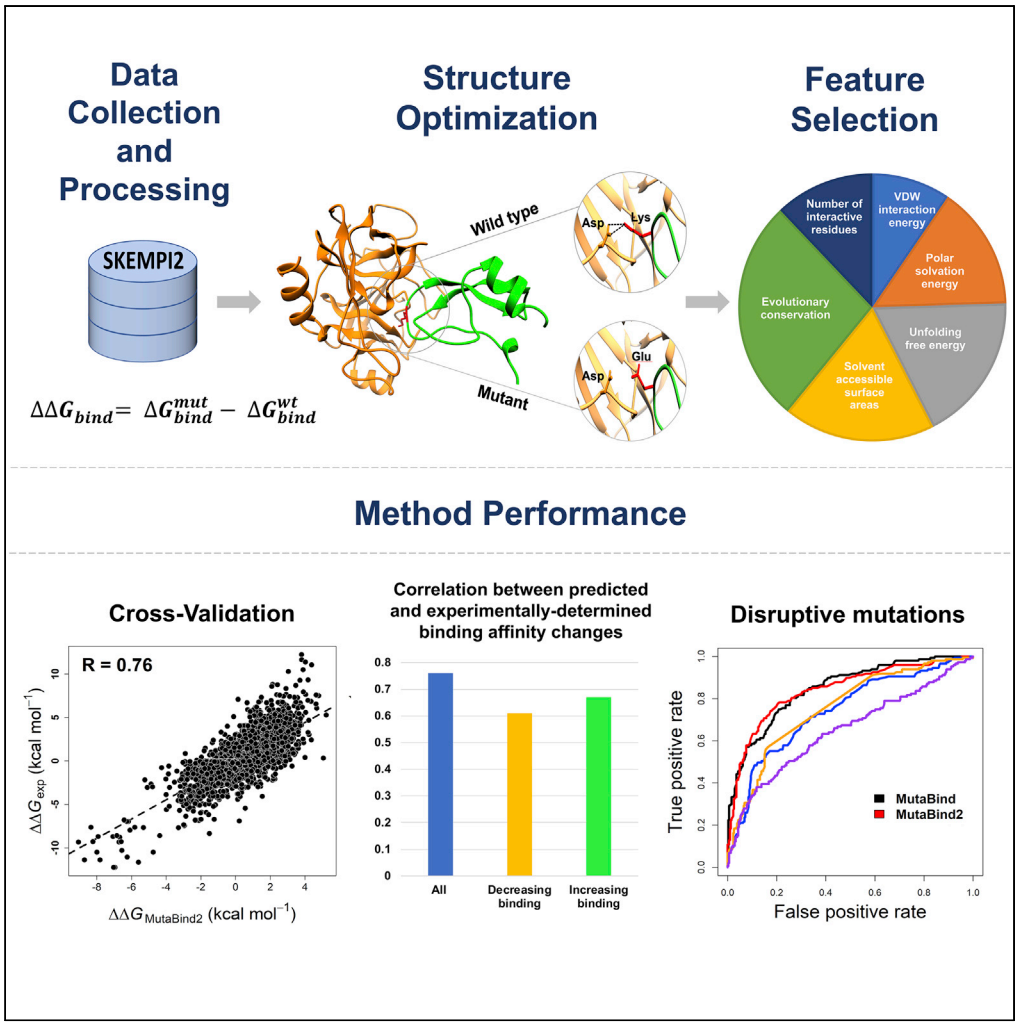
iScience

CellPress

Article

MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions



Ning Zhang, Yuting Chen, Haoyu Lu, ..., Alexander Goncearenco, Anna R. Panchenko, Minghui Li

panch@ncbi.nlm.nih.gov (A.R.P.)
minghui.li@suda.edu.cn (M.L.)

HIGHLIGHTS

A new method to predict binding affinity changes upon single and multiple mutations

Improved performance in evaluating the effects of mutations increasing binding affinity

Generation of the structural model of a mutant complex

Zhang et al., iScience *23*, 100939
March 27, 2020
https://doi.org/10.1016/j.isci.2020.100939

iScience

CellPress

Article

MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions

Ning Zhang,[1,6] Yuting Chen,[1,6] Haoyu Lu,[1] Feiyang Zhao,[1] Roberto Vera Alvarez,[2] Alexander Goncearenco,[2,5] Anna R. Panchenko,[2,3,4,*] and Minghui Li[1,7,*]

SUMMARY

Missense mutations may affect proteostasis by destabilizing or over-stabilizing protein complexes and changing the pathway flux. Predicting the effects of stabilizing mutations on protein-protein interactions is notoriously difficult because existing experimental sets are skewed toward mutations reducing protein-protein binding affinity and many computational methods fail to correctly evaluate their effects. To address this issue, we developed a method MutaBind2, which estimates the impacts of single as well as multiple mutations on protein-protein interactions. MutaBind2 employs only seven features, and the most important of them describe interactions of proteins with the solvent, evolutionary conservation of the site, and thermodynamic stability of the complex and each monomer. This approach shows a distinct improvement especially in evaluating the effects of mutations increasing binding affinity. MutaBind2 can be used for finding disease driver mutations, designing stable protein complexes, and discovering new protein-protein interaction inhibitors.

INTRODUCTION

Protein-protein interactions mediate many biological processes, and missense mutations may affect protein interactions and interaction networks leading to dysfunctional proteins and their complexes, pathway dysregulation, and potentially to diseases (Teng et al., 2009; Nishi et al., 2013; Creixell et al., 2015; Tee et al., 2019). Indeed, several recent studies systematically characterized thousands of disease mutations and found that many of them were located on protein-binding interfaces and induced macromolecular interaction perturbations, whereas neutral variants retained most interactions (Nishi et al., 2013; Teng et al., 2009; Creixell et al., 2015; Sahni et al., 2015; Wang et al., 2012; An et al., 2013; Cukuroglu et al., 2014; Tan et al., 2019; Ozdemir et al., 2018). However, not all mutations have severe damaging impacts, and the majority of mutations produce rather subtle effects with unclear clinical significance. Quantification of these effects on specific protein-protein interactions requires assessing the changes in binding affinity induced by mutations. These effects can be quite accurately measured by low-throughput experiments. However, large-scale rapid experimental assays that would allow the assessment of thousands of variants are still limited. The development of reliable computational approaches to predict the effects of missense mutations on protein complexes would enable the prioritization of functionally disrupting mutations and provide a basis for understanding the molecular mechanisms of their impacts.

Several computational approaches have been proposed so far to calculate the changes in binding affinity by mutations (Li et al., 2014, 2016b; Dehouck et al., 2013; Petukh et al., 2015, 2016; Kruger and Gohlke, 2010; Pires et al., 2014; Xiong et al., 2017; Brender and Zhang, 2015; Zhao et al., 2014; Rodrigues et al., 2019; Geng et al., 2019; Jemimah et al., 2019). In the past we developed two methods to address this pressing need. The first method used the modified MM/PBSA (Molecular Mechanics Poisson–Boltzmann Surface Area) approach and structure optimization protocol with an explicit solvent model (Li et al., 2014). Later we came up with another method, MutaBind (Li et al., 2016b). MutaBind was characterized by higher prediction accuracy and speed, making it possible to implement it as a web server, which has been used to quantify the impacts of mutations in a wide range of protein complexes (https://mutabind.org/v1). For instance, it was successfully applied to assess the effects of cancer mutations on binding between CBL ubiquitin ligase and E2 conjugating enzyme where computationally predicted binding affinity changes were compared with the experiments using cancer and non-cancer cell lines (Li et al., 2016a).

[1]Center for Systems Biology, Department of Bioinformatics, School of Biology and Basic Medical Sciences, Soochow University, Suzhou 215123, China

[2]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA

[3]Present address: Department of Pathology and Molecular Medicine, School of Medicine, Queen's University, ON, Canada

[4]Present address: Ontario Institute of Cancer Research, Toronto, ON, Canada

[5]Present address: Translational and Functional Genomics Branch, National Human Genome Research, National Institutes of Health, Bethesda, MD 20892, USA

[6]These authors contributed equally

[7]Lead Contact

*Correspondence: panch@ncbi.nlm.nih.gov (A.R.P.), minghui.li@suda.edu.cn (M.L.)

https://doi.org/10.1016/j.isci.2020.100939

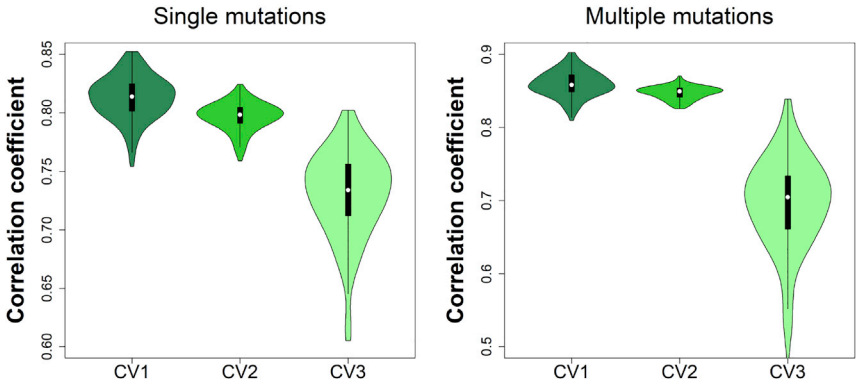


Figure 1. Pearson Correlation Coefficients between Experimental and Calculated $\Delta\Delta G$ for Three Types of Cross-Validation Tests on the S4191 (Single Mutations) and M1707 (Multiple Mutations) Sets

See also Table S1.

The functional effects of mutations decreasing binding affinity are better understood compared with the effects of mutations increasing binding affinity. However, the latter may also have important consequences leading to disruption of proteostasis by over-stabilizing transient protein complexes (Nishi et al., 2013; Stefl et al., 2013; Rutten et al., 2018; Shoichet et al., 1995; Nagatani et al., 2007; Witham et al., 2011; Jubb et al., 2016) and changing the pathway flux. Critically, existing computational methods perform much better for mutations decreasing than for mutations increasing binding affinity. Several studies tried to determine factors contributing to this bias by comparing the methods' performance using experimental data on changes of protein stability (Usmanova et al., 2018; Montanucci et al., 2018; Pucci et al., 2018). These studies concluded that all computational methods produced predictions that were immensely skewed toward higher accuracy for mutations decreasing binding affinity and the amplitude of this bias increased with the number of introduced mutations in a protein (Usmanova et al., 2018). There are several reasons for such predisposition of proteins toward mutations with decreasing effects: proteins and their binding interfaces tend to be optimized in evolution with regard to their stability; the existing experimental sets are enriched with mutations that decrease binding and the methodologies of training procedures employed by many computational methods rely on these experimental sets. Correcting for such bias in performance is demanding as it requires developing new energy functions and enriching the datasets with mutations increasing binding affinity either by using additional experimental data or by calculating the effects of reverse mutations and modeling their mutant structures.

To address this issue, we developed a new method, MutaBind2, with significantly improved performance (https://mutabind.org/). MutaBind2 uses a new minimization protocol and scoring function composed of seven terms. In addition to single mutations, it can predict the effects of multiple mutations on protein binding affinity. MutaBind2 can be applied to a large number of tasks, including, but not limited to, finding disease driver mutations and understanding their molecular mechanisms, assessing the effects of sequence variants on protein fitness, structural modeling of mutant complexes, and designing protein interaction inhibitors (Goncearenco et al., 2017).

RESULTS AND DISCUSSION

By developing a new version of MutaBind2 we tried to achieve the following goals: (1) to improve the overall performance, especially for mutations increasing binding affinity; (2) to avoid overfitting; and (3) to allow for multiple mutations predictions. To do this, first we designed a scoring function with seven terms instead of 10 used in the previous MutaBind version (detailed in Supplemental Information, Transparent Methods). Second, we trained our models on a much larger dataset from SKEMPI2, which encompassed 1.7 times more mutations and 3.3 times more complexes compared with SKEMPI used for training of MutaBind. Third, to enrich the existing dataset with mutations increasing binding affinity and to build a more balanced training dataset, we produced structural models of complexes with reverse mutations and estimated the corresponding values of each term of the scoring function. Finally, we added a new functionality of predicting the effects of multiple mutations (up to 10 mutations) on binding affinity to account for possible cooperative and epistatic effects.

Training/Test Set	Model	All Mutations			Decreasing		Increasing	
		R	RMSE	Slope	R	RMSE	R	RMSE
Single mutations								
Skempi + Reverse/S1748	MutaBind2^	0.63	1.25	0.83	0.45	1.17	0.77	1.52
Skempi/S1748	MutaBind	0.38*	1.51	0.72	0.44	1.11	–	2.43
	BeAtMuSiC	0.30*	1.58	0.55	0.43	1.14	−0.25*	2.57
Test: S1748	FoldX	0.42*	1.57	0.52	0.41	1.37	0.26*	2.12
Test: S4191	MutaBind2 CV4	0.76	1.34	1.11	0.61	1.31	0.67	1.39
	MutaBind2 CV5	0.69	1.50	1.18	0.54	1.41	0.47	1.65
Multiple mutations								
Test: M1707	MutaBind2 CV4	0.74	2.13	1.09	0.51	2.04	0.60	2.26
	MutaBind2 CV5	0.71	2.24	1.00	0.47	2.18	0.56	2.33
Test: M1337	FoldX	0.49	2.43	0.52	0.37	2.49	0.24	2.21

Table 1. Comparison of Methods' Performance for Single and Multiple Mutations

MutaBind2^: MutaBind2 was retrained on "Skempi + Reverse" set.

*Significant difference between MutaBind2 and other methods with p value < 0.01 calculated on a test set S1748 (implemented in R package *cocor*).

R, Pearson correlation coefficient between experimental and predicted ΔΔG values; RMSE (kcal mol^{-1}), root-mean-square error, the standard deviation of the residuals (prediction errors); Slope, the slope of the regression line between experimental and predicted ΔΔG values. All presented values of correlation coefficients are statistically significantly different from zero (p value << 0.01). The details about datasets are shown in Table S1.

Evaluating the Performance of MutaBind2 Using Cross-Validation

Avoiding overfitting is one of our major concerns while developing a computational method that should make predictions with high accuracy for diverse sets of single or multiple mutations. Overfitting of model parameters may occur while minimizing the mean square deviations of predicted from experimental values in the training set, which could indicate the loss of generalization in the model (Dehouck et al., 2013). To overcome this issue, a cross-validation can be applied, which allows to estimate the future performance of the method on previously unseen data. Five types of cross-validation were performed in our study (explained in more detail in the Methods section). Figure 1 shows the Pearson correlation coefficients between experimental and calculated $\Delta\Delta G$ of the first three types of cross-validation procedures. The correlation coefficients of each cross-validation round exceed 0.80 for "CV1" and "CV2" and about 0.70 for "CV3" cross-validation for both single and multiple mutations.

The Pearson correlation coefficient between experimental and computed $\Delta\Delta G$ values using the "leave-one-complex-out" ("CV4") procedure reaches 0.76 for single mutations and 0.74 for multiple mutations (Table 1 and Figure 2). In addition, we performed a validation by leaving one binding site out of the training set ("CV5" cross-validation). According to this validation, the model was parameterized and tested using completely different non-overlapping sets of binding sites. Nevertheless, the correlation coefficient remained statistically significant, being equal to 0.69 for single mutations and 0.71 for multiple mutations (Table 1 and Figure S4D). From the evaluation of the performance of MutaBind2 using cross-validation, we can conclude that the MutaBind2 for single mutations significantly outperforms the previous version of MutaBind, which had R = 0.68 and R = 0.57 for "CV4" and "CV5," respectively (Li et al., 2016b) (see Table 1 for RMSE values).

To better understand MutaBind2's limitations and strengths, we analyzed 5% and 1% outliers and evaluated the performance after removing outliers using leave-one-complex-out validation (CV4) on S4191. Studentized residuals are used in detecting outliers. Figure S11A shows that the performance is improved significantly after removing 5% outliers. Moreover, we found that the outliers are more likely to appear in complexes with higher protein-protein binding affinity and at the mutated sites with higher number of hydrogen bonds (Figure S11B). Consistent with the observation of Pires et al. (Pires et al., 2016), the outliers usually correspond to those mutations with extreme experimental values (highly decreasing/increasing

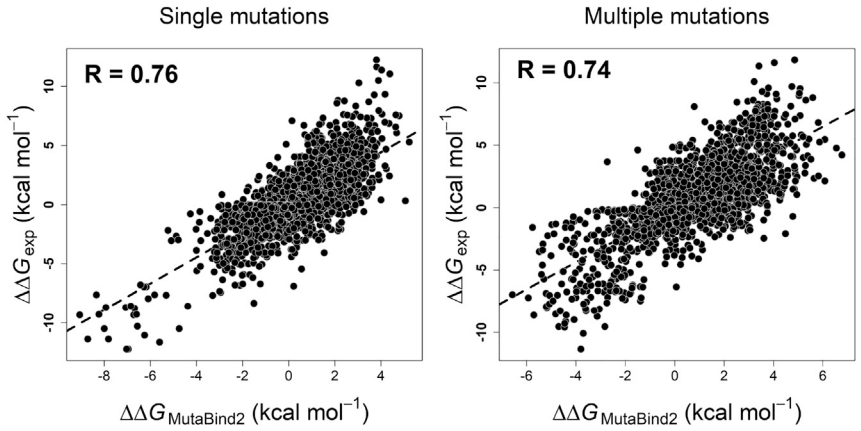


Figure 2. Experimental and Predicted Values of Changes in Binding Affinity for All Mutations in the S4191 (Single Mutations) and M1707 (Multiple Mutations) Sets Using "Leave-One-Complex-Out" (CV4) Cross-Validation
See also Table S1.

binding affinity), and MutaBind2 could correctly classify them as either increasing or decreasing binding affinity (Figures S11C and S11D).

Validation of MutaBind2 on Independent Test Sets

To check if enriching the training set with mutations increasing binding affinity improved the performance, we constructed an independent "unseen" test set consisting of complexes and mutations that were present in SKEMPI2 but were absent from SKEMPI (S1748 set). The original version of MutaBind, which was trained on binding affinity changes dataset from SKEMPI (referred to as the "Skempi" set), yielded a Pearson correlation coefficient of R = 0.44 between predicted and experimentally determined values for mutations decreasing binding affinity on an independent S1748 set, whereas it did not yield statistically significantly predictions for mutations increasing binding affinity (Table 1 and Figure S4A). However, after applying a model trained on the SKEMPI set enriched with mutations increasing binding affinity ("Skempi + Reverse" set) using MutaBind2 features and protocol, the performance on an independent S1748 set improved considerably (the correlation coefficient increased from 0.38 to 0.63, and root-mean-square error decreased from 1.51 to 1.25 kcal mol^{-1}). Moreover, predictions for mutations increasing binding affinity were significantly improved without compromising the accuracy for predicting mutations decreasing binding affinity (Table 1).

For single mutations, we also compared MutaBind2 with four other methods, BeAtMuSiC (Dehouck et al., 2013), FoldX (Guerois et al., 2002), iSEE (Geng et al., 2019), and mCSM-PPI2 (Rodrigues et al., 2019). BeAtMuSiC is a machine learning method, which uses a combination of different statistical potentials to predict $\Delta\Delta G$ values and is parameterized on mutations from SKEMPI. FoldX uses an empirical energy function, which is parametrized on experimental changes of unfolding free energy. iSEE is parameterized on the SKEMPI set and uses several dozens of interface, structure, evolution, and energy-based features. iSEE is not available as a server or a standalone version, so it could not be applied to the S1748 set. mCSM-PPI2 uses several dozens of features such as graph-based signatures, evolutionary conservation, and interaction energy between two partners calculated from FoldX and also incorporates features derived from reverse mutations. It has been trained on 8,338 mutations from the SKEMPI2 dataset, which includes almost all mutations from the MutaBind2 training dataset S4191.

For comparison with iSEE, we used the S487 dataset obtained from the iSEE article (Geng et al., 2019) where the MutaBind2 model was retrained after removing S487 from the S4191 training set. As can be seen in Table 2, the MutaBind2 model parameterized on this training set shows the best performance on S487 compared with other methods (more comparisons can be found in Table S5). We did not have an independent set for comparing the predictions between MutaBind2 and mCSM-PPI2, therefore we used the same training protocol and retrained MutaBind2 on the dataset of S8338 (a training dataset of mCSM-PPI2), even though our feature selection was not based on this dataset. We obtained comparable correlation coefficients with mCSM-PPI2 using the CV4 and CV5 cross-validations (Table 2), which were slightly lower than results reported for the original MutaBind2 model on the S4191 (Table 1). Additional comparisons with mCSM-PPI2 are shown in Table S6, which points to a slightly

Test Set	Method	R	RMSE
S487	MutaBind2	0.41	1.25
	MutaBind	0.29**	1.63
	BeAtMuSiC	0.35	1.28
	FoldX	0.34*	1.53
	iSEE	0.25**	1.32
S8338	MutaBind2 CV4	0.74	1.37
	MutaBind2 CV5	0.66	1.53
	mCSM-PPI2 CV4	0.75	1.30
	mCSM-PPI2 CV5	0.67	1.39

Table 2. Comparison of Methods' Performance on Different Datasets

* and ** indicate statistically significant difference between MutaBind2 and other methods in terms of R with p value < 0.05 and p value < 0.01, respectively, calculated on test set S487 (implemented in R package *cocor*).

R, Pearson correlation coefficient; RMSE, root-mean-square error.

R and RMSE values were taken from the mCSM-PPI2 article (Rodrigues et al., 2019). For testing on S487 set, MutaBind2 was retrained after removing S487 from the training dataset. For testing on S8338 set, MutaBind2 was retrained on S8338. See also Table S6.

better performance for MutaBind2 in terms of the slope of the regression line indicating that predicted and experimental values are on the same scale.

Recently the impacts of 2,009 missense mutations across 2,185 human protein-protein interactions (4,797 mutation-interaction pairs) were measured by yeast two hybrid experiments (Fragoza et al., 2019), and 903 mutations were identified as interaction-disruptive mutations. A mutation was defined as disruptive if it damaged one or more protein-protein interactions and was defined as non-disruptive otherwise. Among 4,797 mutation-interaction pairs, 451 mutations, including 147 interaction disruptive mutations, could be mapped on corresponding protein-protein complexes with the known 3D structures. Then we calculated binding affinity changes for these mutations using different methods. Figures 3 and S5 show excellent performance of MutaBind and MutaBind2 in distinguishing interaction-disruptive from other mutations.

Prediction of Mutations Highly Decreasing and Increasing Binding Affinity

The previous version, MutaBind, could predict single mutations highly decreasing binding affinity relatively well but failed to annotate mutations highly increasing affinity. Table 3 and Figure S6 demonstrate the high performance of MutaBind2 in predicting mutations highly decreasing and highly increasing binding affinity (see Methods for details). MutaBind2 further improves the performance for both interfacial and non-interfacial mutations compared with the previous version and outperforms other methods on the S1748 set (Figure S7). We subdivided complexes with multiple mutations into different categories based on the number of mutations and the number of mutated chains involved (see Table S9 for more details). We found that MutaBind2 performed well for almost all categories; the worst performance was observed for three or more mutations introduced on the same chain (R = 0.61 in CV4 validation), and the best performance was achieved for double mutations on multiple chains (R = 0.85 in CV4 validation).

Next we would like to elucidate the main differences between predictors like BeAtMuSiC, iSEE, and mCSM-PPI2 and methods like FoldX, MutaBind, and MutaBind2. The first group of methods uses powerful machine learning approaches with several dozens of features to calculate the changes in binding affinity and does not provide contribution of each feature for each mutation. On the other hand, methods like FoldX, MutaBind, and MutaBind2 use very few interpretable energy terms and perform structure optimization and energy calculations. This allows the construction of actual molecular models of mutant structures and to evaluate changes in binding affinity for these mutants, potentially accounting for structural changes that cannot be captured by machine learning methods. The molecular models of mutants have been used extensively by researchers to understand the molecular mechanisms of disease mutations, to design drugs, to identify drug targets, to predict driver mutations, and to decipher the mechanisms of drug-resistant mutations (Figure S8). Importantly, MutaBind and

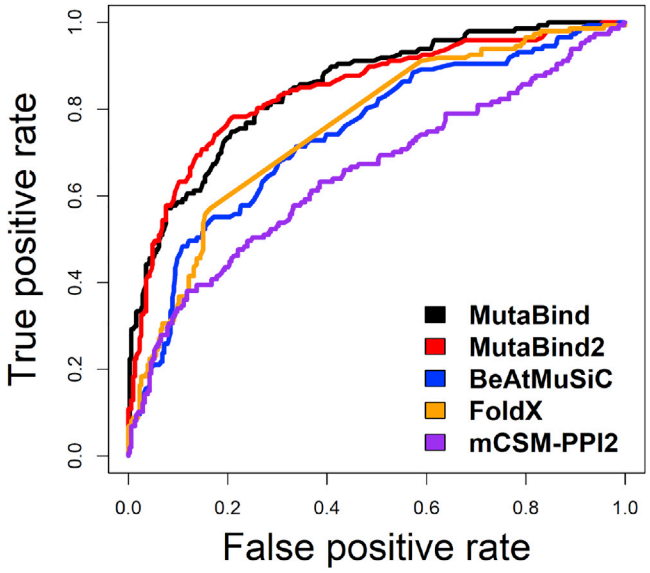


Figure 3. Receiver Operating Characteristic Curves for Predicting Mutations Disrupting Protein-Protein Interactions Using Different Methods

As one mutation/interaction could be mapped to several Protein DataBank structures, the maximum predicted value of each method was used for each interaction-disruptive mutation and the minimum predicted values were used for those mutations that do not disrupt interactions. More details are shown in Figure S5.

MutaBind2, unlike other methods mentioned earlier, estimate interactions of a protein with the solvent, which is one of the most important terms together with the site evolutionary conservation and thermodynamic stability of a protein complex and each binding partner (Table S2). In addition, FoldX and MutaBind2 provide predictions and structural models for multiple mutations introduced at the same time in a protein complex.

Online Web Server

MutaBind2 is available at https://mutabind.org/v2. The main requirement of the webserver is the availability of the 3D structure of a protein-protein complex, which can be provided by the Protein DataBank accession or by a file with the coordinates uploaded by the researcher. In either case, the structure file should contain at least two protein chains. In the next step two interaction partners should be defined. It is possible to assign one chain or multiple chains to either "Partner 1" or "Partner 2," and only assigned chains will be considered during the calculation. If the interface size between assigned partners is smaller than 100 $Å^2$, an error message is displayed. The interface size is calculated as a difference between the solvent-accessible surface areas of assigned chains in a complex and unbound partner. The final step is to select mutations. We provide three options to allow users to do large-scale mutational scanning (Figure S9).

- An option "Upload file" allows to submit a list of mutations specified in the uploaded file
- The "alanine scanning" option allows to perform alanine scanning for all contact residues between interaction partners. Contact residues here are defined as those with inter-atomic distances less than 6 Å between any heavy atom of interaction partners. MutaBind2 provides the contact residues list for download
- Contact residues are shown in orange in the residue list of "Specify One or More Mutations," which allows to view contact residues in the 3D structure

For each mutation on a protein-protein complex, the MutaBind2 server provides the following results:

- ΔΔG (kcal mol^{-1}): predicted change in binding affinity induced by mutations (positive and negative signs correspond to mutations decreasing and increasing binding affinity, respectively)
- The location on interface (yes/no): indicating whether the residue is located on a protein-protein interface in the case when a residue's solvent accessibility in the complex is lower than in the corresponding unbound partners

	MutaBind2	MutaBind	BeAtMuSiC	FoldX
Highly decreasing				
Sensitivity	0.75	0.86	0.73	0.58
Specificity	0.89	0.82	0.87	0.94
MCC	0.63	0.63	0.58	0.57
AUC	0.82	0.82	0.79	0.79
Highly increasing				
Sensitivity	0.55	0	0	0.44
Specificity	0.99	1.00	1.00	0.99
MCC	0.64	−0.01	0	0.51
AUC	0.86	0.65*	0.56*	0.74*

Table 3. Comparative Performance of MutaBind2 and Three Methods for Predicting Mutations Highly Decreasing and Increasing Binding Affinity on the Independent Test Set of S1748

MutaBind2 was retrained on the dataset "Skempi + Reverse."

MCC, Matthews correlation coefficient.

*p value < 0.01 calculated by Delong test (DeLong et al., 1988) comparing AUC (area under the ROC curve) produced by a given method and AUC produced by MutaBind2; points to significant differences in performance. See also Figure S6.

- Coordinates of the minimized mutant structure
- Deleterious (yes/no), a mutation is classified as deleterious if $\Delta\Delta G \geq 1.5$ or $\Delta\Delta G \leq -1.5$ kcal mol^{-1}
- The contribution of each term of the target function for every mutation
- Homologous binding sites: the Inferred Biomolecular Interactions Server (Shoemaker et al., 2012) is used to identify the binding sites in protein-protein complexes homologous to the query

Limitation of the Study

1. Requirement of the 3D structure of a protein-protein complex. Six features out of seven in our model are calculated using 3D structure of a protein-protein complex, which limits the application to those mutations that could not be mapped to the structural complex.
2. Multiple mutations instances with more than 10 mutations. As the number of multiple mutations with more than 10 mutations is small in our training dataset and prediction accuracy for these multiple mutations is low, the upper limit of 10 mutations was used in the study. Therefore, our model cannot be applied to the multiple mutation instances with more than 10 mutations.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

MutaBind2 is available at https://mutabind.org/v2, and the training and test datasets are available for download from the server.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.100939.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant No. 31701136), Natural Science Foundation of Jiangsu Province, China (Grant No. BK20170335), and the Priority Academic Program Development of Jiangsu Higher Education Institutions. A.G. and R.V.A. were supported by the

Intramural Research Program of the National Library of Medicine at the US National Institutes of Health. A.R.P. was in part supported by the Intramural Research Program of the National Library of Medicine at the US National Institutes of Health and by the Department of Pathology and Molecular Medicine, Queen's University, Canada. A.R.P. is the recipient of a Senior Canada Research Chair in Computational Biology and Biophysics and a Senior Investigator award from the Ontario Institute of Cancer Research, Canada. We would like to thank Dr. Thomas Madej for proofreading of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, A.R.P. and M.L.; Methodology, N.Z., Y.C., M.L., and A.R.P.; Software, N.Z., Y.C., and A.G.; Validation, N.Z., Y.C., A.G., and M.L.; Formal Analysis, N.Z. and Y.C.; Investigation, N.Z., Y.C., M.L., and A.R.P.; Data Curation, N.Z. and Y.C.; Writing – Original Draft, M.L. and A.R.P.; Writing – Review & Editing, M.L. and A.R.P.; Visualization, H.L., F.Z., R.V.A., and A.G.; Supervision, A.R.P. and M.L.; Project Administration, M.L.; Funding Acquisition, M.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 15, 2019
Revised: November 21, 2019
Accepted: February 20, 2020
Published: March 27, 2020

REFERENCES

An, O., Gursoy, A., Gurgey, A., and Keskin, O. (2013). Structural and functional analysis of perforin mutations in association with clinical data of familial hemophagocytic lymphohistiocytosis type 2 (FHL2) patients. Protein Sci. *22*, 823–839.

Brender, J.R., and Zhang, Y. (2015). Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. PLoS Comput. Biol. *11*, e1004494.

Creixell, P., Schoof, E.M., Simpson, C.D., Longden, J., Miller, C.J., Lou, H.J., Perryman, L., Cox, T.R., Zivanovic, N., Palmeri, A., et al. (2015). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. Cell *163*, 202–217.

Cukuroglu, E., Engin, H.B., Gursoy, A., and Keskin, O. (2014). Hot spots in protein-protein interfaces: towards drug discovery. Prog. Biophys. Mol. Biol. *116*, 165–173.

Dehouck, Y., Kwasigroch, J.M., Rooman, M., and Gilis, D. (2013). BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res. *41*, W333–W339.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics *44*, 837–845.

Fragoza, R., Das, J., Wierbowski, S.D., Liang, J., Tran, T.N., Liang, S., Beltran, J.F., Rivera-Erick, C.A., Ye, K., Wang, T.-Y., et al. (2019). Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. Nat. Commun. *10*, 4141.

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

Goncearenco, A., Li, M., Simonetti, F.L., Shoemaker, B.A., and Panchenko, A.R. (2017). Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows. Methods Mol. Biol. *1647*, 221–236.

Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. *320*, 369–387.

Jemimah, S., Sekijima, M., and Gromiha, M.M. (2019). ProAffiMuSeq: sequence-based method to predict the binding free energy change of protein-protein complexes upon mutation using functional classification. Bioinformatics *35*, 462–469.

Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa-Montaño, B., Blundell, T.L., and Ascher, D.B. (2016). Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog. Biophys. Mol. Biol. *128*, 3–13.

Kruger, D.M., and Gohlke, H. (2010). DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. Nucleic Acids Res. *38*, W480–W486.

Li, M., Kales, S.C., Ma, K., Shoemaker, B.A., Crespo-Barreto, J., Cangelosi, A.L., Lipkowitz, S., and Panchenko, A.R. (2016a). Balancing protein stability and activity in cancer: a new approach for identifying driver mutations affecting CBL ubiquitin ligase activation. Cancer Res. *76*, 561–571.

Li, M., Petukh, M., Alexov, E., and Panchenko, A.R. (2014). Predicting the impact of missense mutations on protein-protein binding affinity. J. Chem. Theor. Comput. *10*, 1770–1780.

Li, M., Simonetti, F.L., Goncearenco, A., and Panchenko, A.R. (2016b). MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. Nucleic Acids Res. *44*, W494–W501.

Montanucci, L., Savojardo, C., Martelli, P.L., Casadio, R., and Fariselli, P. (2018). On the biases in predictions of protein stability changes upon variations: the INPS test case. Bioinformatics *35*, 2525–2527.

Nagatani, R.A., Gonzalez, A., Shoichet, B.K., Brinen, L.S., and Babbitt, P.C. (2007). Stability for function trade-offs in the enolase superfamily "catalytic module". Biochemistry *46*, 6688–6695.

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B.A., Hashimoto, K., Alexov, E., Wuchty, S., and Panchenko, A.R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. PLoS One *8*, e66273.

Ozdemir, E.S., Gursoy, A., and Keskin, O. (2018). Analysis of single amino acid variations in singlet hot spots of protein-protein interfaces. Bioinformatics *34*, i795–i801.

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

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

Pires, D.E., Ascher, D.B., and Blundell, T.L. (2014). mCSM: predicting the effects of mutations in

proteins using graph-based signatures. Bioinformatics *30*, 335–342.

Pires, D.E.V., Blundell, T.L., and Ascher, D.B. (2016). mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci. Rep. *6*, 29575.

Pucci, F., Bernaerts, K.V., Kwasigroch, J.M., and Rooman, M. (2018). Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics *34*, 3659–3665.

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

Rutten, L., Lai, Y.-T., Blokland, S., Truan, D., Bisschop, I.J.M., Strokappe, N.M., Koornneef, A., van Manen, D., Chuang, G.-Y., and Farney, S.K. (2018). A universal approach to optimize the folding and stability of prefusion-closed HIV-1 envelope trimers. Cell Rep. *23*, 584–595.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647–660.

Shoemaker, B.A., Zhang, D., Tyagi, M., Thangudu, R.R., Fong, J.H., Marchler-Bauer, A., Bryant, S.H., Madej, T., and Panchenko, A.R. (2012). IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. Nucleic Acids Res. *40*, D834–D840.

Shoichet, B.K., Baase, W.A., Kuroki, R., and Matthews, B.W. (1995). A relationship between protein stability and protein function. Proc. Natl. Acad. Sci. U S A *92*, 452–456.

Stefl, S., Nishi, H., Petukh, M., Panchenko, A.R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. J. Mol. Biol. *425*, 3919–3936.

Tan, Z.W., Tee, W.V., Guarnera, E., Booth, L., and Berezovsky, I.N. (2019). AlloMAPS: allosteric mutation analysis and polymorphism of signaling database. Nucleic Acids Res. *47*, D265–D270.

Tee, W.V., Guarnera, E., and Berezovsky, I.N. (2019). On the allosteric effect of nsSNPs and the emerging importance of allosteric polymorphism. J. Mol. Biol. *431*, 3933–3942.

Teng, S., Madej, T., Panchenko, A., and Alexov, E. (2009). Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. Biophys. J. *96*, 2178–2188.

Usmanova, D.R., Bogatyreva, N.S., Arino Bernad, J., Eremina, A.A., Gorshkova, A.A., Kanevskiy, G.M., Lonishin, L.R., Meister, A.V., Yakupova, A.G., Kondrashov, F.A., et al. (2018). Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. Bioinformatics *34*, 3653–3658.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat. Biotechnol. *30*, 159–164.

Witham, S., Takano, K., Schwartz, C., and Alexov, E. (2011). A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. Proteins *79*, 2444–2454.

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

Zhao, N., Han, J.G., Shyu, C.R., and Korkin, D. (2014). Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. PLoS Comput. Biol. *10*, e1003592.

iScience, Volume *23*

Supplemental Information

MutaBind2: Predicting the Impacts of Single and Multiple Mutations on Protein-Protein Interactions

Ning Zhang, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R. Panchenko, and Minghui Li

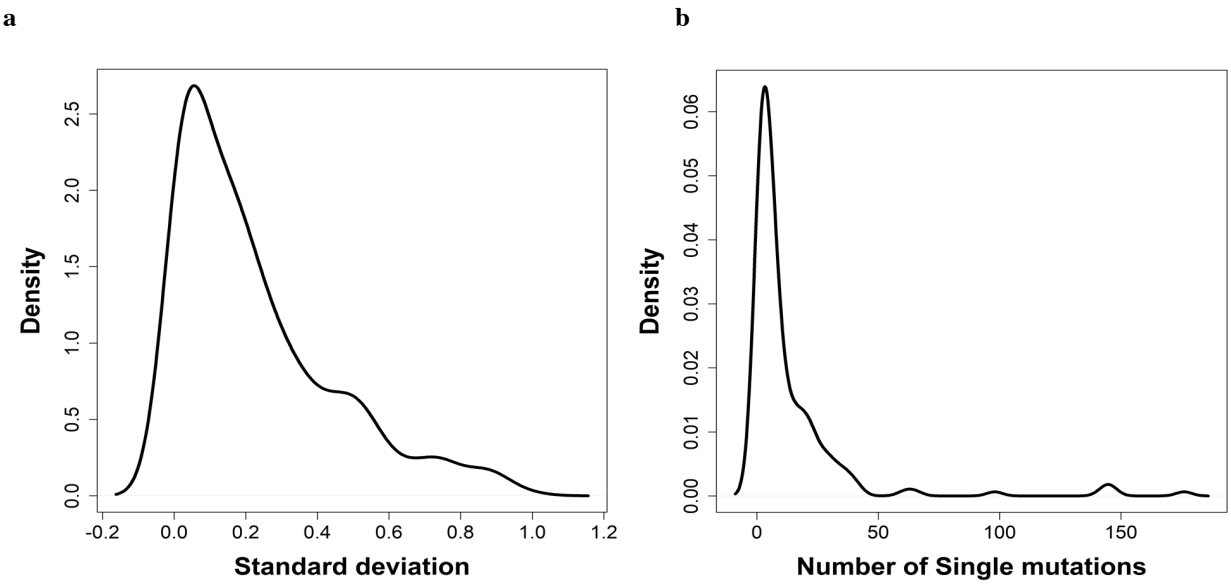
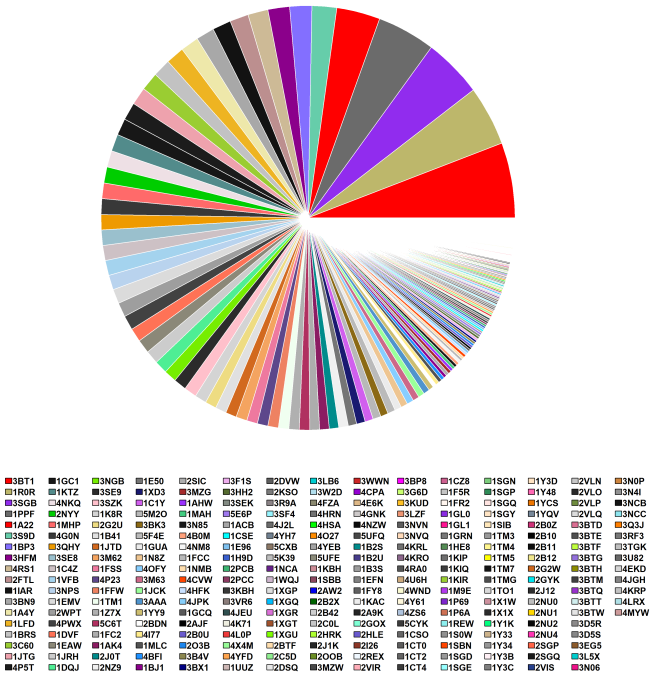


Figure S1. (a) Distribution of the standard deviation for 408 single mutations with multiple experimental measurements of changes in binding affinity in S3310 dataset. (b) Distribution of the number of mutations with $\Delta\Delta G_{exp} > 0$ over protein complexes in S3310 dataset, Related to Figure 1 and Figure 2.

The number of single mutations for each protein-protein complex



The number of multiple mutations for each protein-protein complex

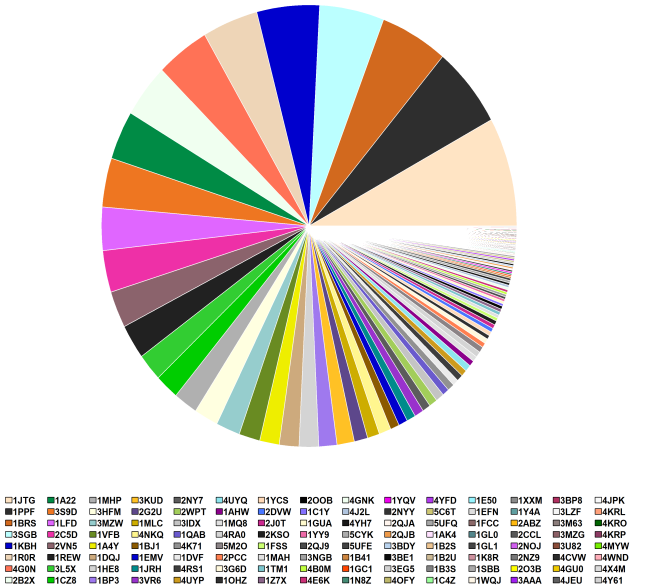


Figure S2. The number of mutations for each protein-protein complex for single and multiple mutation dataset of S3310 and M1337, respectively, Related to Figure 1 and Figure 2.

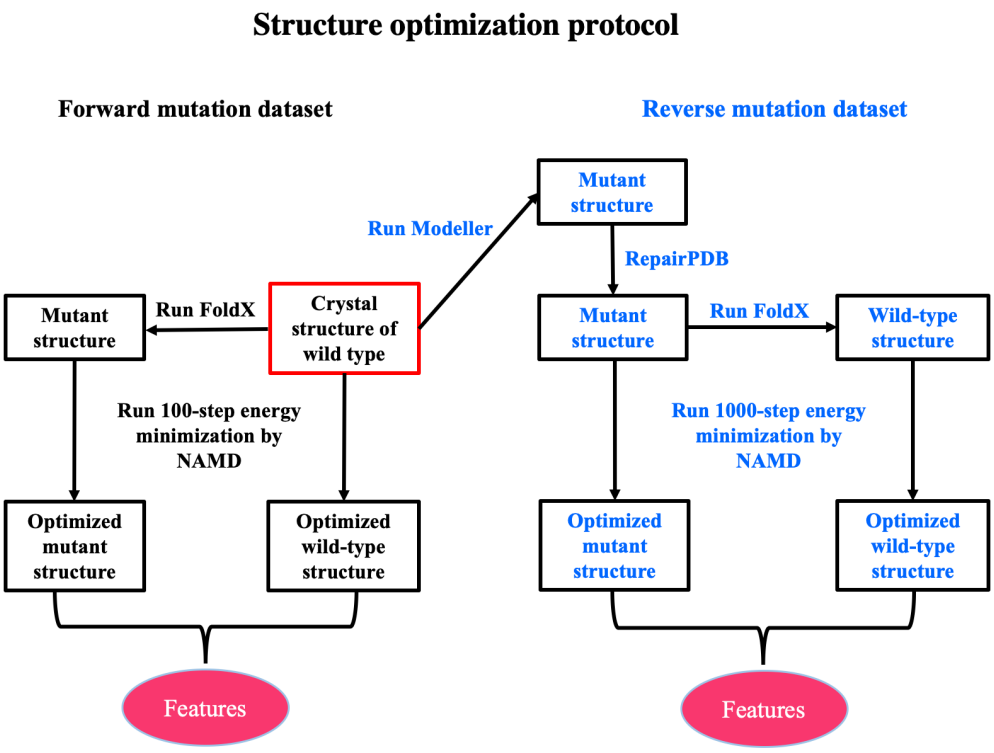
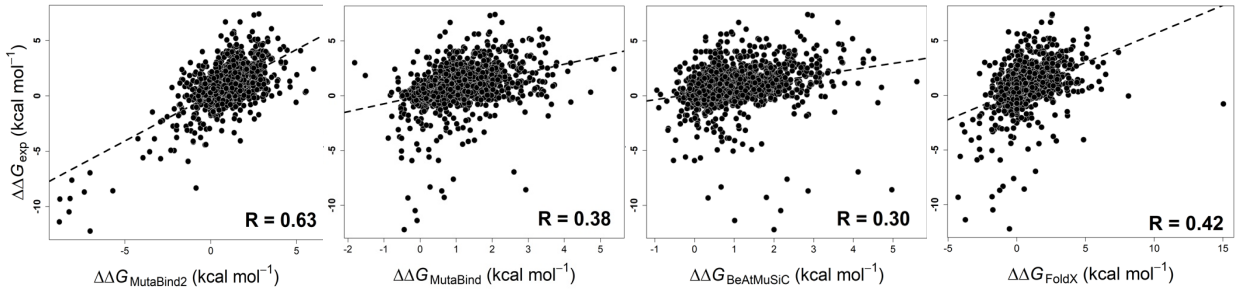
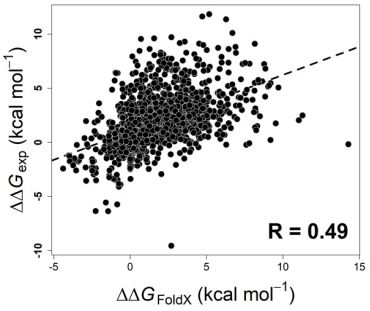


Figure S3. The flowchart of the structure optimization protocol, Related to Figure 1 and Figure 2.

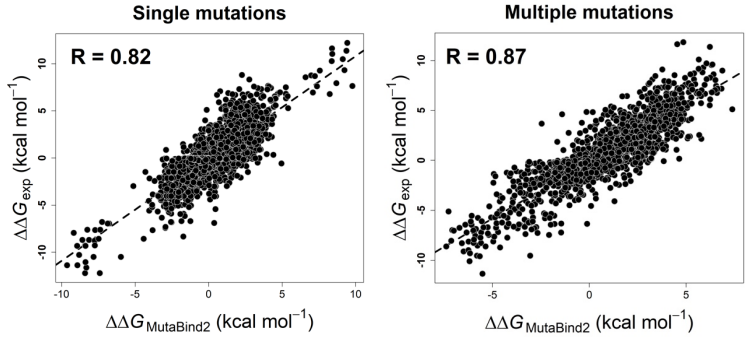
a. Test on S1748 set



b. Test on M1337 set



c. Training and testing on S4191 and M1707 sets



d. "leave-one-binding-site-out" (CV5) cross-validation

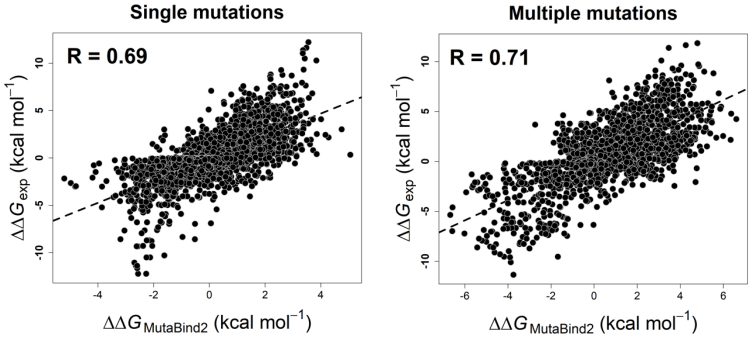
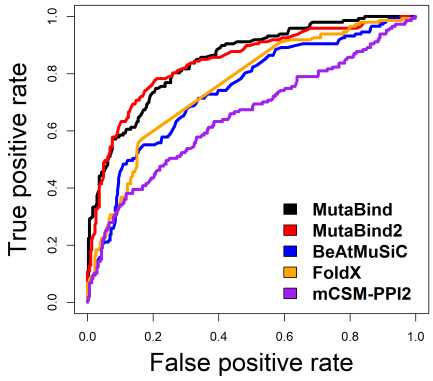


Figure S4. Correlation between experimental and calculated changes in binding free energies ($\Delta\Delta G$) for (a) MutaBind2, MutaBind, BeAtMuSiC and FoldX methods tested on S1748 independent dataset. Here MutaBind2 model is trained on the dataset of "Skempi+Reverse"; (b) FoldX tested on M1337; (c) MutaBind2 trained and tested on S4191 and M1707; (d) MutaBind2 tested on S4191 and M1707 using "leave-one-binding-site-out" (CV5) cross-validation respectively, Related to Table 1.

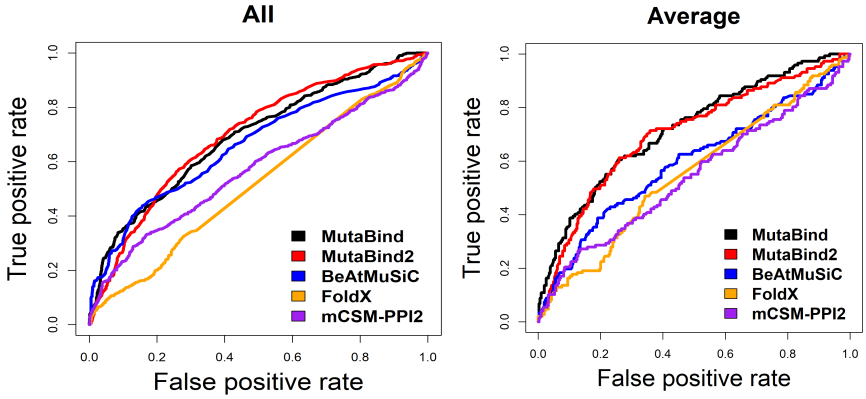
a.

Dataset	Mutations-interaction pairs	Interaction-disruptive mutations	Interaction non-disruptive mutations
Original	4797	903	3894
Mapped to Structure	451(2033)	147(831)	304(1202)

b.



c.

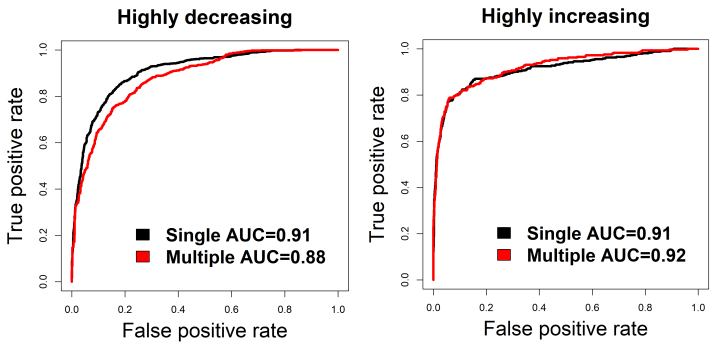


d.

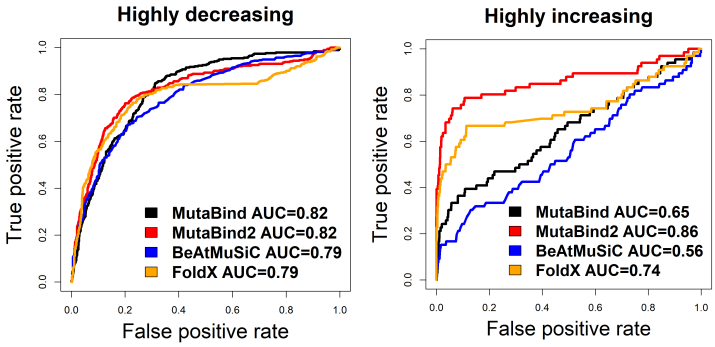
Method	Maximum		All		Average	
	AUC	MCC	AUC	MCC	AUC	MCC
MutaBind2	0.84	0.56	0.70	0.31	0.70	0.34
MutaBind	0.85	0.55	0.69	0.32	0.72	0.34
BeAtMuSiC	0.74*	0.42	0.66*	0.31	0.59*	0.21
FoldX	0.76*	0.42	0.52*	0.10	0.55*	0.12
mCSM-PPI2	0.65*	0.31	0.57*	0.20	0.54*	0.17

Figure S5. Classification performance for predicting of mutations disrupting (decreasing) protein-protein interactions using different methods. (a) The number of mutations used for performance evaluation. "Original": the impact of 2009 missense mutations across 2185 human protein-protein interactions, interaction profiles for 4797 mutations-interaction pairs were measured by yeast two hybrid (2H) experiments (Fragoza et al., 2019), and 903 mutations were identified as interaction-disruptive mutations. "Mapped to Structure": the number of mutations/interactions that could be mapped to protein-protein crystal structures. Since one mutation/interaction could be mapped to several PDB structures, the values in parentheses show the total number of mutations/interactions mapped to different PDB structures. (b) ROC curves, the maximum predicted value of binding affinity changes calculated for all mapped PDB structures for each mutation and minimum predicted values were used for interaction non-disruptive mutation. (c) "All": all predicted values of binding affinity changes calculated using all mapped PDB structures were used for each mutation; "Average": average values of binding affinity changes calculated using all mapped PDB structures were used for each mutation. (d) AUC and MCC values for classification scenarios using different methods. * denotes a statistically significant difference between MutaBind2 and other methods with p-value < 0.01 estimated by Delong test (DeLong et al.), Related to Figure 3.

a.



b.



c.

Category		Definition	# of mutations		
			S4191	M1707	S1748
Highly decreasing	Positive	$\Delta\Delta G_{exp}$ (kcal mol^{-1}) >= 1.5	897	690	375
	Negative	$\Delta\Delta G_{exp}$ (kcal mol^{-1}) <= 0.5	2457	810	924
Highly increasing	Positive	$\Delta\Delta G_{exp}$ (kcal mol^{-1}) <= -1.5	425	289	66
	Negative	$\Delta\Delta G_{exp}$ (kcal mol^{-1}) >= -0.5	3289	1233	1599

d.

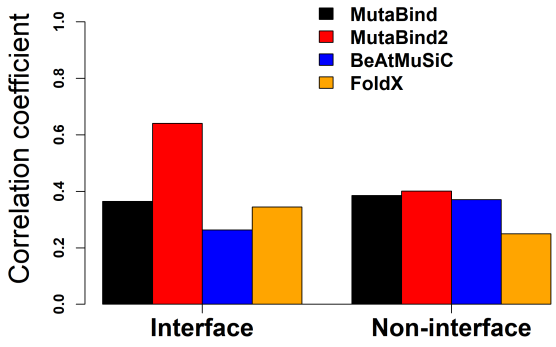
		True	False
Highly decreasing	Positive	$\Delta\Delta G_{MutaBind2}$≥1.5 & $\Delta\Delta G_{exp}$≥1.5 (TP)	$\Delta\Delta G_{MutaBind2}$≥1.5 & $\Delta\Delta G_{exp}$≤0.5 (FP)
	Negative	$\Delta\Delta G_{MutaBind2}$≤0.5 & $\Delta\Delta G_{exp}$≤0.5 (TN)	$\Delta\Delta G_{MutaBind2}$≤0.5 & $\Delta\Delta G_{exp}$≥1.5 (FN)
Highly increasing	Positive	$\Delta\Delta G_{MutaBind2}$≤-1.5 & $\Delta\Delta G_{exp}$≤-1.5 (TP)	$\Delta\Delta G_{MutaBind2}$≤-1.5 & $\Delta\Delta G_{exp}$≥-0.5 (FP)
	Negative	$\Delta\Delta G_{MutaBind2}$≥-0.5 & $\Delta\Delta G_{exp}$≥-0.5 (TN)	$\Delta\Delta G_{MutaBind}$≥-0.5 & $\Delta\Delta G_{exp}$≤-1.5 (FN)

e.

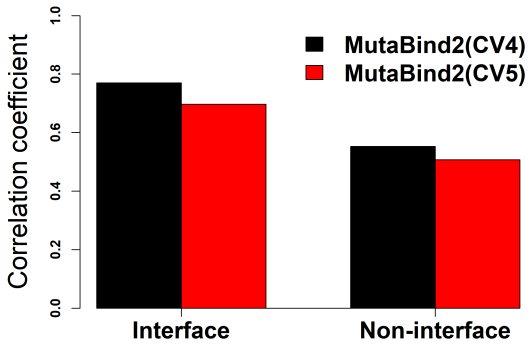
	Highly decreasing		Highly increasing	
	S4191	M1707	S4191	M1707
Sensitivity	0.83	0.84	0.71	0.75
Specificity	0.95	0.84	0.99	0.97
MCC	0.78	0.68	0.76	0.76
AUC	0.91	0.88	0.91	0.92

Figure S6. (a) ROC curves for predicting mutations highly decreasing and increasing binding affinity using "leave-one-complex-out" cross-validation (CV4) results ($\Delta\Delta G$) of S4191 and M1707, (b) ROC curves for predicting mutations highly decreasing and increasing binding affinity by applying different methods on the independent test set S1748, and (c) Definitions of highly decreasing and increasing mutations and the number of mutations for making ROC curves. True positive rate (Sensitivity) = (TP/TP+FN) and False positive rate = (FP/FP+TN). (d) The definition of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for predicting highly decreasing and increasing mutations. (e) Performance of MutaBind2 for predicting mutations highly decreasing and increasing binding affinity using "leave-one-complex-out" cross-validation (CV4) on S4191, Related to Table 3.

a. Test on 1748



b. Test on S4191



c.

Category	**# of mutations**	
	S1748	**S4191**
Interface	1221	3240
Non-interface	527	951

Figure S7. Pearson correlation coefficients between predicted and experimental $\Delta\Delta G$ for (a) mutations from S1748 test set located on interface and non-interface predicted by different methods. MutaBind2 here is trained on the dataset of "Skempi+Reverse"; (b) mutations from S4191 test set located on interface and non-interface predicted by MutaBind2(CV4) and MutaBind2(CV5). (c) The number of interfacial and non-interfacial mutations for two sets. Only statistically significant correlation coefficients (p-value < 0.01, calculated by one-sample *t*-test) are shown, Related to Table 1 and Figure 2.

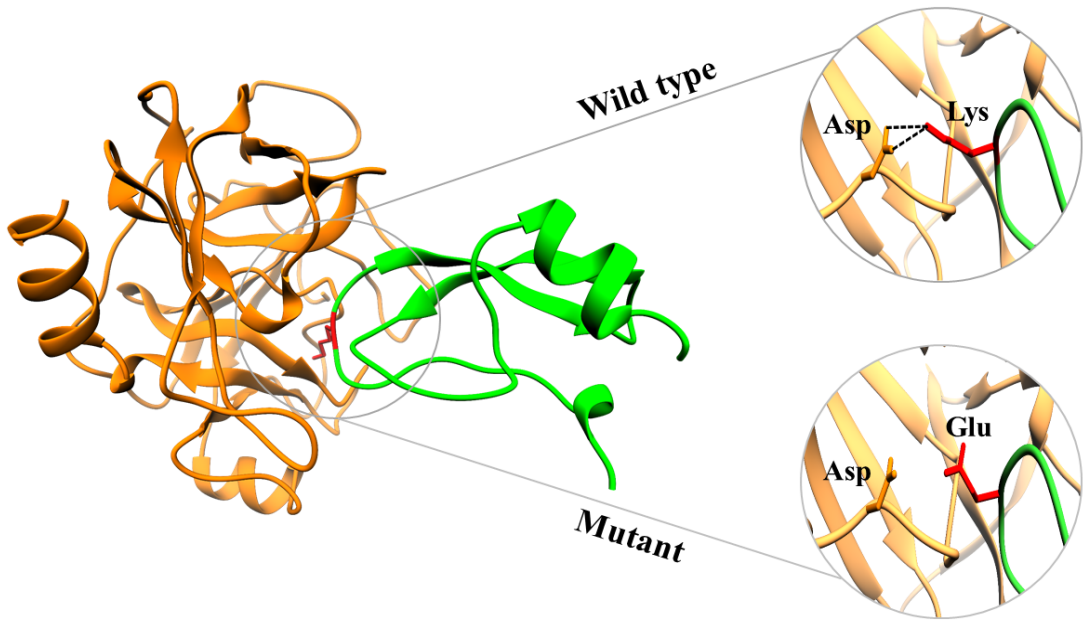


Figure S8. Salt bridges were disrupted after Lys was mutated to Glu in Cationic Trypsin / Pancreatic Trypsin Inhibitor complex (PDB code: 2FTL). The experimental and predicted binding affinity change by MutaBind2 is 9.31 and 9.21 kcal mol^{-1} respectively, Related to Figure 2.

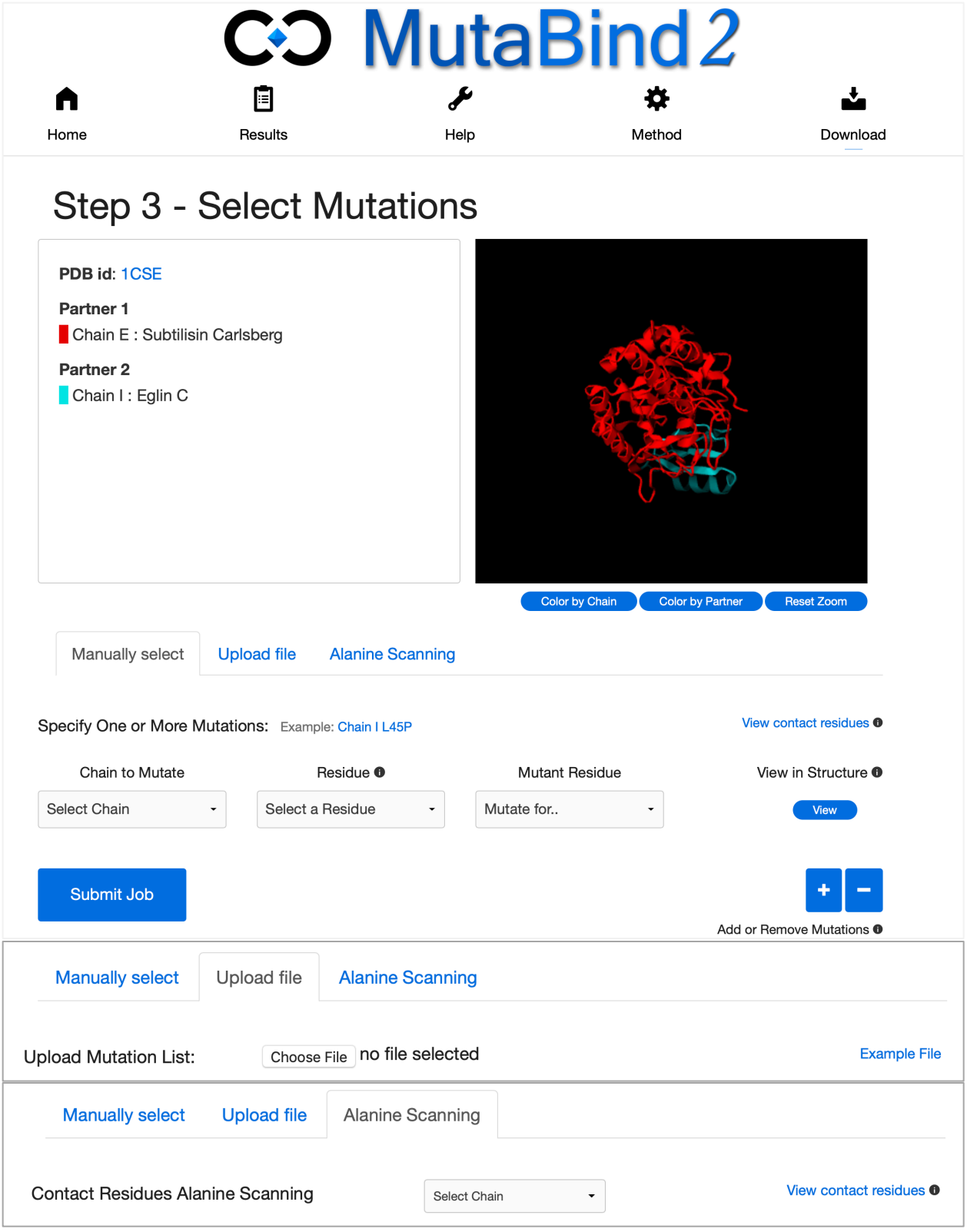


Figure S9. The illustration of the third step of mutation selection, Related to Figure 3.

Results

Job Summary

- **Job Id:** 20191220071202516210853334
- **PDB Id:** 1CSE
- **Partner 1:** E
- **Partner 2:** I
- **Multiple Mutations?:** No
- **Processing time:** 12 hour 1 min

Mutated Chain	Mutation	ΔΔG	Interface?	Deleterious?	Mutant PDB	Homologous binding sites
E	A1C	**0.12**	No	No	Download	Explore
E	A1Q	**0.25**	No	No	Download	Explore
E	A1I	**0.14**	No	No	Download	Explore

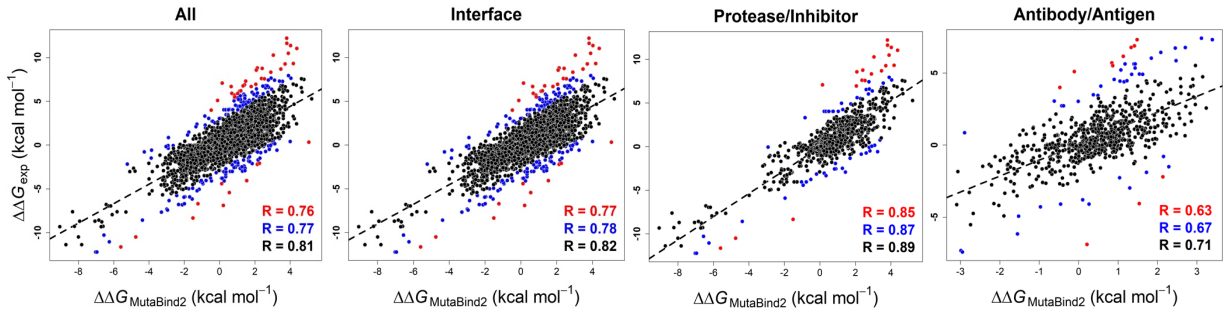
.....

E	G53F	**0.15**	No	No	Download	Explore
E	G53A	**0.26**	No	No	Download	Explore

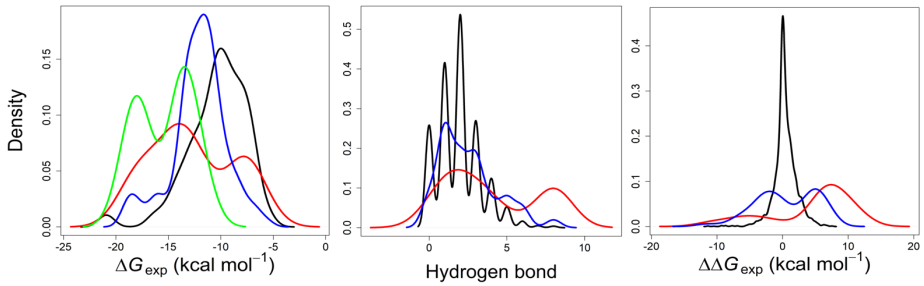
Download Table

Figure S10. Illustration of the time required for MutaBind2 to run predictions for 1000 mutations from a complex with 350 residues, ~12 hours, Related to Figure 3.

a.



b.



c.

Definition	5% outlier	1% outlier
($\Delta\Delta G_{MutaBind2}$≥0 & $\Delta\Delta G_{exp}$≥5.0)/$\Delta\Delta G_{exp}$≥5.0	0.99	0.97
($\Delta\Delta G_{MutaBind2}$<0 & $\Delta\Delta G_{exp}$≤-5.0/$\Delta\Delta G_{exp}$≤-5.0	0.92	0.67
($\Delta\Delta G_{MutaBind2}$≥1.0 & $\Delta\Delta G_{exp}$≥5.0)/$\Delta\Delta G_{exp}$≥5.0	0.88	0.77
($\Delta\Delta G_{MutaBind2}$≤-1.0 & $\Delta\Delta G_{exp}$≤-5.0/$\Delta\Delta G_{exp}$≤-5.0	0.92	0.67

d.

PDB id	Mutation	$\Delta\Delta G_{exp}$	MutaBind2 (CV4)
2FTL	K15G	12.22	3.80
2FTL	K15V	11.63	3.81
2FTL	K15D	11.38	4.02
2FTL	K15A	10.29	3.04
2FTL	K15W	8.57	2.08
2FTL	K15I	11.05	4.38
2FTL	K15T	10.49	3.89
2FTL	K15L	8.77	2.41
2FTL	K15Q	9.27	3.71
2FTL	K15M	7.61	2.37
2FTL	K15H	8.69	3.47
2FTL	K15E	9.32	4.18
2FTL	K15F	6.95	2.05
2FTL	V15K	-11.63	-5.60
2FTL	T15K	-10.49	-4.75

e.

Definition		# of mutations	
		Not-Outlier	5% Outlier
Location	Interface	3044	196
	Non-interface	937	14
Type of protein complex	Pr/PI	716	70
	AB/AG	673	51
	AB/AG & Pr/PI	102	5
	TCR/pMHC	148	2
	unknown	2342	82

Pr/PI: protease/inhibitor; AB/AG: antibody/antigen.

Figure S11. Outlier analysis of MutaBind2. The leave-one-complex-out validation (CV4) results of S4191 dataset were used for these analyses. (a) Experimental and predicted values of changes in binding affinity for all single mutations, mutations on protein-protein binding interfaces, mutations from protease/inhibitor and antibody/antigen complexes, respectively. Black: all mutations except for outliers; blue: 5% outliers; red: 1% outliers. (b) Distibution of experimental binding affinity for wild-type complexes (ΔG_{exp}), the number of hydrogen bonds formed by mutated

sites (Hydrogen bond) and experimental binding affinity changes upon mutations ($\Delta\Delta G_{exp}$), respectively. Black: all mutations with the exception of outliers; blue: 5% outliers; red: 1% outliers; green: mutations from seven complexes (PDB ID: 2FTL, 3HFM, 1PPF, 1BRS, 3QHY, 1DQJ and 1B41) in the 5% outlier, and the reason for showing these seven complexes is that they have more mutations (more than five mutations) included in the 5% outlier. (c) True positive rate (d) Mutations from a complex of bovine pancreatic trypsin inhibitor and bovine β-trypsin included in 1% outlier. (e) The number of mutations in categories of interface, non-interface and different types of protein complexes, Related to Figure 2.

Table S1. Experimental datasets used for training and testing different methods, Related to Figure 1, Figure 2, Table 1, Table 2 and Table 3.

Dataset	Description
Single mutations	
S3310	Compiled from SKEMPI2
S4191	S3310 plus reverse mutations; **training dataset of MutaBind2, single mutation model**
S4169	Compiled from SKEMPI2
S8338	S4169 plus all reverse mutations**; training dataset of mCSM-PPI2**
Skempi	1925 mutations extracted from SKEMPI; training dataset of MutaBind
S1748	Mutations included in S3310 but not in Skempi
S877	Mutations contained in S4169 but not in S3310
S487	Compiled by iSEE including 487 mutations contained in SKEMPI2 but not in SKEMPI
S33	33 mutations of MDM2-P53 complex (PDB 1YCR) that are not included in SKEMPI2
S19	19 mutations from INTERLEUKIN-4 / RECEPTOR ALPHA CHAIN COMPLEX (PDB 1IAR) that are not included in SKEMPI2
Multiple mutations	
M1337	Compiled from SKEMPI2
M1707	M1337 plus reverse mutations; **training dataset of MutaBind2 multiple mutation model**

S33, S19 and S487 datasets were obtained from https://github.com/haddocking/iSee. Protein complexes with more than 10 single mutations with experimental values of binding affinity changes from "Skempi" were used to build its reverse mutation set.

Dataset	All mutations		$\Delta\Delta G_{exp} > 0$		$\Delta\Delta G_{exp} < 0$	
	# of mutations	# of complexes	# of mutations	# of complexes	# of mutations	# of complexes
S3310	3310	265	2504	188	712	173
S3310.R	881	49	0	0	881	49
S4191	4191	265	2504	188	1593	188
M1337	1337	120	1059	100	272	64
M1337.R	370	19	0	0	370	19
M1707	1707	120	1059	100	642	72
S4169	4169	319	3126	238	901	203
Skempi	1925	80	1478	77	410	43
S1748	1748	212	1286	137	390	145
S877	877	63	643	56	186	34
S487	487	65	414	55	66	33
S33	33	1	27	1	6	1
S19	19	1	15	1	3	1

S3310.R: reverse mutations dataset of S3310; M1337.R: reverse mutations dataset of M1337.

Table S2. The importance of each feature for MutaBind2 single and multiple mutation models, respectively. IncNodePurity is used for describing the importance which is the total decrease in node impurities from splitting on the variable, averaged over all trees, Related to Figure 1, Figure 2, Table 1, Table 2 and Table 3.

Model	Feature	Importance
Single	$\Delta\Delta E_{vdw}$	1605
	$\Delta\Delta G_{solv}$	2523
	$\Delta\Delta G_{fold}$	3027
	SA_{com}^{wt}	1522
	SA_{part}^{wt}	1577
	CS	4555
	N_{cont}^{wt}	2032
Multiple	$\Delta\Delta E_{vdw}$	1894
	$\Delta\Delta G_{solv}$	1984
	$\Delta\Delta G_{fold}$	3784
	SA_{com}^{wt}	1614
	SA_{part}^{wt}	1226
	CS	3852
	ΔE_{vdw}^{wt}	2313

Table S3. MutaBind2 performance, Related to Table 1.

Model	Training/Test set	All mutations			$\Delta\Delta G_{exp} \geq 0$		$\Delta\Delta G_{exp} < 0$	
		R	**RMSE**	**Slope**	**R**	**RMSE**	**R**	**RMSE**
Single	S4191/S4191	0.82	1.19	1.09	0.72	1.13	0.72	1.28
	S4191/S3310	0.76	1.16	1.07	0.72	1.13	0.74	1.25
	S4191/S3310.R	0.69	1.30	0.89	-	-	0.69	1.30
Multiple	M1707/M1707	0.87	1.61	1.14	0.73	1.54	0.75	1.72
	M1707/M1337	0.78	1.56	1.14	0.73	1.54	0.44	1.64
	M1707/M1337.R	0.75	1.78	1.01	-	-	0.75	1.78

R: Pearson correlation coefficient between experimental and predicted $\Delta\Delta G$ values. RMSE (kcal mol^{-1}): root-mean square error, is the standard deviation of the residuals (prediction errors). Slope: the slope of the regression line between experimental and predicted ΔΔG values. All reported correlation coefficients are statistically significantly different from zero (p-value << 0.01).

Table S4. The performance for Random Forest regression (RF), Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) methods on single mutation training dataset S4191. CV4: leave-one-complex-out validation; CV5: leave-one-binding site-out validation, Related to Table 1 and Figure 2.

Method	**Validation**	**All mutations**		$\Delta\Delta G_{exp} \geq 0$		$\Delta\Delta G_{exp} < 0$	
		R	**RMSE**	**R**	**RMSE**	**R**	**RMSE**
RF	CV4	0.76	1.34	0.61	1.31	0.67	1.39
	CV5	0.69	1.50	0.54	1.41	0.47	1.65
SVM	CV4	0.72*	1.42	0.55*	1.41	0.66	1.43
	CV5	0.61*	1.62	0.45*	1.54	0.41*	1.74
XGBoost	CV4	0.75*	1.35	0.58*	1.34	0.70*	1.35
	CV5	0.67*	1.53	0.50*	1.45	0.46	1.66

*p-value < 0.01 compared to Random Forest (Hittner2003 test)

Table S5. Comparison of methods' performance for mutations from two independent test sets S33 and S19, Related to Table 2.

Methods	R	RMSE
S33, 33 mutations		
MutaBind2	0.59	1.07
MutaBind	0.59	1.18
iSEE	0.62	0.81
mCSM-PPI2	0.75*	0.63
BeAtMuSiC	0.48	1.02
FoldX	0.50	1.36
S19, 19 mutations		
MutaBind2	0.65	1.33
MutaBind	0.67	1.27
iSEE	0.73	1.37
mCSM-PPI2	0.41[a]	1.61
BeAtMuSiC	0.24[b]*	1.70
FoldX	0.72	1.15

All presented values of correlation coefficients are statistically significantly different from zero (p-value < 0.01) except for [a]p-value = 0.08 and [b] p-value = 0.32. * show statistically significant difference with p-value < 0.05 compared to MutaBind2 (Hittner2003 test implemented in R package *cocor* is used for comparing correlation coefficients (Diedenhofen and Musch, 2015; Hittner et al., 2003)). The majority mutations in S33 and S19 are mutations decreasing binding affinity with experimentally measured $\Delta\Delta G_{exp}$ ≥ -0.38 and -0.24 kcal mol^{-1}, respectively.

Table S6. Comparison of performance between MutaBind2 and mCSM-PPI2, Related to Table 2.

Training/Test	Methods	All mutations			$\Delta\Delta G_{exp} \geq 0$		$\Delta\Delta G_{exp} < 0$	
		R	**RMSE**	**Slope**	**R**	**RMSE**	**R**	**RMSE**
S4191/S3310	MutaBind2	0.76	1.16	1.07	0.72	1.13	0.74	1.25
S4169/S3310	mCSM-PPI2	0.76	1.20	1.29	0.69**	1.20	0.62**	1.17
S4169/S4169	MutaBind2	0.74	1.18	1.13	0.73	1.10	0.62	1.43
	mCSM-PPI2	0.76*	1.19	1.29	0.69**	1.20	0.56*	1.17
S8338/S8338	MutaBind2 CV4	0.74	1.37	1.08	0.64	1.33	0.58	1.41
	MutaBind2 CV5	0.66	1.53	1.17	0.54	1.47	0.43	1.60
	mCSM-PPI2 CV4	0.75	1.30	*NA*	*NA*	*NA*	*NA*	*NA*
	mCSM-PPI2 CV5	0.67	1.39	*NA*	*NA*	*NA*	*NA*	*NA*

**p-value < 0.01 and *p-value < 0.05 compared to MutaBind2 (Hittner2003 test). NA: Data not available. The $\Delta\Delta G$ values for mCSM-PPI2 trained and tested on S4169 (forward mutation dataset for parameterizing mCSM-PPI2 method) were obtained from http://biosig.unimelb.edu.au/mcsm_ppi2/. $\Delta\Delta G$ values for mCSM-PPI2 trained and tested on S8338 (training dataset of mCSM-PPI2 method) were not provided on mCSM-PPI2 website, and the R and RMSE values for CV4 and CV5 were obtained from paper (Rodrigues et al., 2019).
S4191/S3310: MutaBind2 trained on S4191 and tested on S3310.
S4169/S4169: MutaBind2 retrained on S4169 and tested on S4169.
S8338/S8338: CV4 and CV5 validation for MutaBind2 retrained on S8338 and tested on S8338.

Table S7. Performance of MutaBind2 parameterized on different datasets where the reverse mutation dataset was compiled using complexes with a different number of mutations binned from 5 to 100. Related to Table 1 and Table 2.

a. Performance of MutaBind2 retrained on "S3310+Reverse" set and tested on independent datasets of S487 and S877.

Cutoff	All mutations		$\Delta\Delta G_{exp} > 0$		$\Delta\Delta G_{exp} < 0$	
	# of mutations	# of complexes	# of mutations	# of complexes	# of mutations	# of complexes
S3310	3310	265	2504	188	712	173
5 to 30	4458	265	2504	188	1860	211
10 to 30	4191	265	2504	188	1593	188
15 to 30	4023	265	2504	188	1425	184
10 to 50	4517	265	2504	188	1919	188
10 to 100	4741	265	2504	188	2143	188
All	6620	265	3216	265	3216	265

Protein-protein complexes in S3310 with the number of $\Delta\Delta G_{exp} > 0$ mutations binned from 5 to 30, 10 to 30, 15 to 30, 10 to 50 and 10 to 100 were used for building the reverse mutation dataset. All: with every single forward mutation in S3310 being reversed. (The distribution of the number of $\Delta\Delta G_{exp} > 0$ mutations over protein complexes is shown in Figure S1b).

Dataset	Cutoff	R	RMSE
S487	**10 to 30**	**0.41**	**1.25**
	5 to 30	0.39**	1.28
	15 to 30	0.42*	1.23
	10 to 50	0.41	1.26
	10 to 100	0.40	1.26
	All	0.36**	1.31
S877	**10 to 30**	**0.55**	**1.37**
	5 to 30	0.54*	1.38
	15 to 30	0.55	1.37
	10 to 50	0.55	1.37
	10 to 100	0.54	1.38
	All	0.53**	1.40

* show statistically significant difference with p-value < 0.05 and **p-value < 0.01 compared to MutaBind2 trained on S4191 set where reverse mutation dataset was compiled using the cutoff of "10 to 30". For testing on S487 set, MutaBind2 was retrained after removing S487 from the training dataset.

b. Performance of MutaBind2 retrained on "Skempi+Reverse" set and tested on the independent dataset of S1748.

Cutoff	All mutations		$\Delta\Delta G_{exp} \geq 0$		$\Delta\Delta G_{exp} < 0$	
	R	RMSE	R	RMSE	R	RMSE
10 to 30	**0.63**	**1.25**	**0.45**	**1.17**	**0.77**	**1.52**
5 to 30	0.62*	1.27	0.44**	1.21	0.78	1.47
15 to 30	0.63*	1.25	0.46	1.16	0.77	1.52
10 to 50	0.63	1.26	0.45	1.18	0.77	1.50
10 to 100	0.62**	1.27	0.43**	1.21	0.77	1.47
All	0.63	1.26	0.42**	1.21	0.80**	1.40

The protein-protein complexes in Skempi with the number of $\Delta\Delta G_{exp} > 0$ mutations binned from 5 to 30, 10 to 30, 15 to 30, 10 to 50 and 10 to 100 were used for building the reverse mutation dataset. All: with every single forward mutation in Skempi being reversed. *p-value < 0.05 and **p-value < 0.01 compared to MutaBind2 using cutoff of "10 to 30" (Hittner2003 test).

Table S8. Performance of MutaBind2 using Modeller and FoldX to generate initial mutant structures for a reverse mutation dataset, Related to Table 1 and Figure 2.

Training/Test set	Model	All mutations			$\Delta\Delta G_{exp} \geq 0$		$\Delta\Delta G_{exp} < 0$	
		R	RMSE	Slope	R	RMSE	R	RMSE
Modeller								
Test: S1748	MutaBind2	0.63	1.25	0.83	0.45	1.17	0.77	1.52
S4191/S4191	MutaBind2	0.82	1.19	1.09	0.72	1.13	0.72	1.28
	MutaBind2 CV4	0.76	1.34	1.11	0.61	1.31	0.67	1.39
	MutaBind2 CV5	0.69	1.50	1.18	0.54	1.41	0.47	1.65
Test: S487	MutaBind2	0.41	1.25	0.59	0.39	1.20	-	1.52
Test: S877	MutaBind2	0.55	1.37	0.97	0.56	1.34	-	1.48
S4191/S3310	MutaBind2	0.76	1.16	1.07	0.72	1.13	0.74	1.25
FoldX								
Test: S1748	MutaBind2	0.62	1.27	0.84	0.46	1.15	0.73	1.60
S4191/S4191	MutaBind2	0.82	1.18	1.09	0.72	1.13	0.73	1.26
	MutaBind2 CV4	0.76	1.34	1.11	0.60	1.31	0.67	1.38
	MutaBind2 CV5	0.69	1.49	1.17	0.54	1.40	0.47	1.64
Test: S487	MutaBind2	0.43	1.23	0.61	0.42	1.17	-	1.56
Test: S877	MutaBind2	0.55	1.37	0.96	0.56	1.34	-	1.48
S4191/S3310	MutaBind2	0.75	1.17	1.08	0.72	1.13	0.70	1.31

Table S9. Performance of MutaBind2 for different types of multiple mutations, Related to Table 1 and Figure 2**.**

Type	MutaBind2(CV4)		
	# of mutations (# of complexes)	**R**	**RMSE**
All mutations	1707(120)	0.74	2.13
Double mutations	881(99)	0.81	2.22
Triple or higher number of mutations	826(75)	0.62	2.03
Mutations on one chain	853(105)	0.63	2.09
Mutations on multiple chains	854(43)	0.81	2.16
Double mutations on one chain	347(82)	0.66	2.04
Double mutations on multiple chains	534(31)	0.85	2.32
Triple or higher number of mutations on one chain	506(66)	0.61	2.13
Triple or higher number of mutations on multiple chains	320(25)	0.65	1.86

Table S10. Performance of MutaBind2 parameterized on different datasets where mutations with multiple experimental measurements of $\Delta\Delta G_{exp}$ were processed using different ways, Related to Table 1, Table 2 and Figure 2.

The leave-one-complex-out validation results							
Training/Test	**All mutations**			**$\Delta\Delta G_{exp} \geq 0$**		**$\Delta\Delta G_{exp} < 0$**	
	R	**RMSE**	**Slope**	**R**	**RMSE**	**R**	**RMSE**
S4191/S4191 (Std < 1.0)	0.76	1.34	1.11	0.61	1.31	0.67	1.39
S4944/S4944 (All values)	0.76	1.34	1.12	0.61	1.31	0.67	1.38
S4090/S4090 (Std ≤ 0.4)	0.76	1.34	1.11	0.61	1.31	0.67	1.39
S4183/S4183 (Diff < 2)	0.76	1.34	1.11	0.61	1.31	0.67	1.39

No significant difference between MutaBind2 trained and tested on S4191 and other test sets.
In the dataset of S3310, there are 408 mutations with multiple experimental measurements of binding affinity changes and their standard deviations are all less than 1 kcal mol^{-1} (The distribution of standard deviation is shown in Figure S1a). S4191: the average values were used for these 408 mutations; S4944: using all experimental measurements of $\Delta\Delta G_{exp}$ for these 408 mutations; S4090: only mutations with standard deviations with less than or equal to 0.4 kcal mol^{-1} are included in the training set, and the average value was used for these cases; S4183: only mutations with the difference between maximal and minimal $\Delta\Delta G_{exp}$ values of less than 2 kcal mol^{-1} (the cutoff was used by mCSM-PPI2) are included in the training set, and the average value was used for these cases.

Test set	**Training set**	**R**	**RMSE**
S487	S4191	0.41	1.25
	S4944	0.41	1.24
	S4090	0.41	1.25
	S4183	0.42	1.25
S877	S4191	0.55	1.37
	S4944	0.55	1.37
	S4090	0.55	1.37
	S4183	0.54*	1.38

*p-value < 0.05 compared to MutaBind2 trained and tested on S4191 (Hittner2003 test).
For testing on S487 set, MutaBind2 was retrained after removing S487 from the training dataset.

Table S11. The performance for MutaBind2 trained and tested on the dataset including 34 multiple mutations with more than 10 mutations using leave-one-complex-out validation, Related to Table 1 and Figure 2.

Composition of multiple mutations	**# of multiple mutations**	**R**	**RMSE**
2	881	0.79	2.27
3-5	618	0.59	2.07
6-10	208	0.74	1.92
10+	34	-	4.59

All presented values of correlation coefficients are statistically significantly different from zero (p-value << 0.01).

Transparent Methods

Experimental datasets of mutations used for training

The training dataset is compiled from the most recent SKEMPI2.0 database (Jankauskaite et al., 2019), which includes experimentally measured values of dissociation constants for wild-type and mutant proteins with the available crystal structures. Changes in binding affinity are also provided in SKEMPI2.0 and calculated as $\Delta G = RTln(K_D)$. We applied the following criteria to the SKEMPI2.0 data set by removing the following complexes and mutations: (a) complexes with modified residues at the protein-protein binding interface; (b) complexes containing a chain of less than 20 residues long; (c) mutations with mutated sites having missing coordinates; (d) changes in affinity measured by an 'unusual method' as defined in SKEMPI2.0; (e) mutations on metal coordination sites; (f) mutations without binding affinity experimental values; and (g) entries with ten or more multiple mutations (Figure S1). There are 408 mutations with multiple experimental measurements of changes in binding affinity and their standard deviations are all less than 1.0 kcal mol^{-1} (Figure S1a), and the average value was used for these cases. Three additional ways to process these cases were tried but the performance did not change (Table S10). As the number of multiple mutations with more than 10 mutations is small and prediction accuracy for these multiple mutations is low (Table S11), the upper limit of 10 mutations was used in the study. As a result, 3,310 single mutations from 265 wild-type protein-protein complexes (it will be referred to as S3310) and 1,337 multiple mutations from 120 wild-type protein-protein complexes (it will be referred to as M1337) were retained (Table S1 and Figure S2). *Multiple mutations* correspond to cases where several mutations are introduced on one or several chains of a protein complex simultaneously.

The Gibbs free energy ($\Delta\Delta G$) of a system can be represented as a thermodynamic state function where the absolute values of $\Delta\Delta G$ for a *forward mutation* ($\Delta\Delta G_{wt\rightarrow mut}$) and $\Delta\Delta G$ for the *reverse mutation* ($\Delta\Delta G_{mut\rightarrow wt}$) should be approximately equal to each other. In order to prepare a more balanced training dataset, we augmented the existing mutations increasing binding affinity from the forward mutation sets with the modelled reverse mutations (see Table S1). In order to balance the prediction accuracy for both types of mutations: decreasing and increasing binding affinity, we used protein complexes with the number of experimentally characterized mutations from 10 to 30 and 10 to 50 to build the single and multiple reverse mutation dataset respectively. The performance is shown in Table S7. Therefore, the final training set including both forward and reverse mutations comprised 4,191 single mutations from 265 wild-type protein complexes (it will be referred to as S4191) and 1,707 multiple mutations from 120 wild-type protein complexes (it will be referred to as M1707), respectively (Table S1).

Structure optimization protocol

For the forward mutation datasets S3310 and M1337 ($\Delta\Delta G_{wt\rightarrow mut}$), the structure optimization protocol was the same as the one used in MutaBind (the flowchart for the structure optimization protocol is shown in Figure S3). Namely, we used the BuildModel module of FoldX (Guerois et al., 2002) to introduce single or multiple point mutations on the wild-type crystal structure obtained from the Protein Data Bank (PDB) (Berman et al., 2000). Next we added missing heavy side-chain and hydrogen atoms via the VMD program (Humphrey et al., 1996) using the topology parameters of the CHARMM36 force field (MacKerell et al., 1998). After that we performed a 100-step energy minimization in the gas phase for both wild-type and mutant complex structures applying harmonic restraints with the force constant of 5 kcal mol^{-1} Å^{-2} on the backbone atoms of all residues. The energy minimization was carried out by NAMD program version 2.9 (Phillips et al., 2005) using the force field CHARMM36 (MacKerell et al., 1998). A 12 Å cutoff distance for nonbonded interactions was applied to the systems. Lengths of hydrogen-containing bonds were constrained by the SHAKE algorithm (Hoover, 1985).

For the reverse mutation datasets, we modeled the mutant structures with the Modeller software (Sali and Blundell, 1993) using wild-type crystal structures as the templates (Table S8). To minimize the error introduced by structural modelling, only the mutated protein chain was modeled for single mutations, and for multiple mutations on multiple protein chains, the whole complex was modelled. The structural model was discarded if the root-mean-square deviation of all aligned C_α atoms between any of the modelled chains and the template was larger than 2 Å. Then the RepairPDB module was applied to further optimize the structure and mutations were introduced using the BuildModel module from FoldX. After that a 1000-step energy minimization in the gas phase was carried out for both wild-type and mutants using harmonic restraints (with the force constant of 5 kcal mol^{-1} Å^{-2}) applied on backbone atoms of all residues using NAMD. Minimization was performed for the whole protein complex.

Calculating changes in binding affinity

The scoring function of MutaBind2 includes seven distinct terms for single and multiple mutations, it is parameterized using the S4191 and M1707 datasets (Table S1), respectively. The terms that contribute significantly to the quality of the MutaBind2 single and multiple mutation models are shown in Table S2 and described below.

The six terms of the scoring function described below are common for both single and multiple mutation models.

- $\Delta\Delta E_{vdw}$ is the change of van der Waals interaction energy upon a single or multiple mutation(s) ($\Delta\Delta E_{vdw} = \Delta E_{vdw}^{mut} - \Delta E_{vdw}^{wt}$). ΔE_{vdw} is calculated as a difference between van der Waals energies of a complex and each interacting partner using the ENERGY module of CHARMM (Brooks et al., 1983). The minimized structure of the wild-type or mutant complex structure was used for the calculation.

- $\Delta\Delta G_{solv}$ approximates the change of polar solvation energy upon mutation(s) ($\Delta\Delta G_{solv} = \Delta G_{solv}^{mut} - \Delta G_{solv}^{wt}$), ΔG_{solv} is obtained from numerically solving the Poisson-Boltzmann (PB) equation with the PBEQ module (Im et al., 1998) of the CHARMM program using the minimized structure of the wild-type or mutant complex. For the PB calculation, dielectric constants ε = 2 for the protein interior and ε = 80 for the exterior aqueous environment were used.
- $\Delta\Delta G_{fold}$ is the change in stability of the protein complex upon mutation(s) ($\Delta\Delta G_{fold} = \Delta G_{fold}^{mut} - \Delta G_{fold}^{wt}$) where each term is defined as the unfolding free energy of the mutant and wild-type protein complexes. It is calculated with the BuildModel module from the FoldX software (Guerois et al., 2002) which uses an empirical force field. This term may account for those cases where mutated proteins are unfolded in unbound states and can only fold upon binding to its partner.
- SA_{part}^{wt} and SA_{com}^{wt} are solvent accessible surface areas of the mutated residues in the wild type unbound partner and complex structure respectively. These terms are calculated by the DSSP program (Joosten et al., 2011) using the crystal structure of the wild-type complex. For multiple mutations this term is calculated as a sum of the solvent accessible surface areas of all mutated residues.
- CS is the change of evolutionary conservation of a mutated site upon introducing mutations calculated using the PROVEAN program (Choi et al., 2012). This is used to account for the fact that a site can be evolutionary conserved because it is important for interactions with other proteins and any change in this site may affect its function in a detrimental way. For multiple mutations this term is calculated by summing up *CS* for all mutations.

A scoring function for single mutations included an additional term, N_{cont}^{wt}, representing the number of interactive residues between one partner where a mutation was introduced and another partner in a wild type structure. If any heavy atom of a residue in one partner was located within 10 Å from any heavy atom of another partner, we defined this residue as an interactive residue. A scoring function for multiple mutations included an additional term ΔE_{vdw}^{wt} calculated as a difference between van der Waals energies of a complex and each interacting partner for the wild-type structure, as described above.

MutaBind2 predictive models were built using the random forest (RF) regression algorithm implemented in the R package "randomForest" using Breiman's random forest algorithm (Breiman, 2001). Hyperparameter optimization in a balanced "CV2" cross-validation (see the section below) suggested that the number of trees "ntree" parameter should be set to 500 and the number of features/terms randomly sampled as candidates for splitting at each node ("mtry") should be set to 2. Feature importance in RF models is shown in Supplementary Table 2 and all features listed above contribute significantly to the quality of the models. The performance of MutaBind2 trained on S4191 and M1707 sets is shown in Table S3 and Figure S4c. The Pearson correlation coefficient between experimental and calculated changes in binding affinity is R = 0.82 and the corresponding root-mean-square error (RMSE) is 1.19 kcal mol^{-1} for single mutations and R = 0.87 and the RMSE is 1.61 kcal mol^{-1} for multiple mutations. We also tested the performance of other algorithms, including Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost), however, the random forest regression algorithm shows the best performance (Table S4).

MutaBind2 calculations take several minutes for a single mutation in a protein complex of about 350 residues, and require less than a minute for each additional mutation introduced in the same complex and therefore require about 12 hours for calculations of one thousand mutations (Figure S10).

Five types of cross-validation (CV) procedures

We performed five types of cross-validation. In the "CV1" cross-validation, 80% of all mutations from S4191 or M1707 set were randomly selected to train the model and the remaining 20% of the mutations were used for testing; we repeated the procedure 100 times. For "CV2" cross-validation, 50% of the mutations were randomly chosen for training and the remaining mutations for testing, also repeated 100 times. As shown in Figure S2, the distribution of

the number of mutations per protein is not uniform, so to take this bias into account, we performed the third type of cross-validation ("CV3"). First, we randomly sampled up to ten mutations per protein complex from S4191 and M1707, the procedure was repeated 10 times and yielded ten subsets. Then 80 percent of the mutations were randomly selected from each subset for training and the rest for testing, repeated 10 times.

We also performed the "CV4" cross-validation by leaving one complex and its mutations out as a test set and using the rest of the complexes/mutations to train the model, repeating this process for each protein complex. In addition, a "CV5" cross-validation accounted for similarities between binding sites of different complexes (the definition of similar binding sites is taken from (Jankauskaite et al., 2019; Moal and Fernandez-Recio, 2012)). Namely, we used all complexes and corresponding mutations from one cluster/type of binding site for testing and trained the model on the rest of the complexes/mutations, repeated for each type of binding site. During the cross-validation procedures, forward and reverse mutations were kept in the same set, either training or testing.

Assessment of quality of classification

One way to evaluate the performance of MutaBind2 is to assess the quality of classification of mutations into mutations with large amplitudes of their effects on binding affinity: highly decreasing ($\Delta\Delta G \geq$ 1.5 kcal mol^{-1}) and highly increasing ($\Delta\Delta G \leq$ -1.5 kcal mol^{-1}). The explanation is provided in Figure S6c and Figure S6d.

Prediction performance was measured using area under the ROC curve (AUC), accuracy, precision, sensitivity, specificity, negative predictive value (NPV) and Matthews correlation coefficient (MCC). Positives and negatives were defined as those mutations with predicted $\Delta\Delta G$ values within or outside the range specified above for experimental $\Delta\Delta G$ values. The accuracy was defined as a percentage of correctly classified mutations (true positives, TP, and true negatives, TN) out of the total number of mutations (TP + TN)/(TP + TN + FP + FN), where FN are false negatives, and FP are false positives. Sensitivity was defined as TP/(TP + FN), specificity was calculated as TN/(TN + FP) (false negatives, FN and false positives, FP). Additionally, in order to account for imbalances in the labeled dataset, the quality of the predictions was described by the Matthews correlation coefficient (MCC), a performance measure which is known to be more robust on unbalanced datasets:

$$\mathrm{MCC} = \frac{TP * TN - \mathrm{FP} * \mathrm{FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Supplemental References

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE. (2000). The Protein Data Bank. Nucleic Acids Research 28, 235-242.

Breiman L. (2001). Random Forests. Machine Learning 45, 5-32.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S and Karplus M. (1983). Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. Journal of Computational Chemistry 4, 187-217.

Choi Y, Sims GE, Murphy S, Miller JR and Chan AP. (2012). Predicting the functional effect of amino acid substitutions and indels. Plos One 7, e46688.

DeLong ER, DeLong Dm Fau - Clarke-Pearson DL and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837-845.

Diedenhofen B and Musch J. (2015). cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10, e0121945.

Fragoza R, Das J, Wierbowski SD, Liang J, Tran TN, Liang S, Beltran JF, Rivera-Erick CA, Ye K, Wang TY, et al. (2019). Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. Nat Commun 10, 4141.

Guerois R, Nielsen JE and Serrano L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal Of Molecular Biology 320, 369-387.

Hittner JB, May K and Silver NC. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. J Gen Psychol 130, 149-168.

Hoover WG. (1985). Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A 31, 1695-1697.

Humphrey W, Dalke A and Schulten K. (1996). VMD: visual molecular dynamics. J Mol Graph 14, 33-38, 27-38.

Im W, Beglov D and Roux B. (1998). Continuum Solvation Model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. Computer Physics Communications 111, 59-75.

Jankauskaite J, Jimenez-Garcia B, Dapkunas J, Fernandez-Recio J and Moal IH. (2019). SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics 35, 462-469.

Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C and Vriend G. (2011). A series of PDB related databases for everyday needs. Nucleic Acids Research 39, D411-419.

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102, 3586-3616.

Moal IH and Fernandez-Recio J. (2012). SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinformatics 28, 2600-2607.

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L and Schulten K. (2005). Scalable molecular dynamics with NAMD. Journal of Computational Chemistry 26, 1781-1802.

Rodrigues CHM, Myung Y, Pires DEV and Ascher DB. (2019). mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. Nucleic Acids Research 47, W338-W344.

Sali A and Blundell TL. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal Of Molecular Biology 234, 779-815.