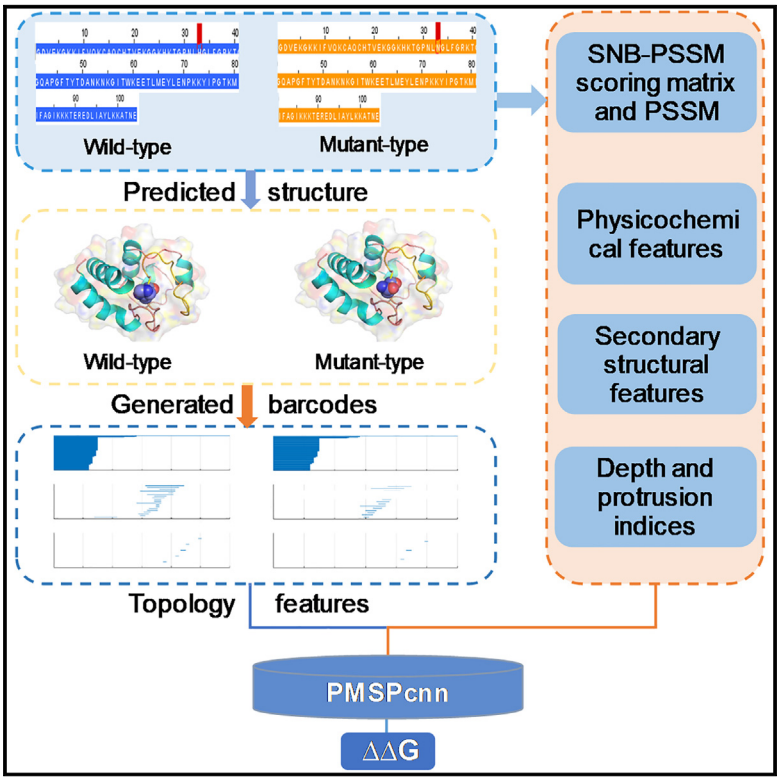
Structure

PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network

Graphical abstract



Highlights

- PMSPcnn is an unbiased CNN-based predictor to predict protein stability changes
- Persistent homology is used to extract topological features
- Includes a strategy of regression stratification cross-validation
- PMSPcnn achieves state-of-the-art performance compared to existing methods

Authors

Xiaohan Sun, Shuang Yang, Zhixiang Wu, Jingjie Su, Fangrui Hu, Fubin Chang, Chunhua Li

Correspondence

chunhuali@bjut.edu.cn

In brief

Sun et al. present the unbiased model PMSPcnn for the prediction of mutation-induced protein stability changes. PMSPcnn considers the anti-symmetry property, utilizes five types of characteristics including PH-based topology features, and is trained via regression stratification cross-validation. PMSPcnn achieves a state-of-the-art performance compared to existing methods on multiple test sets.

Sun et al., 2024, Structure *32*, 1–11
June 6, 2024
https://doi.org/10.1016/j.str.2024.02.016

CellPress

Structure

CellPress

Resource

PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network

Xiaohan Sun,[1] Shuang Yang,[1] Zhixiang Wu,[1] Jingjie Su,[1] Fangrui Hu,[1] Fubin Chang,[1] and Chunhua Li[1,2,*]
[1]College of Chemistry and Life Science, Beijing University of Technology, Beijing 100124, China
[2]Lead contact
*Correspondence: chunhuali@bjut.edu.cn
https://doi.org/10.1016/j.str.2024.02.016

SUMMARY

Protein missense mutations and resulting protein stability changes are important causes for many human genetic diseases. However, the accurate prediction of stability changes due to mutations remains a challenging problem. To address this problem, we have developed an unbiased effective model: PMSPcnn that is based on a convolutional neural network. We have included an anti-symmetry property to build a balanced training dataset, which improves the prediction, in particular for stabilizing mutations. Persistent homology, which is an effective approach for characterizing protein structures, is used to obtain topological features. Additionally, a regression stratification cross-validation scheme has been proposed to improve the prediction for mutations with extreme ΔΔG. For three test datasets: S^{sym}, p53, and myoglobin, PMSPcnn achieves a better performance than currently existing predictors. PMSPcnn also outperforms currently available methods for membrane proteins. Overall, PMSPcnn is a promising method for the prediction of protein stability changes caused by single point mutations.

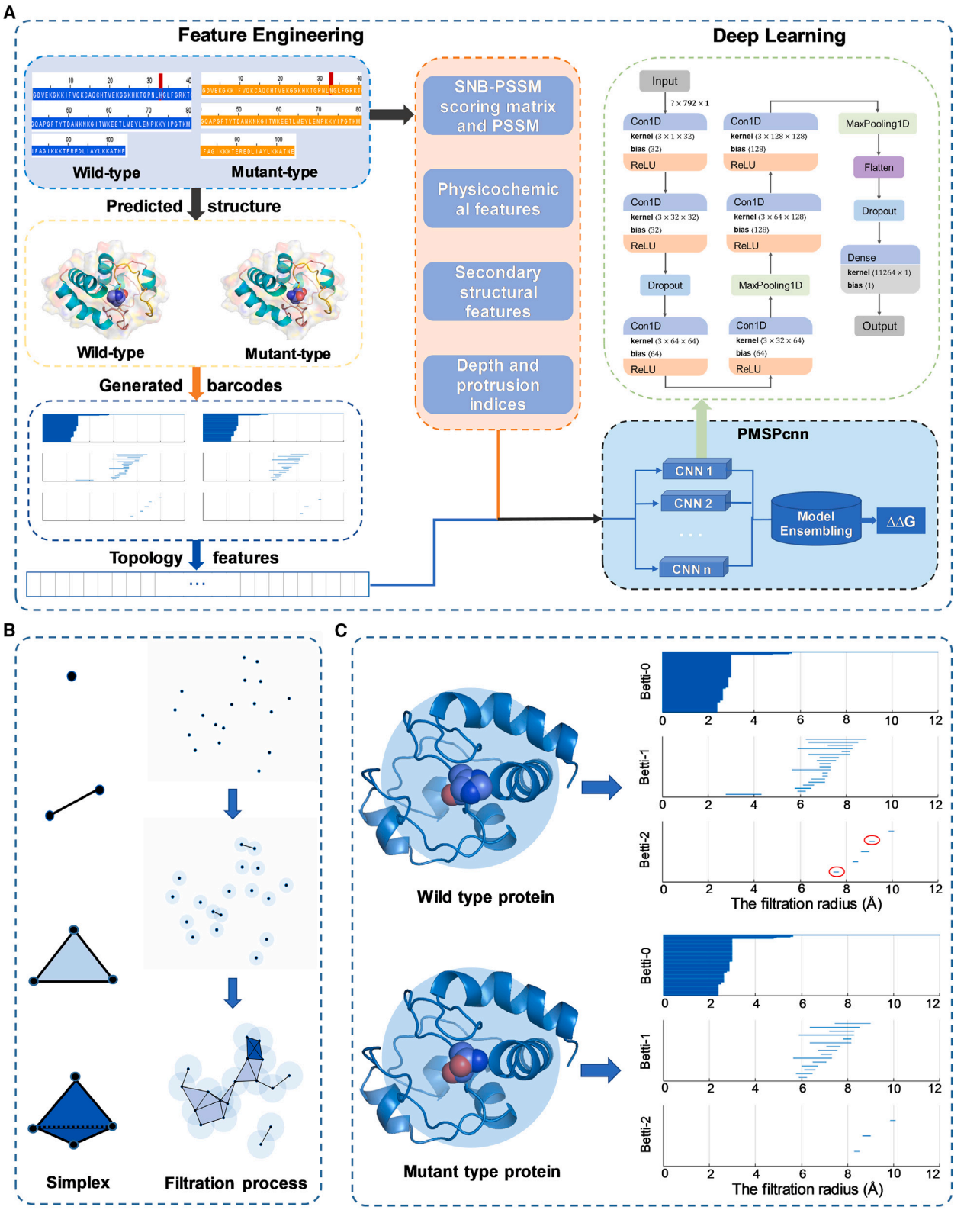
INTRODUCTION

Many human genetic diseases are directly linked to missense mutations in proteins.[1] Missense mutations often cause an increase or decrease in protein stabilities, thereby affecting their functions and bringing about diseases.[1,2] Currently, the widely existing single point mutations underlie a variety of diseases. Accurately predicting protein stability changes caused by single point mutations is an important and challenging issue. Typically, the mutation effect on protein stability is quantified by the mutation-caused change in folding free energy (ΔΔG).[1] Experimental methods to detect the stability changes, such as differential scanning calorimetry and thermal fluorescence spectroscopy are time-consuming and labor-intensive.[3] Thus, it is urgent to develop an efficient and effective theoretical method to predict the stability changes caused by mutations.

Currently many theoretical methods have been developed for predicting the protein stability changes (ΔΔG), which can be mainly divided into three categories: energy-based, knowledge-based, and machine learning-based methods.[4] Free energy perturbation (FEP) is an early developed method based on molecular dynamics simulation and generally acknowledged to have high accuracy.[5] And FoldX is a traditional energy-based method that uses empirical force fields and energy term calculations to estimate the effects of mutations on protein stability.[6] Although the performances of the two methods are commendable, their calculation processes are complex and time-consuming. For knowledge-based methods, there are two well-known tools SDM[7] and DDGun.[8] SDM uses environment-specific substitution tables (ESSTs), and DDGun, a linear parametric model, uses evolutionary information to calculate the stability difference between the wild-type and mutant structures. Due to the complexity of the stability prediction, the two approaches using statistical method and linear model, respectively, have limited generalization ability. In recent years, machine learning has achieved great success, and many machine learning-based methods have been developed, including MAESTRO,[9] mCSM,[10] DUET,[11] I-Mutant,[12,13] iStable,[14] ThermoNet,[15] and so on. Table S1 lists their detailed information. These methods use traditional tree-based models or deep neural network, and take into account a variety of features including sequence, structural, and physicochemical ones, which make them have an improvement in generalization ability to some extent.

The common problems with most machine learning-based methods are the poor predictions for the stabilizing mutations and for the mutations with extreme ΔΔG values. For the first one, the main reason is the use of unbalanced experimental data (more than 75% of mutations are destabilizing)[16] as the training set, which leads to a prediction bias toward destabilizing variations. As for the change in Gibbs free energy, there is an important property called anti-symmetry ($\Delta\Delta G_{reverse} = -\Delta\Delta G_{direct}$) between the direct and reverse mutations.[17] This property, ignored by most methods, can be utilized to construct

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016



(legend on next page)

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

a balanced training set to improve the prediction bias. For the second issue, the main reason lies in that the majority of the mutations in existing data have moderate stability changes and fewer mutations have extreme values. The extreme ΔΔG values (highly destabilizing or stabilizing mutations) are more likely to lead to severe pathological consequences probably by severely affecting protein structures. For the currently available methods, to the best of our knowledge, they do not do any attempt at a solution to the problem. We think that a stratified sampling scheme where the samples with different levels of ΔΔG values need to be covered by the training set may be constructive for improving the prediction of extreme ΔΔG values.

In addition to the data processing mentioned previously that may improve the model's prediction performance, the molecular characterization is also crucial for training an accurate predictor. Besides the sequence information, the further consideration of the topological structure information, we think, can enhance the model's performance. Persistent homology (PH), a branch of algebraic topology, provides an efficient approach to characterize protein topological structures.[18] Different from the traditional methods, the topological signatures from PH consider the connectivity of data points with continuous values rather than a single fixed value. Additionally, PH provides a delicate balance between data simplification and intrinsic structure characterization, and has been successfully applied to various aspects, such as protein-ligand binding affinity prediction, protein-protein interaction prediction, and drug virtual screening.[19–21] To obtain the PH-based topology features, protein structure data are required. However, for a studied protein, its 3D structure is usually not available (the available experimental protein structures only account for about 0.2%).[22] Fortunately, AlphaFold2 has achieved a highly accurate protein structure prediction.[23] Many studies have utilized AlphaFold2 as a tool to extract protein structural information for various predictions. For example, Yuan et al.[24] proposed an accurate predictor for identifying DNA-binding residues based on the structural models predicted by AlphaFold2, and Shohei et al.[25] obtained secondary structure and dihedral angle information from the predicted protein structures with AlphaFold2 to predict protein mononucleotide binding sites. Thus, we want to extract the PH-based topological features from the protein structures obtained by AlphaFold2 for protein stability change prediction.

In this work, we construct an unbiased model PMSPcnn based on convolutional neural network (CNN) for predicting ΔΔG upon single point mutations. The anti-symmetry is considered to construct a balanced training dataset. The topological features are extracted from the topological invariance obtained by PH. An effective scheme named regression stratification cross-validation (RScv) is proposed to improve the model's prediction ability for the mutations with extreme ΔΔG values. The performance comparison of PMSPcnn with other methods is carried out on four different datasets and PMSPcnn shows the state-of-the-art performance.

RESULTS

An overview of PMSPcnn

Our model PMSPcnn is constructed based on the CNN which includes two modules: feature engineering and deep learning, as shown in Figure 1A. For feature engineering, topological features are generated from the barcodes obtained by applying PH analysis on the predicted protein structures by AlphaFold2. And other four types of features are extracted from protein sequences, containing physicochemical features, secondary structure features, depth and protrusion indices, as well as SNB-PSSM and position specific scoring matrix (PSSM) based evolutionary information. A detailed description of the five features is given in STAR Methods. CNN can automatically learn the important features for prediction, different from the traditional machine learning algorithms which often require manual selection and design of features. Then, the features of all the samples are combined and fed into the prediction model. For deep learning, we adopt 15-fold RScv (see STAR Methods) to train the model, and the obtained 15 sub-models with least validation loss on the corresponding validation fold are integrated into the ensemble model PMSPcnn, whose final output is the average value of the outputs from the 15 sub-models. The setting of the model's hyperparameters is described in Methods S1 and Figure S1 in supplemental information.

Effect of RScv on model training

In this work, we propose the scheme of RScv to improve the model training. To detect the effect of RScv, we trained the CNN model via 5-fold RScv on the training set Q6422 (see STAR Methods for more details about Q6422), with the results shown in Figure 2. Figure 2 also shows the corresponding results from the trained CNN model via traditional cross-validation (CV) for comparison. From Figure 2, it is clear that the model trained via the RScv performs much better than the model trained via the traditional one at each epoch in terms of mean and median of Pearson correlation coefficient (PCC, see STAR Methods). The former achieves the highest average PCC of 0.72 when epoch = 1,000, and thus the epoch is set to 1,000 in the CNN model training. The reason for the improvement, we think, is that RScv scheme can make sure the samples from eight bins in Q6422 can be sampled in the training data of each fold, so that the model can well learn the special features in different bins of samples, especially for the mutations with extreme ΔΔG values, which can be seen clearly in the PMSPcnn testing on the test sets.

Figure 1. The architecture of PMSPcnn

(A) PMSPcnn is constructed based on the CNN which includes two modules: feature engineering and deep learning. The network structure of CNN contains six 1D convolutional layers, two Dropout layers, two MaxPooling1D layers, one Flatten layer and one fully connected layer. Each convolutional layer uses Rectified Linear Unit (ReLU) as its activation function. Mean squared error (MSE) is used as the loss function during training.

(B) The filtration process of PH. As the filtration parameter (sphere radius) increases, a series of nested simplicial complexes are generated.

(C) The barcodes from the wild type and mutant protein (PDB ID: 1akk, with mutation H34N). The mutation residues are represented by cartoons. The spheres are the areas within 12 Å of the mutation residues (Cα atoms). The mutation H34N occurs at residue 34 from a large histidine to a small asparagine. The topological feature vectors of the mutation site are obtained based on the point clouds formed by the heavy atoms within the spheres. Due to the fewer atoms in asparagine than in histidine, the number of barcodes from the mutant is less than that from the wild-type. Evidently, the number of mutant Betti-2 bars is two less than that of the wild-type Betti-2 bars.

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

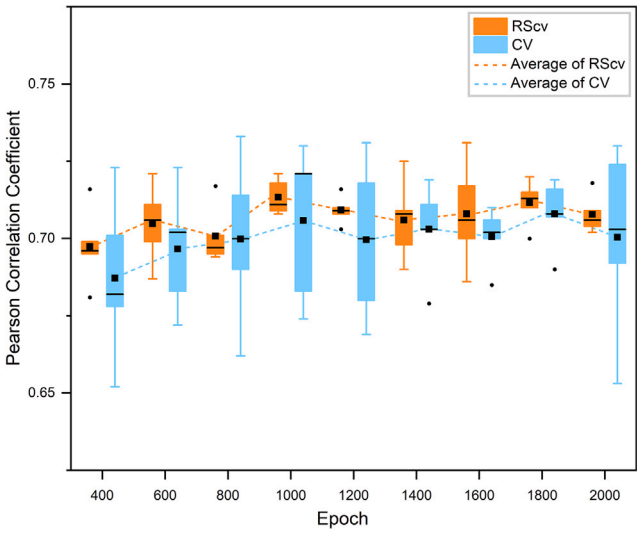


Figure 2. The boxplots of PCCs on the training set Q6422

The values of PCCs trained via 5-fold RScv and 5-fold traditional cross-validation (CV) at each epoch. The mean and median values are represented by a small square and a black line, respectively. The interquartile range is represented by the height of the box in the plot and spans from the lower quartile (25%) to the upper quartile (75%). The whiskers extend from the box and represent the range of non-outlier data points. Outliers are individual data points that fall beyond the whiskers and are represented as small black dots.

Analysis on feature importance

To evaluate the importance of different features for protein stability change prediction, we grouped all the features into three types, and used each type and all the features to separately train a CNN model on the training set Q6422 via 5-fold RScv. The three types of features are sequence-based type (physicochemical properties, as well as SNB-PSSM and PSSM based evolutionary information), structure-based type (secondary structural features, and depth and protrusion indices) and topology-based one (topological features from PH analysis), and the corresponding models are named PMSPcnn_Seq, PMSPcnn_Str, and PMSPcnn_Top, and the model utilizing all the features are denoted as PMSPcnn_All. Figure 3 shows the results from the four models. From Figure 3, the model PMSPcnn_All trained with all the features has the best performance (r = 0.72 and σ = 1.21 kcal/mol), followed by PMSPcnn_Top (r = 0.70 and σ = 1.28 kcal/mol). The detailed explanations of r and σ are given in STAR Methods. The remaining two models PMSPcnn_Seq and PMSPcnn_Str have a similar performance. Thus, the topological features obtained from the PH make important contributions to PMSPcnn_Top's good performance, which can characterize proteins at different scales and extract crucial and useful information for protein stability change prediction.

Performance of PMSPcnn on test set S^{sym}

Based on the results from the aforementioned two sections, with all the features and the RScv strategy considered, we trained our model PMSPcnn based on CNN via 15-fold RScv. The following presents the detailed analyses of the testing results of PMSPcnn on the test set S^{sym} (see STAR Methods for more details about S^{sym}).

On the test set S^{sym}, a balanced set widely used by many methods, we evaluated the prediction bias of our model PMSPcnn, with the results listed in Table 1. Table 1 also shows the corresponding results from the currently available twenty other methods for comparison. For the direct mutations, PMSPcnn outperforms most of the methods with r_{dir} = 0.73 and σ_{dir} = 1.06 kcal/mol (Figure 4A), only a little inferior to the two methods MUPRO[26] (r_{dir} = 0.79 and σ_{dir} = 0.94 kcal/mol) and STRUM[27] (r_{dir} = 0.75 and σ_{dir} = 1.05 kcal/mol). For the reverse mutations, PMSPcnn achieves the best prediction with r_{rev} = 0.73 and σ_{rev} = 1.06 kcal/mol (Figure 4B), significantly superior to all the other methods with r_{rev} increasing by 24% compared with the second best method ThermoNet. MUPRO and STRUM with the best performances on the direct mutations show not good performances on the reverse mutations, with r_{rev} = 0.07 and −0.15 respectively. Thus, generally our method PMSPcnn achieves the state-of-the-art performance on the test set S^{sym}. Additionally, it needs to be pointed out that due to ThermoNet's second best performance (r_{dir} = 0.58 and r_{rev} = 0.59) among all the methods listed in Table 1 and considering anti-symmetry property, the model comparison with ThermoNet was also performed on p53 and myoglobin test sets in the following section.

We note that some datasets used by some methods as the training sets have overlap with S^{sym} set to different extent, which mainly include eight widely used datasets whose duplicate rates with S^{sym} range from 0.11 to 0.47 as shown in Figure S2. Among them, S1615 set which MUPRO was trained on has 125 (about 37%) duplicate samples (direct mutations), and Q3421 set which

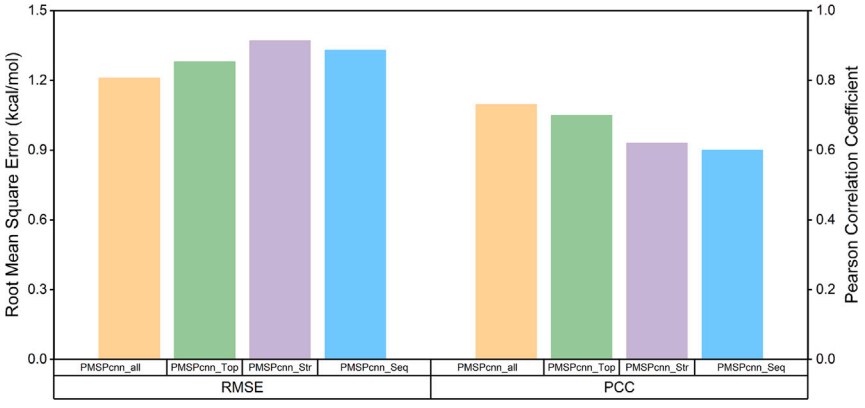


Figure 3. The performances of four models trained by different features

Prediction performances of the four models trained separately on the training set Q6422 via 5-fold RScv, which utilize sequence-based, structure-based, topology-based, and all the features denoted as PMSPcnn_Seq, PMSPcnn_Str and PMSPcnn_Top, and PMSPcnn_All, respectively.

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

CellPress

Table 1. Comparison of PMSPcnn with other methods on the balanced test set S^{sym}

Methods	r_{dir}	σ_{dir}	r_{rev}	σ_{rev}	$r_{dir\text{-}rev}$	$\langle\delta\rangle$
PMSPcnn[a]	0.73	1.06	0.73	1.06	−0.96	0.01
ThermoNet[a]	0.58	1.42	0.59	1.38	−0.95	−0.05
ACDC-NN-Seq[a]	0.55	1.44	0.55	1.44	−0.99	−0.01
DDGun3D[a]	0.56	1.42	0.53	1.46	−0.99	−0.02
DDGun[a]	0.48	1.47	0.48	1.50	−0.99	−0.01
PoPMuSiC^{sym}	0.48	1.58	0.48	1.62	−0.77	0.03
Rosetta	0.69	2.31	0.43	2.61	−0.41	−0.69
FoldX	0.63	1.56	0.39	2.13	−0.38	−0.47
MAESTRO	0.52	1.36	0.32	2.09	−0.34	−0.58
SDM[a]	0.51	1.74	0.32	2.28	−0.75	−0.32
PoPMuSiC 2.1	0.63	1.21	0.25	2.18	−0.29	−0.71
mCSM	0.61	1.23	0.14	2.43	−0.26	−0.91
DUET	0.63	1.20	0.13	2.38	−0.21	−0.84
MUPRO	0.79	0.94	0.07	2.51	−0.02	−0.97
CUPSAT	0.39	1.71	0.05	2.88	−0.54	−0.72
NeEMO	0.72	1.08	0.02	2.35	0.09	−0.60
AUTOMUTE	0.73	1.07	−0.01	2.61	−0.06	−0.99
I-Mutant 3.0	0.62	1.23	−0.04	2.32	0.02	−0.68
iStable	0.72	1.10	−0.08	2.28	−0.05	−0.6
STRUM	0.75	1.05	−0.15	2.51	0.34	−0.87
SAAFEC-SEQ	0.71	1.09	−0.39	2.71	0.58	−1.84

r_{dir}/r_{rev} (Equation 1) and $\sigma_{dir}/\sigma_{rev}$ (Equation 2) are the PCC and RMSE (kcal/mol) between the predicted and experimental ΔΔG values for the direct/reverse mutations (n = 342), respectively. $r_{dir\text{-}rev}$ (Equation 3) is the PCC between the predicted values for direct mutations and those for reverse mutations, and the bias $\langle\delta\rangle$ (Equation 4) is used to quantify prediction bias. The results of ACDC-NN-Seq and SAAFEC-SEQ are from reference,[28] those of ThermoNet from reference,[15] and those of all the other methods from ref. [32] The methods are ranked according to their performance r_{rev} on reverse mutations.

[a]Methods that consider the anti-symmetry property.

STRUM was trained on has 120 (about 35%) duplicate samples (direct mutations), which may contribute to their good performances on the direct mutations in S^{sym} to some extent.

Currently, it has been a challenge to predict the highly destabilizing or stabilizing mutations which may lead to severe pathological consequences. To detect the prediction ability of PMSPcnn on such mutations, the distribution of the predicted ΔΔG values from PMSPcnn (see Table S2 for more details) was calculated with the result shown in Figure 4D. Figure 4D also gives the corresponding results from the experimental data and ThermoNet for comparison. From Figure 4D, compared with the distribution [−3.10, 2.85] kcal/mol obtained by ThermoNet, the one [−7.03, 6.28] kcal/mol obtained by PMSPcnn is more similar to that [−7.50, 7.50] kcal/mol from experimental ΔΔG values. The results suggest that PMSPcnn has a good predictive power for the highly destabilizing or stabilizing mutations, which we think mainly attributes to the application of the RScv strategy in model training.

Analysis on the anti-symmetry property

As for the anti-symmetry property of ΔΔG, there are only six methods that apply it to construct a balanced training set. For evaluating the anti-symmetry property of the results from the predictors, we calculated their PCC ($r_{dir\text{-}rev}$) between the predicted values for direct mutations and those for reverse mutations, and the prediction bias $\langle\delta\rangle$ (see STAR Methods), as shown in Table 1. From Table 1, generally the methods with anti-symmetry operation not applied have a biased performance with much worse results for the reverse mutations than for the direct ones, while the methods with anti-symmetry operation applied have a similar performance on both direct and reverse mutations. PMSPcnn has an excellent unbiased performance ($r_{dir\text{-}rev}$ = 0.96 and δ = 0.01 kcal/mol, shown in Figure 4C). It should be noted that the two metrics ($r_{dir\text{-}rev}$ and δ) can only be used for evaluating the model's balancing power, but not for the prediction power of ΔΔG values.

Besides our model PMSPcnn, five predictors consider the anti-symmetry property, which are ThermoNet,[15] ACDC-NN-Seq,[28] DDGun3D,[8] DDGun,[8] and SDM.[7] ThermoNet uses 3D-convolutional neural network, but only considers seven residue biophysical properties. ACDC-NN-Seq is an Antisymmetric Convolutional Differential Concatenated Neural Network, and only uses the evolutionary information derived from multiple sequence alignments. The DDGun3D, DDGun and SDM are non-machine learning methods, and generally non-machine learning methods have a non-ideal generalization ability in prediction. There are several main reasons, we think, for better performance of PMSPcnn than the others: (1) introduce of new topological features based on PH; (2) selection of CNN as our model architecture; and (3) utilization of the scheme of RScv to train our model. All of these measures are highly effective in improving our model for protein stability change prediction.

Performance of PMSPcnn on mutations between different types of amino acids

We want to know whether there are some rules in protein stability changes caused by mutations between different types of amino acids, and how the performance of PMSPcnn is on the mutations. To detect it, the 20 amino acids denoted by one-letter symbol were categorized into four types: hydrophobic (HYD, including AVILMFYW), polar (POL, including STNQ), charged (CHA, including RHKDE), and special residues (SPC, including CGP).[29] Figure 5 shows the PCC and the root-mean-square error (RMSE) for the predictions of PMSPcnn on the 4 × 4 types of mutations from S^{sym} dataset. From Figure 5, PMSPcnn shows a better prediction performance for the mutations between the residue types with similar physicochemical properties (with small |ΔΔG|) than for those with large differences in physicochemical properties (with large |ΔΔG|). Taking the mutations between POL and SPC types (with similar properties) and between CHA and HYD types (with large differences) for example, the PCC corresponding to them are above 0.95 and below 0.60, respectively. The percentage of the mutations with |ΔΔG| > 3.0 kcal/mol in the latter is three times that in the former, which may be one reason for the not good predictions on the latter. Additionally, the mutations between the two residue types with large differences in physicochemical properties often cause a significant conformational change, and however PMSPcnn does not consider it, which may be another reason for the not good predictions on the latter.

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

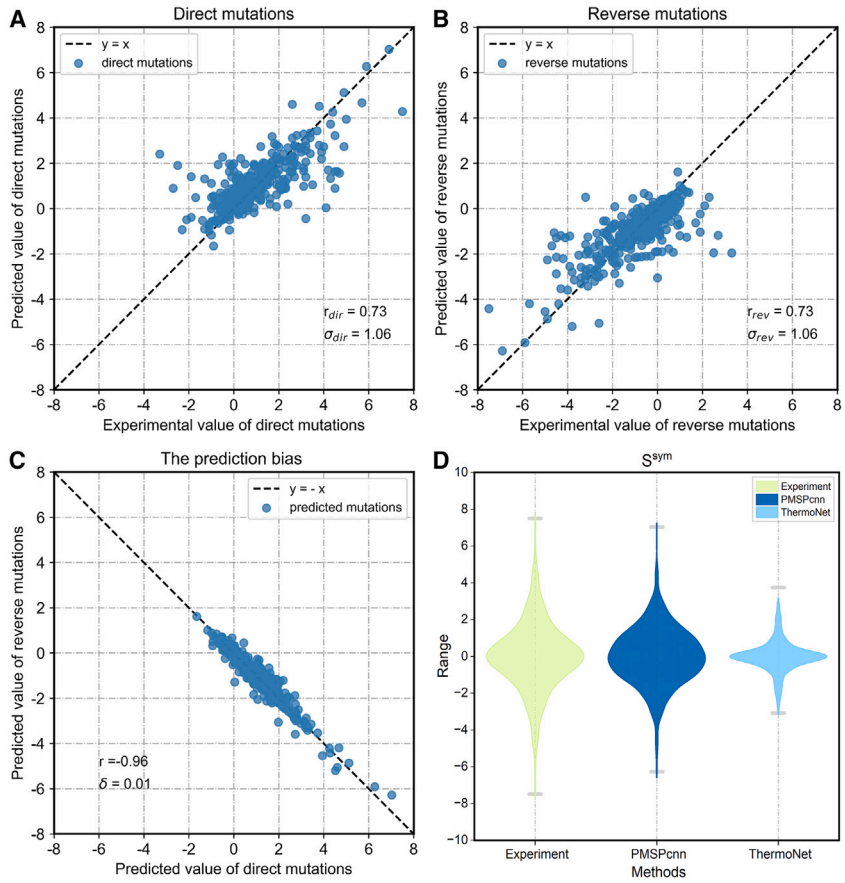


Figure 4. The performances of PMSPcnn on S^{sym} test set

(A and B) The scatterplots of the predicted samples of direct and reverse mutations (n = 342), respectively, both with r = 0.73 and σ = 1.06 kcal/mol.

(C) Bias of PMSPcnn is $r_{dir\text{-}rev}$ = 0.96 and $\langle\delta\rangle$ = 0.01 (n = 342).

(D) Distributions (n = 684) of the predicted ΔΔG, with those from experimental data and ThermoNet also shown for comparison. And the maximum and minimum values are represented by gray lines.

Performances of PMSPcnn on p53 and myoglobin test sets

Besides S^{sym}, the two independent test sets p53 and myoglobin (see STAR Methods) were used to test the ability of PMSPcnn, with the results shown in Table 2. Table 2 also lists the corresponding results from ThermoNet for comparison.

For the test set p53, from Table 2, PMSPcnn gives the r and σ of 0.67 and 1.39 kcal/mol for both direct and reverse mutations, respectively (with 3 outliers in all predicted samples shown in Figure 6A), much better than the corresponding results (0.45 and 2.01 kcal/mol for direct mutations, and 0.56 and 1.92 kcal/mol for reverse mutations) from ThermoNet (with 14 outliers shown in Figure 6C). The PCC has an improvement of 49% and 20% on direct and reverse mutations, respectively. From the distributions of the predicted values (Figure 6E), compared with ThermoNet, PMSPcnn gives the result much closer to that of experimental ΔΔG values. For the test set myoglobin, PMSPcnn gives the r and σ of 0.67 and 0.87 kcal/mol for direct mutations, respectively, and of 0.58 and 0.99 kcal/mol for reverse mutations, respectively (with 2 outliers in all predicted ΔΔG values shown in Figure 6B), much better than the corresponding results (0.38 and 1.16 kcal/mol for direct mutations, and 0.37 and 1.18 kcal/mol for reverse mutations) from ThermoNet (with 7 outliers shown in Figure 6D). The PCC has an elevation of 0.29 and 0.21, respectively. It is worth noting that there are two cases: the direct mutation (L39N) and its reverse mutation (N39L) between leucine (hydrophobic) and asparagine (polar) at residue 39 of myoglobin protein, on which both PMSPcnn and ThermoNet do not show good predictions (Figures 6B and 6D). The large difference of leucine and asparagine in physicochemical properties may cause a large conformational change that has a big effect on the protein stability change. However, the two methods consider the conformational change inadequately, which is the possible reason for the not good predictions. In addition, from the distributions of the predicted values (see Figure 6F), compared with ThermoNet, PMSPcnn gives the result much closer to that of experimental ΔΔG values. In summary, PMSPcnn performs much better than ThermoNet on p53 and myoglobin test sets, suggesting PMSPcnn is a promising predictor for the prediction of protein stability changes. Predicted ΔΔG values of PMSPcnn on the test sets p53 and myoglobin are shown in Table S3 and Table S4.

Performance of PMSPcnn on membrane proteins

The aforementioned text shows our model PMSPcnn has a good performance on S^{sym}, p53, and myoglobin test sets, and we want to know how PMSPcnn performs on membrane proteins. Thus, we tested PMSPcnn on the membrane protein dataset M223 (see STAR Methods), with the results listed in Table 3. Table 3 also shows the results from the existing eleven methods for comparison. It should be pointed out that all these methods are trained on the non-membrane proteins. From Table 3, all the methods present a non-ideal performance. PMSPcnn gives the r and σ of 0.43 and 1.24 kcal/mol, generally much better than the corresponding results from all the other methods, and with an improvement of 59% in terms of r compared with the second best method EASE-MM (r = 0.27 and σ = 1.44 kcal/mol). However, compared with the predictions on S^{sym}, p53 and myoglobin test sets, the performance of PMSPcnn has a big drop on dataset M223. How about the results if a new model is trained via the same process with PMSPcnn on the membrane protein set M223? We constructed such a model PMSPcnn_m (see Method S1), with the results also listed in Table 3. The corresponding results r and σ are 0.59 and 1.06 kcal/mol, which have a significant improvement but are still much worse than those from PMSPcnn (with r and σ of 0.72 and 1.09 kcal/mol, respectively) on the training set Q6842.

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

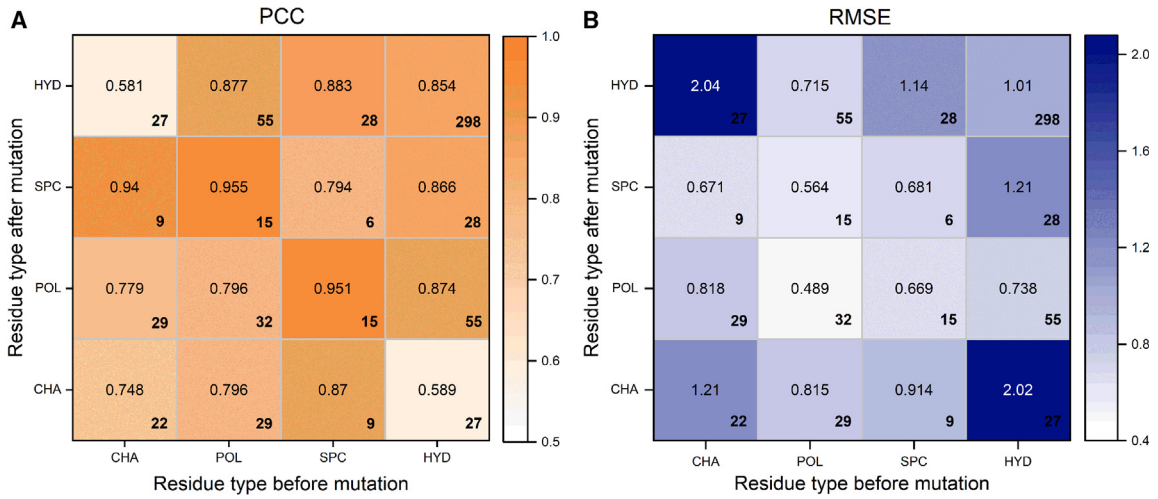


Figure 5. Performance of PMSPcnn on mutations between different types of amino acids

There are two numbers in each square, with the one in the center representing the PCC (A) and RMSE ((B), kcal/mol), and the one in the lower right representing the number of samples.

The aforementioned result implies that there exists evident difference in protein thermodynamic stability mechanisms between the membrane proteins and non-membrane proteins. The former is embedded in the membrane and interact with the hydrophobic environment, while the latter are often in solvent and interact with polar water molecules. Additionally, the membrane proteins tend to be larger and more complex in structure than the non-membrane proteins. The deeper explorations are needed to reveal the intricate differences between them, and also the new features (involved in structure and surrounding environment) applicable to membrane proteins need to be explored to improve the prediction of membrane protein stability changes upon mutations.

DISCUSSION

Accurately predicting the protein stability changes caused by mutations is one of the most attractive and challenging problems. Here, we construct an unbiased CNN-based method PMSPcnn for predicting protein stability changes (ΔΔG) caused by single point mutations. A balanced dataset is constructed using the anti-symmetry property. Then we exploit the useful topological descriptors obtained from PH analysis for characterizing protein structures, which make a significant contribution to the prediction due to the PH's ability to delicately characterize the topological relationship among atoms in protein at different scales. Additionally, we propose an effective scheme named RScv to train the model, which makes the model well learn the special features in the samples with different ΔΔG value ranges. The test results of PMSPcnn on multiple datasets, including S^{sym}, p53, myoglobin and M223 indicate that our model has the state-of-the-art performance. PMSPcnn outperforms the currently existing methods on the widely used dataset S^{sym} with PCC and RMSE between experimental and predicted ΔΔG for direct and reverse mutations both being 0.73 and 1.06 kcal/mol, respectively. For the p53 dataset, PMSPcnn achieves a PCC of 0.67 and RMSE of 1.39 kcal/mol for both direct and reverse mutations, and for the myoglobin dataset, PMSPcnn obtains a PCC of 0.67/0.58 and RMSE of 0.87/0.99 kcal/mol for direct/reverse mutations. In addition, on membrane proteins that are very different from non-membrane proteins, PMSPcnn also outperforms the existing methods. In conclusion, PMSPcnn is a promising method, with excellent abilities in robustness and generalization in protein stability change prediction.

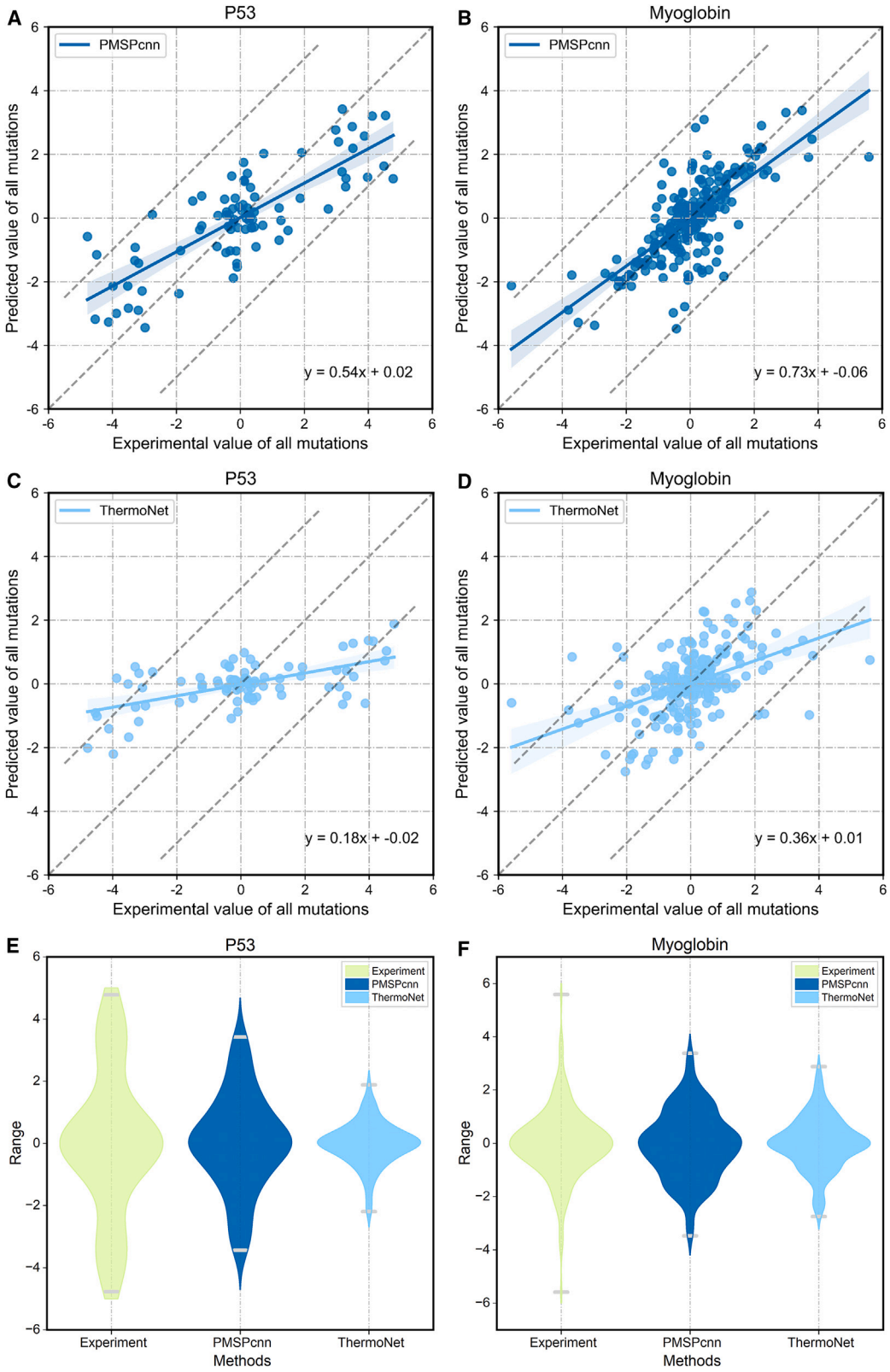
The proposed PMSPcnn solved several challenges to some extent in the development of computational methods for protein stability change prediction. First, due to the consideration of the anti-symmetry property, PMSPcnn improves the model's capacity of predicting stabilizing mutations and reduces the model's prediction bias. Second, by applying our proposed RScv, PMSPcnn enhances the model's ability to predict the mutations with extreme ΔΔG values. Third, AlphaFold2 is used to construct protein 3D structures because of the huge gap between the numbers of protein sequences and structures, so that PMSPcnn can be used on the proteins with structure unknown, which makes it more practical in real-world study. In the future, we can collect the experimental data of stability changes of the proteins without experimental structures to enlarge the training data, and then construct the

Table 2. Performance comparison of PMSPcnn with ThermoNet on test sets p53 and myoglobin

	P53		Myoglobin	
Metrics	PMSPcnn	ThermoNet	PMSPcnn	ThermoNet
r_{dir}	0.67	0.45	0.67	0.38
σ_{dir}	1.39	2.01	0.87	1.16
r_{rev}	0.67	0.56	0.58	0.37
σ_{rev}	1.39	1.92	0.99	1.18
r_{all}	0.81	0.59	0.74	0.47
σ_{all}	1.39	1.96	0.94	1.21

r_{dir}/r_{rev} (Equation 1) and $\sigma_{dir}/\sigma_{rev}$ (Equation 2) are the PCC and RMSE (kcal/mol) between the predicted and experimental ΔΔG values for the direct/reverse mutations, respectively. r_{all} and σ_{all} are the PCC and RMSE (kcal/mol) between the predicted and experimental ΔΔG values for all the mutations in the test sets p53 (n = 42) and myoglobin (n = 134).

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016



(legend on next page)

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

prediction model utilizing the predicted structures by AlphaFold2, which, we think, can further improve the model's prediction power for mutation-induced protein stability changes. Overall, our work can provide a valuable reference for mutation-induced protein stability changes, guiding researchers to choose certain residues to mutate for obtaining stable protein structures.

For the mutation-induced conformational change, considering it reasonably is very important for the accurate prediction. Besides the Scap software used in this work, OPUS-Mut,[30] Rosetta[31] and FoldX[6] can also be used to predict the mutation-induced structural changes. OPUS-Mut is a modified version derived from OPUS-Rota4 that is a deep learning-based predictor for protein side-chain modeling. Rosetta optimizes the mutated conformation using a physics-based energy function that takes into account various physical and chemical properties of protein atoms. FoldX optimizes the mutated conformation with an empirical energy function derived from experimental data and statistical analysis of known protein structures. In the future, these tools can be tried to obtain the mutant conformation for better prediction of protein stability change caused by residue mutation.

As for the membrane protein stability prediction, there is still no accurate and reliable method to predict it, which attributes to some factors such as inadequate data and the complexity of membrane proteins. The folding pathway of membrane protein is considerably more intricate than that of non-membrane protein. It involves a series of coordinated interactions with various cellular components, such as translocons and chaperons. These interactions play crucial roles in guiding membrane protein folding process, ensuring their proper conformation and maintaining their stability. The research into the complexity of membrane protein folding pathways will enhance our understanding for their thermodynamic stability. One feasible direction to improve membrane protein stability change prediction is to incorporate the additional factors that take into account their specific structural characteristics and their surrounding environment, such as the number of transmembrane helices, the position and hydrophobicity of residues in the membrane protein, the lipid composition, pH, and ion concentration. It needs to note that these properties vary across different membrane proteins, and therefore the appropriate features need to be designed based on the type and location of the membrane proteins within the cell. Although PMSPcnn shows improved performance compared to many existing methods on membrane protein stability prediction, there is still considerable room for further enhancement. Thus, this is also an important direction we need continue to study in the follow-up work.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

Table 3. Comparison of PMSPcnn with the existing methods on M223 dataset

Methods	r	σ
PMSPcnn_m	0.59	1.06
PMSPcnn	0.43	1.24
EASE-MM	0.27	1.44
Rosetta	0.26	1.63
FoldX	0.26	2.56
PROVEAN	0.26	4.23
PPSC (M8)	0.23	1.61
I-Mutant 3.0	0.22	1.63
ELASPIC	0.20	1.17
mSCM	0.18	1.51
DUET	0.17	1.59
PPSC (M47)	0.09	1.48
SDM	0.09	2.40

Results from the existing methods are taken from the ref. [33].
r and σ are the PCC and RMSE (kcal/mol) between the predicted and experimental $\Delta\Delta G$ values of the mutations in M223 dataset (n = 223), respectively.

- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data processing
  - Protein stability change ($\Delta\Delta G$) and its anti-symmetry
  - Feature extraction
  - Feature integration
  - Regression stratification cross-validation
  - Performance metrics
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2024.02.016.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China [32271294 and 31971180].

AUTHOR CONTRIBUTIONS

X.S. and C.L. designed the research. X.S. wrote the program of PMSPcnn model. X.S., S.Y., Z.W., J.S., F.H., F.C., and C.L. participated in the construction of datasets and performed data analyses. C.L. validated the results. X.S. and C.L. wrote the manuscript. All authors have given approval to the final version of the manuscript.

Figure 6. The performances of PMSPcnn and ThermoNet on p53 and myoglobin test sets
(A and B) For PMSPcnn's performance on p53 (n = 84) and myoglobin (n = 268) sets, respectively.
(C and D) For ThermoNet's performance on the two sets, respectively.
(E and F) For the distributions of the predicted $\Delta\Delta G$ from PMSPcnn on p53 and myoglobin sets, respectively, with those from the experimental data and ThermoNet also shown for comparison. The maximum and minimum values are represented by gray lines. The outliers are defined as the points that deviate from the sample space within the two dotted lines.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 11, 2023
Revised: December 19, 2023
Accepted: February 22, 2024
Published: March 19, 2024

REFERENCES

1. 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. https://doi.org/10.1016/j.jmb.2013.07.014.
2. Banerjee, A., and Mitra, P. (2020). Estimating the Effect of Single-Point Mutations on Protein Thermodynamic Stability and Analyzing the Mutation Landscape of the p53 Protein. J. Chem. Inf. Model. *60*, 3315–3323. https://doi.org/10.1021/acs.jcim.0c00256.
3. Sanavia, T., Birolo, G., Montanucci, L., Turina, P., Capriotti, E., and Fariselli, P. (2020). Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput Struct Biotec *18*, 1968–1979. https://doi.org/10.1016/j.csbj.2020.07.011.
4. Pan, Q., Nguyen, T.B., Ascher, D.B., and Pires, D.E.V. (2022). Systematic evaluation of computational tools to predict the effects of mutations on protein stability in the absence of experimental structures. Brief. Bioinform. *23*, bbac025. https://doi.org/10.1093/bib/bbac025.
5. Zwanzig, R. (1954). High-temperature equation of state by a perturbation method. J. Chem. Phys. *8*, 1420–1426. https://doi.org/10.1063/1.1740409.
6. 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. https://doi.org/10.1016/S0022-2836(02)00442-4.
7. Pandurangan, A.P., Ochoa-Montaño, B., Ascher, D.B., and Blundell, T.L. (2017). SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res. *45*, W229–W235. https://doi.org/10.1093/nar/gkx439.
8. Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N., and Fariselli, P. (2019). DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinf. *20*, 335. https://doi.org/10.1186/s12859-019-2923-1.
9. Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. (2015). MAESTRO–multi agent stability prediction upon point mutations. BMC Bioinf. *16*, 116. https://doi.org/10.1186/s12859-015-0548-6.
10. Pires, D.E.V., Ascher, D.B., and Blundell, T.L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics *30*, 335–342. https://doi.org/10.1093/bioinformatics/btt691.
11. Pires, D.E.V., Ascher, D.B., and Blundell, T.L. (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. *42*, W314–W319. https://doi.org/10.1093/nar/gku411.
12. Capriotti, E., Fariselli, P., and Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics *20* (*Suppl 1*), i63–i68. https://doi.org/10.1093/bioinformatics/bth928.
13. Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. *33*, W306–W310. https://doi.org/10.1093/nar/gki375.
14. Chen, C.W., Lin, J., and Chu, Y.W. (2013). iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinf. *14* (*Suppl 2*), S5. https://doi.org/10.1186/1471-2105-14-S2-S5.
15. Li, B., Yang, Y.T., Capra, J.A., and Gerstein, M.B. (2020). Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLoS Comput. Biol. *16*, e1008291. https://doi.org/10.1371/journal.pcbi.1008291.
16. Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability. Curr Opin Struc Biol *19*, 596–604. https://doi.org/10.1016/j.sbi.2009.08.003.
17. Thiltgen, G., and Goldstein, R.A. (2012). Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One *7*, e46084. https://doi.org/10.1371/journal.pone.0046084.
18. Zomorodian, A., and Carlsson, G. (2005). Computing persistent homology. Discrete Comput. Geom. *33*, 249–274. https://doi.org/10.1007/s00454-004-1146-y.
19. Cang, Z., and Wei, G.W. (2018). Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. Int J Numer Meth Bio *34*, e2914. https://doi.org/10.1002/cnm.2914.
20. Wang, M., Cang, Z., and Wei, G.W. (2020). A topology-based network tree for the prediction of protein-protein binding affinity changes following mutation. Nat. Mach. Intell. *2*, 116–123. https://doi.org/10.1038/s42256-020-0149-6.
21. Xia, K., and Wei, G.W. (2014). Persistent homology analysis of protein structure, flexibility, and folding. Int J Numer Meth Bio *30*, 814–844. https://doi.org/10.1002/cnm.2655.
22. Li, G., Panday, S.K., and Alexov, E. (2021). SAAFEC-SEQ: A Sequence-Based Method for Predicting the Effect of Single Point Mutations on Protein Thermodynamic Stability. Int. J. Mol. Sci. *22*, 606. https://doi.org/10.3390/ijms22020606.
23. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596. https://doi.org/10.1038/s41586-021-03828-1.
24. Yuan, Q., Chen, S., Rao, J., Zheng, S., Zhao, H., and Yang, Y. (2022). AlphaFold2-aware protein–DNA binding site prediction using graph transformer. Brief. Bioinform. *23*, bbab564. https://doi.org/10.1093/bib/bbab564.
25. Yamaguchi, S., Nakashima, H., Moriwaki, Y., Terada, T., and Shimizu, K. (2022). Prediction of protein mononucleotide binding sites using AlphaFold2 and machine learning. Comput. Biol. Chem. *100*, 107744. https://doi.org/10.1016/j.compbiolchem.2022.107744.
26. Cheng, J., Randall, A., and Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. Proteins *62*, 1125–1132. https://doi.org/10.1002/prot.20810.
27. Quan, L., Lv, Q., and Zhang, Y. (2016). STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics *32*, 2936–2946. https://doi.org/10.1093/bioinformatics/btw361.
28. Pancotti, C., Benevenuta, S., Repetto, V., Birolo, G., Capriotti, E., Sanavia, T., and Fariselli, P. (2021). A Deep-Learning Sequence-Based Method to Predict Protein Stability Changes Upon Genetic Variations. Genes-Basel *12*, 911. https://doi.org/10.3390/genes12060911.
29. Cang, Z., and Wei, G.W. (2017). TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS Comput. Biol. *13*, e1005690. https://doi.org/10.1371/journal.pcbi.1005690.
30. Xu, G., Wang, Q., and Ma, J. (2023). OPUS-Mut: Studying the Effect of Protein Mutation through Side-Chain Modeling. J. Chem. Theory Comput. *19*, 1629–1640. https://doi.org/10.1021/acs.jctc.2c00847.
31. Kellogg, E.H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins *79*, 830–838. https://doi.org/10.1002/prot.22921.
32. 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. https://doi.org/10.1093/bioinformatics/bty348.
33. Kroncke, B.M., Duran, A.M., Mendenhall, J.L., Meiler, J., Blume, J.D., and Sanders, C.R. (2016). Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. Biochemistry-Us *55*, 5002–5009. https://doi.org/10.1021/acs.biochem.6b00537.
34. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021).

Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

35. Xiang, Z., and Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. J. Mol. Biol. *311*, 421–430. https://doi.org/10.1006/jmbi.2001.4865.
36. Singh, J., Paliwal, K., Litfin, T., Singh, J., and Zhou, Y. (2022). Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. Sci. Rep. *12*, 7607. https://doi.org/10.1038/s41598-022-11684-w.
37. Mihel, J., Sikić, M., Tomić, S., Jeren, B., and Vlahovicek, K. (2008). PSAIA - protein structure and interaction analyzer. BMC Struct. Biol. *8*, 21. https://doi.org/10.1186/1472-6807-8-21.
38. Abadi, M.I.N., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A System for Large-Scale Machine Learning. In OSDI'16, secondary_author, pp. 265–283. USA. https://doi.org/10.48550/arXiv.1605.08695.
39. Kumar, M.D.S., Bava, K.A., Gromiha, M.M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006). ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. *34*, D204–D206. https://doi.org/10.1093/nar/gkj103.
40. Bateman, A., Martin, M., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., et al. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.
41. Bullock, A.N., Henckel, J., and Fersht, A.R. (2000). Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. Oncogene *19*, 1245–1256. https://doi.org/10.1038/sj.onc.1203434.
42. Ordway, G.A., and Garry, D.J. (2004). Myoglobin: an essential hemoprotein in striated muscle. J. Exp. Biol. *207*, 3441–3446. https://doi.org/10.1242/jeb.01172.
43. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242. https://doi.org/10.1093/nar/28.1.235.
44. Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., Capriotti, E., and Fariselli, P. (2022). Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. Brief. Bioinform. *23*, bbab555. https://doi.org/10.1093/bib/bbab555.
45. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. *36*, D202–D205. https://doi.org/10.1093/nar/gkm998.
46. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. https://doi.org/10.1002/bip.360221211.
47. Ligeti, B., Vera, R., Juhász, J., and Pongor, S. (2017). CX, DPX, and PCW: Web Servers for the Visualization of Interior and Protruding Regions of Protein Structures in 3D and 1D. Methods Mol. Biol. *1484*, 301–309. https://doi.org/10.1007/978-1-4939-6406-2_20.
48. Vlahovicek, K., Pintar, A., Parthasarathi, L., Carugo, O., and Pongor, S. (2005). CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3D structures. Nucleic Acids Res. *33*, W252–W254. https://doi.org/10.1093/nar/gki362.
49. Cheng, C.W., Su, E.C.Y., Hwang, J.K., Sung, T.Y., and Hsu, W.L. (2008). Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinf. *9* (*Suppl 12*), S6. https://doi.org/10.1186/1471-2105-9-S12-S6.
50. Liu, Y., Gong, W., Yang, Z., and Li, C. (2021). SNB-PSSM: A spatial neighbor-based PSSM used for protein-RNA binding site prediction. J. Mol. Recognit. *34*, e2887. https://doi.org/10.1002/jmr.2887.
51. Liu, Y., Gong, W., Zhao, Y., Deng, X., Zhang, S., and Li, C. (2021). aPRBind: protein-RNA interface prediction by combining sequence and I-TASSER model-based structural features learned with convolutional neural networks. Bioinformatics *37*, 937–942. https://doi.org/10.1093/bioinformatics/btaa747.
52. Cang, Z., Mu, L., and Wei, G.W. (2018). Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. PLoS Comput. Biol. *14*, e1005929. https://doi.org/10.1371/journal.pcbi.1005929.
53. Maria, C., Boissonnat, J., Glisse, M., and Yvinec, M. (2014). In The Gudhi Library: Simplicial Complexes and Persistent Homology, H. Hong and C. Yap, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 167–174.
54. Anand, D.V., Meng, Z., Xia, K., and Mu, Y. (2020). Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis. Sci. Rep. *10*, 9685. https://doi.org/10.1038/s41598-020-66710-6.
55. Máté, G., Hofmann, A., Wenzel, N., and Heermann, D.W. (2014). A topological similarity measure for proteins. Biochim. Biophys. Acta *1838*, 1180–1190. https://doi.org/10.1016/j.bbamem.2013.08.019.
56. Adams, H., Tausz, A., and Vejdemo-Johansson, M. (2014). javaPlex: A Research Software Package for Persistent (Co)Homology. In Mathematical Software – ICMS 2014, H. Hong and C. Yap, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 129–136. https://doi.org/10.1007/978-3-662-44199-2_23.
57. Fasy, B.T., Kim, J., Lecci, F., and Maria, C.E.M. (2014). Introduction to the R Package TDA. Preprint at Arxiv. https://doi.org/10.48550/arXiv.1411.1830.
58. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In KDD '16 (New York, NY, USA: Association for Computing Machinery), pp. 785–794. https://doi.org/10.1145/2939672.2939785.
59. Fang, J. (2020). A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Brief. Bioinform. *21*, 1285–1292. https://doi.org/10.1093/bib/bbz07.

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Analyzed data	This paper	https://github.com/ChunhuaLab/PMSPcnn/tree/main/datasets
Software and algorithms		
PMSPcnn	This paper	https://github.com/ChunhuaLab/PMSPcnn
AlphaFold2	Jumper et al.[34]	https://github.com/deepmind/alphafold
Scap	Xiang et al.[35]	https://honig.c2b2.columbia.edu
SPOT_1D_LM	Singh et al.[36]	https://github.com/jas-preet/SPOT-1D-LM.git
PSAIA	Mihel et al.[37]	https://sourceforge.net/projects/psaia
TensorFlow	Abadi et al.[38]	https://github.com/tensorflow/tensorflow
PyMol	Schrodinger LLC	https://pymol.org/2/
Matlab_R2021a_Linux	MathWorks	https://www.mathworks.com/
Python version 3.8	Python Software Foundation	https://www.python.org
R	N/A	https://www.r-project.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Chunhua Li (chunhuali@bjut.edu.cn)

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Mutation data used in this work is from the Protherm database[39] and the protein sequences are from the publicly database Uniprot.[40] All data reported in this paper is available at (https://github.com/ChunhuaLab/PMSPcnn).
- The PMSPcnn source code is written in python-3.8 of the anaconda. All original code has been deposited at GitHub and is available at (https://github.com/ChunhuaLab/PMSPcnn).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data processing

The training set is from the most widely used dataset Q3421[27] extracted from Protherm database, which contains 3421 distinct single point mutations from 150 proteins, all validated by experiments. Three datasets including S^{sym}, p53 and myoglobin are used as test datasets. S^{sym} is a widely used blind test set consisting of 684 mutations from 15 different proteins.[32] The p53 dataset contains 42 mutations within p53 protein's DNA binding domain,[41] and p53 protein plays a key role in the pathogenesis of human cancers. Myoglobin dataset contains 134 mutations from sperm-whale myoglobin,[42] and myoglobin is a globular protein that helps to store and transport oxygen. Additionally, M223 dataset containing 223 mutations from 7 membrane proteins[33] is also adopted to assess PMSPcnn's performance on membrane protein mutations. Table S5 summarizes the detailed information on these datasets.

For the training set, the redundancy removement and antisymmetric operation were performed. First, the duplicate data with the three test sets (S^{sym}, p53 and myoglobin) were removed from the training set Q3421, and the remaining contain 3211 mutations covering 147 different proteins. Then, a balanced training set was built based on the anti-symmetry property of $\Delta\Delta G$, which consists of 6422 mutations denoted as Q6422 set. Figure S3 shows the histogram distributions of $\Delta\Delta G$ values in the five datasets used in this work, with S^{sym} and Q6422 being the balanced datasets and the others imbalanced ones. The mutation locations in the five datasets are provided based on protein structures from Protein Data Bank.[43] All the mutation locations are mapped onto the corresponding sequences from UniProt database[40] for easier feature extraction (see Methods S2 for more details).

Please cite this article in press as: Sun et al., PMSPcnn: Predicting protein stability changes upon single point mutations with convolutional neural network, Structure (2024), https://doi.org/10.1016/j.str.2024.02.016

Protein stability change (ΔΔG) and its anti-symmetry

Under the assumption that the protein folding process is a reversible, two-state transition—and thus that the protein does not precipitate or aggregate, the thermodynamic stability of a protein can be measured by its folding free energy ΔG.[32] Thus the protein stability change (ΔΔG) caused by a mutation can be determined by the difference in ΔG between its mutant and wild-type proteins: $\Delta\Delta G = \Delta G_{wild} - \Delta G_{mut}$. A positive ΔΔG value indicates that the mutation is stabilizing; otherwise the mutation is destabilizing. For ΔΔG, there is an important anti-symmetry between the direct and reverse mutations.[44] Take a protein for example, its Gibbs free energy change of amino acid A mutated to B is $\Delta\Delta G_{A\to B} = \Delta G_A - \Delta G_B$, and the corresponding change of B mutated to A is $\Delta\Delta G_{B\to A} = \Delta G_B - \Delta G_A$, which is to say $\Delta\Delta G_{A\to B} = -\Delta\Delta G_{B\to A}$.

Feature extraction

Before feature extraction, wild-type protein structures are constructed using AlphaFold2,[34] and their corresponding mutant structures are generated with Scap utility Jackal package[35] based on their predicted wild-type structures. For each mutation sample, we consider its features in both wild-type and mutant proteins, as well as their corresponding differences. A total of 1140 features involved in the following types are extracted from protein sequences and structures. Topology characteristics are calculated from the atoms in the area within 12 Å of the mutated residue (Cα atom), and the other features are extracted from the mutated site.

Physicochemical characteristics: Amino Acid index database (AAindex) is a database of numerical indices representing physicochemical properties of amino acids.[45] A total of 8 physicochemical properties for each amino acid are obtained from the database, including number of atoms, hydrophobicity, hydrophilicity, propensity, isoelectric point, mass, volume and accessible surface area (ASA). Additionally, the electrostatic properties (positive, negative and non-charged) of amino acids are also considered (denoted as 1, -1 and 0, respectively). Here, a total of 27 features are constructed.

Secondary structural features: SPOT-1D-LM is used to predict protein secondary structural features based on sequences, which is an ensemble method based on convolution and Long Short-Term Memory (LSTM).[36] It achieves high-accuracy prediction of both secondary and tertiary structural properties without sequence alignment. Here, a total of 57 features are constructed from SPOT-1D-LM, containing the three states of protein secondary structures (helix (H), strand (E) and coil (C)), eight states of protein secondary structures (3_{10}-helix (G), α-helix (H), π-helix (I), β-strand (E), bridge (B), turn (T), bend (S) and others (C)), relative solvent accessible (RSA), protein backbone angles (ψ, φ, θ, and τ), half-sphere exposures (HSE), and contact number (CN). Here it should be pointed out that we do not use DSSP[46] to identify the secondary structure types from the predicted structures by AlphaFold2 since the performances of SPOT-1D-LM on protein sequences and DSSP on the predicted structures are comparable, and the former can provide highly accurate prediction without homologous sequences (See Methods S3 for more details).

Depth and protrusion indices: The depth index (DPX) and protrusion index (CX) for an atom describe the local concavity and convexity of a protein.[47] They are useful descriptors to characterize the geometry shape at the mutation site.[48] We use PSAIA software[37] to calculate the indexes for a mutation residue including the means of DPXs and CXs of all atoms of the residue and their standard deviations, and the means of DPXs and CXs of side-chain atoms and their standard deviations, which leads to a total of 24 features.

PSSM and SNB-PSSM based evolutionary information: The position specific scoring matrix (PSSM) is generated by sequence similarity alignment, which contains the conserved information of amino acids.[49] Considering that the protein stability change upon residue mutation is related to its adjacent residues in tertiary structure, the spatial neighbor-based PSSM (SNB-PSSM) is adopted, which was proposed in our previous study.[50] SNB-PSSM uses a spatial neighbor-based smooth processing and a window scheme to encode the evolutionary information. Here, the window size is set to 1, and thus the evolutionary score of a target residue is the average value of the evolutionary scores from the standard PSSM over the residues whose Cα atoms are within 7.5 Å from that of the target residue.[51] We extracted the evolutionary score of the mutated site, resulting 3 features from PSSM and SNB-PSSM respectively.

PH-based topology feature calculation: Persistent homology (PH) is a mathematical technique used in the field of algebraic topology to analyze the topological structure of materials science data. Often, the topological persistence is studied in a growing sequence of simplicial complexes. A simplicial complex is a mathematical structure composed of simplexes (such as points, lines, triangles, etc.) and higher dimensional analogs. The persistent homology used for topological data analysis can reduce the geometric complexity, but preserve the critical information directly related to protein stability changes (ΔΔG). The topological features are calculated from the topological invariants which are characterized by the independent components, rings and cavities whose numbers are called Betti-0, Betti-1 and Betti-2, respectively. Topology feature calculation is mainly divided into two steps: simplicial complex construction, and topological feature vector generation. For the first step, the relationship of atoms to each other in protein topological space is all that we care about. The atoms within 12 Å (see Methods S4 and Figure S4 for more details) of a mutated residue (Cα atom) are built into a point cloud, with only three types of heavy atoms C, N and O considered. Topological invariants are quantified by their persistence through a filtration process,[52] during which as the spheres centered at the points in a point cloud grow larger, the sphere centers are connected with an edge when the two spheres mutually touch each other. Thus, a series of nested simplicial complexes[53] are generated to be used for characterizing the first and higher order interactions. Each event of connecting any two previously disjoint components and the change of component numbers can be tracked, with the procedure illustrated in Figure 1B for a particular set of points. For the second step, the conventional output from persistent homology is a barcode graph. Each instance of component, ring and cavity is represented by a bar.[54] The start point, end point and length of a bar correspond to the birth, death and "lifetime" of the instance. Longer bars correspond to more robust topological structures.[55] The bars characterize how the topology of the object changes on different scales. In order to apply the representations as input to machine learning models, they need to

be transformed into vectors. The binning method is adopted to discretize the persistent barcode into a feature vector. In the barcode graph, the horizontal axis represents the filtration time, which can be separated into equal-length bins. The statistical measures of each bin, such as maximum, minimum, mean, sum, and others can be calculated as feature vectors. The upper bound of the filtration parameter is set to 12 Å which is used to build point clouds, with the whole filtration range divided into 12 bins of 1.0 Å. For Betti-0 barcodes, we calculate the numbers of persistence intervals and death events in each bin, giving rise to a total of 648 features. For Betti-1 and Betti-2 barcodes, we calculate the sum, maximum, minimum, mean and standard derivation of bar lengths, birth values and death values, producing a total of 378 features. The final integration results in a total of 1026 features. Betti-0 barcodes are obtained from Javaplex,[56] and Betti-1 and Betti-2 barcodes are generated from TDA package[57] in R. For better understanding the process above, we take the protein with PDB ID 1akk for example to show the topological feature generation from the barcode graph (Figure 1C).

Feature integration

The dimension of the feature vector obtained from Betti-0 barcode is 648, which is quite sparse and has plentiful zero entries, and thus we need to reduce its dimension. The top 300 are selected from the 648 according to the importance ranking with XGBoost which is a sparsity-aware algorithm for sparse data.[58] Next, the 300 features are integrated with all the other 492 features as the inputs to the CNN, resulting in a total of 792 features.

Regression stratification cross-validation

Aiming at the difficult prediction for the mutations with extreme ΔΔG values, we propose a scheme called regression stratification cross-validation (RScv) to improve the model's prediction for such cases. First, the range of ΔΔG values in the training set are divided into eight bins (labeled as 1-8) of different intervals, with more bins produced for the denser data. Then, we assign each sample the corresponding label of the bin where its ΔΔG falls. Finally, the stratified n-fold cross validation is carried out to ensure that the samples in different bins can be sampled in each fold, so that the model can well learn the special features in different bins of samples (see Figure S5 for more details).

Performance metrics

To evaluate the performance of the predictor, the Pearson correlation coefficient (PCC) and the root-mean-square error (RMSE) between experimental and predicted ΔΔG values are calculated as r and σ respectively:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad \text{(Equation 1)}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad \text{(Equation 2)}$$

where x_i and y_i are the experimental and predicted ΔΔG values of mutation site i, respectively (x and y are their average values), and n denotes the number of mutations in the dataset.

To measure the anti-symmetry property of the predictor, the PCC ($r_{dir\text{-}rev}$) is calculated between the predicted values for direct mutations and those for reverse mutations, and the bias $\langle\delta\rangle$ is used to quantify prediction bias.[59] The formulas are as follows:

$$r_{dir-rev} = \frac{\sum_{i=1}^{n}\left(\Delta\Delta G^{i}_{dir} - \overline{\Delta\Delta G_{dir}}\right)\left(\Delta\Delta G^{i}_{rev} - \overline{\Delta\Delta G_{rev}}\right)}{\sqrt{\sum_{i=1}^{n}\left(\Delta\Delta G^{i}_{dir} - \overline{\Delta\Delta G_{dir}}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\Delta\Delta G^{i}_{rev} - \overline{\Delta\Delta G_{rev}}\right)^2}} \quad \text{(Equation 3)}$$

$$\langle\delta\rangle = \frac{\sum_{i=1}^{n}\left(\Delta\Delta G^{i}_{dir} + \Delta\Delta G^{i}_{rev}\right)}{n} \quad \text{(Equation 4)}$$

where $\Delta\Delta G^{i}_{dir}$ and $\Delta\Delta G^{i}_{rev}$ are the predicted values for direct and reverse mutations of site i, respectively ($\overline{\Delta\Delta G_{dir}}$ and $\overline{\Delta\Delta G_{rev}}$ are their average values), and n denotes the number of mutations in the dataset. A perfect antisymmetric predictor should have $r_{dir\text{-}rev} = -1$ and $\langle\delta\rangle = 0$.

QUANTIFICATION AND STATISTICAL ANALYSIS

We performed statistical analysis using in-house Python scripts. Performance metrics reported in Tables 1, 2, and 3 as well as additional results shown in Tables S2, S3 and S4. All experimental details can be found in the STAR Methods section.