JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

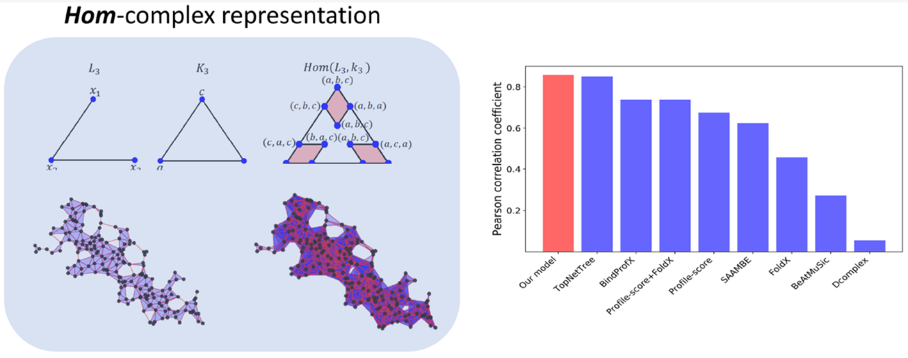
Hom-Complex-Based Machine Learning (HCML) for the Prediction of Protein−Protein Binding Affinity Changes upon Mutation

Xiang Liu, Huitao Feng, Jie Wu, and Kelin Xia*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 3961−3969

Read Online

ACCESS | Metrics & More | Article Recommendations

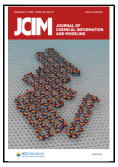


ABSTRACT: Protein−protein interactions (PPIs) are involved in almost all biological processes in the cell. Understanding protein−protein interactions holds the key for the understanding of biological functions, diseases and the development of therapeutics. Recently, artificial intelligence (AI) models have demonstrated great power in PPIs. However, a key issue for all AI-based PPI models is efficient molecular representations and featurization. Here, we propose *Hom*-complex-based PPI representation, and *Hom*-complex-based machine learning models for the prediction of PPI binding affinity changes upon mutation, for the first time. In our model, various *Hom* complexes $Hom(G_1, G)$ can be generated for the graph representation G of protein−protein complex by using different graphs G_1, which reveal G_1-related inner connections within the graph representation G of protein−protein complex. Further, for a specific graph G_1, a series of nested *Hom* complexes are generated to give a multiscale characterization of the PPIs. Its persistent homology and persistent Euler characteristic are used as molecular descriptors and further combined with the machine learning model, in particular, gradient boosting tree (GBT). We systematically test our model on the two most-commonly used data sets, that is, 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 potential of our model for the analysis of PPIs. Our model can be used for the analysis and design of efficient antibodies for SARS-CoV-2.

■ INTRODUCTION

The understanding of protein−protein interaction (PPI) mechanism is highly important to biomedical applications, such as cancer genomics, anticancer therapy, and drug discovery,[17,18] because of the essential role of PPI in various biological processes and mechanisms, including cell metabolism, signaling, protein transport, and immune system.[17,18] Historically, molecular dynamic (MD) based models[11,17,19,25,43] and statistical learning methods[10,29,36,50,58] have been developed for PPI analysis, in particular, the binding affinity change of PPIs upon mutations. Recently, data-driven learning models have gained great momentum. One of the driving force is the ever-increasing data accumulated in various PPI data sets, including ASEdb,[51] PINT[28] SKEMPI,[35] SKEMPI 2.0,[20] DACUM,[15] AB-Bind,[48] and PROXiMATE.[22] Based on them, various learning models have been proposed,[17,47] such as mCSM,[46] ELASPIC,[49] BindProf,[2] MutaBind,[59] iSEE,[16] MuPIPR,[61] ProAffiMuSeq,[21] GeoPPI,[31] etc. These models have demonstrated great promise in the prediction of binding affinity change of PPIs upon mutations. Even with the great progress, to design efficient molecular featurization still remains to be a key issue for learning models.[32,45] More recently, topological data analysis (TDA)[13,62] based advanced mathematical tools have been applied to molecular representation and featurization.[3,6,34,37] Learning models combined with TDA-based features have achieved great success in various steps of drug design[3−8,14,23,41,42,52,53,56,57,60] and D3R Grand challenge.[38−40] More interestingly, Top-

Received: May 8, 2022
Published: August 30, 2022



https://doi.org/10.1021/acs.jcim.2c00580

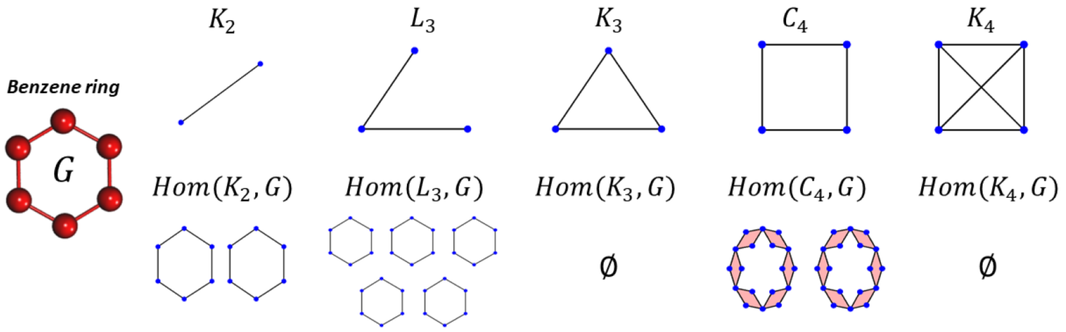


Figure 1. *Hom*-complex-based molecular representations for a benzene ring. The benzene ring is represented as a hexagon graph G with six vertices. Different *Hom* complexes $Hom(G_1, G)$ can be derived by using different graph G_1. We consider five different G_1 graphs, including K_2, L_3, K_3, C_4, and K_4. It can be seen that both $Hom(K_2, G)$ and $Hom(L_3, G)$ are repetition of the benzene hexagon ring, and $Hom(K_3, G)$ and $Hom(K_4, G)$ are empty. A detailed discussion of *Hom*-complexes can be found in Methods section.

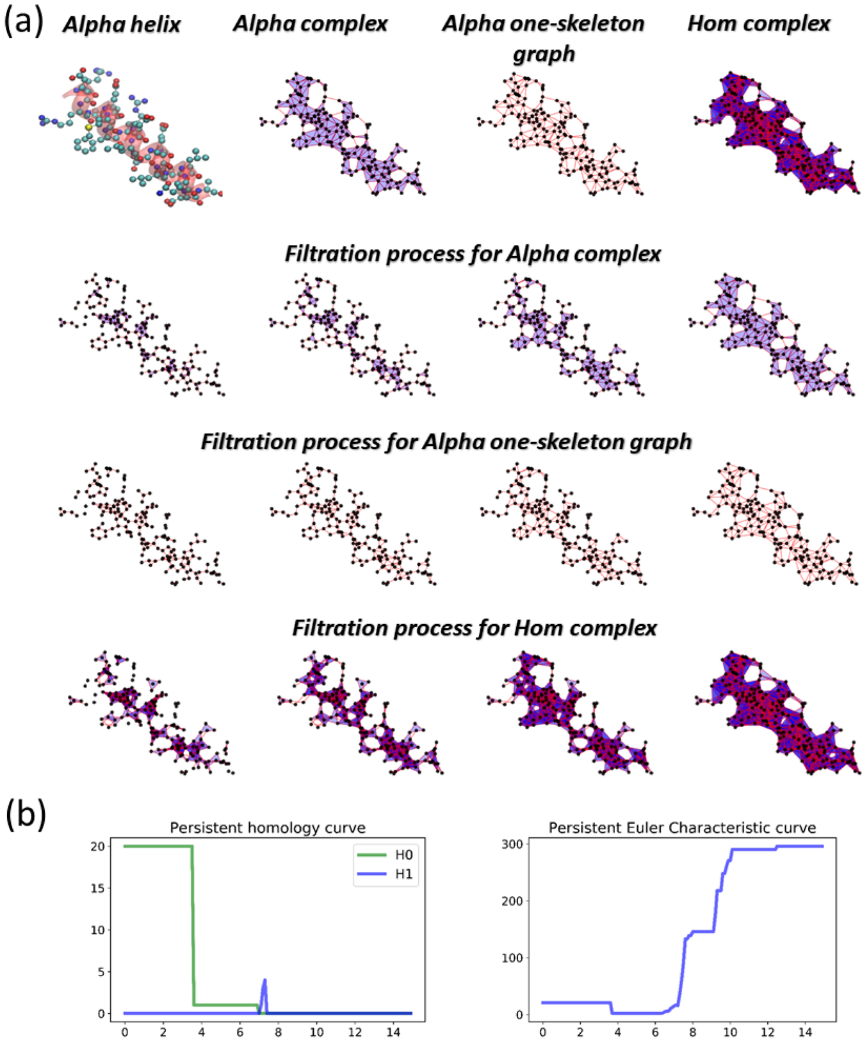


Figure 2. *Hom*-complex-based filtration for the helix of the protein with PDBID 1C26, denoted as G. (a) To generate the protein-based *Hom*-complex $Hom(K_2, G)$, an Alpha complex is constructed from the protein atoms (only the helix part). Its one-skeleton graph is then used for the generation of $Hom(K_2, G)$. Further, the Alpha-complex-based filtration process induces a series of one-skeleton graphs, which further generates a filtration process for *Hom*-complexes. (b) The persistent Betti curve (from persistent homology) and persistent Euler characteristic curve for the *Hom*-complex-based the filtration process of the protein.

NetTree model has outperformed all existing models for the prediction of protein−protein binding affinity changes upon mutations[54] and has demonstrated great power in SARS-CoV-2 virus mutation analysis.[9]

Here, we propose *Hom*-complex-based machine learning models for the prediction of PPI binding affinity changes upon mutations, for the first time. Mathematically, the *Hom* complex is one of the key concepts in combinatorial topology.[26] It is developed by Lovasz in the analysis of graph coloring problem.[1,26,27,33] *Hom* complex is a generalization of neighborhood complex, which is also introduced by Lovasz to analyze the Kneser Conjecture and construct some of the first nontrivial algebra-topological lower bounds for chromatic numbers of graphs.[26,27] In our model, *Hom* complex is used to characterize molecular structures and interactions at atomic level. Further, a multiscale representation of molecules is obtained by the employment of a filtration process, during which a series of nested *Hom* complexes at different scales are systematically generated. Two topological invariants, that is, persistent homology and persistent Euler characteristic, are used as molecular descriptors or fingerprints. Together with a series of auxiliary features from molecular physical properties, they are combined with machine learning models for the analysis of PPI properties. Our model is tested on the two most-commonly used data sets, i.e., SKEMPI and AB-Bind data sets, for the prediction of PPI binding affinity changes upon mutations. The state-of-the-art results can be achieved by our model.

RESULTS

Hom-Complex-Based Molecular Representation and Featurization. Based on a graph, various graph complexes can be generated to characterize different topological and combinatorial properties within the graph.[24] Among them is *Hom* complex, which is one of the most important tools in combinatorial topology.[1,26,27,33] Mathematically, a *Hom* complex $Hom(G_1, G_2)$ can be generated from any two undirected graphs G_1 and G_2. *Hom* complex is a polyhedral complex,[1] which is a collection of cells glued together along with their faces, and the cells are just all the graph multihomomorphisms from one graph to the other. Hence, the *Hom* complex $Hom(G_1, G_2)$ naturally characterizes the connections between G_1 and G_2. A detailed discussion of *Hom* complex can be found in the Methods section.

Figure 1 illustrate some basic *Hom* complexes generated from a benzene graph, denoted as G. By using five different graphs K_2, L_3, K_3, C_4, K_4, five different *Hom* complexes $Hom(K_2, G)$, $Hom(L_3, G)$, $Hom(K_3, G)$, $Hom(C_4, G)$, and $Hom(K_4, G)$, can be derived. Note that both $Hom(K_3, G)$ and $Hom(K_4, G)$ are empty sets, because in graph G there does not exist any three (or four) vertices that form fully connected graph K_3 (or K_4). Obviously, graph G does not have any higher order fully connected subgraphs; thus, all the $Hom(K_n, G)$ $(n > 2)$ are empty sets (Note that K_n are fully connected graph with n vertices). Both $Hom(K_2, G)$ and $Hom(L_3, G)$ are disjoint unions of hexagon graphs. In fact, $Hom(K_2, G)$ is homotopy equivalent to the neighborhood complex of G.[26,27] Note that if we treat the benzene graph G as a bipartite graph by placing the adjacent vertices in two different sets, its neighborhood complex will have two disjoint components generated among the atoms within each set. The homotopy equivalence between $Hom(K_2, G)$ and the neighborhood complex of G indicates that $Hom(K_2, G)$ should also has two disjoint components as we observed. In general, the graph G_1 of $Hom(G_1, G_2)$ works like a convolution kernel, as in convolutional neural networks (CNNs), on graph G_2. Different choices of G_1 will lead to different *Hom* complexes $Hom(G_1, G_2)$, which characterize different topological and combinatorial properties within the graph G_2.

We use *Hom* complex to represent molecular structures and interactions. To balance the computational cost and accurate molecular topological characterization, we propose an Alpha-complex-based *Hom* complex model for molecular representation. The Alpha complex, which is a very popular method in computational topology,[12] is used to generate molecular graph models by using its one-skeleton graph. Further, a filtration process for one-skeleton graphs can be obtained from Alpha complex-based filtration process, and further induce a filtration process for *Hom* complexes. Figure 2a illustrates our *Hom* complex-based filtration process for the protein PDBID 1C26 (only its helix region is considered). Note that a multiscale representation is obtained from the filtration process.

Based on the filtration process, a series of persistent models, including persistent homology/cohomology, can be considered and the corresponding topological and geometric invariants can be used as molecular descriptors or fingerprints.[3,30,34,55] It has been found that these intrinsic invariant based molecular features can be highly efficient in molecular characterization and significantly enhance the accuracy for the machine learning models. Here, we consider two topological invariants, i.e., persistent homology and persistent Euler characteristic, for our *Hom* complexes. Figure 2b illustrates persistent homology and persistent Euler characteristic from *Hom*-complex-based filtration process for protein PDB 1C26 in Figure 2a.

Hom-Complex-Based Machine Learning Models for PPIs. The ability to predict PPIs is crucial to the understanding of a wide range of biological activities and functions. In our model, $Hom(K_2, G)$ complexes (with protein graph G) are used to represent PPIs and $Hom(K_2, G)$ complex-based persistent homology and persistent Euler characteristic are used as molecular features for PPIs. Computationally, the results from persistent homology are represented as persistent barcode or persistent diagrams. Based on them, many discretization models are proposed,[3] including barcode static, algebraic and tropical functions, binning approaches, and others. We use the binning approach, in which the filtration region is divided into equal-sized bins and the total number of barcode within each bin (known as persistent Betti number) is a feature vector. Persistent Euler characteristic can also be discretized into feature vectors in a similar way. These topological features are combined with gradient boosting tree model and used in the prediction of PPI binding energy change upon mutations.

Feature Generation. Protein−protein complexes are usually of very large sizes, while their interactions mainly happen at interface regions. In our $Hom(K_2, G)$ complex model, only protein atoms near binding sites are considered to reduce computational cost at the same time avoid the irrelevant information. Further, to characterize mutation effects, we have also taken into consideration of the topological information directly from the mutation sites and neighborhood regions near the mutation sites.

More specifically, for a protein−protein complex composed of two proteins with atom sets denoted as $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, we assume the mutation happens at first protein. We consider the following four types of atom sets in our *Hom*-complex-based PPI models.

- $\mathcal{P}_1^{BS}$: atoms from $\mathcal{P}_1$, and it is within 10 Å cutoff-distance of the binding site (BS)
- $\mathcal{P}_2^{BS}$: atoms from $\mathcal{P}_2$, and it is within 10 Å cutoff-distance of the binding site (BS)
- $\mathcal{P}_1^{MS}$: atoms from the mutation site (MS) of $\mathcal{P}_1$
- $\mathcal{P}_1^{MN}$: atoms within 10 Å cutoff-distance from the mutation site of $\mathcal{P}_1$, i.e., mutation site neighborhood (MN).

Both protein structures from the wild type and the mutated type are considered, so there are eight types of atom sets in total. Further, we consider element-specific representations.[3] More specifically, for each atom set, we consider six element-specific combinations, including $\{C\}$, $\{N\}$, $\{O\}$, $\{C, N\}$, $\{C, O\}$, and $\{N, O\}$. As a result, there are totally 48 atom combinations for each protein−protein complex. For each atom combination, the corresponding *Hom* complex is generated and its persistent homology (β_0) and persistent Euler characteristic (χ for vertices and edges) are computed. The filtration goes from 0 to 5 Å with step 0.1 Å, so the size of topological feature is $4800 = 48(\textit{atom combination}) \times 50(\textit{persistence}) \times 2(\textit{feature})$. Besides the topological features, we also consider 707 auxiliary features,[54] so the total feature size is 5507 = 4800 + 707.

Benchmark Data Sets. In our benchmark tests, we consider two data sets, i.e., AB-Bind data set and SKEMPI data set. The original AB-Bind data set has 1101 mutational data points with experimentally determined binding affinities.[48] The single-point mutations from the data set, which include 645 data point across 29 antibody−antigen complexes and among them 20% are stabilizing mutations and 80% are destabilizing ones, is considered and is called AB-Bind S645 set.[44,54] Note that there are 27 nonbinders in this data sets (without experimental binding affinities), and their binding affinity changes are set to be 8 kcal/mol.[44,54] The SKEMPI data set contains 3047 binding free energy changes upon mutation;[35] it contains single-point mutations and multipoints 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 nonredundant interface single-point mutations, are selected and called SKEMPI S1131 data set.[58] Both AB-Bind S645 data set and SKEMPI S1131 data set are widely used for in the benchmark of machine learning models for PPIs.[44,54] The SKEMPI 2.0 data set is an updated version of the SKEMPI data set.[20] This data set contains 7085 entries, including single-point and multipoint mutations. David et al. filtered only single-point mutations and selected 4169 variants in 319 complexes, denoted by S4169.[46] S4169 is also considered in our benchmark data sets.

Performance. In our *Hom*-complex-based machine learning model, we use gradient boosting tree (GBT) with parameters as follows, "n_estimators = 40000", "max_depth = 6", "learning_rate = 0.001", "loss = ls", and "subsample = 0.7". We construct two types of models, one only using topological features and the other using both topological and auxiliary features, denoted by Hom-ML-V1 and Hom-ML-V2, respectively. 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 median PCC and RMSE are used as the measurement of the performance of our model.

Performance on SKEMPI 2.0 Data Set. We tested our models on S4169 data set by 10-fold cross-validation. Our model outperforms all existing models on this data set. More specifically, our Hom-ML-V2 model achieved a PCC of 0.80 and RMSE of 1.06 kcal/mol. TopNetTree model has a PCC of 0.79 and RMSE of 1.13 kcal/mol. mCSM-PPI2 model has a PCC of 0.76 and RMSE of 1.19 kcal/mol. Our Hom-ML-V1 model only using Hom-complex-based features also can achieved a PCC of 0.77 and RMSE of 1.12 kcal/mol, which demonstrates the great power of our *Hom*-complex-based molecular representation and featurization.

Performance on SKEMPI-1131 Data Set. Table 1 shows the PCCs for all the machine learning models, as far as we know, on data set SKEMPI S1131 data set using 10-fold cross-validation.

Table 1. Comparison of the Performance between Our Model and Other Models on SKEMPI S1131 Data Set

method	PCC
Hom-ML-V2	**0.857**
TopNetTree	0.850
Hom-ML-V1	**0.792**
BindProfX	0.738
Profile-score+FoldX	0.738
Profile-score	0.675
SAAMBE	0.624
FoldX	0.457
BeAtMuSic	0.272
Dcomplex	0.056

It can be seen that our model has achieved the best results with PCC of 0.857 and RMSE of 1.279 kcal/mol. Detailed information on our prediction results can be found in Figure 3. Further, the comparison of average and variance between predicted binding affinity changes and the experimental ones are demonstrated as in Figure 4. The residue to residue matrix representation is used with x-axis for the wild residue types and y-axis for the mutated residue types. The $\Delta\Delta G$ for a reverse mutation, i.e., from mutated types to wide types, is set to be the opposite values. In this way, the residue to residue matrix is the antisymmetric matrix. A highly consistent pattern between the experimental-based matrix and prediction-based matrix can be observed for both average binding affinity changes (a) and the variance (b), which indicates that our predictions are highly accurate.

Performance on AB-Bind S645 Data Set. Table 2 lists the results for AB-Bind S645 data set. It can be seen that our model ranked second among all the existing models, as far as we know. There are 27 nonbinders that do not follow the general distribution of the other data in the data set. It has been reported that these nonbinders have a strong negative impact on the prediction model accuracy.[54] Our model can rank as first if we exclude these 27 nonbinders from the data set. More specifically, the PCC increases from 0.58 to 0.70 by excluding these 27 nonbinders. Detailed information on our results can be found in Figure 5.

■ DISCUSSION

Efficient molecular representations 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 preserve 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 *Hom*-complex-based models, *Hom* complexes are used in protein−protein interaction representation. For PPI based graph G, various *Hom* complexes $Hom(G', G)$ can be constructed to characterize different inner connection information with graph G, by changing graph G'. Further, a series of nested *Hom* complexes are generated from a specially designed filtration process, and naturally introduce a multiscale representation of the PPIs. Molecular descriptors and finger-

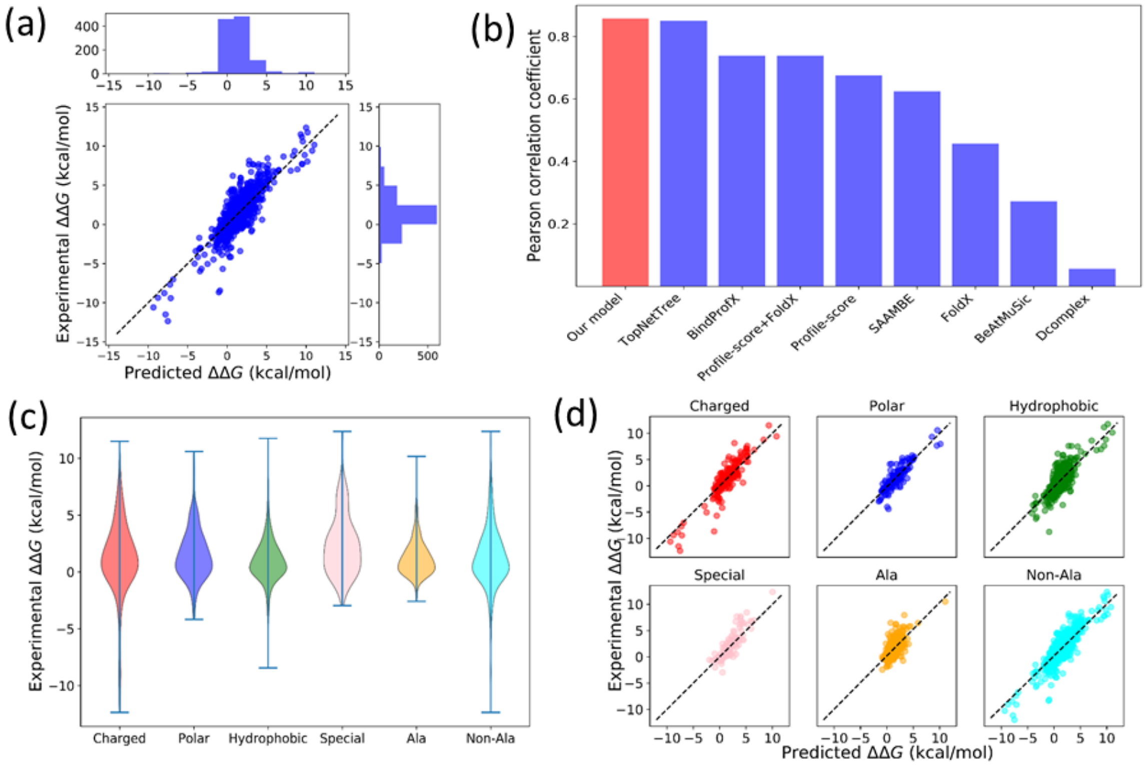


Figure 3. Performance of our model on SKEMPI S1131. (a) The comparison between the experimental binding affinity changes (kcal/mol) and predicted binding affinity changes (kcal/mol). (b) The comparison between our model and other existing models. (c) Distributions of experimental binding affinity changes grouped by charges, polar, hydrophobic, special cases, alanine, and nonalnine. (d) Prediction results in terms of PCC (and RMSE) in different groups. They are 0.908 (1.192), 0.873 (1.107), 0.795 (1.217), 0.866 (1.401), 0.672 (1.158), and 0.884 (1.334) for charged, polar, hydrophobic, special, alanine, and nonalanine, respectively.

prints can be obtained from *Hom*-complex-based molecular representations by using various persistent functions, in particular, persistent Betti curve and persistent Euler characteristics. Our *Hom*-complex-based learning model has achieved great accuracy in the prediction of binding affinity changes upon mutations. To the best of our knowledge, this is the first time that *Hom* complex is used for molecular representation and featurization, and combined with machine learning for PPI analysis.

■ METHODS

Mathematical Background for *Hom* Complex. *Hom* complex $Hom(G_1, G_2)$ is a polyhedral complex defined on any two undirected graphs G_1 and G_2. It is developed to give lower bounds in graph coloring problem.[33] In general, a vertex coloring of a graph G of n vertices can be seen as a graph homomorphism from G to the complete graph K_n. Actually, all the graph homomorphisms from G_1 to G_2 are just the 0-cells of $Hom(G_1, G_2)$, and the graph multihomomorphisms from G_1 to G_2 form the higher dimensional cells of $Hom(G_1, G_2)$.

For any graph G, we denote the vertex set of G as $V(G)$, and the edge set of G by $E(G)$, $E(G) \subset V(G) \times V(G)$.

Definition 0.1 (Graph Homomorphism). For two graphs G_1 and G_2, a graph homomorphism from G_1 to G_2 is a map ϕ: $V(G_1) \to V(G_2)$ such that if $x_1, x_2 \in G_1$ are connected by an edge in $E(G_1)$, then $\phi(x_1)$ and $\phi(x_2)$ are also connected by an edge in $E(G_2)$.

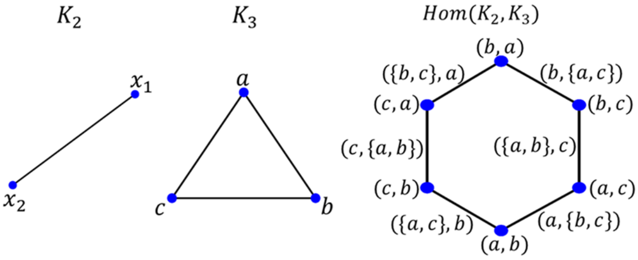
Definition 0.2 (Graph Multihomomorphism). For two graphs G_1 and G_2, a graph multihomomorphism from G_1 to G_2 is a map η: $V(G_1) \to 2^{V(G_2)}\backslash\{\varnothing\}$, such that if $x_1, x_2 \in V(G_1)$ form an edge in G_1, then $\eta(x_1) \times \eta(x_2) \subseteq E(G_2)$.

Note that if G_1 has n vertices $x_1, x_2, ..., x_n$, then a graph multihomomorphism η from G_1 to G_2 can be written as a tuple $(z_1, z_2, ..., z_n)$ where $\eta(x_i) = z_i \subset V(G_2)(i = 1, ..., n)$.

Definition 0.3 (Hom Complex). For two graphs G_1 and G_2, the *Hom* complex $Hom(G_1, G_2)$ is the polyhedral complex with all the graph multihomomorphisms as cells.

We consider two examples for the illustration of *Hom* complex.

Example 1. $Hom(K_2, K_3)$



The cells are as follows:

- 0-cells: (a, b), (a, c), (b, c), (b, a), (c, a), (c, b)
- 1-cells: $(a, \{b, c\})$, $(b, \{a, c\})$, $(c, \{a, b\})$, $(\{b, c\}, a)$, $(\{a, c\}, b)$, $(\{a, b\}, c)$

In *Hom* complex $Hom(K_2, K_3)$, a graph multihomomorphism η maps the vertices in K_2, i.e., x_1 and x_2, into vertices in K_3, i.e., a, b, and c. It is represented as a tuple $(\eta(x_1), \eta(x_2))$, as K_2 has only two vertices. Here both $\eta(x_1)$ and $\eta(x_2)$ are one or more than

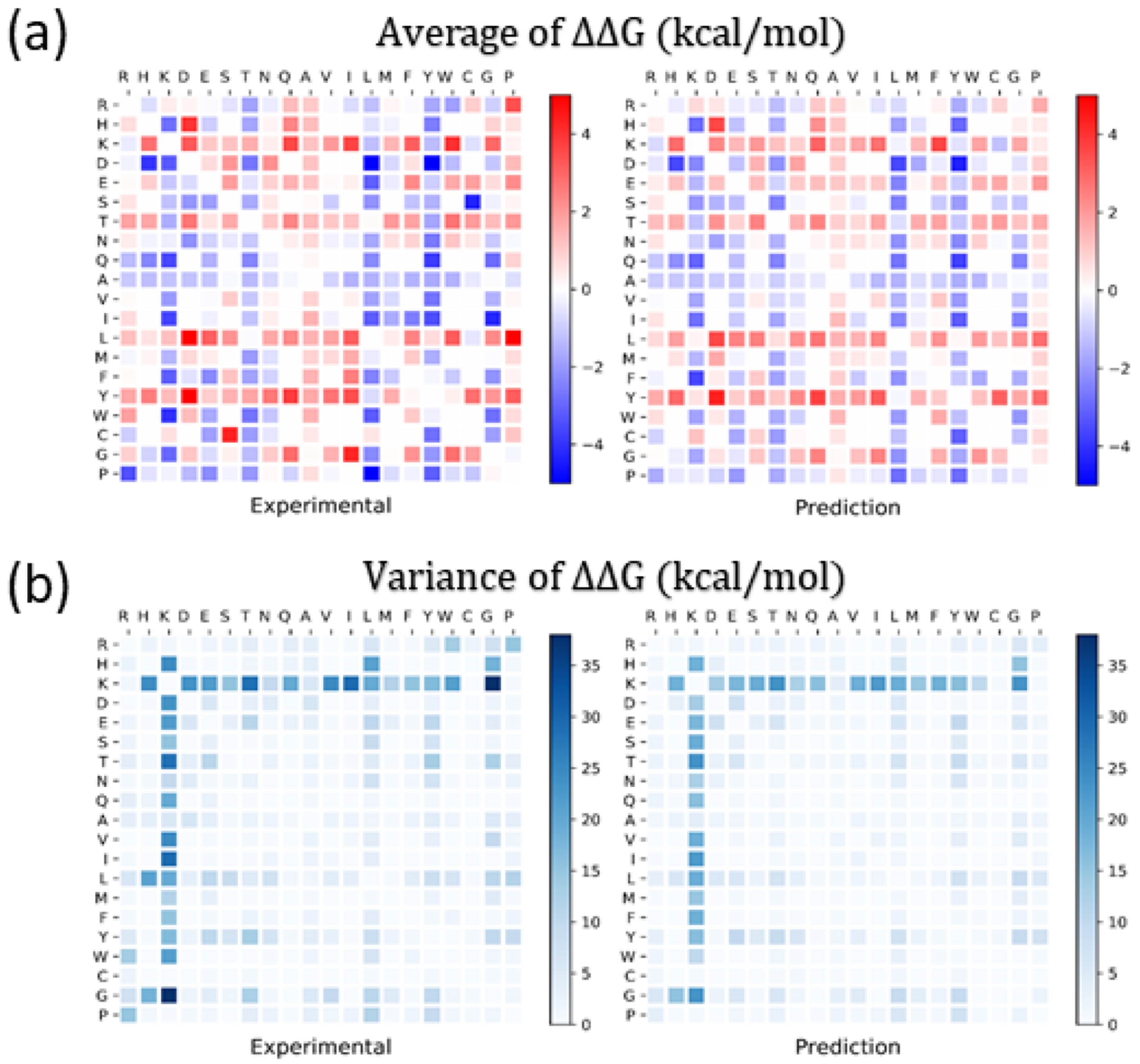


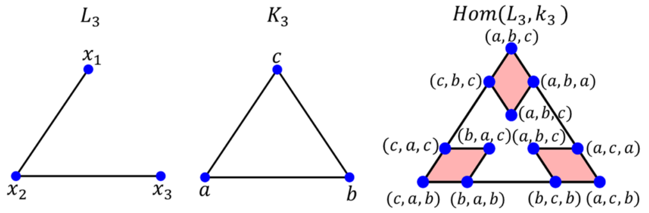
Figure 4. Comparison of the average and variance between the experimental binding affinity changes (kcal/mol) and predicted binding affinity changes (kcal/mol) for data set SKEMPI S1131. The residue to residue mutations are illustrated in a matrix. The *x*-axis represents wild residue types while *y*-axis is for the mutated residue types. The ΔΔG 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).

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
Hom-ML-V2	**0.58**	**0.70**
Hom-ML-V1	**0.58**	**0.68**
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	

one vertices from K_2, and any vertex from $\eta(x_1)$ will form an edge with the vertex in $\eta(x_2)$ in K_3. For instance, tuple $(a, \{b, c\})$ is a graph multihomomorphism that lets $\eta(x_1) = a$ and $\eta(x_2) = \{b, c\}$, so it is an 1-cell in $Hom(K_2, K_3)$. We list in Example 1 all the possible graph multihomomorphisms, which are 0-cells and 1-cells of $Hom(K_2, K_3)$. Note that there are no cells with dimension larger than 2. It can be seen that $Hom(K_2, K_3)$ is just a hexagon with 6 vertices and 6 edges.

Example 2. $Hom(L_3, K_3)$



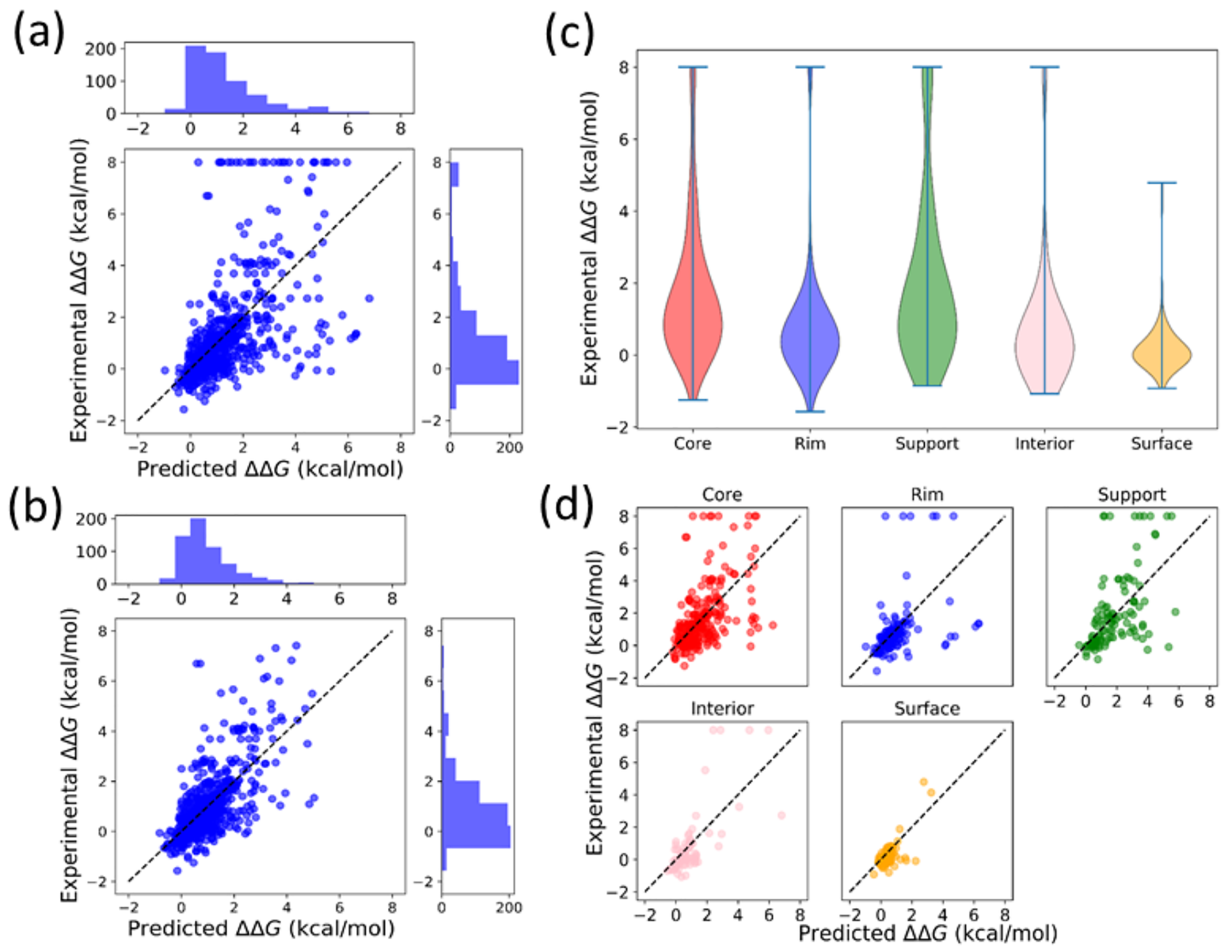


Figure 5. Performance of our model on AB-Bind S645 data set. (a) The comparison between the experimental binding affinity changes (kcal/mol) and predicted binding affinity changes (kcal/mol) with nonbinders. (b) The comparison between the experimental binding affinity changes and predicted binding affinity changes without nonbinders. (c) Distributions of experimental binding affinity changes grouped according to residue region types. (d) Prediction results for different groups, with PCC (RMSE) 0.506 (1.705), 0.393 (1.685), 0.528 (2.019), 0.725 (1.341), and 0.719 (0.583) for core, rim, support, interior, and surface, respectively.

The cells are as follows:

- 0-cells: (a, b, a), (a, b, c), (a, c, a), (a, c, b), (b, a, b), (b, a, c), (b, c, a), (b, c, b), (c, a, b), (c, a, c), (c, b, a), (c, b, c)
- 1-cells: $(\{a, b\}, c, a)$, $(\{a, b\}, c, b)$, $(\{a, c\}, b, a)$, $(\{a, c\}, b, c)$, $(\{b, c\}, a, b)$, $(\{b, c\}, a, c)$, $(c, \{a, b\}, c)$, $(b, \{a, c\}, b)$, $(a, \{b, c\}, a)$, $(a, c, \{a, b\})$, $(b, c, \{a, b\})$, $(a, b, \{a, c\})$, $(c, b, \{a, c\})$, $(b, a, \{b, c\})$, $(c, a, \{b, c\})$
- 2-cells: $(\{a, b\}, c, \{a, b\})$, $(\{a, c\}, b, \{a, c\})$, $(\{b, c\}, a, \{b, c\})$

In $Hom(L_3, K_3)$, each graph multihomomorphism η can be represented as a tuple $(\eta(x_1), \eta(x_2), \eta(x_3))$. We require that edges to be formed between $\eta(x_1)$ and $\eta(x_2)$ and also between $\eta(x_1)$ and $\eta(x_2)$, in graph K_3. Note that there are three 2-cells and each of them consists of four vertices. For instance, the 2-cell $(\{b, c\}, a, \{b, c\})$ is composed of four 1-cells, including (b, a, b), (b, a, c), (c, a, b), and (c, a, c). We list in Example 2 all the possible graph multihomomorphisms, i.e., cells for $Hom(L_3, K_3)$. Only 0-cells are marked in the figure and the four 2-cells are marked as parallelograms.

In general, for *Hom* complex $Hom(G_1, G_2)$, the graph G_1 can be viewed as a "probing" graph that is used to reveal G_1-related intrinsic patterns within G_2. Different types of G_1 graphs can be employed to characterize different inner connection information on graph G_2. Mathematically, $Hom(K_2, G_1)$ is homotopy equivalent to the neighborhood complex $N(G)$ of G.[26,27]

***Hom*-Complex-Based Molecular Representation and Featurization.** Efficient molecular representations and featurization are of essential importance for machine learning models in molecular data analysis. In our model, to balance the computational cost and accuracy of molecular characterization, we propose Alpha-complex-based *Hom* complex representation. The essential idea is to use the one-skeleton graphs from Alpha complexes to generate *Hom* complexes, at the same time use the Alpha-complex-based filtration to induce a filtration process for *Hom* complexes.

Based on molecular structures, is a series of nested Alpha complexes can be generated

$$\mathrm{Alpha}(f_1) \subset \mathrm{Alpha}(f_2) \subset \ldots \subset \mathrm{Alpha}(f_n)$$

where $f_1 < f_2 < \ldots < f_n$ are the filtration values. For each $\mathrm{Alpha}(f_k)$ $(k = 1, \ldots, n)$, its one-skeleton graph, denoted as G_{f_k}, is extracted. Consequently, a sequence of graphs are generated, and this sequence of graphs naturally form a filtration process,

$$G_{f_1} \subset G_{f_2} \subset \ldots \subset G_{f_n}$$

Each graph $G_{f_k}(k = 1, ..., n)$ can be used to generate *Hom* complexes $Hom(G_1, G_{f_k})$ with G_1 be any "probing" graph. And a sequence of *Hom* complexes is obtained, and this sequence of *Hom* complexes also form a filtration process,

$$Hom(G_1, G_{f_1}) \subset Hom(G_1, G_{f_2}) \subset \ldots \subset Hom(G_1, G_{f_n})$$

Computationally, we only consider *Hom* complexes $Hom(K_2, G)$ in this paper. An example for the protein PDBID 1C26 can be found in Figure 2.

From the filtration process of the *Hom*-complexes, a series of persistent functions, in particular persistent homology and persistent Euler characteristics, can be computed and further used as molecular descriptors or fingerprints. The essential idea is to extract mathematical invariants from differential geometry, algebraic topology, combinatorics, and others, from the filtration process. These invariant-based descriptors are more intrinsic and fundamental and have better transferability for machine learning models.[7]

AUTHOR INFORMATION

Corresponding Author

Kelin Xia − *Division of Mathematical Sciences, School of Physical and Mathematical Sciences Nanyang Technological University, Singapore 637371;* orcid.org/0000-0003-4183-0943; Email: xiakelin@ntu.edu.sg

Authors

Xiang Liu − *Chern Institute of Mathematics and LPMC, Nankai University, Tianjin, China 300071; Division of Mathematical Sciences, School of Physical and Mathematical Sciences Nanyang Technological University, Singapore 637371*

Huitao Feng − *Division of Mathematical Sciences, School of Physical and Mathematical Sciences Nanyang Technological University, Singapore 637371; Mathematical Science Research Center, Chongqing University of Technology, Chongqing, China 400054*

Jie Wu − *Yanqi Lake Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, China 101408*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.2c00580

Author Contributions

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

Notes

The authors declare no competing financial interest.

Code can be found from this link https://github.com/LiuXiangMath/Hom-Complex-ML. Data would be available upon reasonable request.

ACKNOWLEDGMENTS

This work was supported in part by Nanyang Technological University Startup Grant M4081842 and Singapore Ministry of Education Academic Research fund Tier 1 RG109/19, 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. 11221091, 11271062, 11571184). The third author was supported by Natural Science Foundation of China (NSFC Grant No. 11971144), High-level Scientific Research Foundation of Hebei Province and the start-up research fund from BIMSA.

REFERENCES

(1) Babson, E.; Kozlov, D. N. Complexes of graph homomorphisms. *Israel Journal of Mathematics* **2006**, *152* (1), 285−312.

(2) Brender, J. R.; Zhang, Y. Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS Computational Biology* **2015**, *11* (10), No. e1004494.

(3) Cang, Z. X.; Mu, L.; Wei, G. W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Computational Biology* **2018**, *14* (1), No. e1005929.

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

(5) Cang, Z. X.; Wei, G. W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering* **2017**, DOI: 10.1002/cnm.2914.

(6) Cang, Z. X.; Wei, G. W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLOS Computational Biology* **2017**, *13* (7), No. e1005690.

(7) Cang, Z. X.; Wei, G. W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int. J. Numer. Meth. Biomed. Eng.* **2018**, *34* (2), No. e2914.

(8) Chen, D.; Gao, K.; Nguyen, D. D.; Chen, X.; Jiang, Y.; Wei, G.-W.; Pan, F. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat. Commun.* **2021**, *12*, 3521.

(9) Chen, J.; Wang, R.; Gilby, N. B.; Wei, G.-W. Omicron variant (b.1.1. 529): infectivity, vaccine breakthrough, and antibody resistance. *J. Chem. Inf. Model.* **2022**, *62* (2), 412−422.

(10) Dehouck, Y.; Kwasigroch, J. M.; Rooman, M.; Gilis, D. BeAtMuSiC: prediction of changes in protein−protein binding affinity on mutations. *Nucleic Acids Research* **2013**, *41* (W1), W333−W339.

(11) Dourado, D. F.; Flores, S. C. A multiscale approach to predicting affinity changes in protein−protein interfaces. *Proteins: Struct., Funct., Bioinf.* **2014**, *82* (10), 2681−2690.

(12) Edelsbrunner, H.; Harer, J. *Computational Topology: An Introduction*; American Mathematical Society, 2010.

(13) Edelsbrunner, H.; Letscher, D.; Zomorodian, A. Topological persistence and simplification. *Discrete Comput. Geom.* **2002**, *28*, 511−533.

(14) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative network complex for the automated generation of drug-like molecules. *J. Chem. Inf. Model.* **2020**, *60* (12), 5682−5698.

(15) Geng, C.; Vangone, A.; Bonvin, A. M. Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Engineering, Design and Selection* **2016**, *29* (8), 291−299.

(16) Geng, C.; Vangone, A.; Folkers, G. E.; Xue, L. C.; Bonvin, A. M. iSEE: interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Struct., Funct., Bioinf.* **2019**, *87* (2), 110−119.

(17) Geng, C.; Xue, L. C.; Roel-Touris, J.; Bonvin, A. M. 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* **2019**, *9* (5), No. e1410.

(18) Gonzalez, M. W.; Kann, M. G. Chapter 4: Protein interactions and disease. *PLoS Computational Biology* **2012**, *8* (12), No. e1002819.

(19) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology* **2002**, *320* (2), 369−387.

(20) Jankauskaitė, J.; Jiménez-García, B.; Dapkūnas, J.; Fernández-Recio, J.; Moal, I. H. SKEMPI 2.0: an updated benchmark of changes in protein−protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **2019**, *35* (3), 462−469.

(21) Jemimah, S.; Sekijima, M.; Gromiha, M. M. ProAffiMuSeq: sequence-based method to predict the binding free energy change of

protein−protein complexes upon mutation using functional classification. *Bioinformatics* **2020**, *36* (6), 1725−1730.

(22) Jemimah, S.; Yugandhar, K.; Michael Gromiha, M. PROXiMATE: a database of mutant protein−protein complex thermodynamics and kinetics. *Bioinformatics* **2017**, *33* (17), 2787−2788.

(23) Jiang, J.; Wang, R.; Wei, G.-W. GGL-Tox: geometric graph learning for toxicity prediction. *J. Chem. Inf. Model.* **2021**, *61* (4), 1691−1700.

(24) Jonsson, J. *Simplicial Complexes of Graphs*; Springer Science & Business Media, 2007.

(25) Kortemme, T.; Baker, D. A simple physical model for binding energy hot spots in protein−protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (22), 14116−14121.

(26) Kozlov, D. *Combinatorial Algebraic Topology*, Vol. *21*; Springer Science & Business Media, 2007.

(27) Kozlov, D. N. Chromatic numbers, morphism complexes, and Stiefel−Whitney characteristic classes. *arXiv* **2005**, No. 0505563, DOI: 10.48550/arXiv.math/0505563.

(28) Kumar, M. S.; Gromiha, M. M. PINT: protein−protein interactions thermodynamic database. *Nucleic acids research* **2006**, *34*, D195−D198.

(29) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Struct., Funct., Bioinf.* **2004**, *56* (1), 93−101.

(30) Liu, X.; Wang, X.; Wu, J.; Xia, K. Hypergraph based persistent cohomology (HPC) for molecular representations in drug design. *Briefings in Bioinformatics* **2021**, *22* (5), No. bbaa411.

(31) Liu, X.; Luo, Y.; Li, P.; Song, S.; Peng, J. Deep geometric representations for modeling effects of mutations on protein−protein binding affinity. *PLoS computational biology* **2021**, *17* (8), No. e1009284.

(32) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23* (8), 1538−1546.

(33) Lovász, L. Kneser's conjecture, chromatic number, and homotopy. *Journal of Combinatorial Theory, Series A* **1978**, *25* (3), 319−324.

(34) Meng, Z.; Xia, K. Persistent spectral-based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science Advances* **2021**, *7* (19), No. eabc5329.

(35) Moal, I. H.; Fernández-Recio, J. SKEMPI: A structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics* **2012**, *28* (20), 2600−2607.

(36) Moal, I. H.; Fernandez-Recio, J. Intermolecular contact potentials for protein−protein interactions extracted from binding free energy changes upon mutation. *J. Chem. Theory Comput.* **2013**, *9* (8), 3715−3727.

(37) Nguyen, D. D.; Cang, Z. X.; Wei, G. W. A review of mathematical representations of biomolecular data. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4343−4367.

(38) Nguyen, D. D.; Cang, Z. X.; Wu, K. D.; Wang, M. L.; Cao, Y.; Wei, G. W. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *Journal of Computer-Aided Molecular Design* **2019**, *33* (1), 71−82.

(39) Nguyen, D. D.; Cang, Z. X.; Wu, K. D.; Wang, M. L.; Cao, Y.; Wei, G. W. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *Journal of Computer-Aided Molecular Design* **2019**, *33* (1), 71−82.

(40) Nguyen, D. D.; Gao, K. F.; Wang, M. L.; Wei, G. W. MathDL: Mathematical deep learning for D3R Grand Challenge 4. *Journal of Computer-Aided Molecular Design* **2020**, *34* (2), 131−147.

(41) Nguyen, D. D.; Wei, G. W. AGL-Score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *J. Chem. Inf. Model.* **2019**, *59* (7), 3291−3304.

(42) Nguyen, D. D.; Xiao, T.; Wang, M. L.; Wei, G. W. Rigidity strengthening: A mechanism for protein−ligand binding. *J. Chem. Inf. Model.* **2017**, *57* (7), 1715−1721.

(43) Petukh, M.; Dai, L.; Alexov, E. Saambe: webserver to predict the charge of binding free energy caused by amino acids mutations. *International Journal of Molecular Sciences* **2016**, *17* (4), 547.

(44) Pires, D. E.; Ascher, D. B. mcsm-ab: a web server for predicting antibody−antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Research* **2016**, *44* (W1), W469−W473.

(45) Puzyn, T.; Leszczynski, J.; Cronin, M. T. *Recent Advances in QSAR Studies: Methods and Applications*, Vol. *8*; Springer Science & Business Media, 2010.

(46) Rodrigues, C. H.; Myung, Y.; Pires, D. E.; Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein−protein interactions. *Nucleic acids research* **2019**, *47* (W1), W338−W344.

(47) Shi, Q.; Chen, W.; Huang, S.; Wang, Y.; Xue, Z. Deep learning for mining protein data. *Briefings in bioinformatics* **2021**, *22* (1), 194−218.

(48) Sirin, S.; Apgar, J. R.; Bennett, E. M.; Keating, A. E. AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Sci.* **2016**, *25* (2), 393−409.

(49) Strokach, A.; Lu, T. Y.; Kim, P. M. ELASPIC2 (EL2): combining contextualized language models and graph neural networks to predict effects of mutations. *Journal of molecular biology* **2021**, *433* (11), 166810.

(50) Szilagyi, A.; Zhang, Y. Template-based structure modeling of protein−protein interactions. *Curr. Opin. Struct. Biol.* **2014**, *24*, 10−23.

(51) Thorn, K. S.; Bogan, A. A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **2001**, *17* (3), 284−285.

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

(53) Wang, B.; Zhao, Z. X.; Wei, G. W. Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies. *J. Chem. Phys.* **2016**, *145* (12), 124110.

(54) Wang, M.; Cang, Z.; Wei, G.-W. A topology-based network tree for the prediction of protein−protein binding affinity changes following mutation. *Nature Machine Intelligence* **2020**, *2* (2), 116−123.

(55) Wee, J.; Xia, K. Forman persistent ricci curvature (FPRC) based machine learning models for protein-ligand binding affinity prediction. *Briefings in Bioinformatics* **2021**, *22* (6), No. bbab136.

(56) Wu, K. D.; Wei, G. W. Quantitative toxicity prediction using topology based multi-task deep neural networks. *J. Chem. Inf. Model.* **2018**, *58*, 520.

(57) Wu, K. D.; Zhao, Z. X.; Wang, R. X.; Wei, G. W. TopP−S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *Journal of computational chemistry* **2018**, *39* (20), 1444−1454.

(58) Xiong, P.; Zhang, C.; Zheng, W.; Zhang, Y. Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology* **2017**, *429* (3), 426−434.

(59) Zhang, N.; Chen, Y.; Lu, H.; Zhao, F.; Alvarez, R. V.; Goncearenco, A.; Panchenko, A. R.; Li, M. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience* **2020**, *23* (3), 100939.

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

(61) Zhou, G.; Chen, M.; Ju, C. J.; Wang, Z.; Jiang, J.-Y.; Wang, W. Mutation effect estimation on protein−protein interactions using deep contextualized representation learning. *NAR genomics and bioinformatics* **2020**, *2* (2), lqaa015.

(62) Zomorodian, A.; Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **2005**, *33*, 249−274.