



OPEN ACCESS

EDITED BY

Laurent Roberto Chiarelli,
University of Pavia, Italy

REVIEWED BY

Julie Carol Mitchell,
Oak Ridge National Laboratory (DOE),
United States
Debsindhu Bhowmik,
Oak Ridge National Laboratory (DOE),
United States

*CORRESPONDENCE

Piero Fariselli,
✉ piero.fariselli@unito.it

SPECIALTY SECTION

This article was submitted to Protein Biochemistry for Basic and Applied Sciences, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 20 October 2022

ACCEPTED 15 December 2022

PUBLISHED 05 January 2023

CITATION

Benevenuta S, Birolo G, Sanavia T, Capriotti E and Fariselli P (2023), Challenges in predicting stabilizing variations: An exploration. *Front. Mol. Biosci.* 9:1075570. doi: 10.3389/fmolb.2022.1075570

COPYRIGHT

© 2023 Benevenuta, Birolo, Sanavia, Capriotti and Fariselli. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Challenges in predicting stabilizing variations: An exploration

Silvia Benevenuta¹, Giovanni Birolo¹, Tiziana Sanavia¹,
Emidio Capriotti² and Piero Fariselli^{1*}

¹Department of Medical Sciences, University of Torino, Torino, Italy, ²Department of Pharmacy and Biotechnology (FaBIT), University of Bologna, Bologna, Italy

An open challenge of computational and experimental biology is understanding the impact of non-synonymous DNA variations on protein function and, subsequently, human health. The effects of these variants on protein stability can be measured as the difference in the free energy of unfolding ($\Delta\Delta G$) between the mutated structure of the protein and its wild-type form. Throughout the years, bioinformaticians have developed a wide variety of tools and approaches to predict the $\Delta\Delta G$. Although the performance of these tools is highly variable, overall they are less accurate in predicting $\Delta\Delta G$ stabilizing variations rather than the destabilizing ones. Here, we analyze the possible reasons for this difference by focusing on the relationship between experimentally-measured $\Delta\Delta G$ and seven protein properties on three widely-used datasets (S2648, VariBench, Ssym) and a recently introduced one (S669). These properties include protein structural information, different physical properties and statistical potentials. We found that two highly used input features, i.e., hydrophobicity and the Blosom62 substitution matrix, show a performance close to random choice when trying to separate stabilizing variants from either neutral or destabilizing ones. We then speculate that, since destabilizing variations are the most abundant class in the available datasets, the overall performance of the methods is higher when including features that improve the prediction for the destabilizing variants at the expense of the stabilizing ones. These findings highlight the need of designing predictive methods able to exploit also input features highly correlated with the stabilizing variants. New tools should also be tested on a not-artificially balanced dataset, reporting the performance on all the three classes (i.e., stabilizing, neutral and destabilizing variants) and not only the overall results.

KEYWORDS

protein stability, single-point mutation, stability predictors, machine learning, stabilizing variants

1 Introduction

Non-synonymous DNA variations can affect the stability of the protein structure, jeopardising its function with potential pathogenic outcomes (Casadio et al., 2011; Yue et al., 2005; Hartl, 2017; Martelli et al., 2016; Cheng et al., 2008; Compiani and Capriotti, 2013; Birolo et al., 2021). For this reason, the impact of these variations on the protein structure and its stability is a widely studied problem, even though its *in silico* prediction is still challenging for bioinformaticians.

The effects of non-synonymous variants on the protein stability are usually expressed as the difference in the Gibbs free energy of unfolding ($\Delta\Delta G$), measured in kcal/mol and defined as the difference between the unfolding free energy of the mutated structure (M) of the protein and its wild-type form (W):

$$\Delta\Delta G_{WM} = \Delta G_M - \Delta G_W. \quad (1)$$

With this notation, destabilizing variants are associated to a negative $\Delta\Delta G$, while this value is positive for the stabilizing ones. For those mutations showing $\Delta\Delta G$ values close to zero, the significant experimental uncertainties (Montanucci et al., 2019b; Benevenuta and Fariselli, 2019) make their $\Delta\Delta G$ signs less reliable. At the same time, these $\Delta\Delta G$ values indicate a minimal variation of folding stability of the variants. To account for this issue, we consider an intermediate class of variants named “neutral”, whose $\Delta\Delta G$ values are in the range between -0.5 and 0.5 kcal/mol. The choice of 0.5 kcal/mol is based on the average experimental error, as reported in Capriotti et al. (2008).

Over the years, several computational tools have been developed to predict the $\Delta\Delta G$, combining machine learning methods, statistical potential energy, physico-chemical properties, sequence features and evolutionary information (Benevenuta et al., 2021; Pancotti et al., 2021; Montanucci et al., 2019a; Pires et al., 2014b; Worth et al., 2011; Samaga et al., 2021; Pires et al., 2014a; Rodrigues et al., 2018; Rodrigues et al., 2021; Schymkowitz et al., 2005; Li et al., 2021; Cheng et al., 2006; Kellogg et al., 2011; Capriotti et al., 2005; Li et al., 2020; Chen et al., 2020; Dehouck et al., 2011; Laimer et al., 2016; Savojardo et al., 2016; Savojardo et al., 2019). As highlighted in Sanavia et al. (2020), most of these tools still provide over-optimistic performance due to sequence similarity between the proteins used in the training and test sets. To provide a more realistic estimate of their performance, we recently compared the predictive ability of 18 popular tools on a novel manually-curated dataset in Pancotti et al. (2022). This dataset, named S669, was extracted from ThermoMutDB (Xavier et al., 2021) and it contains only variants belonging to proteins having less than 25% sequence identity with those of S2648 (Pires et al. (2016)) and VariBench (Nair and Vihinen (2013)), two datasets on which most of the state-of-the-art methods were trained. Our analysis underlined that, across all the methods, the performance is more accurate in predicting destabilizing variants rather than the stabilizing ones.

As possible solutions to this issue, methods had accounted for unbalanced training datasets by “artificially balancing” it or by exploiting the antisymmetry property, which is a relationship imposed on $\Delta\Delta G$ values by thermodynamics. Specifically, given the wild-type W and the mutated M protein structures, the folding free energy $\Delta\Delta G_{WM}$ from W to M is equal and has the opposite sign of the folding free energy $\Delta\Delta G_{MW}$ from M to W , considering identical experimental conditions:

$$\Delta\Delta G_{WM} = \Delta G_M - \Delta G_W = -(\Delta G_W - \Delta G_M) = -\Delta\Delta G_{MW}. \quad (2)$$

Using this property, any unbalanced training dataset can be “artificially balanced” by introducing the reverse variants, which are substitutions created from the experimentally-measured variants, henceforth named “direct”. Considering the mutation from W to M as the “direct” variant, its “reverse” is simply defined as:

$$M \rightarrow W, \text{ with } \Delta\Delta G_{MW} = -\Delta\Delta G_{WM}. \quad (3)$$

By artificially balancing the training dataset or by enforcing the antisymmetric property in the model itself, a method should theoretically be able to predict stabilizing variations with the same accuracy as the destabilizing ones. This statement has been already verified on S669 when considering both direct and reverse variants (Pancotti et al., 2022). Here we showed that, when considering only the direct variants, the performance is highly unbalanced with the direct stabilizing variants, which were predicted much worse than the

direct destabilizing ones. Investigating the reasons for this discrepancy is the main objective of the present work.

For most methods, this imbalance could be due to any reason concerning the tool architecture or to the training phase. It is difficult, in these cases, to isolate where the problem lies. On the other hand, for an untrained method such as DDGun3D Montanucci et al. (2019a), whose prediction is a linear combination of its features, this issue can only arise from the features themselves.

For these reasons, we decided to study the impact on the predictions of the following properties of residue substitutions: the difference in hydrophobicity and volume, the logarithm of the conservation ratio, the Blosum62 evolutionary score, the relative solvent accessibility and the Skolnick and Bastolla-Vendruscolo potentials. These are the most common features considered by the state-of-the-art methods and they include the DDGun3D input features.

In this study, we showed that some of these commonly-used features are only useful to predict destabilizing variants and unhelpful for the stabilizing ones. Our findings highlight an intrinsic difference between these two classes, suggesting the importance of using different properties for the stabilizing variants in order to develop methods with more consistent performance between variant classes.

2 Materials and methods

2.1 Structural information, physical properties and statistical potentials

We considered seven different features that include the DDGun3D inputs and two more properties (conservation and volume of the amino acids involved). All these features are not specific only to DDGun3D, but they are commonly employed by $\Delta\Delta G$ predictors. We analyzed:

1) two physical properties:

- **Difference in hydrophobicity:** the difference in hydrophobic regions between wild-type and mutant residues according to the Kyte-Doolittle scale (Kyte and Doolittle, 1982);
- **Volume difference:** the volume difference between the wild-type and the mutated residue measured in Å^3 (Zamyatin, 1972).

2) three features based on conservation and structural information:

- **Logarithm of the conservation ratio**, defined as $\log\left(\frac{\text{CON}_{SW+\epsilon}}{\text{CON}_{SM+\epsilon}}\right)$, with $\epsilon = 0.01$ to avoid invalid results when any of the conservation frequencies were equal to zero. If the mutation changes an amino acid into a less conserved one, the logarithm value is < 0 ;
- **Blosum evolutionary score:** the difference between the wild-type and mutant residues in the Blosum62 substitution matrix, $B(W, M)$ (Henikoff and Henikoff, 1992);
- **Relative Solvent Accessibility:** a measure of the extent of burial or exposure of the residue in the 3D protein structure, ranging from 0 for completely buried to one for completely exposed. It was computed through the DSSP program (Kabsch and Sander, 1983; Touw et al., 2015).

TABLE 1 Datasets composition. The variants are grouped according to their $\Delta\Delta G$ values into three classes: destabilizing ($\Delta\Delta G \leq -0.5$ kcal/mol), neutral ($|\Delta\Delta G| < 0.5$ kcal/mol) and stabilizing ($\Delta\Delta G \geq 0.5$ kcal/mol). The corresponding percentages are reported into brackets.

	Destabilizing	Neutral	Stabilizing
S2648	1,597 (60%)	755 (29%)	295 (11%)
S669	387 (58%)	195 (29%)	85 (13%)
Ssym	225 (33%)	234 (34%)	225 (33%)
VariBench	800 (56%)	426 (30%)	194 (14%)
S4428	2,461 (55%)	1,311 (30%)	656 (15%)

3) two features based on the statistical potentials:

- **Skolnick potential:** the difference in the interaction energy (measured through the Skolnick et al. (1997) statistical potential) between the wild-type and substituted residues with their sequence neighbours within a 2-residue window

$$\sum_{i=-2}^{i=2} (sk(W, a_i) - sk(M, a_i));$$

- **Bastolla-Vendruscolo potential:** the difference in the interaction energy, measured as the Bastolla statistical potential (Bastolla et al., 2001) between the wild-type and mutant residues with its structural neighbours,

$$\sum_{i \in I} (bv(W, a_i) - bv(M, a_i));$$

where I is the set of amino acid residues in the structural neighbourhood of radius 5\AA around the substituted position.

2.2 Datasets

We divided the analysis in two parts. Firstly, we studied the abilities to predict the stabilizing variations of 18 protein stability predictors on:

- **S669** (Pancotti et al., 2022) a recent manually-curated dataset extracted from ThermoMutDB (Xavier et al., 2021) whose variants belong to proteins having less than 25% sequence identity with those of S2648 and VariBench.

Secondly, we studied the correlation and the predictive ability of each different feature on a dataset that we named **S4428**, given by the combination of S669 and:

- **Ssym** (Pucci et al., 2018) which provides 684 balanced (i.e., half direct and half reverse) variations;
- **S2648** (Dehouck et al., 2011) and **VariBench** (Nair and Vihinen, 2013), two of the most used datasets, both extracted from Protherm (Kumar et al., 2006) database. They contain respectively 2,648 and 1,420 manually curated variants with experimentally measured $\Delta\Delta G$ values.

After merging all the datasets, we excluded 19 variants from the analysis due to errors in their 3D neighbors. The composition of all the

datasets, their intersection and the distribution of their $\Delta\Delta G$ s are reported in Table 1, Figures 1, 2, respectively.

2.3 The unbalanced predictions of the state-of-the-arts methods on never-before-seen variants

As shown in Figure 1, Ssym, S2648 and VariBench, which are the datasets most commonly used to train and test predictive methods, share a large number of variants. On the other hand, S669 has no variants in common with the other three datasets and its variants lie in proteins with less than 25% of sequence identity with the proteins in the other three datasets.

Since these characteristics are required for a proper test set, we will only assess $\Delta\Delta G$ prediction performance on the S669 dataset and not on the much larger S4428. The methods' performance in the different variant classes were evaluated by Pearson's correlation coefficient (ρ) and root mean square error (RMSE) between the experimental and the predicted $\Delta\Delta G$ (Table 2), defined as:

$$\rho = \frac{\text{Cov}(\Delta\Delta G^{\text{exp}}, \Delta\Delta G^{\text{pred}})}{\sigma_{\Delta\Delta G^{\text{exp}}} \sigma_{\Delta\Delta G^{\text{pred}}}} \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (\Delta\Delta G_i^{\text{exp}} - \Delta\Delta G_i^{\text{pred}})^2}{N}} \quad (5)$$

where Cov is the covariance, σ is the standard deviation and N is the number of variants.

2.4 Assessing the impact of the input features in the prediction of destabilizing, neutral and stabilizing variants

In order to assess the relevance of the seven features listed in Section 2.1, we computed the Pearson's correlation coefficient (ρ) between each feature and the experimental $\Delta\Delta G$ s in S4428 (Table 3). To remove the bias in the ρ values on the whole dataset caused by the under-representation of the stabilizing and neutral variants with respect to the destabilizing ones (Table 1), we randomly under-sampled the available data to balance the classes. Specifically, we generated 100 random balanced subsets of 984 variants extracted from the original dataset. For each class we selected 328 elements, half the

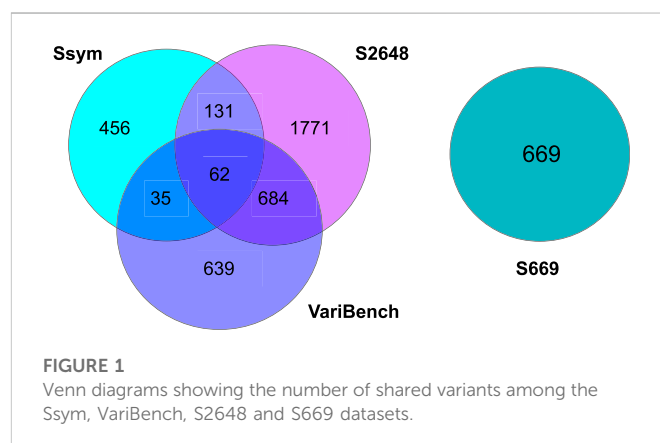


FIGURE 1
Venn diagrams showing the number of shared variants among the Ssym, VariBench, S2648 and S669 datasets.

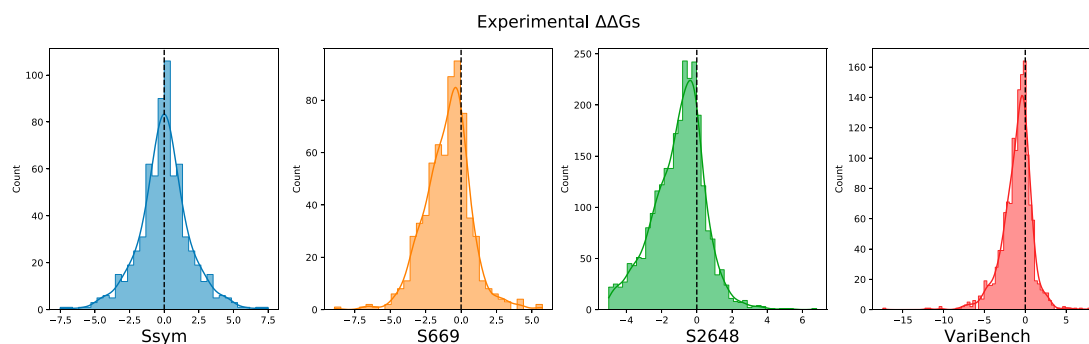


FIGURE 2

Distribution of the experimental $\Delta\Delta G$ values in the Ssym, S669, S2648 and VariBench datasets.

TABLE 2 Pearson's correlations and root mean square error (RMSE) between the experimental and estimated $\Delta\Delta G$ s. The $\Delta\Delta G$ s are predicted by 18 commonly used protein stability prediction tools on the direct variants of the S669 dataset. The correlations and RMSE are calculated on each class separately ("Destabilizing"- "Neutral"- "Stabilizing"), on the whole dataset ("Total") and only on the destabilizing and stabilizing variants, excluding the neutral ("Non-neutral").

Dataset	Pearson/RMSE				
	Total	Destabilizing	Neutral	Stabilizing	Non-neutral
MAESTRO	0.50/1.44	0.42/1.46	-0.01/0.84	0.28/2.26	0.48/1.63
ACDC-NN	0.46/1.49	0.34/1.60	0.09/0.69	0.09/2.14	0.44/1.71
INPS3D	0.43/1.50	0.35/1.40	0.03/1.00	0.02/2.55	0.42/1.67
DDGun3D	0.43/1.60	0.32/1.69	0.13/0.94	0.13/2.22	0.41/1.80
INPS-Seq	0.43/1.52	0.26/1.56	0.10/0.92	0.13/2.25	0.42/1.70
ACDC-NN-Seq	0.42/1.53	0.28/1.64	0.08/0.76	0.07/2.18	0.40/1.75
PremPS	0.41/1.51	0.43/1.48	-0.02/0.84	-0.08/2.53	0.04/1.72
PopMusic	0.41/1.51	0.37/1.40	0.11/0.96	0.09/2.67	0.39/1.69
DUET	0.41/1.52	0.34/1.48	0.02/0.89	0.10/2.54	0.38/1.72
Dynamut	0.41/1.60	0.32/1.81	0.16/0.66	0.29/2.00	0.40/1.85
SDM	0.41/1.67	0.33/1.81	0.16/1.01	0.09/2.14	0.40/1.88
DDGun	0.40/1.75	0.25/1.75	0.12/1.29	0.11/2.46	0.39/1.90
SAAFEC-Seq	0.36/1.54	0.31/1.48	0.03/0.87	0.07/2.60	0.34/1.74
mCSM	0.36/1.54	0.30/1.42	-0.01/0.96	0.06/2.73	0.33/1.73
I-Mutant3.0	0.36/1.54	0.31/1.48	0.03/0.87	0.07/2.60	0.34/1.74
I-Mutant3.0-Seq	0.34/1.56	0.23/1.53	0.05/0.92	0.21/2.53	0.33/1.75
MuPro	0.25/1.61	0.19/1.45	0.08/1.08	-0.01/2.84	0.24/1.78
FoldX	0.21/2.32	0.20/2.25	0.01/2.28	0.17/2.66	0.24/2.33

size of the smallest class of variants on S4428. For each feature, the average Pearson's correlation coefficient and its standard deviation are reported in Table 3 (column "Total balanced").

We also evaluated the impact of the seven features in terms of their discriminative power using the Receiver Operating Characteristic (ROC) curve and its Area-Under-the-Curve (AUC-ROC) metric. We removed the size effect by under-sampling the variants 100 times in order to have 100 subsets of 656 variants for each pair of classes. On these under-sampled datasets we used each feature to calculate three

different ROC curves separating three pairs of classes: Destabilizing-Neutral, Destabilizing-Stabilizing, Neutral-Stabilizing. Here, the assumption is that the higher the AUC scores, the better the separation of the two classes by the variable and, therefore, the more informative the variable is. The results of this analysis are shown in Table 4 and in the Supplementary Materials. The ROC curves and AUC scores of the logarithm of the conservation ratio, the volume difference and the difference in hydrophobicity were calculated using the values with opposite signs to help the interpretation. In Table 4, we also

TABLE 3 Pearson correlation coefficient (ρ) between the experimental $\Delta\Delta G$ s and the seven considered features in the S4428 dataset. On the column “Total”, we have the correlation in the whole unbalanced dataset, while on the column “Total Balanced” we reported the average Pearson’s correlation coefficient and its standard deviation calculated in 100 random subsets with the same number of variants for each class (Stabilizing-Neutral-Destabilizing). The last three columns report the correlation on each class separately, considering all the possible variants.

	Pearson’s correlation coefficient				
	Total	Total balanced	Destabilizing	Neutral	Stabilizing
Accessibility	0.22	0.08 ± 0.02	0.28	0.02	−0.21
Bastolla-Vendruscolo potential	0.45	0.45 ± 0.03	0.36	0.13	0.18
Blosum evolutionary score	0.12	0.02 ± 0.02	0.26	−0.02	−0.18
Difference in hydrophobicity	−0.21	−0.18 ± 0.03	−0.19	−0.04	−0.04
Volume difference	−0.32	−0.34 ± 0.03	−0.22	−0.12	−0.15
Logarithm of the conservation ratio	−0.37	−0.37 ± 0.02	−0.27	−0.11	−0.21
Skolnick potential	0.36	0.35 ± 0.03	0.29	0.07	0.13

TABLE 4 Ability of each feature to separate the different classes and differences in distributions. We extracted one hundred subsets of size 328 from each class and used the variable score to calculate three different ROC curves: Destabilizing-Neutral, Destabilizing-Stabilizing, Neutral-Stabilizing. The assumption, here, is that the higher the AUC-ROC, the better the variable can separate between the two classes and the more informative it is. For each pair of classes we also computed the Mann-Whitney-Wilcoxon test two-sided to establish the statistical significance of the differences in the distributions.

AUC-ROC scores and p -values						
	Destabilizing-neutral		Destabilizing-stabilizing		Neutral-stabilizing	
	AUC ± std	p -values	AUC ± std	p -values	AUC ± std	p -values
Accessibility	0.69 ± 0.02	1.98e-79	0.55 ± 0.02	1.43e-04	0.36 ± 0.01	1.18e-23
Bastolla-Vendruscolo potential	0.66 ± 0.02	1.40e-58	0.75 ± 0.01	8.50e-84	0.63 ± 0.02	1.26e-22
Blosum evolutionary score	0.58 ± 0.02	5.61e-17	0.50 ± 0.02	7.08e-01	0.42 ± 0.02	1.21e-09
Difference in hydrophobicity	0.59 ± 0.02	9.41e-22	0.60 ± 0.02	2.28e-16	0.51 ± 0.02	4.22e-01
volume difference	0.60 ± 0.02	6.90e-24	0.72 ± 0.02	1.24e-63	0.64 ± 0.02	1.02e-22
Logarithm of the ratio of the conservation	0.66 ± 0.02	1.00e-58	0.75 ± 0.01	1.32e-84	0.63 ± 0.02	1.29e-20
Skolnick potential	0.64 ± 0.02	2.56e-41	0.71 ± 0.02	4.27e-63	0.60 ± 0.02	4.97e-14

reported the p -values of the two-sided Mann-Whitney Wilcoxon test computed between the distributions of the scores for each pair of classes.

2.5 Training a predictor with the reduced set of features

To evaluate the combined effect of the Blosum evolutionary score, difference in hydrophobicity and accessibility on the predictions of the different classes, we trained a Random Forest regressor on 100 random balanced subsets of S4428 (excluding S669) and then tested its performance on S669 (Table 5). We used three different sets of features: one with all the seven variables (“full”), a reduced set with all the variables except for Blosum evolutionary score and difference in hydrophobicity (“reduced”) and one where we also excluded the accessibility (“red-no-acc”). Therefore, we tested to which extent the removal of the evolutionary-based component, the hydrophobicity and the accessibility affects DDGun3D predictions. The original DDGun3D score was computed as:

$$\text{DDGun3D}_{\text{full}} = \left(0.20 \cdot S_{Bl} + 0.29 \cdot S_{Sk} + 0.18 \cdot S_{Hp} + 0.33 \cdot S_{BV} \right) \cdot (1.1 - acc), \quad (6)$$

where S_{Bb} , S_{Sk} , S_{BV} , S_{Hp} , acc are, respectively, the components related to the Blosum evolutionary score, the Skolnick and the Bastolla-Vendruscolo potentials, the difference in hydrophobicity and the accessibility, with the same coefficient and definition as in the original paper Montanucci et al. (2019a). We defined two “reduced” scores by dropping the corresponding components:

$$\begin{aligned} \text{DDGun3D}_{\text{reduced}} &= (0.29 \cdot S_{Sk} + 0.33 \cdot S_{BV}) \cdot (1.1 - acc), \\ \text{DDGun3D}_{\text{red-no-acc}} &= (0.29 \cdot S_{Sk} + 0.33 \cdot S_{BV}). \end{aligned} \quad (7)$$

3 Results

We divided the results into four sections: the first shows the methods’ performance on the different variant classes on S669, a blind testing set, the second explores the correlation of the seven features

TABLE 5 Predictions with different sets of features. We tested to which extent the removal of the Blosum component, the hydrophobicity and the accessibility affects the performance of a Random Forest regressor and DDGun3D on S669. We used three different sets of features: one with all the seven variables, one with all the variables except for Blosum evolutionary score and the difference in hydrophobicity (“reduced”) and one where we also excluded the accessibility (“red-no-acc”). Results are reported in terms of Pearson correlation coefficient and RMSE.

	Pearson/RMSE		
	Total	Destabilizing	Stabilizing
DDGun3D _{full}	0.43/1.6	0.32/1.69	0.13/2.22
DDGun3D _{reduced}	0.37/1.69	0.29/1.96	0.17/1.96
DDGun3D _{red-no-acc}	0.34/1.75	0.26/2.02	0.15/2.0
Random Forest _{full}	0.42/1.56	0.28/1.70	0.15/2.09
Random Forest _{reduced}	0.40/1.58	0.27/1.72	0.16/2.15
Random Forest _{red-no-acc}	0.4/1.58	0.26/1.71	0.19/2.13

with the experimental $\Delta\Delta G$ values, the third section investigates their discriminative power and the fourth section analyze their combined impact on the predictions.

3.1 The unbalanced predictions among stabilizing, neutral and destabilizing variants

In a previous study (Pancotti et al., 2022) we assessed the performance of 18 $\Delta\Delta G$ prediction methods on the S669 dataset, a test dataset with no intersection with the methods’ training sets and whose variants are in proteins with less than 25% of sequence identity with those in the training sets. We used both direct and reverse variants (see Introduction) to show that, when dealing with never-before-seen variants, those methods that do not respect the antisymmetry property have a worse performance on stabilizing variants, while the antisymmetric ones performed consistently well on both classes.

In this study, we focused uniquely on direct variants and we showed that the difference in the non-antisymmetric methods’ performance between destabilizing and stabilizing variants is even bigger when we do not consider the reverse variants. In addition, even antisymmetric methods showed a highly uneven performance between the classes (Table 2). The correlations on the whole dataset and on the non-neutral variants (destabilizing and stabilizing) were relatively good, with most methods having $\rho \geq 0.4$, showing that the predictions captured the general trend of the experimental $\Delta\Delta G$ s. Most methods also predicted relatively well the destabilizing variants, with a $\rho \geq 0.3$. None of the methods, however, was able to predict the stabilizing variants with a good correlation. The highest Pearson’s correlation on the stabilizing variants was ≤ 0.3 and the lowest RMSE was ≥ 2 .

This result confirms the difficulty of current methods in adequately recognizing the direct-stabilizing variants.

3.2 Relevance of the seven features: Correlation on S4428

We assessed the relevance of the seven features of interest by computing their Pearson’s correlation coefficient (ρ) with the experimental $\Delta\Delta G$ s

considering the S4428 dataset (Table 3). When considering the whole dataset, all the features showed absolute correlations ranging from $\rho = 0.12$, observed with the Blosum evolutionary score, to $\rho = 0.45$, reached by the Bastolla-Vendruscolo potential (Table 3, column “Total”).

However, when the different classes were balanced by randomly sampling 100 balanced subsets of $\sim 1,000$ variants (column = “Total balanced”), the correlation dropped significantly for the Blosum evolutionary score and for the accessibility, since these two features are correlated with the destabilizing variants ($\rho = 0.26$ and $\rho = 0.28$, respectively) but anti-correlated with the stabilizing ones ($\rho = -0.18$ and $\rho = -0.21$, respectively). The reason for the anti-correlation of the Blosum score with stabilizing variants and the correlation with the destabilizing ones is the symmetry of the score, since $B(W, M) = B(M, W)$. On the other hand, for the accessibility this happens because, in general, the more one amino acid is buried, the greater impact its mutation has on the stability of the protein (either stabilizing or destabilizing). Therefore, this feature should be used to modulate the impact of the mutation, not to predict the sign.

The anti-correlation observed between the experimental $\Delta\Delta G$ and the logarithm of the conservation ratio is coherent with the assumption that the substitution of an amino acid with a less conserved one ($\log(\frac{CONS_{Sw}}{CONS_M}) > 0$) will likely have a disruptive effect, while substituting it with a more conserved one ($\log(\frac{CONS_{Sw}}{CONS_M}) < 0$) will likely have a stabilizing effect. The volume difference and the hydrophobicity between the two amino acids involved were also anti-correlated with the experimental $\Delta\Delta G$, meaning that a big absolute volume difference is associated with a greater effect, with a positive difference (i.e., mutated amino acid larger than the wild-type) being disruptive and a negative difference (i.e., wild-type greater than the mutated) being stabilizing. For hydrophobicity, while there seems to be a general trend (Figure 3, $\rho = -0.21$ on the whole dataset), neutral and stabilizing variants showed little correlations ($\rho = -0.04$ for both).

3.3 Discriminative power of the seven features: AUC-ROC on S4428

We also evaluated the discriminative power of the seven features using the AUC-ROC metric. We assumed that a specific feature is informative and useful for the prediction tools if it is able to separate the three pairs of classes (Destabilizing-Neutral, Destabilizing-Stabilizing, Neutral-Stabilizing). The higher the AUC-ROC scores for these separations, the more informative the feature is.

To remove the size effect, we under-sampled the variants 100 times in order to have 100 subsets of 656 variants for each pair of classes. In addition, due to their anti-correlation with the $\Delta\Delta G$ s, the ROC curves of the logarithm of the conservation ratio, the volume difference and the difference in hydrophobicity were calculated using the values with opposite signs to respect the monotonic assumption that higher AUC-ROC means higher discriminative power, making them more immediately interpretable to the reader. In this way, instead of having a score of e.g., 0.3, the score would become $1-0.3 = 0.7$. All the distributions of the features and the statistical differences between the classes computed using the two-sided Mann-Whitney Wilcoxon test are displayed in Figure 3 and in Table 4. The average AUC scores with their standard deviations are reported in Table 4, while the ROC curves are displayed in the Supplementary Materials.

The boxplots clearly show that none of the features perfectly separates all the three classes (Figure 3).

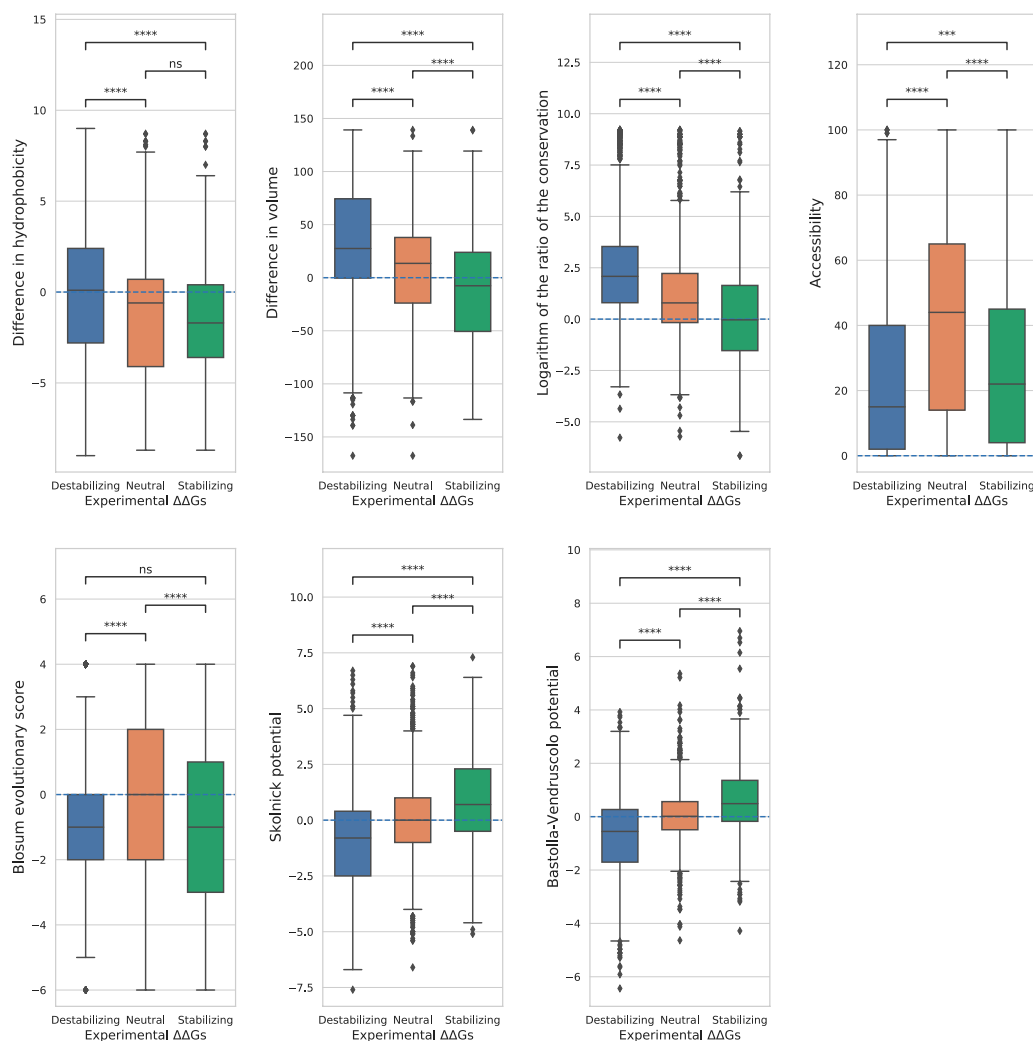


FIGURE 3

Distributions of the features. Boxplots showing the distributions of the features on the three classes. The variations are considered neutral if $\Delta\Delta G \in [-0.5, 0.5]$, stabilizing if $\Delta\Delta G < -0.5$ and destabilizing if $\Delta\Delta G > 0.5$. For each pair of classes we computed the Mann-Whitney-Wilcoxon test two-sided to establish the difference in the distributions. The p -values are reported here in a compact way: "ns" - $p > 0.05$, * - $0.01 < p \leq 0.05$, ** - $1.0e-03 < p \leq 0.01$, *** - $1.0e-04 < p \leq 1.0e-03$, **** - $p \leq 1.0e-04$, the actual values of the p -values are in Tab.4.

Overall, we found that the features with the highest absolute Pearson's correlation coefficients on the balanced datasets (Table 3, "Total balanced") were the best at separating between destabilizing and stabilizing variants, while those with a poor correlation also showed a poor discriminative power. The AUC scores "Destabilizing-Stabilizing" of the Bastolla-Vendruscolo potential, the Skolnick potential, the logarithm of the conservation ratio and the volume difference were all greater than 0.7, reflecting their high correlations on the balanced dataset. On the other hand, the accessibility and the difference in hydrophobicity, which are characterized by a low correlation on the balanced datasets (0.08 and -0.18 , respectively), also showed low AUCs when separating destabilizing variants from stabilizing (0.55 and 0.6, respectively). The Blossum evolutionary score, which was uncorrelated on the balanced dataset, showed also random behaviour ($AUC = 0.5$) in separating the destabilizing from the stabilizing variants and the two distributions were not significantly different (p -value of Mann-Whitney test = 0.71). A non-significant p -value ($p = 0.4$) was also observed for the difference in hydrophobicity between stabilizing and neutral variants. This feature,

while being able to slightly separate the destabilizing variants from the other two classes ($AUC = 0.59$ and $AUC = 0.6$), was not able to separate stabilizing from neutral variants ($AUC = 0.51$).

The accessibility separated fairly well the neutral variants from the other two classes, but not the destabilizing from the stabilizing (Figure 3), suggesting that it should only be used as a modulator of the $\Delta\Delta G$ absolute value. For the other best performing features, the two potentials and the conservation showed slightly higher similarity between neutral and stabilizing variants than between neutral and destabilizing, while the volume difference showed an opposite trend (Figure 3).

3.4 Improving the predictions on the stabilizing variants

We evaluated the effects of including or excluding the Blossum evolutionary score, the difference in hydrophobicity and the accessibility in a Random Forest predictor and in DDGun3D. The

aim of this analysis was not to outperform existing methods, but to analyze how the combination of these variable affects the predictions on the different classes.

Table 5 shows the results obtained by these two predictors when using three possible sets of features: “all variables”, “reduced”, “red-no-acc”. “Reduced” includes all variables except for Blosum evolutionary score and difference in hydrophobicity, while “red-no-acc” also excludes the accessibility.

The results show that, excluding the Blosum evolutionary score and difference in hydrophobicity, the correlation decreases on the destabilizing class for both methods: $\rho = 0.32$ for DDGun3D and $\rho = 0.28$ for Random Forest with “all variables”, compared to $\rho = 0.29$ and 0.27 with the “reduced” set. The same behaviour in the two predictors was observed when the “red-no-acc” set was considered. Given the high unbalance in S669 towards destabilizing mutations, with 387 variants being destabilizing and only 85 being stabilizing, the decreasing performance on the destabilizing variants affects the overall correlation on the whole dataset too.

Removing the Blosum evolutionary score and the difference in hydrophobicity, however, increases the correlation on the stabilizing class, as expected. In addition, the correlation increases when we also remove the accessibility from the Random Forest predictor’s features ($\rho = 0.19$ vs. $\rho = 0.15$), but not when DDGun3D is used ($\rho = 0.15$ vs. $\rho = 0.13$). However, it is worth noticing that, while in the Random Forest predictor the accessibility is used as any other feature, in DDGun3D it is used as modulator of the $\Delta\Delta G$ (see Eqs 6, 7). Indeed, using the Random Forest predictor, the additional removal of the accessibility improves the performance with respect to the removal of only the Blosum evolutionary score and the difference in hydrophobicity ($\rho = 0.19$ vs. $\rho = 0.16$), while for DDGun3D this additional removal negatively affects the correlation ($\rho = 0.15$ vs. $\rho = 0.17$).

4 Discussion and conclusion

Our study is based on the observation that none of the tools available to predict the $\Delta\Delta G$ is very accurate on the stabilizing variants and, in general, the predictions are skewed towards neutral and destabilizing variations. The existing datasets for $\Delta\Delta G$ prediction share a large number of variants and are strongly unbalanced towards the destabilizing variants (Table 1; Figure 2). For this reason, choosing features that favour only the most abundant class can lead to a good overall performance at the cost of penalizing the prediction of the less abundant class.

To better understand the weakness in correctly predicting the stabilizing variants, we evaluated seven properties commonly used by the computational $\Delta\Delta G$ predictors. We considered two features based on physical properties (hydrophobicity and volume), three structural information and conservation-based features (Blosum evolutionary score, conservation and relative solvent accessibility) and two statistical potentials-based features (Skolnick potential and Bastolla-Vendruscolo potential). For each of them, we analyzed the ability to predict the experimental $\Delta\Delta G$ by computing the Pearson’s correlation coefficient (ρ) between them and the experimental $\Delta\Delta G$ s of three of the most used dataset (S2648, VariBench and Ssym) and a recently-released one (S669), all combined in the S4428 dataset. We also computed the AUC-ROCs for each variant to judge how these features separate the different classes of variations. The results showed that the volume difference, the logarithm of the conservation ratio, and the statistical

potentials are better than random in each possible separation, i.e., destabilizing vs neutral, destabilizing vs stabilizing and neutral vs stabilizing variants. On the other hand, the difference in hydrophobicity, the Blosum evolutionary score and the accessibility likely showed random results in at least one of the separations. Although hydrophobicity difference has an anti-correlation trend in the balanced dataset ($\rho = -0.18$), it cannot separate the neutral variants from the stabilizing ones (AUC-ROC = 0.51 ± 0.02), and it has marginal ability to separate destabilizing from neutrals (AUC-ROC = 0.59 ± 0.02). Thus, hydrophobicity difference alone can contribute to entangling stabilizing and neutral variants at prediction time. In turn, this may lead to enforce the role of the destabilizing variants during the method training.

Due to the symmetry of the Blosum evolutionary score, it is not surprising that the AUC-ROC in separating destabilizing and stabilizing variants was 0.5 ± 0.02 . However, the score also failed to separate the other two classes from the neutral ones (AUC-ROCs = $0.58/0.42 \pm 0.02$), which could lead to predictions skewed towards the neutral class.

The accessibility is actually fairly good at separating the neutral from the other two classes (AUC-ROCs = $0.69/0.36 \pm 0.02$), but not the destabilizing from the stabilizing (AUC-ROC = 0.55 ± 0.02). Accessibility is then a good marker of the general impact of a variation (being them stabilizing or destabilizing) since its absolute value of ρ on the stabilizing variants ($|\rho| = 0.21$) is one of the highest among the considered features.

Among the seven features, the logarithm of the conservation ratio best described the stabilizing variants ($\rho = -0.21$), consolidating the known important role of the conservation in the impact of the genetic variants. Moreover, the logarithm of the conservation ratio and the volume difference showed the smallest drop in correlation between destabilizing and stabilizing variants (Tab.3). The other two best performing features, the Bastolla-Vendruscolo potential and the Skolnick potential, showed a bigger drop in correlation, even though the Bastolla-Vendruscolo potential by itself reached a correlation of $\rho = 0.45$ on the full dataset, which is not that far off from the results that many predictors get using way more features and information.

We tested the impact of removing Blosum62, hydrophobicity and accessibility (the three worst features at separating the stabilizing variants from the rest) on prediction by examining two models: DDGun3D, a mostly linear $\Delta\Delta G$ predictor with limited interaction between features and a Random Forest, a fully non-linear generic method. Independently from the predictor used, dropping out Blosum62 and hydrophobicity improves the performance on stabilizing variants at the cost of a loss on destabilizing ones.

This is not surprising for a linear model where each feature has a fixed and independent effect on all variants. On the other hand, a more powerful method like Random Forest, that can model interactions and varying non-linear effects, could be able to use these features (Blosum62 and hydrophobicity) selectively only for the non-stabilizing variants for which they are informative. However, our tests suggest that, in practice, including them can be harmful for stabilizing variant prediction regardless of the model.

On the other hand, the behaviour of the accessibility is more complex since the less accessible residues tend to have a higher impact than the accessible ones. In some way, accessibility can modulate the stability effect as, without it, the performance on the stabilizing variants improves in the Random Forest model but worsens in DDGun3D when it is not used as a shrinking effect (Table 5). This

result also agrees with previous studies that led to the tool PopMusic2 (Dehouck et al., 2011).

Our results suggest not incorporating the substitution matrix score (e.g., Blosom62) and the difference in hydrophobicity in future predictive computational tools, while using the difference in accessibility only to modulate the impact of the variant. Furthermore, we suggest using the features that better separate the stabilizing from the neutral variants (such as the Bastolla-Vendruscolo potential, the logarithm of the conservation ratio and the volume difference) to avoid the compression of the predictions towards the neutral values. Finally, new tools should also be tested on not-artificially balanced datasets, reporting the performance specifically for each variant class.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://academic.oup.com/bib/article/23/2/bbab555/6502552#supplementary-data>.

Author contributions

TS, PF, EC, and GB conceptualized the research. SB and GB prepared the data. SB ran the analyses and drafted the manuscript. All the authors contributed to the final manuscript.

Funding

This work was supported by the PRIN project, “Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant

References

- Bastolla, U., Farwer, J., Knapp, E. W., and Vendruscolo, M. (2001). How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins Struct. Funct. Bioinforma.* 44, 79–96. doi:10.1002/prot.1075
- Benevenuta, S., and Fariselli, P. (2019). On the upper bounds of the real-valued predictions. *Bioinform Biol. Insights* 13, 1177932219871263. doi:10.1177/1177932219871263
- Benevenuta, S., Pancotti, C., Fariselli, P., Birolo, G., and Sanavia, T. (2021). An antisymmetric neural network to predict free energy changes in protein variants. *J. Phys. D Appl. Phys.* 54, 245403. doi:10.1088/1361-6463/abedfb
- Birolo, G., Benevenuta, S., Fariselli, P., Capriotti, E., Giorgio, E., and Sanavia, T. (2021). Protein stability perturbation contributes to the loss of function in haploinsufficient genes. *Front. Mol. Biosci.* 8, 620793. doi:10.3389/fmolb.2021.620793
- 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. doi:10.1093/nar/gki375
- Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinforma.* 9, S6–S9. doi:10.1186/1471-2105-9-S2-S6
- Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Luigi Martelli, P. (2011). Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. *Hum. Mutat.* 32, 1161–1170. doi:10.1002/humu.21555
- Chen, Y., Lu, H., Zhang, N., Zhu, Z., Wang, S., and Li, M. (2020). Premps: Predicting the impact of missense mutations on protein stability. *PLoS Comput. Biol.* 16, e1008543. doi:10.1371/journal.pcbi.1008543
- Cheng, J., Randall, A., and Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Bioinforma.* 62, 1125–1132. doi:10.1002/prot.20810
- Cheng, T. M., Lu, Y.-E., Vendruscolo, M., Liò, P., Blundell, T. L., et al. (2008). Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput. Biol.* 4, e1000135. doi:10.1371/journal.pcbi.1000135
- Compiani, M., and Capriotti, E. (2013). Computational and theoretical methods for protein folding. *Biochemistry* 52, 8601–8624. doi:10.1021/bi4001529
- Dehouck, Y., Kwasigroch, J. M., Gilis, D., and Rooman, M. (2011). Popmusic 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinforma.* 12, 151. doi:10.1186/1471-2105-12-151
- Hartl, F. U. (2017). Protein misfolding diseases. *Annu. Rev. Biochem.* 86, 21–26. doi:10.1146/annurev-biochem-061516-044518
- Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89, 10915–10919. doi:10.1073/pnas.89.22.10915
- Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi:10.1002/bip.360221211
- 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 Struct. Funct. Bioinforma.* 79, 830–838. doi:10.1002/prot.22921
- Kumar, M. D., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., et al. (2006). ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* 34, D204–D206. doi:10.1093/nar/gkj103
- Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132. doi:10.1016/0022-2836(82)90515-0
- Laimer, J., Hiebl-Flach, J., Lengauer, D., and Lackner, P. (2016). Maestroweb: A web server for structure-based protein stability prediction. *Bioinformatics* 32, 1414–1416. doi:10.1093/bioinformatics/btv769

Interpretation” of the Ministero Istruzione, Università e Ricerca (PRIN201744NR8S).

Acknowledgments

We thank the EU projects GenoMed4All (H2020 RIA n.101017549) and BRAINTEASER (H2020 RIA n.101017598). We also thank PNRR M4C2 HPC—1.4 “CENTRI NAZIONALI”- Spoke 8 for fellowship support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2022.1075570/full#supplementary-material>

- 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. doi:10.1371/journal.pcbi.1008291
- 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. doi:10.3390/ijms22020606
- Martelli, P. L., Fariselli, P., Savojardo, C., Babbi, G., Aggazio, F., and Casadio, R. (2016). Large scale analysis of protein stability in omim disease related human protein variants. *BMC genomics* 17, 397–247. doi:10.1186/s12864-016-2726-y
- Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N., and Fariselli, P. (2019a). Ddgun: An untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinforma.* 20, 335. doi:10.1186/s12859-019-2923-1
- Montanucci, L., Martelli, P. L., Ben-Tal, N., and Fariselli, P. (2019b). A natural upper bound to the accuracy of predicting protein stability changes upon mutations. *Bioinformatics* 35, 1513–1517. doi:10.1093/bioinformatics/bty880
- Nair, P. S., and Vihinen, M. (2013). V ari b ench: A benchmark database for variations. *Hum. Mutat.* 34, 42–49. doi:10.1002/humu.22204
- Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., et al. (2022). Predicting protein stability changes upon single-point mutation: A thorough comparison of the available tools on a new dataset. *Briefings Bioinforma.* 23, Bbab555. doi:10.1093/bib/bbab555
- Pancotti, C., Benevenuta, S., Repetto, V., Birolo, G., Capriotti, E., Sanavia, T., et al. (2021). A deep-learning sequence-based method to predict protein stability changes upon genetic variations. *Genes* 12, 911. doi:10.3390/genes12060911
- Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014a). Duet: A server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids Res.* 42, W314–W319. doi:10.1093/nar/gku411
- Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014b). mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342. doi:10.1093/bioinformatics/btt691
- Pires, D. E., Chen, J., Blundell, T. L., and Ascher, D. B. (2016). *In silico* functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.* 6, 19848. doi:10.1038/srep19848
- 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. doi:10.1093/bioinformatics/bty348
- Rodrigues, C. H. M., Pires, D. E. V., and Ascher, D. B. (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.* 30, 60–69. doi:10.1002/pro.3942
- Rodrigues, C. H., Pires, D. E., and Ascher, D. B. (2018). Dynamut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic acids Res.* 46, W350–W355. doi:10.1093/nar/gky300
- Samaga, Y. B., Raghunathan, S., and Priyakumar, U. D. (2021). Scones: Self-consistent neural network for protein stability prediction upon mutation. *J. Phys. Chem. B* 125, 10657–10671. doi:10.1021/acs.jpcc.1c04913
- 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. Biotechnol. J.* 18, 1968–1979. doi:10.1016/j.csbj.2020.07.011
- Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2016). Inps-md: A web server to predict stability of protein variants from sequence and structure. *Bioinformatics* 32, 2542–2544. doi:10.1093/bioinformatics/btw192
- Savojardo, C., Martelli, P. L., Casadio, R., and Fariselli, P. (2019). On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform* 22, 601–603. doi:10.1093/bib/bbz168
- Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The foldx web server: An online force field. *Nucleic acids Res.* 33, W382–W388. doi:10.1093/nar/gki387
- Skolnick, J., Godzik, A., Jaroszewski, L., Kolinski, A., and Godzik, A. (1997). Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct? *Protein Sci.* 6, 676–688. doi:10.1002/pro.5560060317
- Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P., et al. (2015). A series of pdb-related databanks for everyday needs. *Nucleic acids Res.* 43, D364–D368. doi:10.1093/nar/gku1028
- Worth, C. L., Preissner, R., and Blundell, T. L. (2011). Sdm—A server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids Res.* 39, W215–W222. doi:10.1093/nar/gkr363
- Xavier, J. S., Nguyen, T. B., Karmarkar, M., Portelli, S., Rezende, P. M., Velloso, J. P. L., et al. (2021). ThermoMutDB: A thermodynamic database for missense mutations. *Nucleic Acids Res.* 49, D475–D479. doi:10.1093/nar/gkaa925
- Yue, P., Li, Z., and Moulton, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353, 459–473. doi:10.1016/j.jmb.2005.08.020
- Zamyatnin, A. (1972). Protein volume in solution. *Prog. biophysics Mol. Biol.* 24, 107–123. doi:10.1016/0079-6107(72)90005-3