 2020), iSEE (Geng et al., 2019a), MuPIPR (Zhou et al., 2020), ProAffiMuSeq (Jemimah et al., 2020), GeoPPI (Liu et al., 2021), etc.

Recently, advanced mathematics, in particular topological data analysis (TDA) (Edelsbrunner et al., 2002; Zomorodian & Carlsson, 2005), are used in molecular representation and featurization (Cang & Wei, 2017c; Nguyen et al., 2020; Cang et al., 2018; Meng & Xia, 2021). Their combination with learning models have achieved great success in various steps of drug design, including protein-ligand binding affinity prediction (Cang & Wei, 2017c;b; Nguyen et al., 2017; Cang & Wei, 2018; Nguyen & Wei, 2019), protein stability change upon mutation prediction (Cang & Wei, 2017a; Cang et al., 2018), toxicity prediction (Wu & Wei, 2018; Chen et al., 2021; Jiang et al., 2021), solvation free energy prediction (Wang et al., 2016; 2018), partition coefficient and aqueous solubility (Wu et al., 2018), binding pocket detection (Zhao et al., 2018), and drug discovery (Gao et al., 2020). Outstanding performance has been consistently achieved in D3R Grand challenge (Nguyen et al.,

2019b;c;a). In particular, TopNetTree has demonstrated great power in predicting binding affinity changes upon mutations (Wang et al., 2020a). It outperformed all existing models and provided great insights for the SARS-CoV-2 mutations (Chen et al., 2020; Wang et al., 2020b).

Similar to homology, Tor-algebra is another homotopy invariant. Topologically, two simplicial complexes share the same Tor-algebra if they can deform to each other (Buchstaber & Pano, 2015). Tor-algebra of a simplicial complex $\mathcal{K}$ can be seen as a combination of the reduced simplicial cohomology of its subcomplexes. Tor-algebra is from the Tor functor and the Tor functor is one of the central concepts of homological algebra. Tor-algebra is highly related to the moment-angle complex, the key concept in toric topology. More specifically, for a simplicial complex $\mathcal{K}$, its moment-angle complex can be constructed, and the Tor algebra of $\mathcal{K}$ is isomorphic to the integral cohomology of the moment-angle complex of $\mathcal{K}$.

Here we propose persistent Tor-algebra (PTA), PTA-based molecular characterization and featurization, and PTA-based stacking ensemble learning (PTA-SEL) model for PPIs, for the first time. The Vietoris-Rips complex is used to characterize the protein-protein structure and its persistent Tor-algebra is computed to form the input features for our stacking model. More specifically, 72 persistent Tor-algebra features are fed into 72 1D convolutional neural network models separately. Besides the topological features, some precalculated auxiliary features from molecular physical properties are combined with the 72 topological features to form another 72 features, and these features are inputs for 72 gradient boosting tree models. Then all 144 predictions are fed into the meta learner to make the prediction. Our model is systematically tested on the two most-commonly used datasets, SKEMPI and AB-Bind datasets, for PPI binding affinity changes upon mutations. It has been found that our model can outperform all existing models, as far as we know.

2 RESULTS

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 molecules.

Given a simplicial complex $\mathcal{K}$ with vertex set $\{v_1, v_2, ..., v_m\}$, its face ring (or the Stanley-Reisner ring) $\mathbb{F}[\mathcal{K}]$ over the coefficient field $\mathbb{F}$ can be naturally constructed. The face ring $\mathbb{F}[\mathcal{K}]$ has the $\mathbb{F}$-vector space basis consisting of monomials $v_{j_1}^{\alpha_1}...v_{j_i}^{\alpha_i}...v_{j_k}^{\alpha_k}$ where $\alpha_i = 1, 2, ...$ and $\{v_{j_1}, ..., v_{j_k}\}$ is a simplex of $\mathcal{K}$. The face ring can uniquely determine its underlying simplicial complex while the face ring is an algebraic object and simplicial complex is a topological object, which means they are equivalent although they are different types of objects.

For the face ring $\mathbb{F}[\mathcal{K}]$ of simplicial complex $\mathcal{K}$, a Tor module $Tor_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F})$ 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 so two simplicial complexes share the same Tor-algebra if they can deform to each other. For simplicity, we denote $Tor_{\mathbb{F}[m]}(\mathbb{F}[\mathcal{K}], \mathbb{F})$ as $Tor(\mathcal{K})$. The Tor-algebra $Tor(\mathcal{K})$ is highly related to simplicial cohomology of $\mathcal{K}$. Formally, there is a decomposition $Tor(\mathcal{K}) = \bigoplus_{i,j>=0} Tor^{-i,2j}(\mathcal{K})$ where $Tor^{-i,2j}(\mathcal{K})$ is the $(-i, 2j)$-grade component of $Tor(\mathcal{K})$. And the simplicial cohomology of the full subcomplexes with j vertices of $\mathcal{K}$ fully determine the $(-i, 2j)$-grade component $Tor^{-i,2j}(\mathcal{K})$. More specifically, $Tor^{-i,2j}(\mathcal{K}) = \bigoplus_{J\subset[m],|J|=j} \hat{H}^{j-i-1}(\mathcal{K}_J, \mathbb{F})$ where $[m]$ is the vertex set $\{v_1, ..., v_m\}$, J is a vertex subset, $\mathcal{K}_J$ is a full subcomplex of $\mathcal{K}$ obtained by restricting to $J \subset [m]$ and $\hat{H}^{j-i-1}(\mathcal{K}_J)$ is the reduced simplicial cohomology of $\mathcal{K}_J$. From the decomposition, the Tor-algebra of a simplicial complex can be seen as a combination of the reduced simplicial cohomology of its subcomplexes.

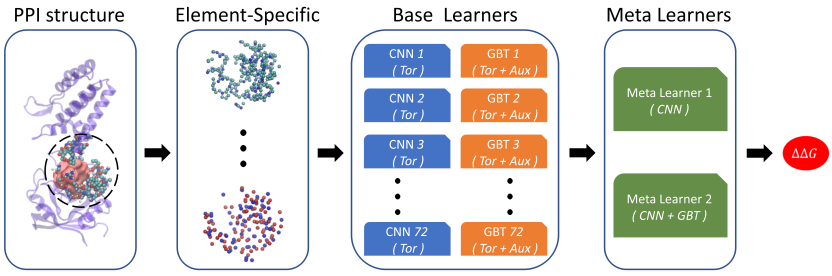

Figure 1: Illustration of our persistent Tor-algebra based stacking ensemble learning models. Tor-algebra features (Tor) are fed into the first type base learner, GBT model, and together with some auxiliary features (Aux) to form inputs for the second type base learner, 1D CNN model. Then the predictions of the base learners are stacked with a meta model to make the prediction. The first type base learners are stacked with a meta learner to form the PTA-SM(M1) model. And all the base learners are stacked with a meta learner to form the PTA-SM(M2) model.

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

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

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

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

This is called the persistent Tor-algebra of $\mathcal{K}$ in the $(-i, 2j)$-th graded component. These Tor-algebra 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.

2.1 Persistent Tor algebra based stacking models for protein-protein binding affinity prediction

2.1.1 Persistent Tor-algebra based stacking models

Our Tor-algebra based stacking models consist of two types of base learners, 1D convolutional neural network and gradient boosting tree, and two meta learners, which can be seen in Figure 1. 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 are combined with these topological features to form another 72 features, and these features are inputs of 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-SEL(M2). We also studied another stacking model PTA-SEL(M1) which only considers the first type base learners (1D CNN models).

2.1.2 FEATURE GENERATION

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.

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 $\mathcal{K}$ is constructed from the atom coordinates. Then six persistent Tor-algebra components are computed, the indexes are $(1, 2)$, $(2, 3)$, $(n-1, 1)$, $(n-2, n)$, $(n-2, n-1)$ and $(n-3, n-1)$ as (i, j) for $Tor^{-i,2j}(\mathcal{K})$ 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. Barcode statistic is used to discretize these barcodes into feature vectors. 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 the 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, a totally 72 topological features of size 240 are generated. Besides these topological features, based on Wang et al. (2020a), 707 auxiliary features are also considered in our model.

2.1.3 PERFORMANCE

Two most commonly used datasets, AB-Bind S645 and SKEMPI S1131 (Pires & Ascher, 2016; Wang et al., 2020a; Xiong et al., 2017), are considered in our benchmark tests. 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. Ten independent regressions are performed and the average PCC and RMSE are used as the measurement of the performance of our model.

Figure 2 shows the comparison of performance between our model and other existing models on AB-Bind S645 and SKEMPI S1131 datasets. It can be seen that our model can achieve the best result with PCC of 0.866 and RMSE of 1.216 kcal/mol on SKEMPI S1131 and rank second on AB-Bind S645. There are 27 nonbinders that do not follow the general distribution of the other data in the AB-Bind dataset. It has been reported that these nonbinders have a strong negative impact on the prediction model accuracy (Wang et al., 2020a). 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. Note that our PTA-SEL(M1) model also achieves state-of-the-art results, which demonstrates the great power of our persistent Tor-algebra theory in molecular representation and featurization.

3 CONCLUSION

In this paper, we propose persistent Tor-algebra (PTA), PTA-based molecular representation and featurization, PTA-based stacking ensemble learning (PTA-SEL) models for PPI binding affinity prediction, for the first time. In our stacking model, the Vietoris-Rips complex is used to characterize the PPI structure, then its persistent Tor-algebra features are computed. These topological features are inputs of 72 1D CNN base models, and together with some auxiliary features form the inputs of 72 GBT base models. Then all the 144 predictions of the 144 base learners are fed into a meta learner to make the prediction. It is found that our model outperforms all the existing models for the protein-protein binding affinity prediction on both SKEMPI S1131 and AB-Bind S645 datasets, as far as we know.

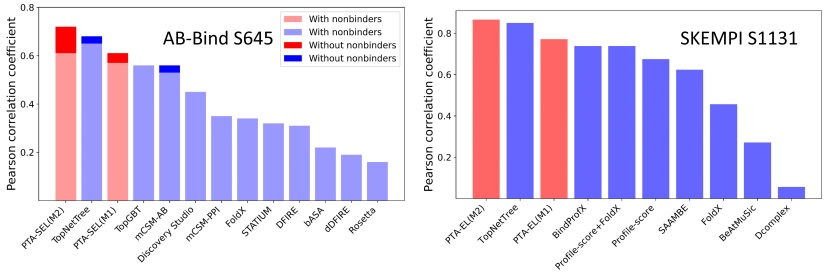

Figure 2: Comparison of the preformance between our model and other existing models on SKEMPI S1131 and AB-Bind S645 datasets. It can be seen that our model outperforms all existing models in SKEMPI S1131. Note that There are 27 nonbinders that have strong negative impact on the performance of learning models in AB-Bind S645 dataset. And our model can rank first if we excluding these nonbinders.

REFERENCES

Jeffrey R Brender and Yang Zhang. Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS computational biology*, 11(10): e1004494, 2015.

Victor M Buchstaber and Taras E Pano. *Toric topology*, volume 204. American Mathematical Soc., 2015.

Z. X. Cang and G. W. Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics*, 33(22):3549–3557, 2017a.

Z. X. Cang and G. W. Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering*, pp. 10.1002/cnm.2914, 2017b.

Z. X. Cang and G. W. Wei. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLOS Computational Biology*, 13(7):e1005690, 2017c.

Z. X. Cang and G. W. Wei. 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, 2018.

Z. X. Cang, L. Mu, and G. W. Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology*, 14(1):e1005929, 2018.

Dong Chen, Kaifu Gao, Duc Duy Nguyen, Xin Chen, Yi Jiang, Guo-Wei Wei, and Feng Pan. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nature Communications*, 12(1):1–9, 2021.

Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations strengthened SARS-CoV-2 infectivity. *Journal of molecular biology*, 432(19):5212–5226, 2020.

H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

Kaifu Gao, Duc Duy Nguyen, Meihua Tu, and Guo-Wei Wei. Generative network complex for the automated generation of drug-like molecules. *Journal of chemical information and modeling*, 60 (12):5682–5698, 2020.

Cunliang Geng, Anna Vangone, Gert E Folkers, Li C Xue, and Alexandre MJJ Bonvin. 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, 2019a.

Cunliang Geng, Li C Xue, Jorge Roel-Touris, and Alexandre MJJ Bonvin. 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*, 9(5):e1410, 2019b.

Mileidy W Gonzalez and Maricel G Kann. Chapter 4: Protein interactions and disease. *PLoS computational biology*, 8(12):e1002819, 2012.

Sherlyn Jemimah, Masakazu Sekijima, and M Michael Gromiha. 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, 2020.

Jian Jiang, Rui Wang, and Guo-Wei Wei. GGL-Tox: geometric graph learning for toxicity prediction. *Journal of chemical information and modeling*, 61(4):1691–1700, 2021.

Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, and Jian Peng. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS computational biology*, 17(8):e1009284, 2021.

Zhenyu Meng and Kelin Xia. Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science Advances*, 7(19):eabc5329, 2021.

D. D. Nguyen and G. W. Wei. 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, 2019.

D. D. Nguyen, T. Xiao, M. L. Wang, and G. W. Wei. Rigidity strengthening: A mechanism for protein–ligand binding. *Journal of chemical information and modeling*, 57(7):1715–1721, 2017.

D. D. Nguyen, Z. X. Cang, K. D. Wu, M. L. Wang, Y. Cao, and G. W. Wei. 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, 2019a.

D. D. Nguyen, Z. X. Cang, K. D. Wu, M. L. Wang, Y. Cao, and G. Wei. Wei. 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, 2019b.

D. D. Nguyen, K. F. Gao, M. L. Wang, and G. W. Wei. MathDL: Mathematical deep learning for D3R Grand Challenge 4. *Journal of computer-aided molecular design*, pp. 1–17, 2019c.

D. D. Nguyen, Z. X. Cang, and G. W. Wei. A review of mathematical representations of biomolecular data. *Physical Chemistry Chemical Physics*, 2020.

Douglas EV Pires and David B Ascher. mcsm-ab: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic acids research*, 44(W1): W469–W473, 2016.

Carlos HM Rodrigues, Yoochan Myung, Douglas EV Pires, and David B Ascher. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47 (W1):W338–W344, 2019.

Qiang Shi, Weiya Chen, Siqi Huang, Yan Wang, and Zhidong Xue. Deep learning for mining protein data. *Briefings in bioinformatics*, 22(1):194–218, 2021.

Alexey Strokach, Tian Yu Lu, and Philip M Kim. ELASPIC2 (EL2): combining contextualized language models and graph neural networks to predict effects of mutations. *Journal of molecular biology*, 433(11):166810, 2021.

B. Wang, Z. X. Zhao, and G. W. Wei. Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies. *The Journal of chemical physics*, 145 (12):124110, 2016.

B. Wang, C. Z. Wang, K. D. Wu, and G. W. Wei. Breaking the polar-nonpolar division in solvation free energy prediction. *Journal of computational chemistry*, 39(4):217–233, 2018.

Menglun Wang, Zixuan Cang, and Guo-Wei Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2 (2):116–123, 2020a.

Rui Wang, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Mutations on COVID-19 diagnostic targets. *Genomics*, 112(6):5204–5213, 2020b.

K. D. Wu and G. W. Wei. Quantitative toxicity prediction using topology based multi-task deep neural networks. *Journal of chemical information and modeling*, pp. 10.1021/acs.jcim.7b00558, 2018.

K. D. Wu, Z. X. Zhao, R. X. Wang, and G. W. Wei. 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, 2018.

Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology*, 429(3):426–434, 2017.

Ning Zhang, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R Panchenko, and Minghui Li. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience*, 23(3):100939, 2020.

R. D. Zhao, Z. X. Cang, Y. Y. Tong, and G. W. Wei. Protein pocket detection via convex hull surface evolution and associated Reeb graph. *Bioinformatics*, 34(17):i830–i837, 2018.

Guangyu Zhou, Muhao Chen, Chelsea JT Ju, Zheng Wang, Jyun-Yu Jiang, and Wei Wang. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR genomics and bioinformatics*, 2(2):lqaa015, 2020.

A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33: 249–274, 2005.

A Appendix

A.1 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 Tor-algebra of the simplicial complex, more detailed descriptions can be found refer to (Buchstaber & Pano, 2015). By considering a filtration process of the simplicial complex, we propose the persistent Tor-algebra for simplicial complex.

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 example is the polynomial algebra $\mathbb{F}[v_1, v_2, ..., v_m]$ with $degree(v_i) = 2$. 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 by our convention. The second, internal, grading of R 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 $Tor_A^i(M, N)$. We can also write

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

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

A.1.1 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 $\{j_i, j_2, ..., j_k\} \in \mathcal{K}$.*

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 $\mathcal{K}_2$.*

A.1.2 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

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

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

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 $Tor_{\mathbb{F}[v_1, v_2, ..., v_m]}(\mathbb{F}[\mathcal{K}], \mathcal{F})$. The bigraded Betti number of $\mathbb{F}[\mathcal{K}]$ is defined as*

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

For simplicity, we denote the Tor-algebra of $\mathcal{K}$ as $Tor(\mathcal{K}) = \bigoplus_{i,j>=0} Tor^{-i,2j}(\mathcal{K})$. 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*

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

where $\mathcal{K}_J$ is a 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 $\hat{H}^{-1} = F$.

The Tor-algebra is highly related to a famous concept, moment-angle complex, which is also the central concept in toric topology. For a simplicial complex $\mathcal{K}$, the moment-angle complex $\mathcal{Z}_{\mathcal{K}}$ can be constructed. Actually, the integral cohomology ring of the moment-angle complex $\mathcal{Z}_{\mathcal{K}}$ is isomorphic to the Tor-algebra of the simplicial complex $\mathcal{K}$.

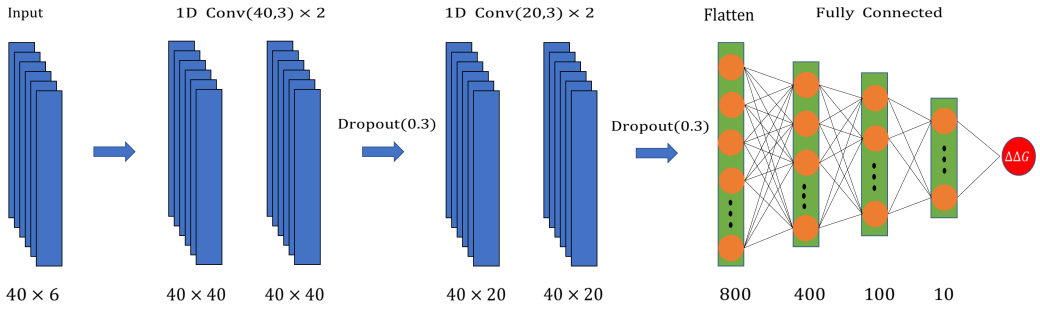

Figure 3: Details of the 1D CNN models in the first-layer base learners. Each of the 72 persistent Tor-algebra features is fed into a base learner CNN model. Each of the CNN model will predict a binding affinity change that will be combined with auxiliary features to form the input for the second-layer base learners in our stacking model.

A.1.3 Persistent Tor-algebra of the simplicial complex

Now we construct the persistent Tor-algebra for a simplicial complex. Assume we have a filtration of simplicial complex $\mathcal{K}$. That is, 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 of the filtration. 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

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

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

A.2 Persistent Tor-algebra based stacking model

Our 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 72 GBT models also share the same parameters. The detailed parameter setting of GBT is illustrated in Table 1. The general architecture for 1D CNN is demonstrated in Figure 3. The CNN

Table 1: 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

hyperparameters are as follows:

- ReLU on all 4 convolutional layers.
- Dropout(0.3)
- First layer weight initialised by he_normal.
- Rest of weights initialised by lecun_uniform
- Batchsize=16, 2000 epochs.
- Adam Optimizer with lr=1e-5.
- ReLU on fully connected layers.

A.3 PERFORMANCE

Tenfold corss validation is used to do the regression. Pearson correlation coefficient (PCC) and root-mean-square error (RMSE) are used to assess the quality of prediction. Ten independent regressions are performed and the average PCC and RMSE are used as the measurement of the performance of our model. The detailed performance between our models and other existing models on SKEMPI S1131 and AB-Bind S645 can be found in Table 2 and Table 3 respectively.

Method	PCC
PTA-SEL(M2)	**0.866**
TopNetTree	0.850
PTA-SEL(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

Table 2: Comparison of the performance between our model and other models on SKEMPI S1131.

Table 3: 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-SEL(M2)	**0.61**	**0.72**
PTA-SEL(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	-