A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation

Menglun Wang[1], Zixuan Cang [ID][1] and Guo-Wei Wei [ID][1,2,3]*

The ability to predict protein–protein interactions is crucial to our understanding of a wide range of biological activities and functions in the human body, and for guiding drug discovery. Despite considerable efforts to develop suitable computational methods, predicting protein–protein interaction binding affinity changes following mutation ($\Delta\Delta G$) remains a severe challenge. Algebraic topology, a champion in recent worldwide competitions for protein–ligand binding affinity predictions, is a promising approach to simplifying the complexity of biological structures. Here we introduce element- and site-specific persistent homology (a new branch of algebraic topology) to simplify the structural complexity of protein–protein complexes and embed crucial biological information into topological invariants. We also propose a new deep learning algorithm called NetTree to take advantage of convolutional neural networks and gradient-boosting trees. A topology-based network tree is constructed by integrating the topological representation and NetTree for predicting protein–protein interaction $\Delta\Delta G$. Tests on major benchmark datasets indicate that the proposed topology-based network tree is an important improvement over the current state of the art in predicting $\Delta\Delta G$.

P rotein–protein interactions (PPIs) are crucial to a wide range of biological activities and functions in the human body, including cell metabolism, signal transduction, muscle contraction and immune systems. The antibody–antigen system is one of the most essential among all PPIs and plays a unique role in the study of PPIs. Antibodies are large proteins that serve important roles in the immune system by counteracting antigens—chemicals recognized as alien by the human body. On the tip of an antibody, there is an antigen-binding fragment that contains a paratope for recognizing a unique antigen via its epitope; more specifically, a paratope consists of a set of complementarity-determining regions that have the highest conformational flexibility among sites on an antibody[1]. The high selectivity of antibody–antigen recognition mechanism and the flexibility of antibodies as large proteins make antibodies a suitable platform for designing counteractants of target molecules. Antibodies have been widely used as therapeutic agents to treat human diseases. Antibody therapy has several advantages over traditional therapy, including longer serum half-life, higher avidity and selectivity, and the ability to invoke desired immune responses[2–4]. Antibody therapy also brings hope of curing several previously incurable diseases and there are ongoing efforts in the direction of HIV vaccine development[5] and cancer therapeutic antibodies[6,7].

Three-dimensional (3D) structural information and thermodynamic measurements are two essential components for understanding the molecular mechanism of PPIs. Many experimental methods have been developed to determine the structure of protein–protein complexes. Among them, X-ray crystallography, NMR and cryo-electron microscopy are the main workhorses[8]. The Protein Data Bank[9], one of the largest protein structure databases, includes tens of thousands of protein–protein complex structures and is expanding at an unprecedented rate.

Site-directed mutation is a key technology for probing the thermodynamic properties of PPIs, including binding affinities of antibody–antigen interactions. Sirin et al.[10] collected an AB-Bind database of mutation-induced antibody–antigen complex binding free energy changes. This database contains 1,101 mutation data entries, including 645 single-point mutations on 32 different antibody–antigen complexes. SKEMPI is a more general database for protein–protein binding affinity changes following mutation ($\Delta\Delta G$)[11], it contains 3,047 mutation data entries for protein–protein heterodimeric complexes with experimentally determined structures.

The aforementioned databases have been widely used as benchmark tests for evaluating the predictive power of computational methods, which are indispensable for the investigation of PPIs, especially for the systematic screening of mutations[12,13]. There are many reliable computational methods that can predict mutant structures on the wild-type, such as Rosetta[14] and Jackal[15]. Computational methods for generating protein structures from sequences (for example, MODELLER[16]) and predicting docking poses for protein–protein complexes (for example, BioLuminate[17]) are also available.

The thermodynamic properties of PPIs are usually interpreted as the binding affinity or binding free energy, ΔG. Given the importance of computational methods, a variety of them have been developed that use structures to predict antibody–antigen binding affinities. DFIRE[18] relies on an all-atom, distance-scaled, pairwise potential that is derived using a database of high-quality diverse protein structures, whereas STATIUM uses a pairwise statistical potential that scores how well a protein complex can accommodate different pairs of residues in the parent complex geometry. Force-fields for proteins can also be used to compute the binding free energy, representing van der Waals interactions, hydrophobic packing, electrostatics and solvation effects. These approaches

[1]Department of Mathematics, Michigan State University, East Lansing, MI, USA. [2]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. [3]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA. *e-mail: weig@msu.edu

include FoldX (FOLDEF)[19], Discovery Studio (CHARMMPLR)[20] and Rosetta[14]. Typically, physics-based methods provide mechanistic interpretations but are not designed for handling large and diverse datasets.

Pires et al. optimized their graph-based cut-off scanning matrix (CSM) method for predicting antibody–antigen affinity changes following mutation given in the AB-Bind database[21]. This method (named mCSM-AB) was shown to outperform the aforementioned physical methods yet only achieve a Pearson's correlation coefficient (R_p) of 0.53 with tenfold cross-validation on a set of 645 single-point mutations. The limited performance of the current methods therefore highlights a pressing need for a new generation of $\Delta\Delta G$ predictors that are constructed with entirely new design principles and/or innovative machine learning algorithms. Although the physics-based methods assume potential functions of certain forms and the graph-based method only considers pairwise interactions, we seek an approach that makes fewer assumptions and allows a systemic description of PPIs.

Persistent homology[22–25]—a new branch of algebraic topology—is able to bridge geometry and topology, leading to a new efficient approach for the simplification of biological structural complexity[26–31]; however, it neglects critical chemical/biological information when it is directly applied to complex biomolecular structures. Element-specific persistent homology can retain critical biological information during the topological abstraction. Paired with advanced machine learning, such as a convolutional neural network (CNN), this new topological method gives rise to some of the best predictions for protein–ligand binding affinities[32], protein folding free energy changes following mutation[33,34] and drug virtual screening[35]. This approach has won many contests in the D3R Grand Challenges, a worldwide competition series in computer-aided drug design[36]; however, the techniques designed for protein–ligand binding analysis could not be directly applied to PPIs due to biological differences and the different characteristics of available datasets.

In the present work we introduce site-specific persistent homology that is tailored for PPI analysis. We explore the utility of site-specific persistent homology and machine learning algorithm for characterizing PPIs that are associated with site-specific mutations. We hypothesize that a topological approach that generates intrinsically low-dimensional representations of PPIs could dramatically reduce the dimensionality of antibody–antigen complexes, leading to a reliable high-throughput screening in searching for valuable mutants in protein design. To validate our hypothesis, we integrate topological descriptors with a machine learning algorithm (CNN-assisted gradient-boosting trees (GBTs)) to predict PPI $\Delta\Delta G$. The resulting topology-based network tree (TopNetTree) method is found to outperform other methods on two major benchmark datasets, AB-Bind[10] and SKEMPI[11], by a large margin. Our TopNetTree offers an accurate and reliable tool for studying PPIs.

TopNetTree model for PPI binding energy change following mutation prediction

This section describes the TopNetTree model and its application to PPI $\Delta\Delta G$ prediction. As shown in Fig. 1, the proposed TopNetTree consists of two major modules: topology-based feature generation and a CNN-assisted GBT model (Fig. 1). For the feature generation, we mainly used element- and site-specific persistent homology to capture structural characteristics, which was enhanced by chemical–physical descriptors, whereas for the learning model we used a GBT fed with inputs from a CNN as a predictor. We demonstrate the performance of the proposed TopNetTree by three commonly used PPI benchmark datasets.

Topological representation of PPIs. The pairwise interactions between atoms are characterized by the zeroth homology group (H_0, also known as the size function[37]). The higher-dimensional

homology groups encode higher-order patterns in PPI complexes. The first homology group (H_1), which is generated with Euclidean distance (D_e)-based filtration, characterizes loop or tunnel-like structures, as shown in Fig. 2, whereas the second homology group (H_2) describes cavity structures in PPI complexes. We obtain a comprehensive topological description of PPIs by combining various dimensions.

A topological representation should be able to extract patterns of different biological or chemical aspects (for example, hydrogen bonds between oxygen and nitrogen atoms, hydrophobicity, polarizability and so on) from a PPI system that is represented by a set of atomic coordinates (that is, a point cloud). We construct simplicial complexes using selected subsets of atomic coordinates and modified distance matrices to achieve this goal.

For the construction of an element- and site-specific persistent homology, we classify the atoms in a PPI complex into various subsets:

(1) $\mathcal{A}_m$: atoms of the mutation site.
(2) $\mathcal{A}_{mn}(r)$: atoms in the neighbourhood of the mutation site within a cut-off distance, r.
(3) $\mathcal{A}_{Ab}(r)$: antibody atoms within r of the binding site.
(4) $\mathcal{A}_{Ag}(r)$: antigen atoms within r of the binding site.
(5) $\mathcal{A}_{ele}(E)$: atoms in the system that has atoms of element type, E. When characterizing interactions between atoms a_i and a_j in set $\mathcal{A}$ and/or set $\mathcal{B}$, we use a modified distance matrix to exclude the interactions between the atoms from the same set. In the following formula, D_{mod} is defined as the modified distance and D_e is defined as the Euclidian distance.

$$D_{mod}(a_i, a_j) = \begin{cases} \infty, & \text{if } a_i, a_j \in \mathcal{A}, \text{ or } a_i, a_j \in \mathcal{B}, \\ D_e(a_i, a_j), & \text{if } a_i \in \mathcal{A} \text{ and } a_j \in \mathcal{B}, \end{cases} \quad (1)$$

Specific designations for sets $\mathcal{A}$ and $\mathcal{B}$ are given in Supplementary Table 1, which summarizes various topological barcodes.

Vectorization of topological barcodes. Using persistent homology, the original 3D point-cloud data are characterized by topological barcodes that are represented as collections of intervals that capture geometric patterns, topological patterns and PPIs while dramatically simplifying complicated structural representations of a PPI-complex. The upper bound of the filtration parameter corresponds to the distance cut-off of interactions of interest, which is set to be the same for different samples in the dataset. Instead of having bounding cubes of different sizes around the binding and mutation sites, topological barcodes for different samples are in the same range of filtration values, which improves the scalability in comparison with the direct use of the original 3D data. We construct feature vectors from these sets of intervals for machine learning models.

One method of vectorization is to discretize the range of the filtration parameter into bins and record the behaviour of the barcodes in each bin[35]. In this work we subdivide a filtration range (for example, [0, 12] Å) into bins of length 0.5 Å; namely, [0, 0.5], (0.5, 1], ⋯ , (11.5, 12] Å. For each bin, we count the numbers of persistence intervals, birth events and death events (see Fig. 3 for an illustration of filtration and persistence). This approach gives us three feature vectors for each topological barcode. Note that this characterization of birth and death might not be stable against different discretizations. As such, only H_0 barcodes obtained from the Vietoris–Rips filtration are used in our approach.

One advantage of binned barcode vectorization is that it keeps the distance information that reflects the strength of hydrogen bonds, van der Waals interactions and so on. The bin representation of barcode features can be easily incorporated into a CNN, which captures and discriminates local patterns; that is, the impact of mutations.

Another method of vectorization is to summarize barcode statistics, including the sum, maximum, minimum, mean and standard
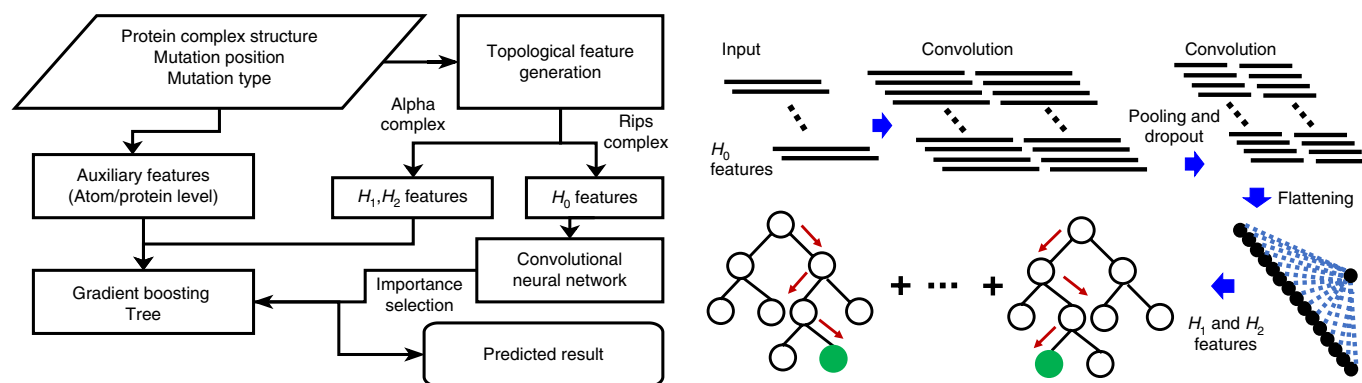
Fig. 1 | An illustration of the proposed TopNetTree model. The H_0 features are processed by a CNN whose flatten layer outputs—together with H_1,H_2 and auxiliary features—are fed into a GBT model for the final prediction.
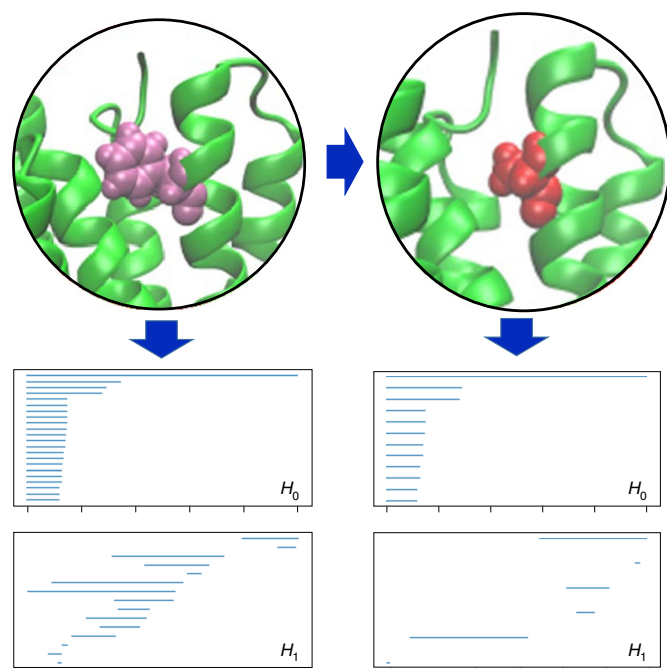

Fig. 2 | Topological barcode change associated with a mutation. Residue leucine in the wild-type is mutated into alanine. Barcodes are generated for carbon atoms within a cut-off of 12 Å of the mutant residue.

derivation of bar lengths, birth values and death values. We use this method to vectorize H_1 and H_2 barcodes obtained from alpha complex filtration as these higher-dimensional barcodes are sparser than the zero-dimensional ones[23].

Machine learning models. A major challenge in the prediction of binding affinity changes following mutation for PPIs is that the data is highly complex due to 3D structures, whereas the datasets are relatively small. We designed a hybrid machine learning algorithm that combines a CNN and GBT to overcome this difficulty. The topologically simplified description of the 3D structures are further converted into concise features by the CNN module; the GBT module then builds robust predictors with effective control of overfitting.

TopGBT model. An ensemble method is a class of machine learning algorithms that builds a powerful model from weak learners. It improves the performance on the weak learners with the assumption

that the individual learners are likely to make different mistakes and thus summing the weak learners will reduce the overall error. In this work we use GBTs that add a tree to the ensemble according to the current prediction error on the training data. This method (a toplogy-based GBT or TopGBT) performs well when there is a moderate number of features and is relatively robust against hyperparameter tuning and overfitting. The implementation provided by the scikit-learn package (v.0.18.1)[38] is used.

TopCNN model. CNNs are some of the most successful deep learning architectures, a regular CNN is a special case of a multilayer artificial neural network where only local connections are allowed between convolution layers and the weights are shared across different locations. We use a topology-based CNN (TopCNN) as an intermediate model; specifically, we feed vectorized H_0 features into CNNs to extract higher level features for the downstream model (detailed parameters and prepossessing process of our model can be found in the Supplementary Information).

TopNetTree model. CNNs can automatically extract high-level features from H_0. These CNN-extracted features are combined with features constructed from high-dimensional topological barcodes, H_1 and H_2, as the inputs of the GBTs; specifically, we build a supervised CNN model with the PPI $\Delta\Delta G$ as labels. After the model is trained, we feed the flatten layer neural outputs into a GBT model to rank their importance. Based on the importance, a subset of CNN features is combined with other features, such as the statistics of H_1 and H_2 barcodes, for the final GBT model as shown in Fig. 1. The GBT is used for its robustness against overfitting, good performance for moderately small data sizes and its model interpretability (further details on TopNetTree are given in the Supplementary Information).

Model performance for PPIs. We consider three datasets: the AB-Bind dataset[10], the SKEMPI dataset[11] and the SKEMPI 2.0 dataset[39] to validate the proposed TopNetTree model. Two evaluation metrics (R_p and the root-mean-square error, r.m.s.e.) are used to assess the quality of prediction. Detailed information of evaluation metrics can be found in the Supplementry Information.

The prediction of AB-Bind free energy changes following mutation. The AB-Bind dataset includes 1,101 mutational data points with experimentally determined binding affinities[10]. We follow Pires et al.[21] by considering only 645 single mutations across 29 antibody–antigen complexes. Among them, 87 mutations are on five complexes with homology structures. This dataset, called the AB-Bind S645 set, consists of about 20% stabilizing mutations and 80%
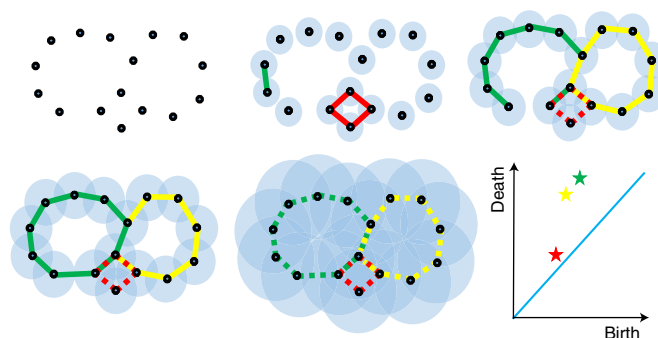
Fig. 3 | Filtration and persistence. An illustration of filtration and H_1 persistence diagram of a set of points on a plane.

destabilizing ones; there are 27 non-binders in the whole dataset, which are variants determined not to bind within the sensitivity of the assay. The binding affinity changes following mutation of these non-binders were set to $-8 \, \mathrm{kcal \, mol^{-1}}$. These non-binders could be regarded as outliers in the database and have a strongly negative impact on the prediction model accuracy.

Our model achieved an R_P of 0.65 on the AB-Bind S645 dataset, which is significantly better than those of other existing methods as shown in Table 1. In comparison with non-machine learning methods such as Rosetta and bASA, our method is over 100% more accurate in terms of R_P, indicating that our topology-based machine learning methods have a better predictive power for PPI systems. Our method is about 22% more accurate than the best-existing score of $R_P = 0.53$ (given by mCSM-AB), indicating the power of our TopNetTree.

Both GBTs and neural networks are quite sensitive to system errors as the training of a model is based on optimizing the mean-square error of the loss function. The $\Delta\Delta G$ of 27 non-binders ($-8 \, \mathrm{kcal \, mol^{-1}}$) did not follow the distribution of the whole dataset. Pires et al.[21] found that excluding non-binders from the dataset would significantly increase the performance of a prediction model. In our case, the R_P increased from 0.65 to 0.68 for the same treatment as shown in Fig. 4. We also applied a blind test on homology structures using the rest of the samples as the training set, achieving an R_P of 0.55, as shown in Fig. 4.

The performance on the SKEMPI dataset. The SKEMPI dataset[11] contains 3,047 binding free energy changes following mutation, which are assembled from the scientific literature for protein–protein heterodimeric complexes with experimentally determined structures; it includes single-point mutations and multipoint mutations. There are 2,317 single point mutation data entries among the whole database, which are referred to as the SKEMPI S2317 set.

Xiong et al. recently selected a subset of 1,131 non-redundant interface single-point mutations (denoted set S1131) from SKEMPI set S2317[40]. The same authors applied several methods to the SKEMPI S1131 set[40], including BindProfX[40], Profile-score[41,42] FoldX[19] BeAtMuSiC[43], SAMMBE[44] and Dcomplex[45].

Table 2 shows the Pearson correlation coefficients on tenfold cross-validations. It is found that the proposed TopNetTree is about 15% more accurate than the best-existing method.

The performance on the SKEMPI 2.0 dataset. The SKEMPI 2.0 (ref. [39]) database is an updated version of the SKEMPI database, containing new mutations collected after its first version, including data from three other databases: AB-Bind[10], PROXiMATE[46] and dbMPIKT[47]. This dataset contains 7,085 entries, including single- and multi-point mutations. By selecting only single-point mutations and excluding mutation entries without energy-change

values, 4,947 data points were chosen from SKEMPI 2.0 (denoted set S4947). David et al. recently applied their updated mCSM-PPI2 method[48] to the SKEMPI2 dataset. They filtered only single-point mutations and selected 4,169 variants in 319 different complexes (denoted set S4169). Set S8338 was derived from set S4169 by setting the reverse mutation energy changes to the negative values of its original energy changes. We applied our TopNetTree model to sets S4947, S4169 and S8338. We tested set S4947 with the regular tenfold cross-validation, achieving an average R_P of 0.82 and an r.m.s.e. of $1.11 \, \mathrm{kcal \, mol^{-1}}$ for the tenfold cross-validation. We followed the method of tenfold stratified cross-validation used in mCSM-PPI2 paper for sets S4169 and S8338[48]. For set S4169, we obtained an average R_P of 0.79 and r.m.s.e. of $1.13 \, \mathrm{kcal \, mol^{-1}}$, compared with the average R_P of 0.76 and r.m.s.e. of $1.19 \, \mathrm{kcal \, mol^{-1}}$ achieved by mCSM-PPI2. Finally, for set S8338, our method attained an average R_P of 0.85 and r.m.s.e. of $1.11 \, \mathrm{kcal \, mol^{-1}}$, whereas mCSM-PPI2 reported the average R_P 0.82 and r.m.s.e. of $1.18 \, \mathrm{kcal \, mol^{-1}}$ (ref. [48]).

We further validated our method by the blind prediction of another subset of the AB-Bind database. As SKEMPI 2.0 contains entries in the AB-Bind dataset, we chose 24 protein complexes that appear in both AB-Bind and SKEMPI 2.0 datasets as the test set for 787 mutations (denoted as the S787 set). The S4947 set, excluding the S787 set, was used as the training set. We achieved an average R_P of 0.53 and r.m.s.e. of $1.45 \, \mathrm{kcal \, mol^{-1}}$ on this blind test (further details are given in the Supplementary Information).

Table 1 | Comparison of the R_P of various methods for the AB-Bind S645 set

Method	R_P
TopNetTree	0.65/0.68[a]
TopGBT	0.56
mCSM-AB	0.53/0.56[a]
TopCNN	0.53
Discovery Studio	0.45
mCSM-PPI	0.31
FoldX	0.34
STATIUM	0.32
DFIRE	0.31
bASA	0.22
dDFIRE	0.19
Rosetta	0.16

Aside from those from present TopNetTrees and TopGBTs, the results are adopted from ref. [21].
[a]Results exclude 27 non-binders (their $\Delta\Delta G$ values were set to $-8 \, \mathrm{kcal \, mol^{-1}}$; ref. [10]).
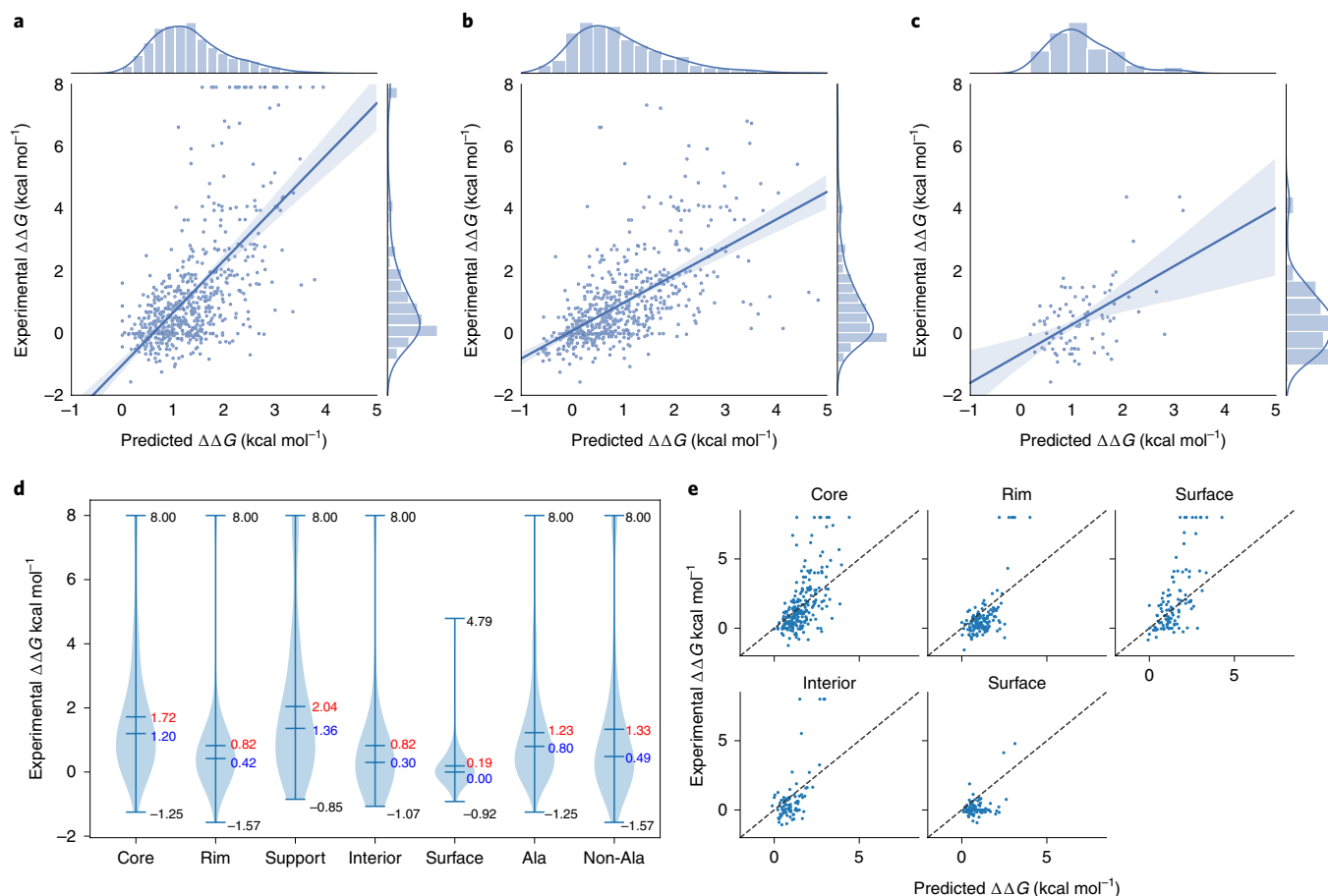
Fig. 4 | Data and results of TopNetTree model. a, A tenfold cross-validation on the AB-Bind S645 set that shows an R_p of 0.65 with a P-value of 5.948×10^{-12} (r.m.s.e. $= 1.57$ kcal mol^{-1}, s.d. $= 0.002$ kcal mol^{-1} for ten repeat experiments). **b**, A tenfold cross-validation on an AB-Bind dataset excluding 27 non-binders that shows an R_p of 0.68 with a P-value of 9.797×10^{-13} (r.m.s.e. $= 1.06$ kcal mol^{-1}, s.d. $= 0.0017$ kcal mol^{-1} for ten repeat experiments). **c**, A blind prediction of the AB-Bind subset associated with homology structures that shows an R_p of 0.55 with a P-value of 8.372×10^{-12} (r.m.s.e. $= 1.68$ kcal mol^{-1}). **d**, Distributions of binding affinity changes following mutation of the AB-Bind dataset that are grouped concerning residue region types and alanine mutations. The maximum, minimum, mean and median values of each group are cited in the violin plot. Mean values of each group are cited in red whereas median values of each group are cited in blue. **e**, Prediction results for different residue region types, with an R_p of 0.60, 0.66, 0.66, 0.65 and 0.48 for the core, rim, support, interior and surface, respectively.

Table 2 | A comparison of the R_P values of various methods for the single-point mutation in the SKEMPI dataset of 1,131 mutations

Method	R_P
TopNetTree	0.850
BindProfX	0.738
Profile-score + FoldX	0.738
Profile-score	0.675
SAAMBE[44]	0.624
FoldX	0.457
BeAtMuSic	0.272
Dcomplex	0.056

Aside from those from TopNetTree and SAMBE, the results are adopted from ref. [40].

Discussion

The quality of machine learning predictions typically depends on model inputs. In our case, the inputs consist of three crucial components: protein structures, the mutation position and mutation type.

In this section we discuss the influence of each component to the prediction quality.

Prediction result analysis for different protein complexes. For the AB-Bind S645 set, mutations can be separated into 24 different protein–protein complexes (we merged the complex with its homology model as one category). We did intra- and inter-protein cross-validations to further analyse the prediction quality across different protein complexes.

Inter-protein-level cross-validation. To perform inter-protein-level cross-validation for 24 different protein–protein complexes, the samples in one protein complex are taken as the test set, whereas the rest of the dataset is used as the training set (see Supplementary Table 2 for more details). For this test, our model reached an average R_p of 0.508 and a median R_p of 0.541. This performance is comparable with the result of blind test on homology models (see Fig. 4); however, the performance of the model varies among different protein families. Models trained on some protein families could extrapolate to other families; for example, the two protein families with the best results, 1KTZ and 2JEL, can reach R_p of 0.866 and 0.818, respectively, whereas the two families with the poorest
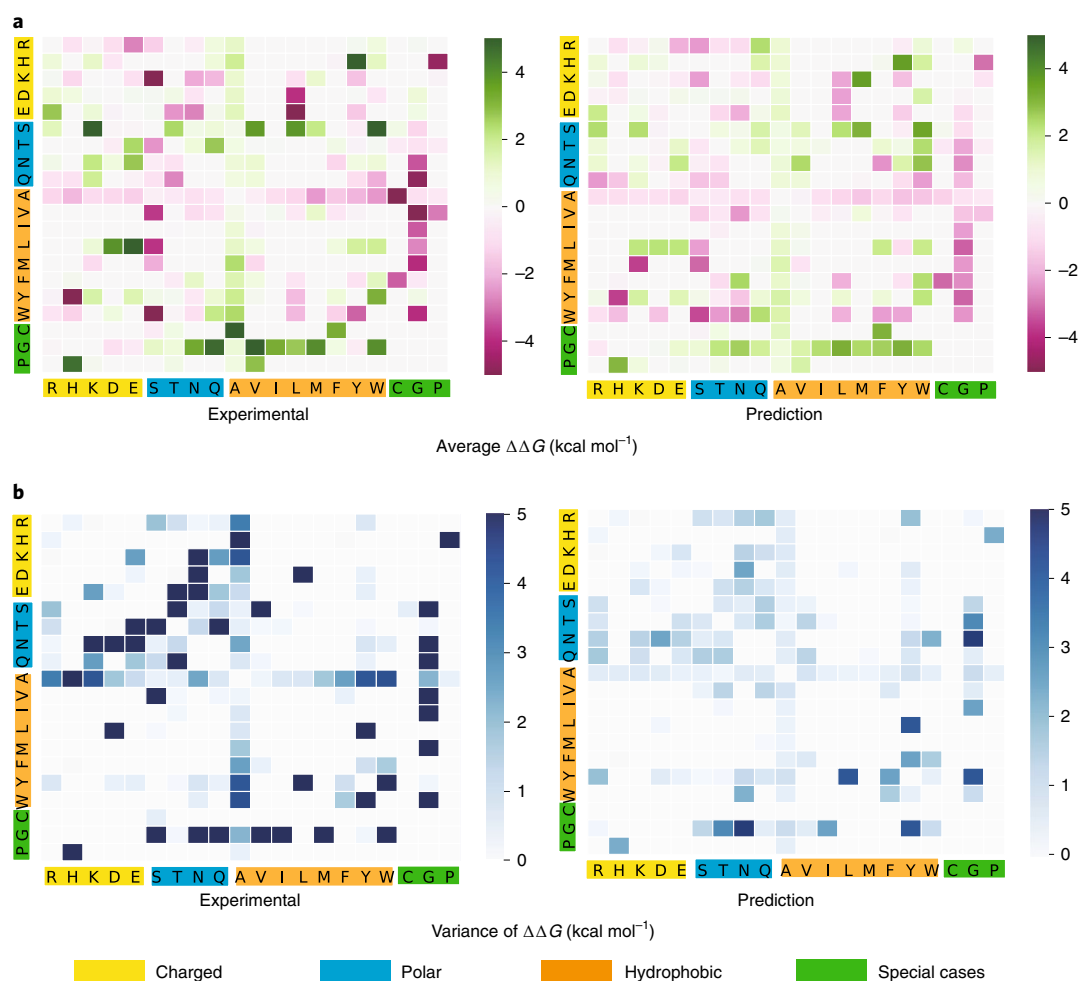
Fig. 5 | A comparison of average experimental and prediction binding affinity changes following mutations associated with different amino acid types for the AB-Bind dataset. The *x*-axis labels the residue type of the original, whereas the *y*-axis labels the residue type of the mutant. For a reverse mutation, its $\Delta\Delta G$ is taken to be the same magnitude as the original value with an opposite sign. **a**, Average binding affinity changes following mutation (kcal mol^{-1}). **b**, Variance of binding affinity changes following mutation (kcal mol^{-1}).

Table 3 | Criteria of residue regions[50], $\Delta rASA = rASA_m - rASA_c$

Region	$\Delta rASA$	$rASA_c$	$rASA_m$
Interior	0	<25%	—
Surface	0	>25%	—
Rim	>0	>25%	—
Support	>0	—	<25%
Core	>0	<25%	>25%

results, 1FFW and 1YY9, have R_P values of −0.043 and −0.068, respectively.

Intra-protein-level leave-one-out cross-validation. Cross-validation was carried out within each protein complex. For this test, our model reached average/median R_P values of 0.170/0.215, which are significantly lower than the tenfold cross-validation result over the entire dataset. One possible reason for this behaviour is that the training set for each complex is too small with only an average of 27 samples per complex. This result also implies that our model needs a diversity of training samples to achieve stable and consistent prediction quality (see Supplementary Table 3 for more details).

Prediction result analysis for different mutation regions. The locations of the site mutations could be categorized into five different regions: interior, surface, rim, support and core (a detailed definition can be found in the Methods). In experimental data, mutations at the core or support region have a higher average energy change of around 1.8 kcal mol^{-1} (1.72 kcal mol^{-1} and 1.91 kcal mol^{-1}, respectively), whereas mutations at the rim or interior region have an average energy change of around 0.8 kcal mol^{-1} (0.82 kcal mol^{-1} and 0.83 kcal mol^{-1}, respectively), as shown in Fig. 4. On the other hand, the surface mutations have an average energy change of less than 0.2 kcal mol^{-1}. Similar patterns regarding mutation sites and energy changes were reported in the literature[49]. A possible reason for these patterns is that different mutation regions vary in their accessibility to water; in general, surface, interior and rim regions have greater access to water than the core and support regions.

Figure 4 shows our predictions concerning different mutation regions. Average R_P values of 0.60, 0.66, 0.66, 0.65 and 0.48 were achieved for the core, rim, support, interior and surface regions, respectively. This result shows that the performance is consistent among different mutation regions except for the surface region. We believe that the relative inferior performance for surface mutations is due to its small data size and that the energy disturbance caused by surface mutations is small on average.

Prediction result analysis for different mutation types. The pattern of PPI binding affinity changes over different mutation types is important for protein design. We test how well can the model prediction resemble the distribution in experimental data. Here we investigate the behaviour of our model for 20 different amino acids types in the AB-Bind S645 set. A reverse mutation from 'B' to 'A' is considered to be the same mutation type as from 'A' to 'B', and the associated energy change admits an opposite sign (the mutations count for each mutation type can be found in Supplementary Fig. 1).

Overall, our predicted patterns are remarkably similar to those of experimental data in terms of both average binding energy changes and variance of binding energy changes, as shown in Fig. 5. It is interesting to note that all the mutations to alanine have a positive energy change—a possible reason is that mutations from a large residue to a small one could lead to a stabilizing effect to the whole system. Aside from the size of amino acids, we also categorized them into charged, polar, hydrophobic and special-case groups. In terms of binding affinity changes, we find that most mutations from polar to hydrophobic residues have a positive free energy change (for example, S to M), which means mutations from polar residues to hydrophobic residues would make the whole PPI system more stable. We also observed that a mutation from charged residues to uncharged polar residues could lead to a negative energy change; for example, lysine to serine (K to S), which means such mutations might have broken some charge–charge interaction pairs.

Although our model shares a similar pattern in the variance of energy changes with experimental data, the variance of the model predictions is generally lower than the experimental data as shown in Fig. 5. It remains a challenging task to come up with predictions with a diversity level the same as that of experimental data.

Conclusion

The importance of PPIs is evident from the intensive efforts to study them from many perspectives, including quantum mechanics, molecular mechanics, biochemistry, biophysics and molecular biology; for example, the R_p value between predicted $\Delta\Delta G$ values and experimental data in cross-validations of a commonly used PPI database, AB-Bind[10], is only 0.53.

Topology has recently been shown to be surprisingly effective in simplifying biomolecular structural complexity[26,27,29]. It has been devised to win worldwide competitions in computer-aided drug design[36]. It is therefore of enormous importance to exploit topology for understanding PPIs. In this work, we propose TopNetTrees for $\Delta\Delta G$ predictions; specifically, an element- and site-specific persistent homology is introduced to characterize PPIs. Furthermore, we propose machine learning algorithms—CNN-assisted GBTs—to pair with the topological method for the prediction of PPI $\Delta\Delta G$. We demonstrate that the proposed TopNetTree achieves an R_p of 0.65, which is about 22% better than the previous best result for the AB-Bind dataset. For another benchmark PPI dataset, SKEMPI, the present method significantly outperforms the state-of-the-art in the literature.

Methods

Simplicial complex and filtration. An abstract simplicial complex is a finite collection of sets of points (that is, atoms) $K = \{\sigma_i\}_i$, where the elements in σ_i are called vertices and σ_i is called a k-simplex if it has $k+1$ distinct vertices. If $\tau \subseteq \sigma_i$ then τ is called a face of σ_i. A simplicial complex, K, is valid if $\tau \subseteq \sigma_i$ for $\sigma_i \in K$ indicates that $\tau \in K$, and that the non-empty intersection of any two simplices σ_1, $\sigma_2 \in K$ is a face of both σ_1 and σ_2.

In practice, it is favourable to characterize point clouds or atomic positions in various spatial scales rather than in a fixed scaled simplicial complex representation. To construct a scale-changing simplicial complex, consider a function $f : K \to \mathbb{R}$ that satisfies $f(\tau) \leq f(\sigma)$ whenever $\tau \subseteq \sigma$. Given a real value, x, f induces a subcomplex of K by constructing a sub-level set, $K(x) = \{\sigma \in K | f(\sigma) \leq x\}$. As K is finite, the range of f is also finite and the induced subcomplexes, when ordered, form a filtration of K,

$$\varnothing \subset K(x_1) \subset K(x_2) \subset \cdots \subset K(x_\ell) = K \qquad (2)$$

There are many constructions of K and one that is widely used for point clouds is the Vietoris–Rips complex. Given K as the collection of all possible simplices from a set of atomic coordinates until a fixed dimension, the filtration function is defined as $f_{Rips}(\sigma) = \max\{d(v_i, v_j)|v_i, v_j \in \sigma\}$ for $\sigma \in K$, where d is a predefined distance function between the vertices; for example, D_e. In practice, an upper bound of the filtration value is set to avoid an excessively large simplicial complex. Another efficient construction called the alpha complex[23] is often used to characterize geometry. and we denote the filtration function by $f_\alpha : DT(X) \to \mathbb{R}$, where $DT(X)$ is the simplicial complex that is induced by the Delaunay triangulation of the set of atomic coordinates, X (ref. [23]). The filtration function is defined as $f_\alpha(\sigma) = \max\{\frac{1}{2} D_e(v_i, v_j)|v_i, v_j \in \sigma\}$ for $\sigma \in DT(X)$. Back to molecular structures, the filtration of simplicial complexes describes the topological characteristics of interaction hypergraphs under various interaction range assumptions.

Homology and persistence. A homology group (in singular homology) of a simplicial complex topologically depicts hole-like structures of different dimensions. Given a simplicial complex, K, a k-chain is a finite formal sum of k-simplices in K; that is, $\sum_i a_i \sigma_i$. There are many choices for coefficients, a_i, and we choose $a_i \in \mathbb{Z}_2$ for simplicity. The kth chain group (denoted $C_k(K)$) comprises all of the k-chains under the addition that is induced by the addition of coefficients. A boundary operator $\partial_k: C_k(K) \to C_{k-1}(K)$ connects chain groups of different dimensions by mapping a chain to the alternating sum of codimension-1 faces. It suffices to define the boundary operator on simplices,

$$\partial_k(\{v_0, \cdots, v_k\}) = \sum_{i=0}^{k} (-1)^i \{v_0, \cdots, \mathbf{v}_i, \cdots, v_k\}, \qquad (3)$$

where $\mathbf{v}_i$ indicates the absence of vertex v_i. The kth cycle group (denoted $Z_k(K)$) is defined to be the kernel of ∂_k, whose members are called k-cycles. The kth boundary group is the image of ∂_{k+1} and is denoted $B_k(K)$. It follows that $B_k(K)$ is a subgroup of $Z_k(K)$ based on the property of boundary maps, $\partial_k \circ \partial_{k+1} = 0$. The kth homology group, $H_k(K)$, is defined to be the quotient group $Z_k(K)/B_k(K)$. The equivalent classes in $H_k(K)$ correspond to k-dimensional holes in K that cannot be deformed to eachother by adding or subtracting the boundary of a subcomplex.

Given a filtration as in equation (2), in addition to characterizing the homology group at each frame $H_k(K(x_i))$, we also want to track how topological features persist along the sequence. Viewing $H_k(K(x_i))$ as vector spaces together with inclusion map induced linear transformations gives a persistence module,

$$H_k(K(x_1)) \to H_k(K(x_2)) \to \cdots \to H_k(K(x_\ell)). \qquad (4)$$

An interval module with respect to $[b, d)$ denoted $\mathbb{I}_{[b,d)}$ is defined as a collection of vector spaces $\{V_i\}$ that are connected by linear maps, $f_i: V_i \to V_{i+1}$, where $V_i = \mathbb{Z}_2$ for $i \in [b, d)$ and $V_i = 0$ elsewhere and f_i is identity map when possible and zero otherwise. The persistence module in equation (4) can be decomposed as a direct sum of interval modules $\oplus_{[b,d) \in B} \mathbb{I}_{[b,d)}$. Each $\mathbb{I}_{[b,d)}$ corresponds to a homology class that appears at filtration value b and disappears at filtration value d (the values b and d are usually called the birth and death values). The collection of these pairs, B, encodes the evolution of k-dimensional holes when varying the filtration parameter and thus records the topological configuration of the input point cloud under different interactions ranges if a distance based filtration is used. Figure 3 illustrates filtration and persistence.

Mutation regions. Mutant residue locations were classified into interface and non-interface regions. Interface residues were further classified as the rim, support or core, and non-interface residues were also further classified as surface or interior, based on the classification approach by Levy[50].

Residue classification is mainly based on the change of relative residue accessible surface area (rASA) between protein–protein complex (rASA$_c$) and individual protein components of complex (rASA$_m$), as shown in Table 3. The accessible surface area was calculated with AREAIMOL from the CCP4 suite[51] and relative solvent accessibility was obtained by normalizing the absolute value with that of the same amino acid in a G–X–G peptide[52].

Data availability

All the data are available through the original papers cited or through our Code Ocean capsule (https://doi.org/10.24433/CO.0537487.v1).

Code availability

All source codes and models are publicly available through a Code Ocean compute capsule (https://doi.org/10.24433/CO.0537487.v1).

References

1. Chothia, C. et al. Conformations of immunoglobulin hypervariable regions. *Nature* **342**, 877–883 (1989).

2. Carter, P. J. Potent antibody therapeutics by design. *Nat. Rev. Immunol.* **6**, 343–357 (2006).

3. Demarest, S. J. & Glaser, S. M. Antibody therapeutics, antibody engineering, and the merits of protein stability. *Curr. Opin. Drug Discov. Dev.* **11**, 675–687 (2008).

4. Shire, S. J., Shahrokh, Z. & Liu, J. Challenges in the development of high protein concentration formulations. *J. Pharm. Sci.* **93**, 1390–1402 (2004).

5. Barouch, D. H. et al. Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* **503**, 224–228 (2013).

6. Glennie, M. J. & van de Winkel, J. G. Renaissance of cancer therapeutic antibodies. *Drug Discov. Today* **8**, 503–510 (2003).

7. Ben-Kasus, T., Schechter, B., Sela, M. & Yarden, Y. Cancer therapeutic antibodies come of age: targeting minimal residual disease. *Molecular Oncology* **1**, 42–54 (2007).

8. Geng, C., Xue, L. C., Roel-Touris, J. & Bonvin, A. M. Finding the $\Delta\Delta G$ spot: are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *WIREs Comput. Mol. Sci.* **9**, e1410 (2019).

9. Berman, H. M. et al. The protein data bank. *Nucl. Acids Res.* **28**, 235–242 (2000).

10. Sirin, S., Apgar, J. R., Bennett, E. M. & Keating, A. E. AB-bind: antibody binding mutational database for computational affinity predictions. *Protein Sci.* **25**, 393–409 (2016).

11. Moal, I. H. & Fernández-Recio, J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics* **28**, 2600–2607 (2012).

12. Patil, S. P., Ballester, P. J. & Kerezsi, C. R. Prospective virtual screening for novel p53–MDM2 inhibitors using ultrafast shape recognition. *J. Comput. Aided Mol. Des.* **28**, 89–97 (2014).

13. Demerdash, O. N. A., Daily, M. D. & Mitchell, J. C. Structure-based predictive models for allosteric hot spots. *PLOS Comput. Biol.* **5**, e1000531 (2009).

14. Kortemme, T., Morozov, A. V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *J. Mol. Biol.* **326**, 1239–1259 (2003).

15. Xiang, J. Z. & Honig, B. Jackal: *A Protein Structure Modeling Package.* (Columbia University and Howard Hughes Medical Institute: 2002.

16. Webb, B. & Sali, A. Comparative protein structure modeling using modeller. *Curr. Protoc. Bioinformatics* **47**, 5–6 (2014).

17. Zhu, K. et al. Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins Struct. Funct. Bioinformatics* **82**, 1646–1655 (2014).

18. Zhang, C., Liu, S. & Zhou, Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Science* **13**, 391–399 (2004).

19. Schymkowitz, J. et al. The foldx web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).

20. *Discovery Studio Modeling Environment* (Biovia, 2017).

21. 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 Res.* **44**, W469–W473 (2016).

22. Frosini, P. & Landi, C. Size theory as a topological tool for computer vision. *Pattern Recognition Image Anal.* **9**, 596–603 (1999).

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

24. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274 (2005).

25. Zomorodian, A. & Carlsson, G. Localized homology. *Comput. Geom.* **41**, 126–148 (2008).

26. Xia, K. L. & Wei, G. W. Persistent homology analysis of protein structure, flexibility and folding. *Int. J. Numer. Methods Biomed. Eng.* **30**, 814–844 (2014).

27. Gameiro, M. et al. Topological measurement of protein compressibility via persistence diagrams. *Japan J. Industr. Appl. Math.* **32**, 1–17 (2014).

28. Xia, K. L. & Wei, G. W. Persistent topology for cryo-EM data analysis. *Int. J. Numer. Methods Biomed. Eng.* **31**, e02719 (2015).

29. Cang, Z. X. et al. A topological approach to protein classification. *Mol. Based Math. Biol.* **3**, 140–162 (2015).

30. Yao, Y. et al. Topological methods for exploring low-density states in biomolecular folding pathways. *J. Chem. Phys.* **130**, 04B614 (2009).

31. Kovacev-Nikolic, V., Bubenik, P., Nikolić, D. & Heo, G. Using persistent homology and dynamical distances to analyze protein binding. *Stat. Appl. Genet. Mol. Biol.* **15**, 19–38 (2016).

32. Cang, Z. & Wei, G.-W. Integration of element specific persistent homology and machine learning for protein–ligand binding affinity prediction. *Int. J. Numerical Methods Biomed. Eng.* **34**, e2914 (2018).

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

34. Cang, Z. & Wei, G.-W. Topologynet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **13**, e1005690 (2017).

35. Cang, Z., Mu, L. & Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **14**, e1005929 (2018).

36. Nguyen, D. D. et al. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R grand challenges. *J. Comput. Aided Mol. Design* https://doi.org/10.1007/s10822-018-0146-6 (2018).

37. Frosini, P. A distance for similarity classes of submanifolds of a euclidean space. *Bull. Australian Math. Soc.* **42**, 407–415 (1990).

38. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Machine Learning Res.* **12**, 2825–2830 (2011).

39. 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* **35**, 462–469 (2018).

40. Xiong, P., Zhang, C., Zheng, W. & Zhang, Y. Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J. Mol. Biol.* **429**, 426–434 (2017).

41. Lensink, M. F. & Wodak, S. J. Docking, scoring, and affinity prediction in CAPRI. *Proteins Struct. Funct. Bioinformatics* **81**, 2082–2095 (2013).

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

43. Dehouck, Y., Kwasigroch, J. M., Rooman, M. & Gilis, D. Beatmusic: prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Research* **41**, W333–W339 (2013).

44. Petukh, M., Dai, L. & Alexov, E. SAAMBE: webserver to predict the charge of binding free energy caused by amino acids mutations. *Int. J. Mol. Sci.* **17**, 547 (2016).

45. 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. Bioinformatics* **56**, 93–101 (2004).

46. Jemimah, S., Yugandhar, K. & Michael Gromiha, M. Proximate: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics* **33**, 2787–2788 (2017).

47. Liu, Q., Chen, P., Wang, B., Zhang, J. & Li, J. dbMPIKT: A database of kinetic and thermodynamic mutant protein interactions. *BMC Bioinformatics* **19**, 455 (2018).

48. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. mCSM-PPI2: Predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338– W344 (2019).

49. Petukh, M., Li, M. & Alexov, E. Predicting binding free energy change caused by point mutations with knowledge-modified MM/PBSA method. *PLoS Comput. Biol.* **11**, e1004276 (2015).

50. Levy, E. D. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* **403**, 660–670 (2010).

51. Collaborative, C. P. et al. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760 (1994).

52. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656 (1987).

Acknowledgements

Author contributions

G.W.-W. was responsible for conceptualization, supervision and funding acquisition. M.W. and Z.C. designed the project. M.W. curated the data. All authors carried out the investigations and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-0149-6.

Correspondence and requests for materials should be addressed to G.-W.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.