nature machine intelligence

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s42256-020-0149-6

In the format provided by the authors and unedited.

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

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

[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

Supplementary material to the Paper,"A topology-based network tree for the prediction of protein-protein binding affinity changes upon mutation"

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

This document contains additional information about methods and discussion in the paper TopNetTree which were not necessary to include in the central part of paper but might be of interest to readers. This supplementary material contains the following sections:

Contents

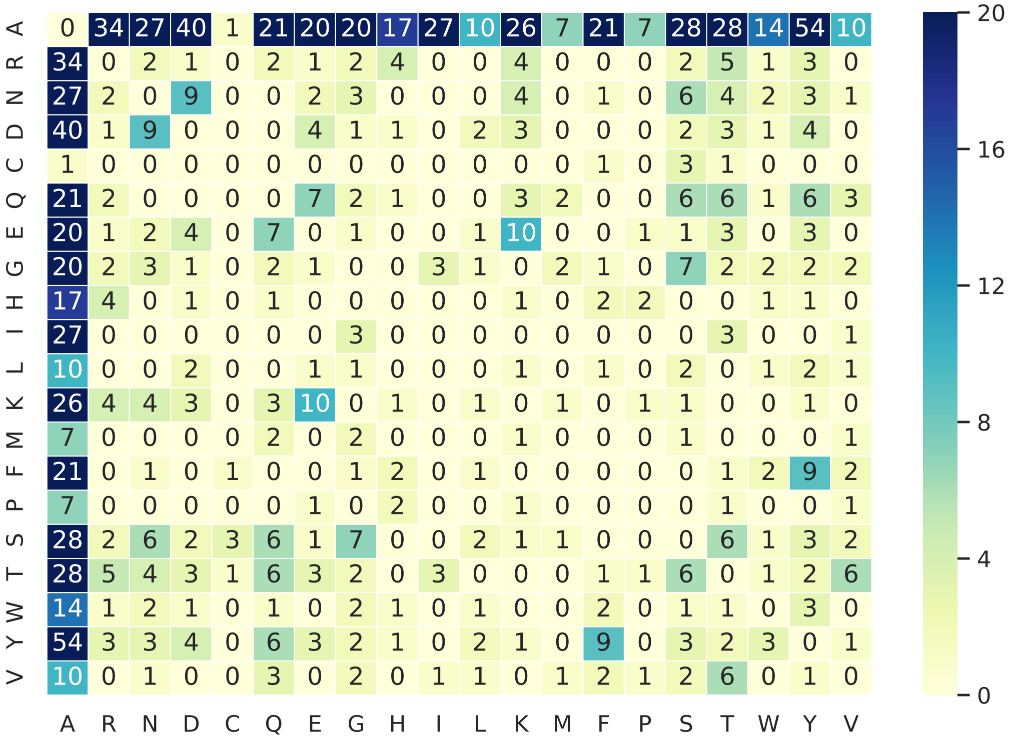
[*1] Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA
[2] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA.
[3] Department of Electrical and Computer Engineering Michigan State University, East Lansing, MI 48824, USA.
To whom correspondence should be addressed. Email: weig@msu.edu

Supplementary Tables

$\mathcal{A}$	$\mathcal{B}$	Dis.	Complex	Dim.
$\mathcal{A}_{\rm m} \cap \mathcal{A}_{\rm ele}({\rm E}_1)$	$\mathcal{A}_{\rm mn}(r) \cap \mathcal{A}_{\rm ele}({\rm E}_2)$	$D_{\rm mod}$	Rips	H_0
$\mathcal{A}_{\rm m} \cap \mathcal{A}_{\rm ele}({\rm E}_1)$	$\mathcal{A}_{\rm mn}(r) \cap \mathcal{A}_{\rm ele}({\rm E}_2)$	$D_{\rm e}$	alpha	H_1, H_2
$\mathcal{A}_{\rm Ab}(r) \cap \mathcal{A}_{\rm ele}({\rm E}_1)$	$\mathcal{A}_{\rm Ag}(r) \cap \mathcal{A}_{\rm ele}({\rm E}_2)$	$D_{\rm mod}$	Rips	H_0
$\mathcal{A}_{\rm Ab}(r) \cap \mathcal{A}_{\rm ele}({\rm E}_1)$	$\mathcal{A}_{\rm Ag}(r) \cap \mathcal{A}_{\rm ele}({\rm E}_2)$	$D_{\rm e}$	alpha	H_1, H_2

Supplementary Table 1: Summary of topological descriptors. Choices for $\rm E_1$ and $\rm E_2$ are $\{\rm C\}$, $\{\rm N\}$, and $\{\rm O\}$. The barcodes are generated upon mutant and wild type complexes.

Supplementary Figures



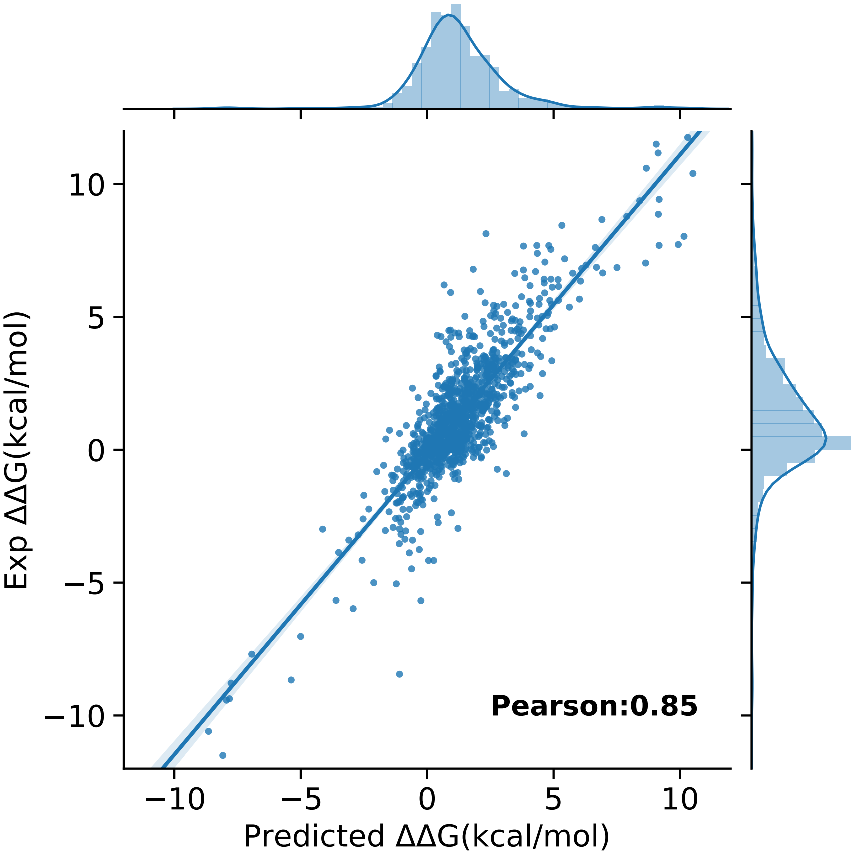
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	34	27	40	1	21	20	20	17	27	10	26	7	21	7	28	28	14	54	10
R	34	0	2	1	0	2	1	2	4	0	0	4	0	0	0	2	5	1	3	0
N	27	2	0	9	0	0	2	3	0	0	0	4	0	1	0	6	4	2	3	1
D	40	1	9	0	0	0	4	1	1	0	2	3	0	0	0	2	3	1	4	0
C	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0	0	0
Q	21	2	0	0	0	0	7	2	1	0	0	3	2	0	0	6	6	1	6	3
E	20	1	2	4	0	7	0	1	0	0	1	10	0	0	1	1	3	0	3	0
G	20	2	3	1	0	2	1	0	0	3	1	0	2	1	0	7	2	2	2	2
H	17	4	0	1	0	1	0	0	0	0	0	1	0	2	2	0	0	1	1	0
I	27	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	3	0	0	1
L	10	0	0	2	0	0	1	1	0	0	0	1	0	1	0	2	0	1	2	1
K	26	4	4	3	0	3	10	0	1	0	1	0	1	0	1	1	0	0	1	0
M	7	0	0	0	0	2	0	2	0	0	0	1	0	0	0	1	0	0	0	1
F	21	0	1	0	1	0	0	1	2	0	1	0	0	0	0	0	1	2	9	2
P	7	0	0	0	0	0	1	0	2	0	0	1	0	0	0	0	1	0	0	1
S	28	2	6	2	3	6	1	7	0	0	2	1	1	0	0	0	6	1	3	2
T	28	5	4	3	1	6	3	2	0	3	0	0	0	1	1	6	0	1	2	6
W	14	1	2	1	0	1	0	2	1	0	1	0	0	2	0	1	1	0	3	0
Y	54	3	3	4	0	6	3	2	1	0	2	1	0	9	0	3	2	3	0	1
V	10	0	1	0	0	3	0	2	0	1	1	0	1	2	1	2	6	0	1	0

Supplementary Figure 1: Counts of mutation types in AB-Bind database. Reverse mutations are also counted in the matrix.

Supplementary Discussion

S1131 dataset

The SKEMPI dataset[1] contains 3047 binding free energy changes upon mutation, while 2317 of them are single-point mutation data entries, denoted the SKEMPI S2317 set. Xiong *et al.* selected a subset of 1131 non-redundant interface single-point mutations from the SKEMPI S2317 set[2] , called set S1131. We applied our model to set S1131 by the 10-fold cross-validation. Result is shown below.



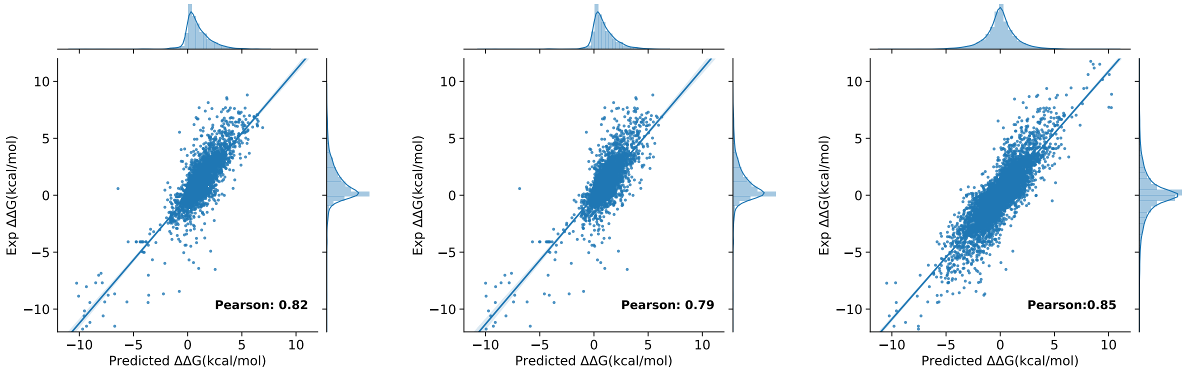
Supplementary Figure 2: Performance evaluation on the 10-fold cross-validation on set S1131. TopNetTree model was able to achieve the R_p of 0.85 and RMSE of 1.55 kcal/mol.

S4947, S4169 and S8338 datasets

The SKEMPI 2.0[3] database is the updated version of the SKEMPI database and contains new mutations collected after the first version was released. There are 7085 mutations in the SKEMPI 2.0 dataset. We choose only single-point mutations with full energy change information, called set S4947. Since binding energy changes upon mutation ($\Delta\Delta G$) are not directly given in the SKEMPI 2.0 database, the following formula is used to obtain the $\Delta\Delta G$ value for each mutation with a given kd value:

$$\begin{aligned}\Delta G &= \frac{8.314}{4184} \times (273.15+25) \times log(kd)\\ \Delta\Delta G &= \Delta G_{MT} - \Delta G_{WT}.\end{aligned}$$

Set S4169 is directly adopted from mCSM-PPI2[4] paper, which is also derived from the SKEMPI 2.0 dataset. Set S8338 is derived from the S4169 set by setting the reverse mutation energy change with a negative sign[4] . We tested our model on S4947 , S4169 and S8338 datasets. For set S4947, we carry out the regular 10-fold cross-validation 10 times. For S4169 and S8338 sets, we follow the 10-fold stratified cross-validation used in mCSM-PPI2 paper.[4] The following is the result of our test.

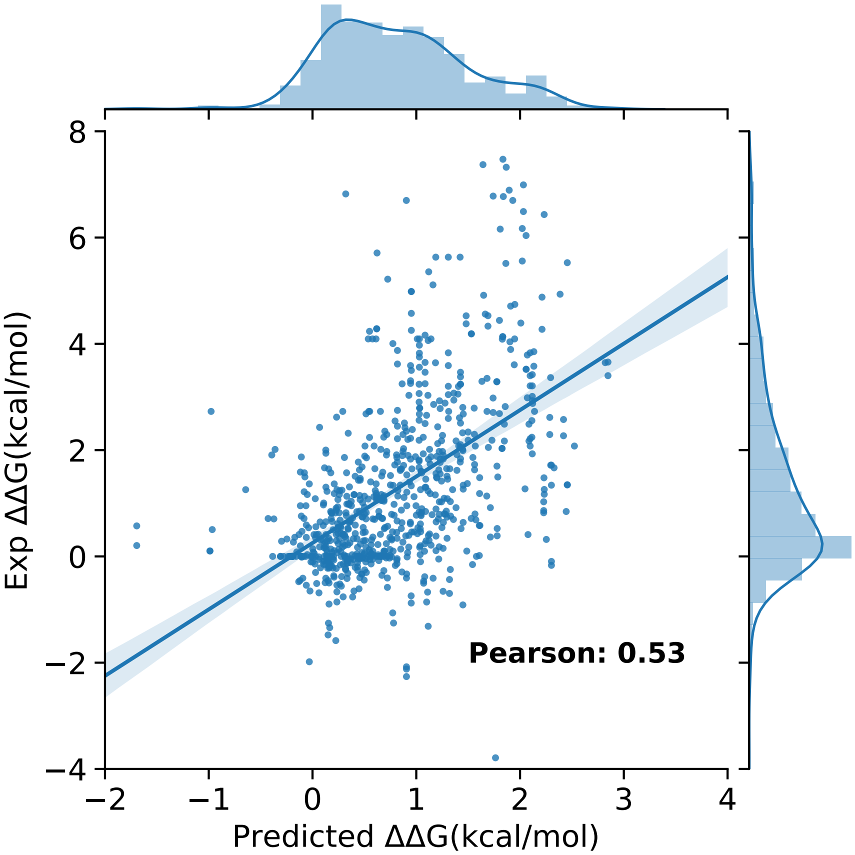


(a) (b) (c)

Supplementary Figure 3: Performance evaluation using 10-fold cross-validations. (a) Set S4947 with the R_p of 0.82 and RMSE of 1.11 kcal/mol. (b) Set S4168 with the R_p of 0.78 and RMSE of 1.13 kcal/mol. (c) Set S8338 with R_p of 0.85 and RMSE of 1.09 kcal/mol.

Blind test of an AB-Bind dataset based on the S4947 dataset

Since the SKEMPI 2.0 dataset includes entries from the AB-Bind dataset, we design a blind test based on AB-Bind and SKEMPI 2.0 sets. From 24 protein complexes existing in both the AB-Bind dataset and the SKEMPI 2.0 dataset, we collect 787 single-point mutations, denoted set S787, as our test set. We construct a training set by excluding set S787 from the SKEMPI 2.0 S4947 set. The results of this training and test are given below.



Supplementary Figure 4: Performance evaluation of a blind prediction. The AB-Bind S787 set is the test. The training is constructed from the SKEMPI 2.0 S4947 dataset, excluding the AB-bind S787 set. TopNetTree model was able to achieve the average R_p of 0.53 and $RMSE$ of 1.45 kcal/mol.

Protein level non overlapping test on the AB-Bind S645 set

To further test the predictive power of our model, we applied protein level non overlapping test on the AB-Bind S645 set. All the 645 mutations in the dataset could be separated into 24 different protein-protein complexes (we merged the complex and its homology model as one category since they are very similar). To perform a non-overlapping test, all the mutations in one specific protein complex are designed as the test set and all other mutations are designed as the training set. By doing this training test splitting, protein complex in the test set is guaranteed to be not in the training set. The result of non-overlapping test of 24 protein complexes is shown in Table 2

Name	Counts	R_p	RMSE(kcal/mol)
1AK4	16	0.528	0.837
1BJ1	19	0.103	1.502
1CZ8	19	0.506	1.077
1DQJ	21	0.568	1.877
1DVF	26	0.553	1.163
1FFW	9	-0.043	1.052
1JRH	2	1	0.812
1JTG	5	0.757	0.549
1KTZ/HM_1KTZ	44	0.866	0.496
1MHP	68	0.505	3.216
1MLC	11	0.397	1.508
1N8Z	34	0.626	2.273
1VFB	41	0.689	1.653
1YY9/HM_1YY9	21	-0.068	1.531
2JEL	43	0.818	0.954
2NYY/HM_2NYY	53	0.589	1.238
2NZ9/HM_2NZ9	35	0.665	1.367
3BDY	34	0.615	0.692
3BE1	34	0.474	0.941
3BN9/HM_3BN9	43	0.368	1.743
3HFM	22	0.262	2.69
3K2M	7	0.705	1.416
3NGB	11	0.459	1.147
3NPS	27	0.242	0.953
Total	645		
Average	27	0.508	1.362
Median	24	0.541	1.201

Supplementary Table 2: Result of non-overlapping protein level test on AB-Bind S645 set, including Pearson correlation coefficient and RMSE in kcal/mol.

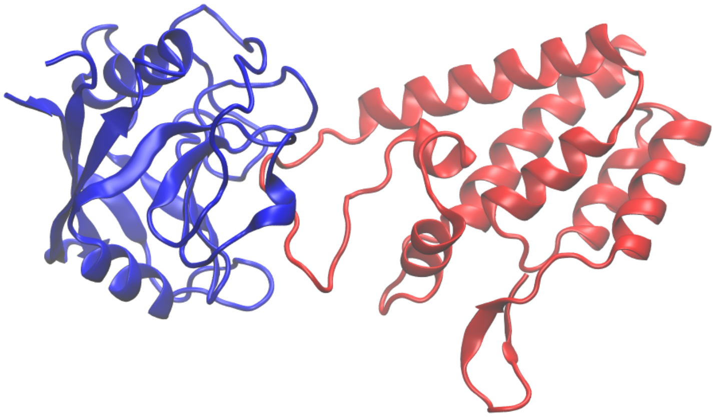
Protein level leave-one-out validation test

To further test the predicting power of our model, we applied protein-level leave-one-out cross-validation test on the AB-Bind S645 set. All the 645 mutations in the dataset are separated into 24 different protein-protein complexes (we merged the complex and its homology model as one category since they are very similar). Then mutations in each protein complex are used in leave-one-complex-out cross-validation. The result of the test of 24 protein complexes is shown in Table 3.

Name	Counts	R_p	RMSE(kcal/mol)
1AK4	16	0.139	0.977
1BJ1	19	0.392	1.350
1CZ8	19	0.671	0.921
1DQJ	21	0.318	1.885
1DVF	26	-0.103	1.345
1FFW	9	0.149	0.512
1JRH	2	-1	0.13
1JTG	5	-0.998	0.689
1KTZ/HM_1KTZ	44	0.973	0.222
1MHP	68	0.145	3.432
1MLC	11	0.606	0.418
1N8Z	34	0.029	3.023
1VFB	41	0.533	1.755
1YY9/HM_1YY9	21	0.547	0.225
2JEL	43	0.707	0.968
2NYY/HM_2NYY	53	0.514	1.175
2NZ9/HM_2NZ9	35	0.121	1.732
3BDY	34	0.604	0.548
3BE1	34	0.054	1.077
3BN9/HM_3BN9	43	0.281	1.938
3HFM	22	-0.121	2.726
3K2M	7	-0.993	1.234
3NGB	11	0.411	1.186
3NPS	27	0.099	0.766
Total	645		
Average	27	0.170	1.218
Median	24	0.215	1.027

Supplementary Table 3: Result of protein-level leave-one-complex-out-validation test on AB-Bind S645 set, including average Pearson correlation coefficient and RMSE in kcal/mol.

Alanine mutation test of 1AK4



Supplementary Figure 5: Structure of protein complex 1AK4, chain A in blue and chain D in red

	Interior		Surface		Rim		Support		Core		All	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
Arg	0.8435	0	1.7463	0.5962	1.6302	0	–	–	1.5676	0	1.5466	0.4017
Asn	1.0064	0.1031	1.3790	1.2943	1.4870	0	1.6727	0	–	–	1.2581	0.5352
Asp	1.0726	0.1059	0.8942	0.0837	–	–	–	–	–	–	0.9707	0.1010
Cys	0.8236	0.0172	0.6246	0	–	–	–	–	–	–	0.7739	0.0203
Gln	–	–	0.9425	0	–	–	2.8500	0.0041	–	–	2.2142	0.8114
Glu	0.8466	0.0174	1.2533	0.3848	–	–	–	–	–	–	1.1794	0.3426
Gly	1.0956	0.4595	0.6091	0.1135	–	–	–	–	1.3921	0	0.9322	0.3761
His	1.2941	0	1.1008	0	–	–	1.7063	0.0410	–	–	1.4519	0.0899
Ile	1.0031	0.1280	0.1595	0	–	–	0.8719	0	–	–	0.9056	0.1658
Leu	1.2601	0.1447	1.8473	0.8087	–	–	1.5133	0	–	–	1.4641	0.3798
Lys	–	–	0.6113	0.1443	–	–	–	–	–	–	0.6113	0.1443
Met	2.2145	0.4070	0.9892	0	–	–	1.9440	0.2721	–	–	1.9153	0.4696
Phe	2.1307	0.9728	1.0778	0.0046	–	–	1.7958	0.2721	–	–	1.9457	0.8788
Pro	0.8306	0	0.7735	0.3486	–	–	–	–	–	–	0.7831	0.2909
Ser	1.0374	0.1150	0.3301	0.0010	–	–	–	–	–	–	0.8606	0.1803
Thr	1.1284	0.1641	0.8129	0.0432	1.1802	0	–	–	–	–	0.9898	0.1205
Trp	–	–	–	–	1.3124	0	–	–	–	–	1.3124	0
Tyr	2.5878	0	1.0924	0	–	–	–	–	–	–	1.8401	0.5590
Val	1.0002	0.0229	0.4613	0	–	–	–	–	–	–	0.9403	0.0490

Supplementary Table 4: Alanine mutation test on 1AK4 chain A, using TopNetTree model with AB-Bind training data. All the $\Delta\Delta G$ values are in kcal/mol. In total there are 165 residues in the chain A of 1AK4. Results of the alanine mutation are also separated into 5 region groups, interior, surface, rim, support, and core, respectively.

Supplementary Methods

Auxiliary features

As we mentioned in the feature generation part of in the main part of paper, element-specific and site-specific persistent homology is able to embed chemical information into topological representations. However, there are other important chemical and physical information that has not been incorporated into element-specific persistent homology but could improve the predictive power of the present topological model. In this work, all none topological features are named as auxiliary features. These features are appended into the machine learning model at the last step of GBT training or the dense layer of a neural network. In general, auxiliary features are categorized into atom-level features and residue-level ones.

Atom-level features

According to different criteria, atoms can be categorized into different groups for feature generation. First, with respect to atom types, we divide atoms into 7 groups, i.e., $\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}, \mathrm{H}$, all heavy atoms, and all atoms. Additionally, with respect to distance to mutation site, atoms are grouped into 3 groups, namely, mutation site atoms, near mutation site atoms (within 10Å of mutation site), and all atoms. Finally, similar to the treatment in topological feature generation, 3 cases, i.e., wild type, mutant type, and their difference are considered, respectively.

- **Surface areas** Atom-level solvent excluded surface areas are computed through our in-house software ESES.[5] All atom areas within the same group are summed as one feature. In this manner, a total of 7*3*3 = 63 features is generated.

- **Partial charges** Partial charge of each atom is generated from pdb2pqr software[6] using the amber force field. After the procedure, the radius and the partial charge of each atom are calculated. The sum of the partial charges and the sum of absolute values of partial charges for each atomic group are counted as partial charge features. In this manner, a total of 7*3*3*2 =126 features is generated.

- **Coulomb interactions** Coulomb energy of the ith single atom is calculated as the sum of pairwise coulomb energy with every other atom.

  $$C_i = \sum_{j,j\neq i} k_e \frac{q_i q_j}{r_{ij}}. \tag{1}$$

  Here, k_e is the Coulomb's constant. Since multiplying the constant coefficient has no effect on machine learning result, we use $k_e = 1$ in our calculation.

  In coulomb interaction feature generation, only 5 groups ($\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}$, and all heavy atoms) are counted. Both coulomb interaction energy and absolute value are counted. In this manner, a total of 5*3*3*2 = 90 features is generated.

- **van der Waals interaction** The van der Waals energy of the ith atom is modeled as the sum of pairwise Lennard-Jones potentials with every other atom. Only 5 groups ($\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}$, and all heavy atoms) are counted.

  $$V_i = \sum_{j,j\neq i} \epsilon \left[\left(\frac{r_i + r_j}{r_{ij}} \right)^{12} - 2 \left(\frac{r_i + r_j}{r_{ij}} \right)^{6} \right]. \tag{2}$$

  Here, ϵ is the depth of the potential well. Since multiplying the constant coefficient has no effect on machine learning result, we use $\epsilon = 1$ in our calculation. In this manner, a total of 5*3*3 = 45 features is generated.

- **Electrostatic solvation free energy** Electrostatic solvation free energy of each atom is calculated using Poisson-Boltzmann model through our in-house software MIBPB.[7–9] By summing up all the solvation free energies in same atom groups, 7*3*3 = 63 features are generated.

Residue-level features

- **Mutation site neighborhood amino acid composition** The residues within 10 Å of the mutation site are regarded as neighbor residues. Distances between residues are calculated using their alpha carbon atoms. Amino acid residues are divided into 5 groups as hydrophobic, polar, positively charged, negatively charged and special cases. The count and percentage of the 5 groups of amino acids in neighbor site are regarding as the environment composition features of the mutation site, which leads to 5*2 = 10 features. Also, the sum, average and variance of residue volumes, surface areas, weights and hydropathy scores are generated as the environment chemical and physical features of a mutation site, which leads to 3*4 = 12 features. In this manner, 10+12 = 22 features are generated.

- **pK_a shifts** The pK_a values of 7 ionizable amino acids, namely, ASP, GLU, ARG, LYS, HIS, CYS, and TYR, are calculated using the PROPKA software.[10] The difference of pK_a values between a wild type and its mutant type are calculated as pK_a shifts. The maximum, minimum, sum, the sum of absolute values, the minimum of absolute value of total pK_a shifts are calculated, which leads to 5 features. Also, besides the shifts of all groups, the sum and the sum of absolute value of pK_a shifts based on the 7 ionizable amino acid groups are calculated, which leads to 2*7=14 features. In this manner, 5+14 = 19 features are generated.
- **Secondary structures** Using SPIDER2[11] software, the probability score of mutation site residues to be coil, helix or strand are calculated as well as torsion angles. The wild type, the mutant type and their difference are calculated as secondary structure features. In this manner, 4*3 =12 features are generated.

Preprocessing of dataset

For the aforementioned databases, crystal structures of the wild type, mutation type, and binding affinity change are given for each data entry. To calculate our structure-based topological feature, the structures of mutant type are also needed. Scap utility in the Jackal package[12] is used to generate mutant structures. This utility predicts side-chain conformations on a given backbone. To fix the missing atoms and residues, the profix utility in the Jackal package[12] is applied to all raw pdb files.

Model parametrization and software used

The details of model parameters and software packages are given below.

TopGBT: Topology based GBT model

- H_1 and H_2 features. Element-specific persistent homology H_1 and H_2 barcodes are constructed as described in Table 1 with cutoff value $r = 12$Å. We consider a wide type and mutant complexes. For each barcode, we extract birth death and persistence information. Statistical values, namely sum, min, max, mean, and standard deviation are computed from these barcodes to generate H_1 and H_2 features, giving rise to a total of 540 features.

TopCNN: Topology based CNN model

- H_0 feature. The same as what described above, except for a finer bin size of 0.25 Å, which leads to a total of 1296 features for CNN.
- Four 1D convolutional layers and one dropout layer have been used in the CNN model.

TopNetTree: Topology based network tree model

- H_0 features. Top 300 high-level CNN features are selected according to their feature importance.
- H_1 and H_2 auxiliary features are the same as those in the TopGBT model.

Model parameters

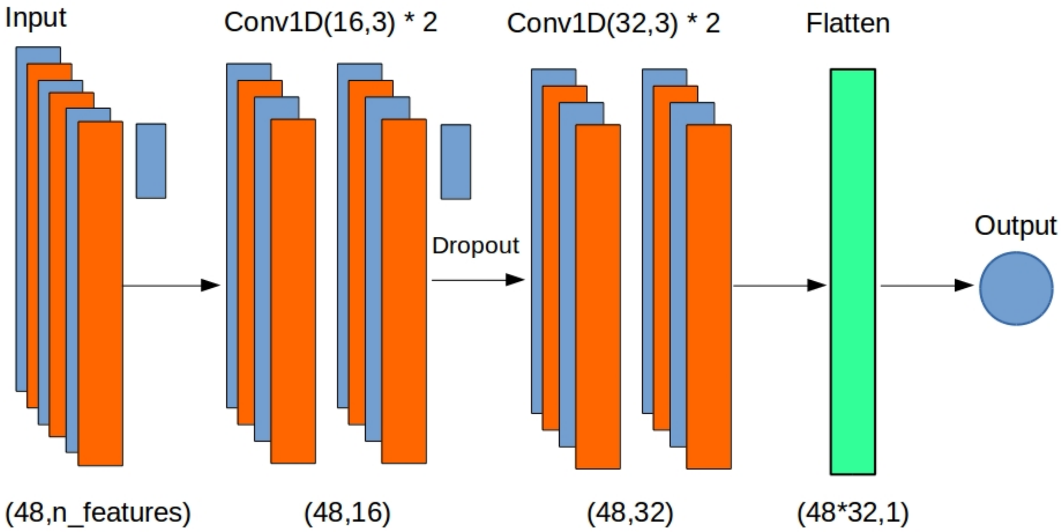
- CNN network structure and parameters are shown in Supplementary Figure 6
- GBT parameters $n\_estimators = 20000, max\_depth = 6, min_samples\_split = 3$, and $learning\_rate = 0.001$.

Software used

- GBT. The scikit-learn (version 0.18.1)[13] is used for the gradient boost regressor function.
- CNN. The Keras (version 2.0.2)[14] package is used for convolutional neural network model.
- Persistent homology feature: Javaplex[15] is used to generate H_0 barcodes and TDA package in R[16] is used to generate H_1 and H_2 barcodes.

Time and memory cost

- All the models are generated and tested on computer facilities at Michigan State University's High performance computing center (HPCC). 8 GB of memory and 5 cpu cores are requested for each feature generation job.
- Average running time for generating topological features for one sample is 1.01 min (time for generating mutant structure is included).
- Average running time for generating auxiliary features is 9.21 min .



Supplementary Figure 6: Illustration of CNN parameters. This CNN network structure contains two 1D convolutional layer of 64 channels and two 1D convolutional layer of 128 channels and 1 flatten layers. On the convolutional dimension, 12 Åcut off and 0.25 Åbin size was chosen, so 48 bins are the size for that dimension. Other parameters for the CNN are listed as follow: $kernal_initializer =' lecun_uniform'$, $optimizer = adam$ and $epochs = 2000$

Evaluation Criteria

- Two evaluation metrics, Pearson's correlation coefficient (R_p) and root-mean-squared error (RMSE), are used to assess the quality of predictions. Let x and y be the vector of predicted values and the ground truth of the n samples, respectively. The definition of R_p is given by

  $$R_p = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3}$$

  where $\bar{x}$ and $\bar{y}$, the means of x and y, respectively. RMSE is computed as

  $$\text{RMSE} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2/n} \tag{4}$$

  For cross validation, the R_p and RMSE of all folds are averaged.

- We use 10-fold cross validation on the AB-bind database for all of our models. Reported values are the averages of 50 individual trials with different random seeds.

Supplementary Data Guide

Supplementary data are given in "SupplementaryData.xlsx". This supplementary file contains 5 spread sheets of detailed dataset information mentioned in the paper, including AB_bind_6454, S1131, S4947, S4169, S8338 and S787. PDBID, mutation type, mutation site, experiment ddg and predicted ddg are given in each spread sheet.

Supplementary Reference

[1] 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).

[2] 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* **429**, 426–434 (2017).

[3] 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).

[4] 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 Research* (2019).

[5] Liu, B., Wang, B., Zhao, R., Tong, Y. & Wei, G. W. ESES: software for Eulerian solvent excluded surface. *Journal of Computational Chemistry* **38**, 446–466 (2017).

[6] Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic acids research* **32**, W665–W667 (2004).

[7] Zhou, Y. C., Zhao, S., Feig, M. & Wei, G. W. High order matched interface and boundary method for elliptic equations with discontinuous coefficients and singular sources. *J. Comput. Phys.* **213**, 1–30 (2006).

[8] Zhou, Y. C., Feig, M. & Wei, G. W. Highly accurate biomolecular electrostatics in continuum dielectric environments. *Journal of Computational Chemistry* **29**, 87–97 (2008).

[9] Geng, W., Yu, S. & Wei, G. W. Treatment of charge singularities in implicit solvent models. *Journal of Chemical Physics* **127**, 114106 (2007).

[10] Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pka values for protein–ligand complexes. *Proteins: Structure, Function, and Bioinformatics* **73**, 765–783 (2008).

[11] Yang, Y. *et al.* Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In *Prediction of protein secondary structure*, 55–63 (Springer, 2017).

[12] Xiang, J. Z. & Honig, B. Jackal: A protein structure modeling package. *Columbia University and Howard Hughes Medical Institute, New York* (2002).

[13] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).

[14] Chollet, F. Keras. `https://github.com/fchollet/keras` (2015).

[15] Adams, H., Tausz, A. & Vejdemo-Johansson, M. Javaplex: A research software package for persistent (co) homology. In *International Congress on Mathematical Software*, 129–136 (Springer, 2014).

[16] Fasy, B. T., Kim, J., Lecci, F. & Maria, C. Introduction to the r package tda. *arXiv preprint arXiv:1411.1830* (2014).