Computational and Structural Biotechnology Journal 21 (2023) 354–364

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Prediction of protein stability changes upon single-point variant using 3D structure profile

Jianting Gong [a,b,1], Juexin Wang [c,b,1], Xizeng Zong [e], Zhiqiang Ma [a,d,*], Dong Xu [b,*]

[a] School of Information Science and Technology, and Institution of Computational Biology, Northeast Normal University, Changchun 130117, China
[b] Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA
[c] Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University Indianapolis, Indianapolis, IN, USA
[d] Department of Computer Science, College of Humanities & Sciences of Northeast Normal University, Changchun 130117, China
[e] School of Computer Science and Engineering, Changchun University of Technology, Changchun 130117, China

ARTICLE INFO

Article history:
Received 1 September 2022
Received in revised form 4 December 2022
Accepted 5 December 2022
Available online 08 December 2022

ABSTRACT

Identifying protein thermodynamic stability changes upon single-point variants is crucial for studying mutation-induced alterations in protein biophysics, genomic variants, and mutation-related diseases. In the last decade, various computational methods have been developed to predict the effects of single-point variants, but the prediction accuracy is still far from satisfactory for practical applications. Herein, we review approaches and tools for predicting stability changes upon the single-point variant. Most of these methods require tertiary protein structure as input to achieve reliable predictions. However, the availability of protein structures limits the immediate application of these tools. To improve the performance of a computational prediction from a protein sequence without experimental structural information, we introduce a new computational framework: MU3DSP. This method assesses the effects of single-point variants on protein thermodynamic stability based on point mutated protein 3D structure profile. Given a protein sequence with a single variant as input, MU3DSP integrates both sequence-level features and averaged features of 3D structures obtained from sequence alignment to PDB to assess the change of thermodynamic stability induced by the substitution. MU3DSP outperforms existing methods on various benchmarks, making it a reliable tool to assess both somatic and germline substitution variants and assist in protein design. MU3DSP is available as an open-source tool at https://github.com/hurraygong/MU3DSP.

1. Introduction

Protein thermodynamic stability changes are associated with heritable diseases [1–6] and drug resistance [7–11]. Nearly-one-third of non-synonymous single-nucleotide variants (nsSNVs) are deleterious to human health [12]. In fact, many disease-causing variants are single-nucleotide variants (SNV) [12–15]. For example, sickle cell anemia (OMIM [16], #603903) is caused by a single-point variant as a result of a Valine (Val) to Glutamic acid (Glu) substitution in the hemoglobin beta-subunit (HBB) [17]. Therefore, when designing and developing new compounds, a given protein's thermodynamic stability should be considered. However, performing biological experiments to detect all possible variants on a specific protein is costly and time-consuming. Computational methods can, thus, provide a powerful tool to speed up the screening of variant proteins, running as effective initial steps in such studies.

In general terms, the protein thermodynamic stability change of a single-point substitution of a protein represents a change of the Gibbs free energy difference in protein folding before and after such a single-residue change [18]. A quantified Gibbs free energy change between a protein's unfolding (G_u) and folding (G_f) states is usually represented $\Delta G = G_u - G_f$ [19]. When a residue is substituted in a protein, the original protein would be the "reference state" or "wild-type protein", whereas the substituted protein is called "variant protein" [20,21]. ΔG_W stands for the difference of Gibbs free energy between the folding and unfolding states for a wild-type protein whilst ΔG_M represents the same difference for variant protein (Fig. 1). The change of the Gibbs free energy between the wild-type and variant proteins is $\Delta\Delta G_{W\to M} = \Delta G_M - \Delta G_W$. $\Delta\Delta G$ is used for short to represent $\Delta\Delta G_{W\to M}$.

Protein stability changes driven by mutations have received great attention in the past two decades (Supplementary Figure S1),

* Corresponding authors.
E-mail address: xudong@missouri.edu (D. Xu).
[1] These authors contributed equally: Jianting Gong, Juexin Wang.

https://doi.org/10.1016/j.csbj.2022.12.008
2001-0370/

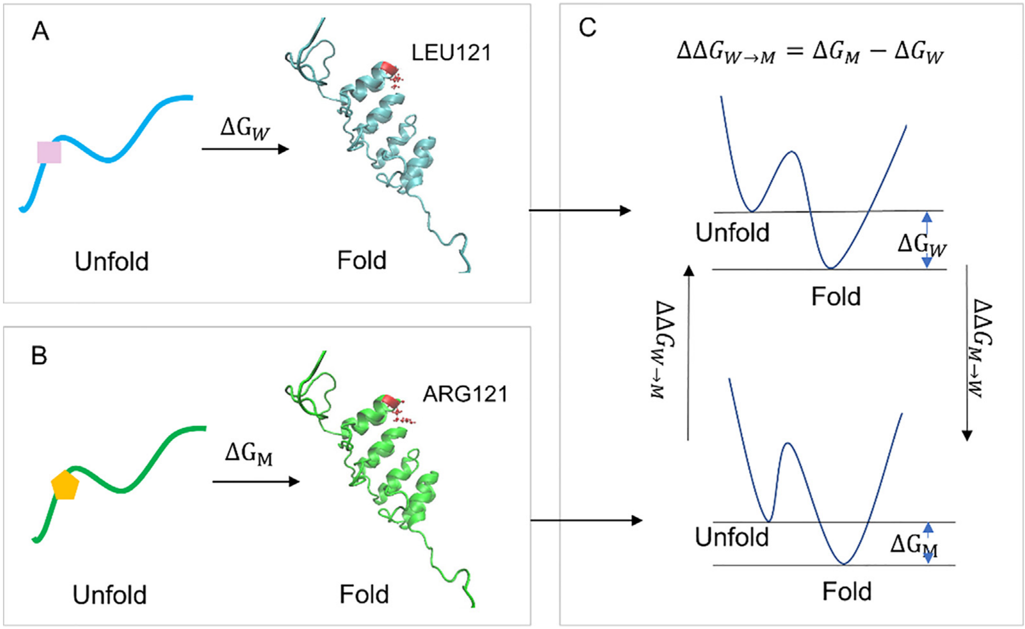

Fig. 1. Definition of a two-state model of stability change upon single-point substitution. Consider protein 1A5E (Protein Data Bank (PDB) ID [22]) as an example, where the amino acid Leucine (Leu) is substituted to Arginine (Arg) at position 121. **A.** Protein folding to the tertiary structure before single amino acid substitution (ΔG_W). **A.** Protein folding to the tertiary structure after single amino acid substitution (ΔG_M). **C.** Change of Gibbs free energy ($\Delta\Delta G_{W\to M}$) between A and B states.

fueled by the significant increase in experimental data. Several computational methods [20,23–50] for predicting thermodynamic stability changes of proteins upon variants were proposed or updated. These computational methods can be categorized into three main types by input data types: 1) predicting from protein sequence (sequence-based approaches) [23–30]; 2) predicting from three-dimensional (3D) structure (structure-based approaches) [20,29,31–43]; and 3) both protein sequence or 3D structure can be inputted to predict (sequence- and structure-based approaches) [18,44–47]. We summarized computational methods with freely accessible webservers or standalone tools for predicting protein stability changes in Supplementary Tables S1-S3.

The majority of current computational methods use 3D protein structure as input, for example, SDM [31], SDM2 [31], DUET [34], mCSM [2], INPS3D [37], AUTO-MUTE2.0 [36], MAESTRO [43], PoPMuSiC [33,38], Pro-Maya [48], TopologyNet [49], ProTSPoM [50], DynaMut [39], DynaMut2 [40], DDGun3D [29], DeepDDG [41], iDeepDDG [41], ThermoNet [20], and PremPS [42] (Supplementary Table S1). Usually, they use experimental 3D structures to calculate statistical potentials, secondary structure (SS), accessible surface area (ASA), and the structural environment of the mutated residue. A small part of computational predictors, such as MUpro [23], SAAFEC-SEQ [28], DDGun [29], EASE-MM [25], BoostDDG [30], PON-tstab [27], INPS [26], iPTREE-STAB [24] (Supplementary Table S2), usually use amino acid properties, evolutionary information from protein families, statistical potentials, and neighbor amino acid information to predict protein stability changes. They are used when experimental protein structures are unavailable and only protein sequences are provided to study the impact of variants. The rest part of predictors, such as ELASPIC [47], STRUM [18], iStable2.0 [46], iStable [45] and I-Mutant2.0 [44], predicted stability changes from protein sequence and protein structure (Supplementary Table S3).

As shown in previous studies, methods that integrate protein structural information with their sequence profiles can enhance the accuracy and robustness of predictions than those employing sequence characteristics [18,25]. However, 3D structures in the Protein Data Bank [22] (PDB) are known for only about 2 % of the proteins available in the UniProt database [28]. Though the success of protein structure prediction tools such as AlphaFold2 [51] and RoseTTAFold [52] are available, *in-silico* predicted variant protein structures may not be accurate [53] and are usually time-consuming to attain.

Based on the consideration of balancing time and accuracy requirements, we introduce MU3DSP (MUtation using 3D Structure Profile), a protein stability change prediction tool based on point mutated 3D structure profiles, which annotate genomic variants from a list of experimental structures using G2S (Genome to Structure) [54]. Importantly, MU3DSP starts from protein sequences and inherently fuses information from protein 3D structures profile into its predictions. Unlike STRUM [18], which predicts protein structure, MU3DSP retrieves a list of experimental protein structures from PDB to generate a 3D structure profile of the querying variants. MU3DSP not only extracts multiple descriptors from wild-type structures but also uses the same descriptors from mutated structures. MU3DSP is able to use wild-type structures and mutated structures either from homology models of query variants if they are available, or otherwise from the annotated genomic variants database G2S. In either case, predictions can be achieved in real-time, and it only takes less than 1 min for one variant when running multiple single-point variants on one protein. To our knowledge, this is the first tool that uses the annotated genomic variants' 3D structure to study protein stability changes. We demonstrate that MU3DSP achieves state-of-the-art performance on two independent testing datasets. Finally, we show the application of MU3DSP in disease-related contexts, including the tumor suppressor protein P53 coded by *TP53* gene in humans, as well as in proteins without structures, such as thiopurine S-methyltransferase (*TPMT*) protein from the Critical Assessment of Genome Interpretation (CAGI) challenge.

2. Materials and methods

2.1. Overview of MU3DSP

MU3DSP includes variant-based structure preparation, feature extraction, and prediction (Fig. 2). First, we queried the substituted position and the corresponding protein sequence to search in G2S [54] (https://g2s.genomenexus.org) to get the 3D structure profiles of annotated genomic variants. Then we calculated variant-based structure features, including changes in secondary structure (ΔSS), amino acid frequencies in spheres space ($\Delta AAFS$) on wild-type and variant protein structures, and relative accessible surface area (RASA) on the matched protein structures. Next, we extracted sequence-based features, including changes in amino acid properties (ΔAAP), position-specific scoring matrix (*PSSM*) [55], changes of the wild-type and mutant residue's evolutionary score in *PSSM* ($\Delta PSSM$), HHblits profiles (HMM) [56], and changes of the wild-type and mutant amino acid emission frequencies from HMM (ΔHMM). Finally, we predicted $\Delta\Delta G$ through the ensemble algorithm LightGBM [57].

2.2. Datasets

Most $\Delta\Delta G$ prediction approaches used training datasets are from ProTherm [58], a thermodynamic database collecting experimentally determined protein stabilities. We used the following three datasets to investigate the prediction of stability changes in proteins (Supplementary Table S4):

S1676 dataset as a benchmark [25,59], which includes 1676 single-point variants in 67 different proteins with 925 destabilizing variants ($\Delta\Delta G < 0.5$), 220 stabilizing variants ($\Delta\Delta G > -0.5$) and 531 neutral variants ($\Delta\Delta G$ in [-0.5, 0.5]). S1676 was used as the training dataset of MASE-MM [25].

S2648 dataset includes 2648 non-redundant single-point variants from 131 proteins compiled by [33]. S2648 is a commonly used dataset for predicting stability changes upon single-point variants. S2648 was used as the training dataset to find direct variants and reverse variants in methods INPS [37], DynaMut2 [40], PoPMuSiC [60], DDGun [29], and PremPS [42]. It includes 1598 destabilizing variants, 295 stabilizing variants, and 775 neutral variants.

S236 [25] dataset, which is curated from EASE-MM as an independent testing data set. S236 comprises 141 destabilizing variants, 20 stabilizing variants, and 75 neutral variants. S236 has no more than 25 % sequence identity with the training dataset S1676.

S543 [25,33,60] dataset, which is a subset of the S2648 dataset [60]. S543 provides PDB IDs [22] and we manually aligned and parsed the S543 dataset to obtain sequences. S543 comprises 342 destabilizing variants, 52 stabilizing variants and 149 neutral variants. S543 has no more than 25 % sequence identity with the training dataset S1676 and testing dataset S236.

S350 dataset is a subset of S2648 randomly selected from S2648 [33]. This dataset is widely used to compare the performance of different methods and the rest of the variants in S2298 ($S2648 - S350$) are used as a training dataset. The dataset contains 192 destabilizing variants, 54 stabilizing variants, and 105 neutral variants.

S^{sym} dataset was manually curated by Pucci et al. [38]. It contains 342 direct variants and 342 reverse variants with available experimental structures for the corresponding variant proteins. It is a common dataset used to assess the antisymmetric property and measure the bias of the predictor.

In addition, we used protein **P53** [35] and **TPMT** coded protein by *TPMT* gene (thiopurine S-methyl transferase) from the CAGI challenge [61] as case studies. Dataset P53 contains experimentally screened protein stability data of the tumor suppressor protein P53 [35], including 21 destabilizing variants, 2 stabilizing variants, and 19 neutral variants. Dataset TPMT includes a total of 3627 variants with 4 destabilizing variants, 2926 stabilizing variants and 654 neutral variants.

2.3. 3D structure profile and variant-based structures

Our in-house G2S [54] provides a real-time web Application Programming Interface (API) that automatically maps genomic variants on 3D protein structures. Giving a protein sequence and the position of a variant as the query, G2S searches similar sequence fragments (covering the surrounding regions of the changed residue in PDB to get a list of protein structures with similar local sequences which contain the same single-point substitution. The list of protein structures is defined as the variant-based 3D structure profile. G2S then categorize protein structure as wild-type residue-based structure (WRBS) or mutant residue-based structure (MRBS) based on containing either wild-type amino acid or mutant amino acid at the aligned position of the queried residue.

According to the availability of PDB structures, four different strategies (Q1-Q4) may be adopted to construct a variant-based structure set (Fig. 2A). Q1: WRBS and MRBS are available. Q2: WRBS is available, and MRBS is unavailable. Q3: WRBS is unavailable, and MRBS is available. Q4: Neither WRBS nor MRBS is available.

When either WRBS or MRBS is available, WRBS and MRBS are two types of variant-based structures. When neither WRBS nor MRBS is available, assessment of genomic variants in G2S can make up the structures. G2s collects the variants information, including their structures from one residue to another residues. For example, there are 3881 substitutions from Val to Glu in G2S and 4927 substitutions from Glu to Val in G2S (Supplementary Table S5 shows the entries of single variants in G2S). Mutant residue structures (3881 in total) based on G2S annotation of genomic variants can be constructed by the entries from Val (wild-type) to Glu (mutant). Wild-type residue structures (4927 in total) based on G2S annotation of genomic variants can be built by the entries from Glu (mutant) to Val (wild-type). They are two other types of variant-based structures. Then, we can obtain variant-based structure features using variant-based structures (Supplementary Figure S2).

2.4. Variant-based structure features

We extracted secondary structure (SS), RASA, and amino acid frequencies in spheres space (AAFS) from the variant-based structures set as base features. The SS feature is a matrix with the size of $L \times 3$, calculated by the DSSP program [62,63], in which L represents the length of the queried protein. SS can be categorized into three types: helix, sheet, and coil. We used the one-hot encoded secondary structure with three states [62,63]. RASA was calculated as follows:

$$RASA = \frac{ASA}{MaxASA}$$

where *ASA* represents the accessible surface area obtained from the DSSP output and *MaxASA* represents the maximum possible solvent accessible surface area for the amino acid [64]. The AAFS feature was calculated as follows:

$$AAFS = \left\{ \frac{C_j}{M} \right\}$$

We select a sphere space on protein 3D structure with a radius of 20 Å centered on the queried residue. C_j represents the number of observed j in the selected sphere space, in which j is one of 20 standard amino acids. M represents the number of amino acids in the sphere space.

Therefore, the prepared features of variant-based structure features for wild-type or variant proteins were achieved by the function:

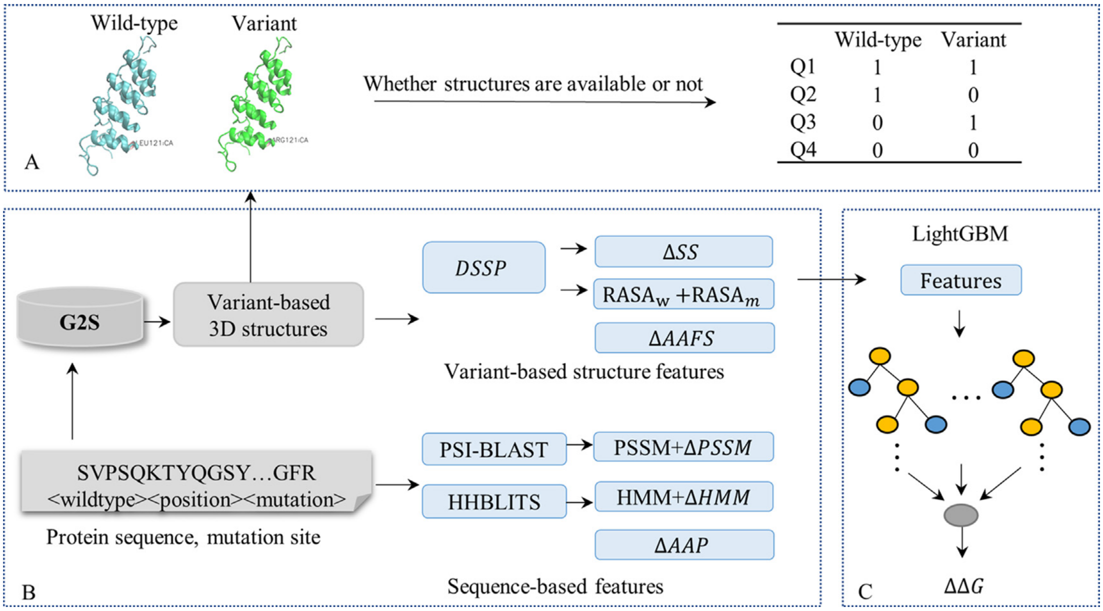

Fig. 2. Overview of feature extraction and feature processing. **A.** Availability of matched protein tertiary structures with wild-type or mutant amino acid. Available tertiary structures are labeled as 1, otherwise as 0. **B.** Preprocessing of the query protein and feature extraction. **C.** Regression algorithm output using extracted features.

$$B_x = \left\{ SS_{max}, \frac{\sum_{i=0}^{N} RASA_i}{N}, \frac{\sum_{i=0}^{N} AAFS_i}{N} \right\}$$

where $x \in \{m, w\}$; m is the variant protein; w is the wild-type protein. When MRBS or WRBS is available (such as Q1, Q2, Q3), N represents the number of WRBS or MRBS. When MRBS is unavailable (such as Q2 and Q4), N represents the number of w to m mutants' entries in G2S. When WRBS is unavailable (such as Q3 and Q4), N represents the number of m to w mutants' entries in G2S (Supplementary Table S5 shows the entries of single variants from G2S). SS_{max} defines the type of secondary structure with the top occurrence on the substituted residue position in the protein tertiary structures; $RASA_i$ and $AAFS_i$ represent the RASA feature and AAFS feature for protein structure i, respectively. To consider the effect of variants, variant-based structure features can be represented by the combination of changes in SS_{max} (ΔSS), changes of AAFS features between wild-type and variant proteins, and RASA features from wild-type and variant proteins.

2.5. Sequence-based features

Amino acid properties (*AAP*) and two evolutionary conversation profiles are derived as follows. Thirteen amino acid properties, including hydrophobicity, volume, helix tendency, sheet tendency, polarizability, isoelectric point and so on have been used, following Folkman et al. [25]. The values of thirteen amino acid properties for twenty standard amino acids are shown in Supplementary Tables S6 and S7. For each mutant residue, we extracted Δ*AAP*, which represents the changes of the corresponding *AAP* from wild-type residue to mutant residue.

Many disease-related mutations, usually in residues that are conserved or conservatively varied during evolution, affect different protein functions, including thermodynamic stability [65]. Therefore, two evolutionary conversation profiles, PSI-BLAST profile [55] and HHblits profile [56], are usually used to study thermodynamic stability changes derived from single-point variants. The PSI-BLAST and HHblits profiles are complementary since their algorithms and searched databases differ.

PSI-BLAST profile. We use position-specific scoring matrix (PSSM) results generated by the alignment tool PSI-BLAST to search the NCBI's SwissProt database for homologous sequences with three iterations and *E*-value cutoff of 0.001 [55]. The size of the generated PSSM is $L \times 20$, where L represents the length of the queried protein and 20 corresponds to 20 standard amino acids. Each element X in the PSSM was normalized to the range (0, 1) by a sigmoid function:

$$PSSM_{(p,j)} = \frac{1}{1 + e^{-X_{(p,j)}}}$$

$PSSM_{(p,j)}$ denotes the normalized result of $X_{(p,j)}$, in which p represents the residue position in the protein sequence and j represents one of 20 standard amino acids. $\Delta PSSM$ is extracted to apply to predict stability changes by the following function:

$$\Delta PSSM = PSSMm - PSSMw$$

HHblits profile. The multiple sequence alignments tool HHblits based on hidden Markov models (HMMs) is applied to search against the Uniclust30 database to get multiple sequence alignments for the queried protein sequence with default parameters [56]. The dimension of HHblits profile features is 30×1, which represents the values from the original HMM matrix consisting of 20 columns of the match state amino acid emission frequencies, seven columns of transition frequencies from the beginning state to the first Match state, Insert state, and Delete state, and three columns of local diversities. Δ*HMM* represents the change of amino acid emission frequencies in the HMM matrix between the wild-type residue and mutant residue. Each score is converted to the range [0, 1]:

$$HMM_{(p,j)} = \frac{Y}{10000}$$

$$\Delta HMM = HMMm - HMMw$$

2.6. Evaluation of protein stability changes and model optimization

In our method, we choose a decision tree algorithm implemented in the package LightGBM [57] to predict ΔΔ*G*. To devise a robust estimate of the prediction performance, we used a 10-fold cross-validation approach to tune parameters using the cross-validation function in the gradient boosting framework LightGBM. First, initial parameters are set as $boosting\_type = gbdt$, $objective = regression$, and $metric = rmse$. Then, to improve the performance, a grid search on parameters is performed for num_leaves and $\max\_depth$. To avoid over-fitting, parameters $feature\_fraction$, $bagging\_freq$, $lambda\_l1$, $lambda\_l2$, $\min\_split\_gain$, and $\min\_data\_in\_bin$ are used to grid search to get the best parameters. After numerous iterations, we get a list of parameters (Supplementary Table S8). Those parameters are used to build the training model of MU3DSP. MU3DSP's training process and prediction path for stability changes of two samples are shown in Supplementary Figures S3 and S4.

The metrics used to evaluate our model and other predictors in comparison were Pearson correlation coefficient (PCC), root mean square error (RMSE) and mean absolute error (MAE) between predicted energy and experimental energy. The evaluated metrics are shown as follows:

$$PCC = \frac{cov\left(\widehat{Y}, Y\right)}{\sigma_{\widehat{Y}}\sigma_Y}$$

$$RMSE = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)$$

where $\widehat{Y}$ represents predicted $\Delta\Delta G$ and Y represents experimental $\Delta\Delta G$; $cov\left(\widehat{Y}, Y\right)$ represents the covariance of $\widehat{Y}$ and Y; $\sigma_{\widehat{Y}}$ represents the standard deviation of the variable $\widehat{Y}$; σ_Y represents the standard deviation of the variable Y; and $\left(Y_i - \widehat{Y}_i\right)$ represents the error between predicted $\Delta\Delta G$ and experimental $\Delta\Delta G$; N denotes the total number of instances.

To assess the antisymmetric property of our method, we used PCC between the direct and corresponding reverse variations. $\widehat{Y}$ represents direct $\Delta\Delta G$ and Y represents reverse ΔΔG. To measure the bias of the predictor and check whether the predictor is toward the destabilizing group of variants or not, we adopted the bias score $\langle\delta\rangle$:

$$\langle\delta\rangle = \frac{\sum_{i=0}^{N}\left(\Delta\Delta G_i^{dir} + \Delta\Delta G_i^{rev}\right)}{2N}$$

A perfectly antisymmetric and unbiased predictor should have PCC between the direct and corresponding reverse variations equal to -1, whereas $\langle\delta\rangle$ equals 0.

3. Results

3.1. Performance on multiple models according to variant-based structure information

To build a robust model for predicting protein stability changes, datasets S1676 and S2648 were used to train a model using 10-fold cross-validation [25]. The results of our methods were replicated 100 times of the 10-fold cross-validation by shuffling the training data. As shown in Table 1, our model MU3DSP with S1676 as the training dataset got an averaged PCC of 0.73 (RMSE of 1.26 Kcal/mol), which constituted a relative improvement of 30.4 % compared to EASE-MM with the same training dataset with an averaged PCC of 0.56 (Table 1). The model MU3DSP-S2648 with S2648 as the training dataset achieved PCC of 0.73 and RMSE of 1.00 kcal/mol. The model MU3DSP-S5296 with S2648 and their reverse variants as a training dataset achieved PCC of 0.82 and RMSE of 1.03 kcal/mol. These results demonstrate that the selected features are reliable and reasonable for predicting stability changes.

MU3DSP^{seq} used sequence descriptors including ΔAAP, $PSSM$, $\Delta PSSM$, HMM, and ΔHMM to train. MU3DSP^{str} used variant-based structure features in the model. MU3DSP used multiple descriptors from models MU3DSP^{str} and MU3DSP^{seq}. MU3DSP* fused sequence-based features and variant-based structure features in which wild-type and mutant 3D structures are unavailable. The model with * used variant-based structure features when WRBS and MRBS are unavailable.

We found the model using a 3D structure profile improved the performance significantly. Based on the availability of the PDB structure, we proposed four models to conduct comparative tests: MU3DSP^{str}, MU3DSP^{seq}, MU3DSP, and MU3DSP* (Table 1, Fig. 3A), to demonstrate the importance of variant-based structure features. MU3DSP^{str} only selected the variant-based structure features we initially proposed. MU3DSP^{seq} selected sequence-based features, including ΔAAP, $PSSM$, $\Delta PSSM$, HMM, and ΔHMM. MU3DSP fused both features that constructed model MU3DSP^{seq} and MU3DSP^{str}. MU3DSP* also used both sequence-based and variant-based structure features. Still, it assumed that WRBS and MRBS are unavailable for all proteins in the training dataset. Variant-based structure features would be calculated by the structures that it gets from m to w mutants' entries and w to m mutants' entries in G2S, respectively.

Notably, the performance of PCC (0.56) on model MU3DSP^{str} (using variant-based structure features only) shows that our proposed variant-based structure features carry pertinent information for the prediction. Furthermore, when adding variant-based structure features (versus the PCC of 0.66 in MU3DSP^{seq}), the model improved its PCC to 0.73. If structures for wild-type and variant proteins are unavailable for the training dataset, adding the baseline variant-based structure features from G2S (i.e., in model MU3DSP*) can change PCC to 0.71. MU3DSP and MU3DSP* demonstrated a significant improvement as shown in Fig. 3A. We found that 40 % of the top 20 features are variant-based structure features (RASA-m, AAFS-I, AAFS-N, AAFS-D, AAFS-L, AAFS-W, AAFS-F, AAFS-C) (Fig. 3B). Using 3D structure profile models such as MU3DSP-S2648 and MU3DSP-S5296 achieved better performance than MU3DSP-S2648* and MU3DSP-S5296* that only used annotated genomic variants from G2S. Taken together, these results suggest that our proposed variant-based structure features contribute to a performance improvement of the model.

3.2. MU3DSP achieves state-of-the-art performance on testing sets

To evaluate the robustness of our model, we used S236 and S543 as independent testing datasets. MU3DSP compared with nine commonly used methods, including EASE-AA [66], EASE-MM [25], MUpro [23], I-Mutant2.0 [44], INPS [37], DynaMut2 [40], PremPS [42], SAAFEC-SEQ [28], DDGun [29], PoPMuSiC [60], and MAESTRO [43] on the testing datasets. Among these, EASE-AA, EASE-MM (EASE-MM-web), MUpro, sequence-based version of I-Mutant2.0 (I-MutantA), INPS, SAAFEC-SEQ, and DDGun predicted $\Delta\Delta G$ starting from sequence whereas the structure-based version of I-Mutant2.0 (I-MutantB), PoPMuSiC, PremPS, DynaMut2 and MAESTRO required structures as input (Table 2).

Methods: <method>#, where "#" represents that S543 is a subset of their training dataset. We shown the perfromacne of these methods on S543 in Supplementary Table S9. I-MutantB, a structure-based version of I-Mutant2.0; I-MutantA, a sequence-based version of I-Mutant2.0; EASE-MM, predicted $\Delta\Delta G$ values taken from the supplementary material in research [25], which used S1676 dataset to train the model. EASE-MM-web, predicted

Table 1
Performance of 10-fold cross-validation on datasets S1676 and S2648.

Method	*PCC*	*RMSE*	*MAE*
Dataset S1676			
EASE-MM	0.56	1.52	1.00
MU3DSP^{str}	0.56	1.50	1.04
MU3DSP^{seq}	0.66	1.47	1.00
MU3DSP	0.73	1.26	0.82
MU3DSP*	0.71	1.29	0.85
Dataset S2648			
MU3DSP-S2648	0.73	1.00	0.75
MU3DSP-S2648*	0.70	1.05	0.78
MU3DSP-S5296	0.82	1.03	0.77
MU3DSP-S5296*	0.78	1.11	0.83

$\Delta\Delta G$ from webserver, which used a joint S1676 + S236 dataset to train the model. The predicted $\Delta\Delta G$ values from PremPS webserver and MAESTRO are negatively correlated on datasets S236 and S543.

The MU3DSP-S5296 had the best performance on the S236 dataset among 15 methods in Table 2 with PCC of 0.73, RMSE of 0.85 Kcal/mol and 0.62 Kcal/mol**.** MU3DSP had a second-best performance on the S236 dataset with a PCC of 0.66 and RMSE of 0.96 Kcal/mol, outperforming other stability change predictors (PCC = −0.02 to 0.62, RMSE = 0.98 to 1.58 Kcal/mol), including sequence/structure-based I-Mutant2.0, EASE-AA, EASE-MM, MUpro, INPS, SAAFEC-SEQ, DDGun, PoPMuSiC, DynaMut2, PremPS, and MAESTRO. Only the MAE of EASE-MM from webserver (EASE-MM-web) performed better than our method. However, this model used a joint S1676 + S236 dataset to train the model. Despite the median values for the prediction errors from methods DDGun, INPS, DynaMut2 and MAESTRO being closer to 0 than that of our methods MU3DSP and MU3DSP-S5296, the data distribution of prediction errors for our method is more centralized (Fig. 3C and 3D). Dataset S543 is used for evaluating our proposed method MU3DSP, as S543 has no more than 25 % sequence identity with the training dataset of MU3DSP. Among sequence-based predictors, without considering reverse variants, MU3DSP showed a good performance for predicting direct variants, such as S236, S^{sym} direct variants (Table 3), the same as SAAFEC-SEQ. In addition, our method (MU3DSP) had the best performance regarding the RMSE (1.19 Kcal/mol) when compared to others (RMSE = 1.37 Kcal/mol to 1.21 Kcal/mol), whose training datasets do not include the testing dataset S543 (Table 2).

We empirically checked for factors in MU3DSP that most influenced Pearson correlation coefficients between experimental $\Delta\Delta G$ and predicted $\Delta\Delta G$. The first observation was that PCC is usually correlated with MAE and RMSE, i.e., the lower the RMSE and MAE values, the higher PCC. However, PCC, MAE and RMSE of EASE-MM-web were 0.48, 0.96 Kcal/mol, and 1.33 Kcal/mol, respectively, while those of MU3DSP were 0.44, 0.92 Kcal/mol, and 1.28 Kcal/mol, respectively, on the Q1 of S543 dataset (Supplementary Table S11). In this case, although RMSE and MAE are lower, PCC in our model is not higher. In parallel, we randomly sampled 10 sub-datasets of S236 and 10 sub-datasets of S543, which were half of the corresponding datasets as replacements. The results show that this assumption is controversial on randomized datasets 1, 3, 7, 8, and 9 of S543 (Supplementary Table S12). The second observation was that the distribution of the dataset

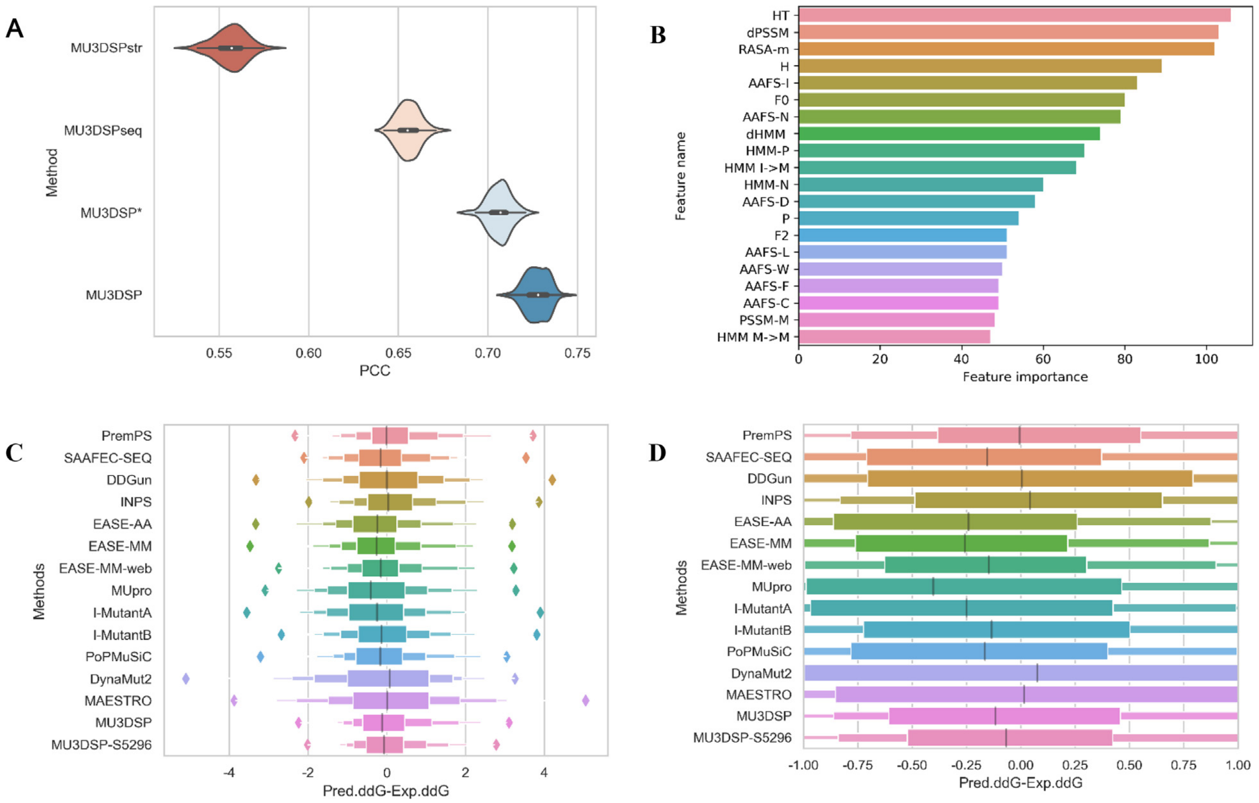

Fig. 3. Performance on comparative test according to variant-based structure information and different testing datasets. **A.** Performance on multiple models according to variant-based structure information. The violin plot for 100 times PCC of 10-fold cross-validation for four models MU3DSP^{str}, MU3DSP^{seq}, MU3DSP, and MU3DSP* on the training dataset. **B.** Top 20 important features for the MU3DSP training used in LightGBM, including variant-based structural features RASA-m, AAFS-I, AAFS-N, AAFS-D, AAFS-L, AAFS-W, AAFS-F, and AAFS-C. Abbreviations of features are shown in Supplementary Table S10. **C.** Prediction errors for dataset S236 calculated by predicted $\Delta\Delta G$ subtracted experimental $\Delta\Delta G$. Box plots show the distribution of errors and black lines represent median values. Outliers are plotted as individual points. **D.** Partial extended view of prediction errors of panel C from −1 to 1.

Table 2
Comparative performance of MU3DSP and MU3DSP-S5296 across testing datasets S236 and S543 with other stability predictors.

Datasets	S236			S543		
Method	*PCC*	*RMSE*	*MAE*	*PCC*	*RMSE*	*MAE*
Structure-based						
I-MutantB	0.52	1.07	0.81	0.36	1.34	1.03
PoPMuSiC#	0.57	1.05	0.79	–	–	–
MAESTRO#	−0.02	1.58	1.23	–	–	–
PromPS#	0.59	1.02	0.74	–	–	–
DynaMut2#	−0.01	1.51	1.21	–	–	–
Sequence-based						
INPS#	0.59	1.02	0.74	–	–	–
SAAFEC-SEQ#	0.64	0.96	0.74	–	–	–
DDGun#	0.49	1.22	0.95	–	–	–
EASE-AA	0.53	1.10	0.83	0.48	1.25	0.94
EASE-MM	0.59	1.03	0.77	0.53	1.22	0.90
EASE-MM-web	0.62	0.98	0.71	0.53	1.21	0.89
MUpro	0.36	1.20	0.97	0.33	1.32	1.04
I-MutantA	0.44	1.18	0.92	0.32	1.37	1.06
MU3DSP-S5296	**0.73**	**0.85**	**0.62**	**–**	**–**	**–**
MU3DSP	**0.66**	**0.96**	**0.72**	**0.52**	**1.19**	**0.90**

plays a role in calculating PCC. When randomly selecting 10 subdatasets of S236 and S543, the PCCs were changed based on different datasets (Supplementary Tables S12 and S13). These results show that the Pearson correlation coefficients between experimental $\Delta\Delta G$ and predicting $\Delta\Delta G$ calculated with different methods are affected by the distribution of each dataset.

3.3. Performances achieved on antisymmetric datasets S^{sym}

S^{sym} is one of the most balanced datasets containing the experimental structures, which includes both direct variants and corresponding reverse variants. We tested the biases of MU3DSP with S2648 dataset and its reverse variants (MU3DSP-S5296) on this balanced dataset and compared its performance with 12 existing methods. Our model MU3DSP-S5296 was among top three (PCC = 0.75) on the direct dataset, top two (PCC = 0.56) on the S^{sym} reverse datasets and top two on the S^{sym} direct and reverse variants among the 12 existing methods. As shown in Table 3, MU3DSP-S5296 got a better performance in the S^{sym} reverse dataset than all listed sequence-based methods. These results have not been achieved previously by other published sequence-based predictors.

3.4. Predicting the impact of single-point variants on P53 thermodynamic stability

Tumor suppressor protein P53 is “the guardian of the genome”, strongly related to cancer, participating in the control of cell survival and division [67]. Over half of all cancers in humans carry loss of function mutations in the transcription factor P53 [68]. Assessment and prediction of stability changes in P53 can aid the interpretation of the association of P53 variants with tumorigenesis.

Protein P53 comprises an *N*-terminal transactivation domain (residues position 1–45), a DNA binding domain (residues position 102–292), and a C-terminal oligomerization domain (residues position 319–359) [69,70]. From UniProt annotation, there are 83 natural variants, 91 % of which are located in the range 102 to 192 in the DNA binding domain. Protein P53 shows an unexpectedly high frequency of mutations in four residues: 175, 245, 248, and 273. Accordingly, mutations in P53 disrupt the wild-type stability and conformation of the protein, thus, interfering with its function [71]. Position 282 is related to a network of interactions underpinning the loop-sheet-helix major groove DNA binding motif. An Arg to Trp (Tryptophan) substitution in this position results in protein unfolding and, consequently, its inactivation [72]. Pires et al. [35] assembled experimentally characterized stabilities of 42 P53 mutants, which provided a benchmark for our study.

We tested MU3DSP and MU3DSP-S5286 on protein P53 with 42 variants. MU3DSP-S5286 is the model that removed the overlapping instances from S2648 and their reverse variants. To test on sequence-based methods, we parsed the P53 structure to sequence and checked its mutations manually. We used six structure-based methods (PoPMuSiC, MAESTRO, DynaMut2, I-MutantB, ProTSPoM, and PremPS) and six sequence-based methods (EASE-MM, I-MutantA, MUpro, INPS, SAAFEC-SEQ, and DDGun) as comparative methods. $\Delta\Delta G$ values from MU3DSP and MU3DSP-S5286 versus experimentally measured stability changes (Exp.ddG) from the P53 dataset are shown in Fig. 4A and 4B. Importantly, MU3DSP and MU3DSP-S5286 had PCC values of 0.72 and 0.70, respectively. MU3DSP-S5286 (PCC = 0.70) is better than other comparative sequence-based methods (from 0.23 to 0.69) on the PCC performance. Structure-based methods ProTSPoM (PCC = 0.88), ProTSPoM-noevo (PCC = 0.81), PremPS (PCC = 0.73), and (PCC = 0.75) are better than our method MU3DSP-S5286 in performance of PCC. However, dataset P53 overlaps with PremPS, ProTSPoM-noevo, ProTSPoM, and DynaMut2′s training dataset. The predictor PremPS without the overlapping mutations got a PCC of 0.72 [42], the same as our proposed MU3DSP. The performance on the P53 dataset demonstrates that MU3DSP is able to give an improved prediction without using the query protein experimental structure.

3.5. Performance on the protein TPMT from CAGI challenge

We further tested MU3DSP on thiopurine S-methyl transferase (TPMT) from CAGI challenge [61]. Due to the TPMT structure is not provided, we used the full-length sequence (UniProt ID: P51580) as input and compared it with seven other leading sequence-based predictors (EASE-MM, I-MutantA, MUpro, INPS, BoostDDG, SAAFEC-SEQ, and DDGun). For further comparisons, we calculated the PCC, RMSE and MAE between experimental $\Delta\Delta G$ and the predictions of the seven methods (Supplementary Table S14). Fig. 5 illustrates the experimental $\Delta\Delta G$ versus predicted $\Delta\Delta G$ from seven comparative methods. MU3DSP achieved the best performance with a PCC of 0.44. The second-best methods SAAFEC-SEQ and BoostDDG yielded a PCC of 0.42 for the TPMT dataset in Fig. 5. The TPMT dataset further demonstrates that MU3DSP is able to improve prediction when experimental structures are not provided.

Table 3
Comparative performance of MU3DSP and MU3DSP-S5296 on antisymmetric datasets S^{sym} with other stability predictors.

Method	S^{sym} direct			S^{sym} reverse			S^{sym} direct + S^{sym} reverse			Anti-symmetry	
	PCC	RMSE	MAE	PCC	RMSE	MAE	PCC	RMSE	MAE	PCC	⟨δ⟩
Structure-based											
MAESTRO	0.57	1.31	0.91	0.27	2.16	1.66	0.43	1.79	1.29	−0.33	−0.62
DDGun3D	0.57	1.40	1.00	0.54	1.43	1.03	0.65	1.41	1.01	−0.99	−0.02
PremPS	0.81	0.96	0.66	0.73	1.13	0.78	0.85	1.05	0.72	−0.93	−0.02
mCSM	0.61	1.23	0.91	0.14	2.43	1.93	0.40	1.93	1.42	−0.26	−0.91
INPS3D	0.61	1.24	0.89	0.29	1.94	1.45	0.56	1.63	1.17	−0.51	−0.51
Dynamut2	0.63	1.21	0.90	0.05	2.39	1.87	0.38	1.90	1.38	−0.11	−0.78
PoPMuSiC	0.63	1.21	0.86	0.25	2.18	1.66	0.50	1.76	1.26	−0.28	−0.71
Sequence-based											
SAAFEC-SEQ	0.73	1.05	0.73	−0.43	2.75	2.11	0.26	2.08	1.42	0.67	−0.97
STRUM	0.75	1.05	–	−0.15	2.51	–	–	–	–	0.34	−0.87
INPS-Seq	0.48	1.47	1.07	0.49	1.45	1.07	0.62	1.46	1.07	−0.99	0.00
DDGun	0.49	1.46	1.09	0.49	1.46	1.09	0.63	1.46	1.09	−1.00	−0.01
MUpro	0.79	0.94	0.53	0.07	2.51	2.03	0.48	1.89	1.28	−0.02	−0.97
MU3DSP_S1676	**0.64**	**1.25**	**0.94**	**0.25**	**2.03**	**1.56**	**0.59**	**1.68**	**1.25**	**−0.56**	**−0.61**
MU3DSP_S5296	**0.75**	**1.06**	**0.76**	**0.56**	**1.53**	**1.13**	**0.76**	**1.31**	**0.95**	**−0.82**	**−0.32**

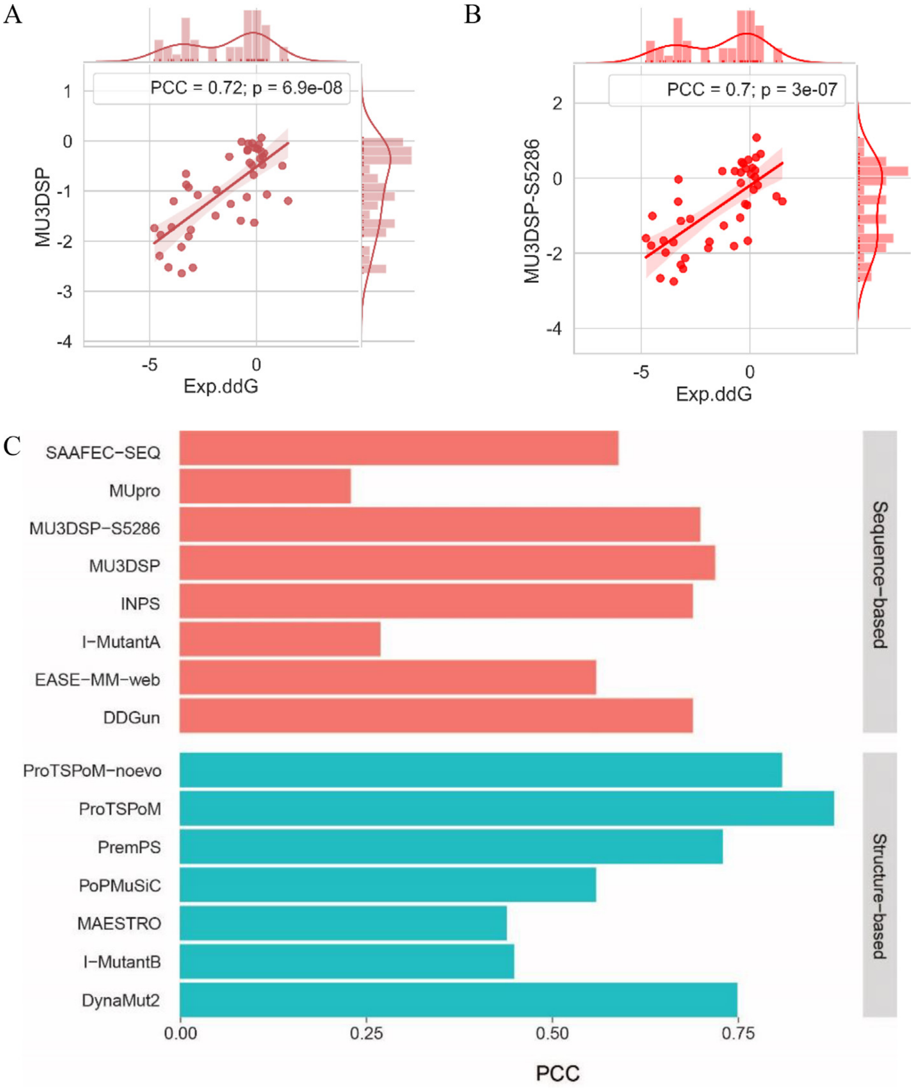

Fig. 4. The performance of our method on predicting the impact of single-point variants on protein P53. **A.** ΔΔG predicted with MU3DSP as a function of Exp.ddG from the P53 dataset. **B.** ΔΔ*G* predicted with MU3DSP-S5286 as a function of Exp.ddG from the P53 dataset. Lines represent linear regression fits. **C.** The bar plot for PCCs between predicted ΔΔ*G* and experimental ΔΔ*G* of MU3DSP, MU3DSP-S5286, SAAFEC-SEQ, DDGun, INPS, MAESTRO, EASE-MM, PoPMuSiC, PremPS, ProTSPoM, DynaMut2, MUpro, and I-Mutant2.0 (A for sequence-based and B for structure-based) for protein P53.

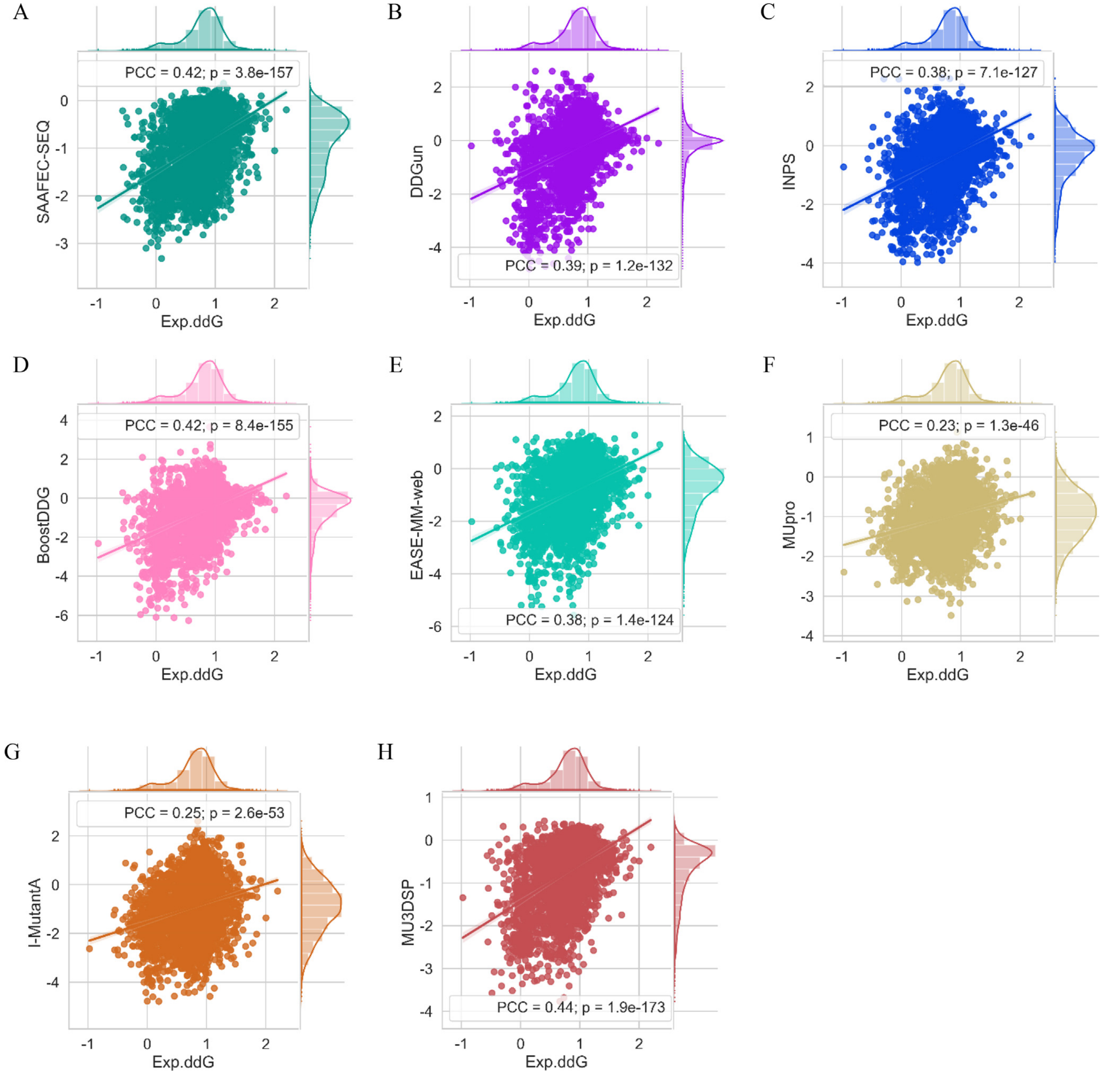

Fig. 5. Multiple bivariate plots for seven comparative methods and MU3DSP with marginal histograms. **A-G.** ΔΔG predicted with seven different sequence-based methods as a function of Exp.ddG from the TPMT dataset. **H.** ΔΔG predicted with MU3DSP as a function of Exp.ddG from the TPMT dataset. Lines represent linear regression fits. PCC, Pearson correlation coefficient.

4. Discussion

Predicting mutation-induced stability changes is essential for protein design and precision medicine. The loss of protein stability can be a main driver of disease; hence, predicting the effects of single-point variants on protein stability facilitates identifying relationships with pathogenicity. For this purpose, *in-silico* predictors can help narrow down the mutational landscape of several studies addressing these questions. Recent advances in machine learning accelerated the development and improvement of these computational methods [20,23–29,31–47].

In this study, we present a novel sequence-based method named MU3DSP, which can efficiently predict protein stability changes upon single-point variant starting from the sequence while using 3D structure information available at PDB. Following the conclusion of existing prediction approaches and tools on protein stability changes, we found sequence-based methods to occupy only a small portion of these methods. This finding is intriguing and especially important for practical applications, considering the fact that most protein 3D structures are unavailable. MU3DSP successfully computed the effect of nsSNVs on protein stability when the protein 3D structure was unavailable. In fact, there were considerable advantages of using 3D structure profiles (variant-based structure features), rather than only using sequence features, to predict protein stability changes upon single-point variants. Furthermore, when compared with a series of computational experiments, MU3DSP outperformed some widely used methods, demonstrating its ability to study the impact of single-point variants. Finally, we successfully applied our MU3DSP method to predict $\Delta\Delta G$ resulting from variants in the tumor suppressor protein P53 and no 3D structure protein TPMT.

Similar to other machine learning methods, the performance of MU3DSP is highly dependent on the quality of training datasets, feature extraction and selection of training algorithm. In this work, we mainly improved and optimized the model in the stage of feature extraction by fusing 3D structure features. Although we achieved increased performance in predicting stability changes based on 3D structure profiles, there is still room for improvement, especially regarding precision. Considering our data, the prediction of stabilizing variants was more difficult because the number of destabilizing variants in the training dataset was around three times higher than the number of stabilizing variants. However, the performance of the model is expected to improve, considering the increase of data concerning experimental stability changes that will allow the construction of a balanced benchmark dataset. Nevertheless, this may be further explored in the future using more advanced machine learning methods. Additionally, we chose the LightGBM gradient boosting framework to predict $\Delta\Delta G$ because it displays many advantages compared to others, such as faster training speed, higher efficiency, and memory usage. In this regard, we explored some deep learning methods, but their performance was not ideal in the setting of this study. In future studies, more advanced deep learning approaches, such as graph neural networks, may be employed to utilize the structure information around the mutant residue to improve protein stability prediction. Furthermore, given the success of protein structure prediction tools such as AlphaFold2 [51] and RoseTTAFold [52], high-quality predicted 3D structures of wild-type and variant proteins may be applied to help identify pathogenic variants in humans [73] and to better predict protein stability changes. Although these models have not currently succeeded in predicting stability changes [53], this will greatly accelerate the identification of pathogenic variants in humans in the near future.

Some limitations of computational methods used for prediction purposes revolve around the quality of training datasets used in predicting stability changes as well the variability of experimental $\Delta\Delta G$ datasets, which usually originate from different experiments, authors, and articles [74]. Moreover, the ΔG depends on several factors, such as experimental conditions, temperature, concentrations of salt, pH values, organic solvents, and other chemical agents, which may be difficult to control and may cause data dispersion [75]. Nevertheless, the continuous increase in data generation will greatly accelerate improvements in the performance of predictive methods, both in terms of reliability and consistency, which may have major practical implications in the future.

CRediT authorship contribution statement

Jianting Gong: Methodology, Software, Writing – original draft, Data curation, Formal analysis, Validation, Visualization. **Juexin Wang:** Methodology, Validation, Formal analysis, Investigation, Writing – review & editing. **Xizeng Zong:** Validation. **Zhiqiang Ma:** Conceptualization, Supervision, Writing – review & editing. **Dong Xu:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the China Scholarship Council [201906620047 to J.G.], and the US National Institute of General Medical Sciences of the National Institutes of Health [R35-GM126985 to D.X.].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.12.008.

References

[1] Jafri M, Wake NC, Ascher DB, et al. Germline Mutations in the CDKN2B Tumor Suppressor Gene Predispose to Renal Cell Carcinoma. Cancer Discov 2015;5(7):723–9.
[2] Pires DEV, Rodrigues CHM, Ascher DB. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. Nucleic Acids Res 2020;48(W1):W147–53.
[3] Trezza A, Bernini A, Langella A, et al. A Computational Approach From Gene to Structure Analysis of the Human ABCA4 Transporter Involved in Genetic Retinal Diseases. Invest Ophthalmol Vis Sci 2017;58(12):5320–8.
[4] Hildebrand JM, Kauppi M, Majewski IJ, et al. A missense mutation in the MLKL brace region promotes lethal neonatal inflammation and hematopoietic dysfunction. Nat Commun 2020;11(1):3150.
[5] Xavier JS, Nguyen TB, Karmarkar M, et al. ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Res 2021;49(D1):D475–9.
[6] Stefl S, Nishi H, Petukh M, et al. Molecular mechanisms of disease-causing missense mutations. J Mol Biol 2013;425(21):3919–36.
[7] Portelli S, Olshansky M, Rodrigues CHM, et al. Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. Nat Genet 2020;52(10):999–1001.
[8] Karmakar M, Rodrigues CHM, Horan K, et al. Structure guided prediction of Pyrazinamide resistance mutations in pncA. Sci Rep 2020;10(1):1875.
[9] Karmakar M, Rodrigues CHM, Holt KE, et al. Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS One 2019;14(5):e0217169.
[10] Phelan J, Coll F, McNerney R, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 2016;14:31.
[11] Hawkey J, Ascher DB, Judd LM, et al. Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 2018;4(3).
[12] Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 2009;19(5):596–604.
[13] Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 2005;353(2):459–73.
[14] Stein A, Fowler DM, Hartmann-Petersen R, et al. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. Trends Biochem Sci 2019;44(7):575–88.
[15] Kucukkal TG, Petukh M, Li L, et al. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Curr Opin Struct Biol 2015;32:18–24.
[16] Hamosh A, Scott AF, Amberger J, et al. Online Mendelian Inheritance in Man (OMIM). Hum Mutat 2000;15(1):57–61.
[17] Zhu C, Miller M, Zeng Z, et al. Computational approaches for unraveling the effects of variation in the human genome and microbiome. Annu Rev Biomed Data Sci 2020;3:411–32.
[18] Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016;32(19):2936–46.
[19] Jacobs DJ, Dallakyan S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. Biophys J 2005;88(2):903–15.
[20] Li B, Yang YT, Capra JA, et al. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLoS Comput Biol 2020;16(11):e1008291.
[21] den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat 2016;37(6):564–9.
[22] Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–42.
[23] Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 2006;62(4):1125–32.
[24] Huang LT, Gromiha MM, Ho SY. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics 2007;23(10):1292–3.
[25] Folkman L, Stantic B, Sattar A, et al. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. J Mol Biol 2016;428(6):1394–405.

[26] Fariselli P, Martelli PL, Savojardo C, et al. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics 2015;31(17):2816–21.

[27] Yang Y, Urolagin S, Niroula A, et al. PON-tstab: Protein Variant Stability Predictor. Importance of Training Data Quality. Int J Mol Sci 2018;19(4).

[28] Li G, Panday SK, Alexov E. SAAFEC-SEQ: A Sequence-Based Method for Predicting the Effect of Single Point Mutations on Protein Thermodynamic Stability. Int J Mol Sci 2021;22(2).

[29] Montanucci L, Capriotti E, Frank Y, et al. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinf 2019;20(Suppl 14):335.

[30] Lv X, Chen J, Lu Y, et al. Accurately Predicting Mutation-Caused Stability Changes from Protein Sequences Using Extreme Gradient Boosting. J Chem Inf Model 2020;60(4):2388–95.

[31] Worth CL, Preissner R, Blundell TL. SDM–a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 2011;39 (Web Server issue):W215–22.

[32] Pandurangan AP, Ochoa-Montano B, Ascher DB, et al. SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 2017;45 (W1):W229–35.

[33] Dehouck Y, Kwasigroch JM, Gilis D, et al. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinf 2011;12:151.

[34] Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 2014;42(Web Server issue):W314–9.

[35] Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2014;30(3):335–42.

[36] Masso M, Vaisman II. AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. Adv Bioinf 2014;2014:278385.

[37] Savojardo C, Fariselli P, Martelli PL, et al. INPS-MD: a web server to predict stability of protein variants from sequence and structure. Bioinformatics 2016;32(16):2542–4.

[38] Pucci F, Bernaerts KV, Kwasigroch JM, et al. Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics 2018;34(21):3659–65.

[39] Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 2018;46(W1):W350–5.

[40] Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci 2021;30(1):60–9.

[41] Cao H, Wang J, He L, et al. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J Chem Inf Model 2019;59 (4):1508–14.

[42] Chen Y, Lu H, Zhang N, et al. PremPS: Predicting the impact of missense mutations on protein stability. PLoS Comput Biol 2020;16(12):e1008543.

[43] Laimer J, Hofer H, Fritz M, et al. MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinf 2015;16(1):116.

[44] Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33(Web Server issue):W306–10.

[45] Chen CW, Lin J, Chu YW. iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinf 2013;14(Suppl 2):S5.

[46] Chen CW, Lin MH, Liao CC, et al. iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules. Comput Struct Biotechnol J 2020;18:622–30.

[47] Witvliet DK, Strokach A, Giraldo-Forero AF, et al. ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinformatics 2016;32(10):1589–91.

[48] Wainreb G, Wolf L, Ashkenazy H, et al. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. Bioinformatics 2011;27(23):3286–92.

[49] Cang Z, Wei GW. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS Comput Biol 2017;13(7):e1005690.

[50] Banerjee A, Mitra P. Estimating the Effect of Single-Point Mutations on Protein Thermodynamic Stability and Analyzing the Mutation Landscape of the p53 Protein. J Chem Inf Model 2020;60(6):3315–23.

[51] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.

[52] Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373(6557):871–6.

[53] Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol 2022;29(1):1–2.

[54] Wang J, Sheridan R, Sumer SO, et al. G2S: a web-service for annotating genomic variants on 3D protein structures. Bioinformatics 2018;34 (11):1949–50.

[55] Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25 (17):3389–402.

[56] Remmert M, Biegert A, Hauser A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2011;9(2):173–5.

[57] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Proces Syst 2017;30:3146–54.

[58] Kumar MD, Bava KA, Gromiha MM, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 2006;34(Database issue):D204–6.

[59] Folkman L, Stantic B, Sattar A. Feature-based multiple models improve classification of mutation-induced stability changes. BMC Genomics 2014;15 (Suppl 4):S6.

[60] Dehouck Y, Grosfils A, Folch B, et al. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 2009;25(19):2537–43.

[61] Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods 2014;11(8):801–7.

[62] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637.

[63] Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res 2015;43(Database issue):D364–8.

[64] Tien MZ, Meyer AG, Sydykova DK, et al. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 2013;8(11):e80635.

[65] Pandurangan AP, Ascher DB, Thomas SE, et al. Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 2017;45(2):303–11.

[66] Folkman L, Stantic B, Sattar A. Towards sequence-based prediction of mutation-induced stability changes in unseen non-homologous proteins. BMC Genomics 2014;15(Suppl 1):S4.

[67] Kang R, Kroemer G, Tang D. The tumor suppressor protein p53 and the ferroptosis network. Free Radic Biol Med 2019;133:162–8.

[68] Leroy B, Fournier JL, Ishioka C, et al. The TP53 website: an integrative resource centre for the TP53 mutation database and TP53 mutant analysis. Nucleic Acids Res 2013;41(Database issue):D962–9.

[69] Sionov RV, Haupt Y. The cellular response to p53: the decision between life and death. Oncogene 1999;18(45):6145–57.

[70] Vousden KH, Lu X. Live or let die: the cell's response to p53. Nat Rev Cancer 2002;2(8):594–604.

[71] Olivier M, Eeles R, Hollstein M, et al. The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat 2002;19 (6):607–14.

[72] Bullock AN, Henckel J, Fersht AR. Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. Oncogene 2000;19(10):1245–56.

[73] Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. Nat Med 2021;27(10):1666–9.

[74] Sanavia T, Birolo G, Montanucci L, et al. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput Struct Biotechnol J 2020;18:1968–79.

[75] Paleyes A, Urma RG, and Lawrence ND, *Challenges in deploying machine learning: a survey of case studies.* arXiv preprint arXiv:2011.09926, 2020.