OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–12
https://doi.org/10.1093/bib/bbab555
Problem Solving Protocol

Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset

Corrado Pancotti†, Silvia Benevenuta†, Giovanni Birolo†, Virginia Alberini, Valeria Repetto, Tiziana Sanavia, Emidio Capriotti and Piero Fariselli

Corresponding authors: Emidio Capriotti, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Via F. Selmi 3, 40126 Bologna, Italy. Tel.: +39 051 20 9 4303; Fax: +39 051 20 9 4286; E-mail: emidio.capriotti@unibo.it; Piero Fariselli, Department of Medical Sciences, University of Torino, Via Santena 19, 10126, Torino, Italy. Tel.: +39 011 6705871; E-mail: piero.fariselli@unito.it
†These authors contributed equally to this work.

Abstract

Predicting the difference in thermodynamic stability between protein variants is crucial for protein design and understanding the genotype-phenotype relationships. So far, several computational tools have been created to address this task. Nevertheless, most of them have been trained or optimized on the same and 'all' available data, making a fair comparison unfeasible. Here, we introduce a novel dataset, collected and manually cleaned from the latest version of the ThermoMutDB database, consisting of 669 variants not included in the most widely used training datasets. The prediction performance and the ability to satisfy the antisymmetry property by considering both direct and reverse variants were evaluated across 21 different tools. The Pearson correlations of the tested tools were in the ranges of 0.21–0.5 and 0–0.45 for the direct and reverse variants, respectively. When both direct and reverse variants are considered, the antisymmetric methods perform better achieving a Pearson correlation in the range of 0.51–0.62. The tested methods seem relatively insensitive to the physiological conditions, performing well also on the variants measured with more extreme pH and temperature values. A common issue with all the tested methods is the compression of the $\Delta\Delta G$ predictions toward zero. Furthermore, the thermodynamic stability of the most significantly stabilizing variants was found to be more challenging to predict. This study is the most extensive comparisons of prediction methods using an entirely novel set of variants never tested before.

Keywords: protein stability, single-point mutation, stability change, antisymmetry, machine learning

Introduction

The problem of predicting protein stability changes upon variation of a single residue is not a trivial task, and it is still affected by several experimental limitations and computational issues [1–5]. Understanding the impact of non-synonymous (or missense) DNA variations leading to the disruption or the enhancement of the protein function was shown to be fundamental for describing the molecular mechanisms of several human diseases [6–9]. Specific protein stability perturbations have already been associated with pathogenic missense variants [8, 10–12]. Variations were shown to contribute to the loss of function in haploinsufficient genes [13] and to modulate drug resistance in several diseases [14]. In this context, tools that can robustly predict the effects of variants on protein stability are crucial to infer their pathogenicity correctly.

The effects of non-synonymous variants on the protein stability are quantified in terms of the Gibbs free energy of unfolding (ΔG). The stability change from a mutated (M) protein to its wild-type (W) form is defined as the

Corrado Pancotti is a PhD student at the University of Torino, Italy. His research activity concerns machine learning in biomedical applications.
Silvia Benevenuta is a PhD student at the University of Torino, Italy. Her research activity is focused on machine learning methods in biomedical applications.
Giovanni Birolo is a postdoc at the Department of Medical Sciences of the University of Torino, Italy. His research interests are in bioinformatics, machine learning and biomedical data analysis.
Virginia Alberini is an undergraduate student in Medical Biotechnology at the University of Torino, Italy.
Valeria Repetto is a graduate student in Physics of Complex Systems at the University of Torino, Italy.
Tiziana Sanavia is an Assistant Professor in Applied Physics at the Department of Medical Sciences of the University of Torino, Italy. Her research activity is based on the development of statistical and machine learning methods for biomedical applications and the analysis of omics data from high-throughput sequencing platforms.
Emidio Capriotti is an Associate Professor at the University of Bologna, Italy. His research activities focuses on the development of methods for the analysis of macromolecular sequence/structure/function and for variant interpretation with biomedical applications.
Piero Fariselli is Full Professor at the Department of Medical Sciences of the University of Torino, Italy. His research interests include bioinformatics, machine learning, software development and modeling of biological systems.

Received: July 28, 2021. **Revised:** November 29, 2021. **Accepted:** December 5, 2021

© The Author(s) 2022. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

difference of the corresponding unfolding free energy $\Delta\Delta G$ of two proteins:

$$\Delta\Delta G_{MW} = \Delta G_M - \Delta G_W,$$

it is commonly measured in kcal/mol and the sign indicates whether the variation decreases (i.e. destabilizing variant) or increases (i.e. stabilizing variant) the protein stability. Contributions based on statistical potentials have been identified to highlight strong and weak residues in proteins on the basis of three types of statistical energy functions describing local interactions along the protein chains [15]. An important property of the $\Delta\Delta G$ is the antisymmetry, for which the change in Gibbs free energy for the corresponding reverse variation (i.e. the folding free energy from the wild-type to the mutated form) is equal and has the opposite sign with respect to the change for the direct variation: $\Delta\Delta G_{MW} = -\Delta\Delta G_{WM}$. However, most of the current prediction tools violate this property and they are highly biased toward predictions of destabilizing variants [2, 16–21]. Several computational tools have been developed so far, showing different levels of prediction performance [16, 18, 22]. However, a fair comparison of these tools has been problematic because they have been optimized on the same (or similar) manually cleaned datasets extracted from ProTherm [23, 24], the main repository collecting published experimental $\Delta\Delta G$ data. Another more recent and useful resource is ThermoMutDB, a new resource for thermodynamic data from protein mutations [25]. From the latest version (v1.3) of ThermoMutDB, we were able to identify 669 novel variants never seen by the state-of-the-art prediction tools. These variants only belong to proteins with less than 25% sequence identity with respect to those found in the most widely used datasets, such as S2648 [26] and VariBench [27]. The obtained results highlighted a group of methods performing equally well on both direct and reverse variants, which outperformed the other methods and they were able to achieve, on average, Pearson correlations slightly lower than 0.5. In addition, the performance of the sequence-based tools tended to be slightly lower than the other methods. Overall, we found that the stabilizing variants are harder to predict, indicating a potential issue affecting all the current approaches, including those that are antisymmetric by construction.

In this paper, we excluded the analysis of membrane protein variations because they reside in a different chemical environment. However, recent methods have been specifically developed to predict the stability change upon mutation for these proteins (CSM-membrane [28], MPTherm[29]) and they should be used in these cases.

Materials and methods

Datasets

From ThermoMutDB, we extracted 900 variants belonging to proteins having less than 25% sequence identity with those of S2648 [26] and VariBench [27], whose union includes variants from almost all the thermodynamic studies available in the literature. We therefore revised all the papers related to these 900 variants, and we decided to exclude about ∼24% of them due to different annotation inconsistencies in the dataset (e.g. free energies measured in terms of transition state kinetics, affinity binding, multiple variants, etc.). Of the remaining 669 variants, we changed the $\Delta\Delta G$s for ∼20% of the reported values, either because the sign of the $\Delta\Delta G$s was not coherent or because the values were imprecise. The final manually cleaned set S669 is released (Supplementary material).

In addition, in order to evaluate the antisymmetry of the different methods, we used an extended version of Ssym [18], which includes 10 more variants that we collected from ThemoMutDB [25], adding 10 new protein structures of the mutants. The new version of Ssym, here refereed as Ssym+ is also available (Supplementary material).

Finally, we generated the reverse variants for S669 to assess the antisymmetry in a context where no variants were seen by the methods. When required, the reverse structures were generated using the Robetta server [30] with default parameters by using the comparative modeling technique. This procedure led to an entirely balanced dataset of 1338 protein variants.

Evaluated methods

We predicted the $\Delta\Delta G$s on the S669 dataset with 21 different tools. Either web server (when available) or stand-alone versions were used with default parameters, as indicated in the following:

- **ACDC-NN** [31] and its sequence-based version **ACDC-NN-Seq** [32] (stand-alone tool): neural network-based methods whose architectures satisfy the antisymmetry properties by construction. They both take as input the local information from the amino acids in the neighbourhood of the mutation and they both use multiple sequence alignments considering the two amino acids involved in the mutation.
- **DDGun3D** and **DDGun** [33] (stand-alone tool): untrained methods that combine evolutionary information and statistical potentials to predict the $\Delta\Delta G$. Compared with the sequence-based DDGun, DDGun3D includes the structural information scored by the Bastolla–Vendruscolo statistical potential [34] and weights the linear combination through the accessibility of the mutated amino acid. They both include antisymmetric features and provide an easy extension to the prediction of multiple variations.
- **mCSM** [35] (web server): considers graph-based structural signatures, encoding for the distance patterns between atoms and used to represent the protein residue environment, to study and predict the impact of single-point mutations on the protein stability.

- **SDM** [36] (web server): statistical potential energy function that uses environment-specific amino acid substitution frequencies within homologous protein families to calculate a stability score as proxy of the free energy difference between the wild-type and mutant protein.
- **DUET** [37] (web server): web server implementing a meta-classifier based on the combined results from mCSM and SDM using support vector machines (SVM).
- **Dynamut and Dynamut2** [38, 39] (web server): machine learning methods implementing a consensus prediction. They combine the effects of mutations on protein stability and dynamics calculated by DUET, Bio3D and ENCoM to generate an optimized and more robust predictor.
- **FoldX** [40] (stand-alone tool): empirical force field-based method predicting the effect of a single-point variation through a linear combination of empirical free energy terms, including entropy contribution, Van der Walls forces, hydrogen bonds and electrostatic interactions.
- **SAAFEC-SEQ** [41] (web server): gradient boosting decision-tree machine learning method that uses physico-chemical properties, sequence features and evolutionary information to predict the $\Delta\Delta G$ values.
- **MUpro** [42] (web server): sequence-based SVM-based approach which considers the local mutation environment encoding the residues in a window centered on the target residue. The input corresponding to the deleted residue is set to −1 and the newly introduced residue to 1; all other inputs are set to 0.
- **Rosetta** [43] (stand-alone tool): a method based on structural modeling that computes the difference in Rosetta energy between the simulated wild-type versus the mutated structures.
- **ThermoNet** [44] (stand-alone tool): deep 3D-convolutional neural network designed for structure-based prediction of the $\Delta\Delta G$ values. Input protein structures are treated as if they were multi-channel 3D images, therefore by using multi-channel voxel grids based on biophysical properties derived from raw atom coordinates.
- **PremPS** [45] (web server): random forest regression-based method that uses evolutionary and structure-based features to make $\Delta\Delta G$ predictions. It has been trained on a balanced dataset with an equal number of stabilizing and destabilizing mutations to obtain unbiased predictions.
- **PoPMuSiC** [26] (web server): energy function-based method providing a linear combination of 13 statistical potentials, two volume-dependent terms of the wild-type and mutant amino acids, and an independent term. The coefficients depend on the solvent accessibility of the mutated residue, based on a sigmoid function whose parameters are optimized through a neural network.
- **MAESTRO** [46] (stand-alone tool): multi-agent prediction method based on statistical scoring functions (SSFs) and exploiting an ensemble of neural networks, support vector and multiple linear regressors, combined into a consensus model.
- **INPS3D** [47] and its sequence-based version **INPS** [21] (stand-alone tool): SVM-based methods using radial basis function kernel. Specifically, INPS uses the substitution score derived from the BLOSUM62 matrix, the difference in the alignment score between the native and variant sequences, hydrophobicity, evolutionary information and others; INPS3D also considers the relative solvent accessibility of the native residue and the difference between wild-type and mutated structures, scored by the Bastolla–Vendruscolo statistical potential [34].
- **I-Mutant** and its sequence-based version **I-Mutant-Seq** [48] (web server): SVM-based methods using radial basis function kernel with 42 features as input, including temperature, Ph, 20 features encoding for the mutations and 20 features encoding for the spatial residue environment when the protein structure is available or the nearest sequence neighbors when only the protein sequence is available.

The usage and availability of the different tools can be found in the Supplementary Materials.

Performance evaluation

Pearson correlation (indicated by r), root mean square error (RMSE) and mean absolute error (MAE) were estimated between predicted and observed $\Delta\Delta G$ values.

To assess the antisymmetric property of $\Delta\Delta G$ predictors, we adopted three previously defined index: r_{d-r}. r_{d-r} is the Pearson correlation coefficient between the direct and the corresponding reverse variations:

$$r_{d-r} = \frac{Cov(\Delta\Delta G^{dir}, \Delta\Delta G^{rev})}{\sigma_{dir}\sigma_{rev}}, \quad (1)$$

where Cov is the covariance and σ is the standard deviation.

Most of the predictors were trained using strongly unbalanced data toward the destabilizing group of variants. To measure the average bias toward a specific class, we adopted the **bias** score $\langle\delta\rangle$:

$$\langle\delta\rangle = \frac{\sum_{i=1}^{N}(\Delta\Delta G_i^{dir} + \Delta\Delta G_i^{rev})}{2N}. \quad (2)$$

A perfectly antisymmetric and unbiased method should have r_{d-r} equal to −1, whereas $\langle\delta\rangle$ equals to 0.

Results

Method performance on model versus experimental structures

The current data repositories and the derived datasets are skewed toward the destabilizing variants. Using the thermodynamic property of antisymmetry, we can

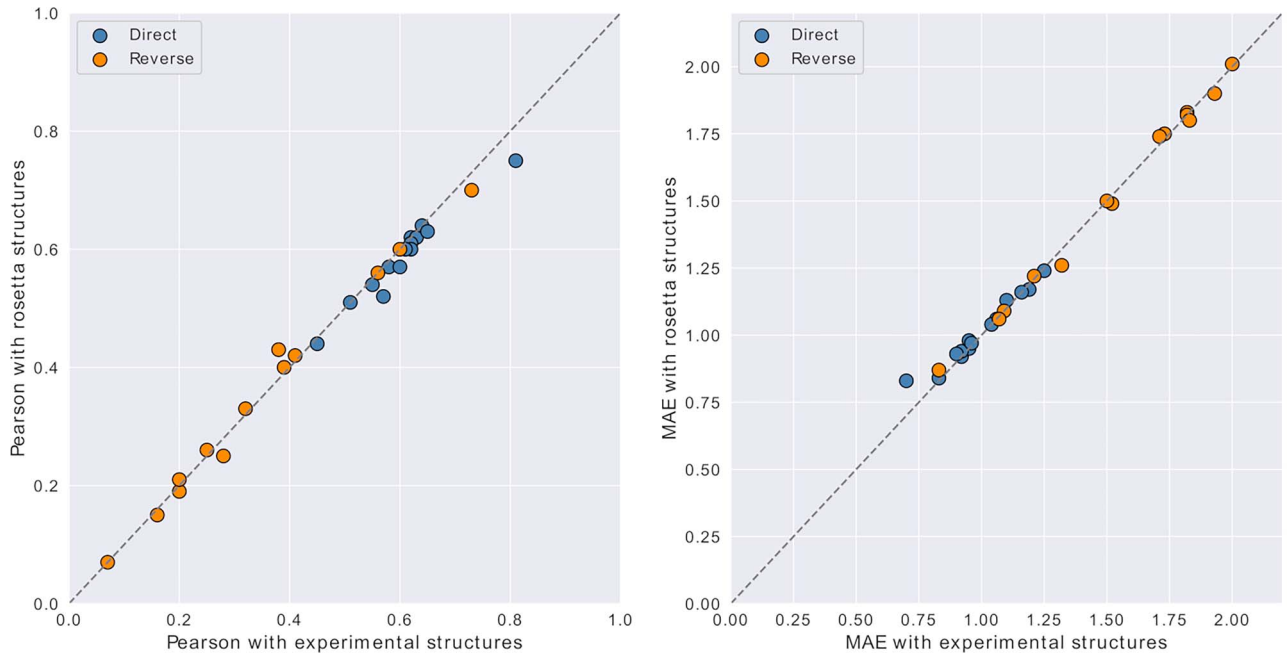

Figure 1. Comparison of method performance on real and modeled structure on the Ssym+ dataset Pearson correlation coefficients (*r*) and MAE are displayed in the left and in the right figure, respectively. The prediction performance obtained from the experimental structures (x-axis) is plotted against those from the Rosetta-simulated structures (y-axis). Performance calculated with real or modeled structures is very consistent, with correlations of 0.995 (P value $< 10^{-13}$) and 0.993 (P value $< 10^{-12}$) for the Pearson and MAE, respectively.

double the data and perfectly balance the distribution by adding the reverse variants. This procedure works smoothly for sequence-based methods, but structure-based methods require the atomic coordinates, and unfortunately, very few pairs of wild-type and mutated protein structures with experimental ΔΔGs are available.

The most significant effort in this direction has produced the Ssym dataset, including 684 variants (342 direct and 342 reverse) with 19 experimental structures for the direct variants and 342 experimental structures for each of the reverse variants [18]. From ThermoMutDB [25], we extracted 10 more reverse structures, which slightly increased the Ssym dataset (Ssym+ consists of 704 variants).

Another way to generate the reverse structure when the experimental one is not available is through comparative modeling. However, it is not clear if using a predicted model can hamper the predictive performance of the methods. To test the possibility of using single-point mutation models as reverse structures, we generated 704 protein models for each Ssym+ structure. Thus, a model of the direct protein is obtained from the corresponding reverse PDB structure (and *vice versa*). To compute the model structures, we used Rosetta/Robetta server.

To assess the generated models regarding the PDB structures, we performed 1408 predictions (704 for the experimental structures and 704 for the modeled ones) for each structure-based method. The comparison of the performance obtained for each method in the two scenarios (experimental versus modeled structures) is reported in Fig. 1 and in the Supplementary Materials. The results indicate that there is no performance degradation using the models as a proxy for the experimental structure. This finding supports the idea of balancing datasets by adding the reverse of all variants and modeling the missing mutated structures with Rosetta/Robetta.

Method performance on the new S669 dataset

To assess the generalization capability of different prediction methods of protein stability changes, we performed the analysis on S669, a dataset of never seen proteins and variants from proteins with less than 25% of sequence identity to previously studied proteins in manually curated datasets (S2648 and Varibench).

Figure 2 and Table 1 report the obtained performance (Pearson correlation, RMSE and MAE), the bias and the antisymmetry metric of each method. The highest correlations observed across all the methods are in the range 0.4–0.6 depending on the group of considered variants.

When direct variants are considered, the Pearson correlation of all the methods ranges from 0.2 to 0.5 (Fig. 2, circles in the central plot). These values are lower than those reported in the original papers but close to the expected performance for methods developed avoiding proteins with high sequence similarity in the training and testing sets to avoid overfitting. It is worth noticing that S669 can be considered an external validation set for the tested methods; thus, a performance drop can be expected.

When the reverse variants are considered (Fig. 2, crosses in the central plot), there is a first group of methods built to be antisymmetric (INPS-Seq, ACDC-NN-Seq, ACDC-NN, DDGun3D, DDGun, ThermoNet, PremPS),

Table 1. Assessment of the protein stability prediction tools on s669. Performance reported in terms of Pearson correlation coefficient (r), RMSE and MAE. The antisymmetry property was assessed in terms of Pearson correlation coefficient(r_{d-r}) and bias ($\langle\delta\rangle$), as described in Section 2.3. RMSE, MAE, $\langle\delta\rangle$, and $\langle\gamma\rangle$ are expressed in kcal/mol. The methods are ordered by r_{d-r}

	Total			Direct			Reverse			Antisimmetry/Bias	
Method	***r***	**RMSE**	**MAE**	***r***	**RMSE**	**MAE**	***r***	**RMSE**	**MAE**	r_{d-r}	$\langle\delta\rangle$
Structure-based											
ACDC-NN	0.61	1.5	1.05	0.46	1.49	1.05	0.45	1.5	1.06	–0.98	–0.02
DDGun3D	0.57	1.61	1.13	0.43	1.6	1.11	0.41	1.62	1.14	–0.97	–0.05
PremPS	0.62	1.49	1.07	0.41	1.5	1.08	0.42	1.49	1.05	–0.85	0.09
ThermoNet	0.51	1.64	1.2	0.39	1.62	1.17	0.38	1.66	1.23	–0.85	–0.05
Rosetta	0.47	2.69	2.05	0.39	2.7	2.08	0.4	2.68	2.02	–0.72	–0.61
Dynamut	0.5	1.65	1.21	0.41	1.6	1.19	0.34	1.69	1.24	–0.58	–0.06
INPS3D	0.55	1.64	1.19	0.43	1.5	1.07	0.33	1.77	1.31	–0.5	–0.38
SDM	0.32	1.93	1.45	0.41	1.67	1.26	0.13	2.16	1.64	–0.4	–0.4
PoPMuSiC	0.46	1.82	1.37	0.41	1.51	1.09	0.24	2.09	1.64	–0.32	–0.69
MAESTRO	0.44	1.8	1.3	0.5	1.44	1.06	0.2	2.1	1.655	–0.22	–0.57
FoldX	0.31	2.39	1.53	0.22	2.3	1.56	0.22	2.48	1.5	–0.2	–0.34
DUET	0.41	1.86	1.39	0.41	1.52	1.1	0.23	2.14	1.68	–0.12	–0.67
I-Mutant3.0	0.32	1.96	1.49	0.36	1.52	1.12	0.15	2.32	1.87	–0.06	–0.81
mCSM	0.37	1.96	1.49	0.36	1.54	1.13	0.22	2.3	1.86	–0.05	–0.85
Dynamut2	0.36	1.9	1.42	0.34	1.58	1.15	0.17	2.16	1.69	0.03	–0.64
Sequence-based											
INPS-Seq	0.61	1.52	1.1	0.43	1.52	1.09	0.43	1.53	1.1	–1	0
ACDC-NN-Seq	0.59	1.53	1.08	0.42	1.53	1.08	0.42	1.53	1.08	–1	0
DDGun	0.57	1.74	1.25	0.41	1.72	1.25	0.38	1.75	1.25	–0.96	–0.05
I-Mutant3.0-Seq	0.37	1.91	1.47	0.34	1.54	1.15	0.22	2.22	1.79	–0.48	–0.76
MUpro	0.32	2.03	1.58	0.25	1.61	1.21	0.2	2.38	1.96	–0.32	–0.95
SAAFEC-SEQ	0.26	2.02	1.54	0.36	1.54	1.13	–0.01	2.4	1.94	–0.03	–0.83

which perform significantly better, followed by Rosetta, Dynamut and INPS3D. On the other hand, the not-antisymmetric predictors (I-Mutant3.0-Seq, SDM, MUpro, PoPMuSiC, MAESTRO, FoldX, DUET, I-Mutant3.0, mCSM, Dynamut2, SAAFEC-SEQ) performed remarkably worse for the reverse variants, showing a strong bias toward the destabilizing class (negative values), as highlighted by the values reported in the last columns of Table 1 and in right-most bar plot of Fig. 2.

The majority of the methods improve when we consider the complete and balanced dataset (Fig. 2, squares in the central plot). This improvement is partially due to the increase of the $\Delta\Delta G$ distribution variance [49, 50]. However, the learnt thermodynamic property allows the antisymmetric methods to increase the performance.

Effect of the experimental technique on the method performance

One interesting point that has been recently studied is the possible dependence of the method performance in the choice of the protein structure [4]. Caldararu et al. [4] showed that some methods, such as FoldX, are more sensitive to the change of the three-dimensional protein structures. To test whether different experimental strategies have an impact on the performance of the structure-based $\Delta\Delta G$ predictors, we divided the S669 into variants from structures that were obtained by nuclear magnetic resonance (NMR) spectroscopy and structures obtained by X-ray diffraction. The protein structure solved using the NMR technique usually presents several models in the corresponding PDB file that are all compatible with the experimental constraints. As usually done, we selected the first model as representative. Figure 3 displays the method performance on the two subsets of variants. Largely overlapping error bars show that most methods are quite insensitive to experimental strategy, even though a general trend of slightly increased performance for NMR-derived structures can be observed. Only FoldX and PremPS showed a clear preference for X-ray- and NMR-derived structures, respectively. However, the observed differences are probably due to the variations in the NMR and X-ray sets rather than to the specific experimental technique.

Furthermore, the overall performance of the methods seems mostly unaffected by the X-ray resolution, at least in the range from 1.2 to 3.2 Angstrom seen in S669. Figure 4 displays the results obtained by splitting the X-ray structures in those that have been crystallized at a resolution above or below Ångstrom. The only methods that seem sensitive to the resolution are PremPS and Dynamut2 for the Pearson correlation and Rosetta for the MAE.

Surface accessibility, pH and Temperature

As already observed in previous studies [1, 26, 51], the residue accessibility impacts the method performance. Figure 5 shows the results for the variants classified by

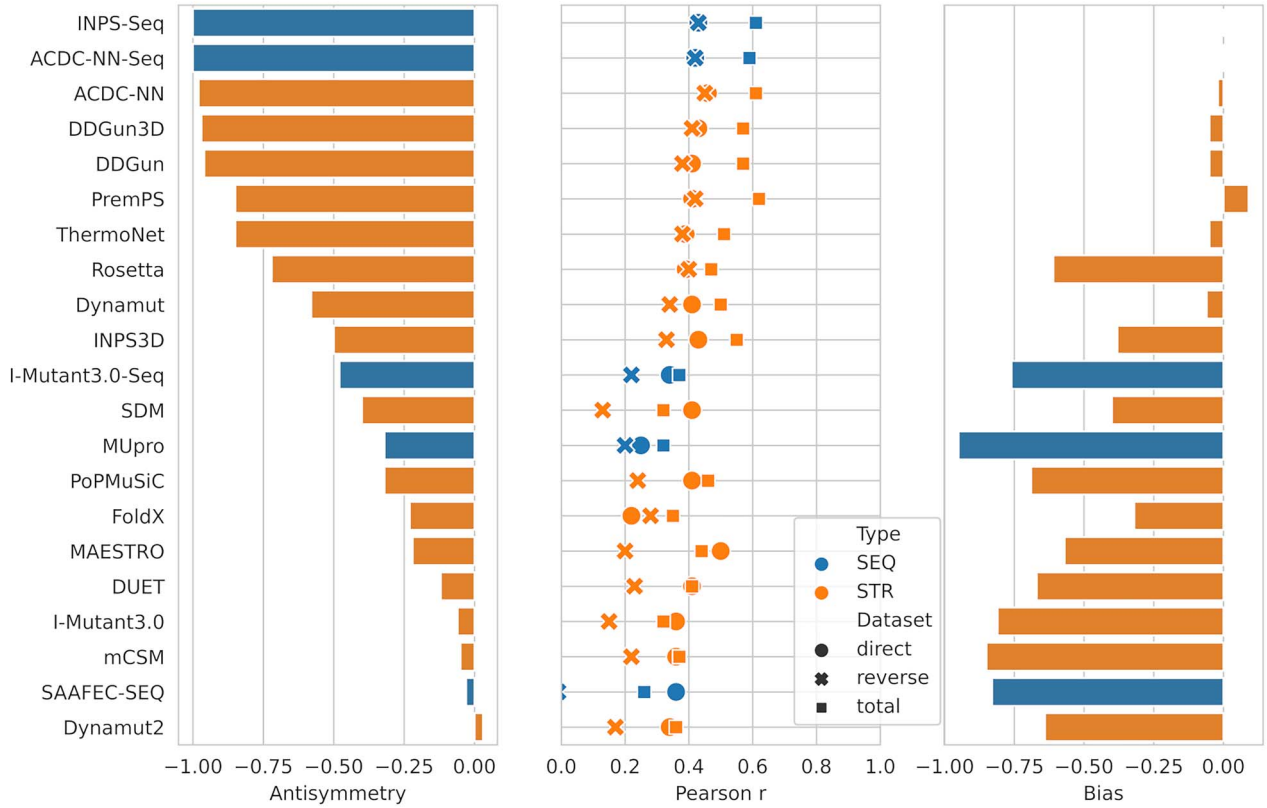

Figure 2. Antisymmetry, Pearson correlation and Bias for all predictors. From the left: antisymmetry expressed as the Pearson correlation between direct and reverse $\Delta\Delta G$ predictions, where perfect antisymmetry corresponds to −1; Pearson correlation of predicted with experimental $\Delta\Delta G$ values for the sets of direct, reverse and total (both direct and reverse) variants; bias expressed as the average of the predicted $\Delta\Delta G$ on the total (direct and reverse) dataset: since the average experimental $\Delta\Delta G$ on the total dataset is zero, unbiased predictors have also a bias of 0, whereas predictors biased toward destabilization have negative values. Color show which predictors need structural (3D) data and which use only sequence data. Predictors are sorted from the most antisymmetric (top) to the least (bottom).

their relative accessibility (RA) median value (buried = 0–24%), superficial = 24–100%). Most predictors (even sequence-based) show much lower Pearson correlations on surface residues, with the exception of FoldX, and to a lower extent PremPS and INPS3D. However, the MAE, which measures the distance between the predicted and observed $\Delta\Delta G$ values, are lower (better) on the surface residues. This means that the methods are able to recognize that the surface residues have a lower impact on stability and coherently predict $\Delta\Delta G$ values closer to zero. However, when values are close to 0, the noise is higher, reducing the Pearson correlation performance.

Another very relevant point is to which extent the methods are affected by $\Delta\Delta G$ measures obtained outside physiological conditions. A recent paper [5] showed that there are some predictors in some extreme ranges of pH and temperature that decreases the performance. S669 dataset was divided into two parts: the former group containing variants whose temperature and pH are in physiological ranges [293.15, 313.15] K (20–40°C) and [6.0, 8.0], respectively. This physiological group consists of 443 variants, whereas the non-physiological one of 226 variants. The results reported in Fig. 6 show that there is not a clear indication of the fact that non-physiological conditions induce more errors in the predictions. The Pearson correlation is slightly better for variants in the group of physiological conditions; however, the MAE has an opposite trend (Fig. 6). In the future, when a far larger set of clean data will hopefully be available, a more thorough study should be carried out.

Classification performance

In many applications, the identification of destabilizing and stabilizing variations is more relevant than the prediction of the exact $\Delta\Delta G$ value. In Fig. 7, we evaluated the classification accuracy of the different methods. The figure shows three broad groups with similar accuracy in the various stability classes and variant subsets. The first group is represented by the most antisymmetric and unbiased predictors: PremPS, ACDC-NN, ACDC-NN-Seq, DDGun3D, DDGun, Dynamut, ThermoNet, INPS-Seq, INPS3D and FoldX. They showed good performance in both stabilizing and destabilizing classes, especially PremPS, ACDC-NN, DDgun and INPS-Seq. However, all these predictors showed a lower accuracy in the underrepresented direct-stabilizing variants and their reverse class, i.e. the reverse-destabilizing variants. This is especially true for the best performing PremPS, ACDC-NN and

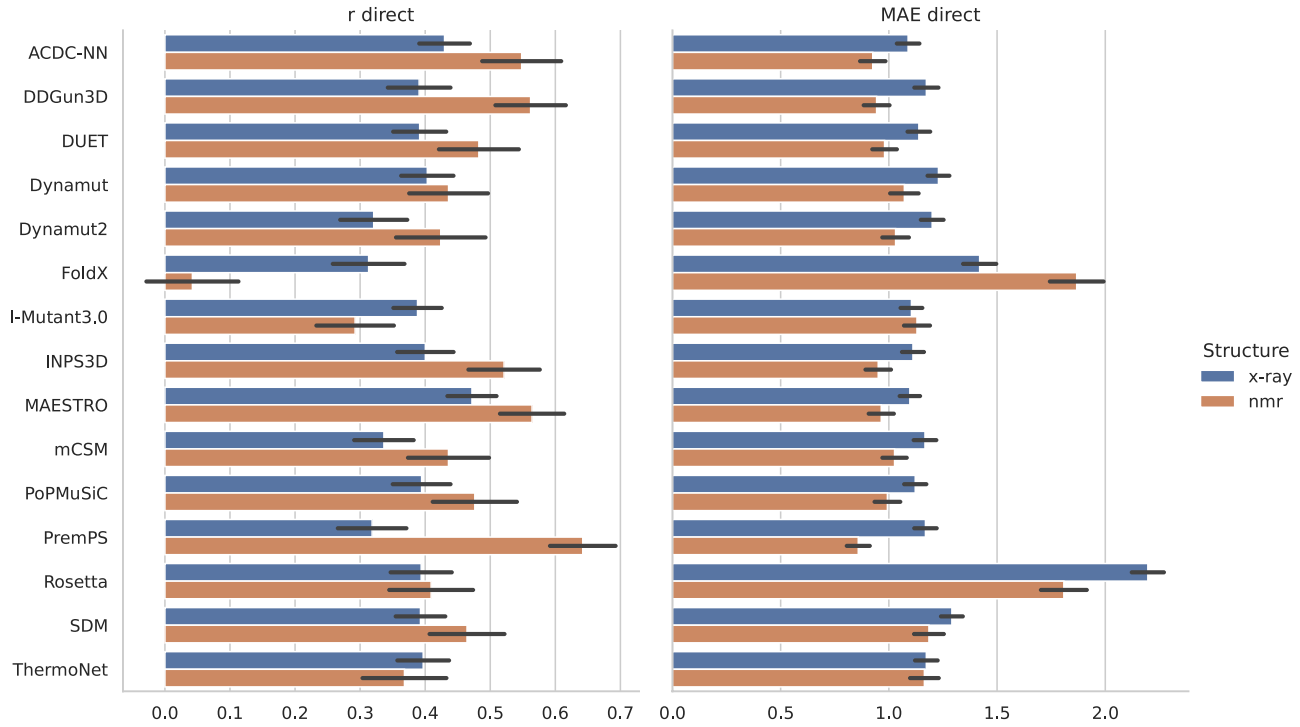

Figure 3. Effect on the method performance of the experimental technique: NMR versus X-ray. After splitting the S669 dataset into NMR and X-ray derived structures (with 196 and 473 variants in 23 and 71 proteins, respectively), Pearson correlation coefficients (*r* direct, on the left) and MAE (MAE direct, on the right) for the direct variants are shown for all structure-based methods. The black error bars represent the bootstrap estimated standard error.

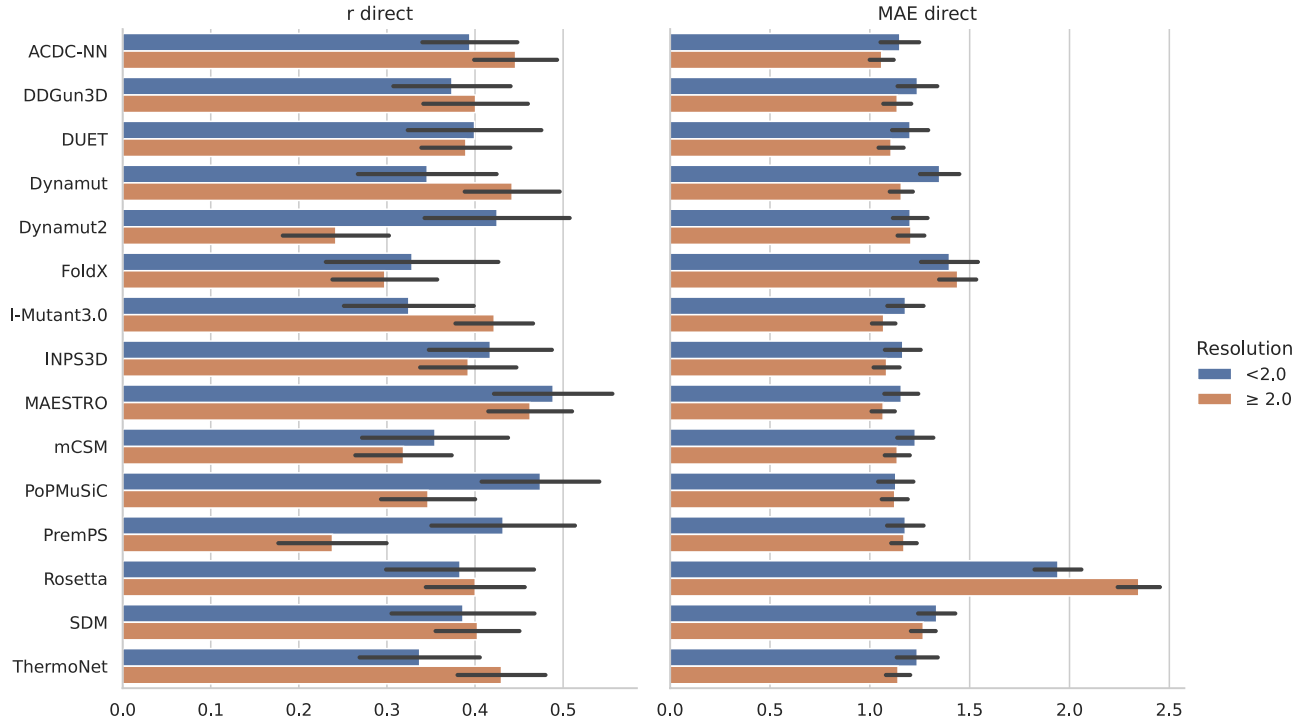

Figure 4. Effect on the method performance of the different X-ray resolution. Evaluation is made on the direct variants in S669 whose structures were obtained by X-ray diffraction. The dataset is split in two classes using 2.0 Årmstrong as a threshold for the resolution, with 177 variants in 34 proteins with resolution < 2.0 and 296 variants in 37 proteins with a resolution ≥ 2.0. Pearson correlation coefficients (*r* direct, on the left) and MAE (MAE direct, on the right) are shown for all the structure-based methods. The black error bars represent the bootstrap estimated standard error.

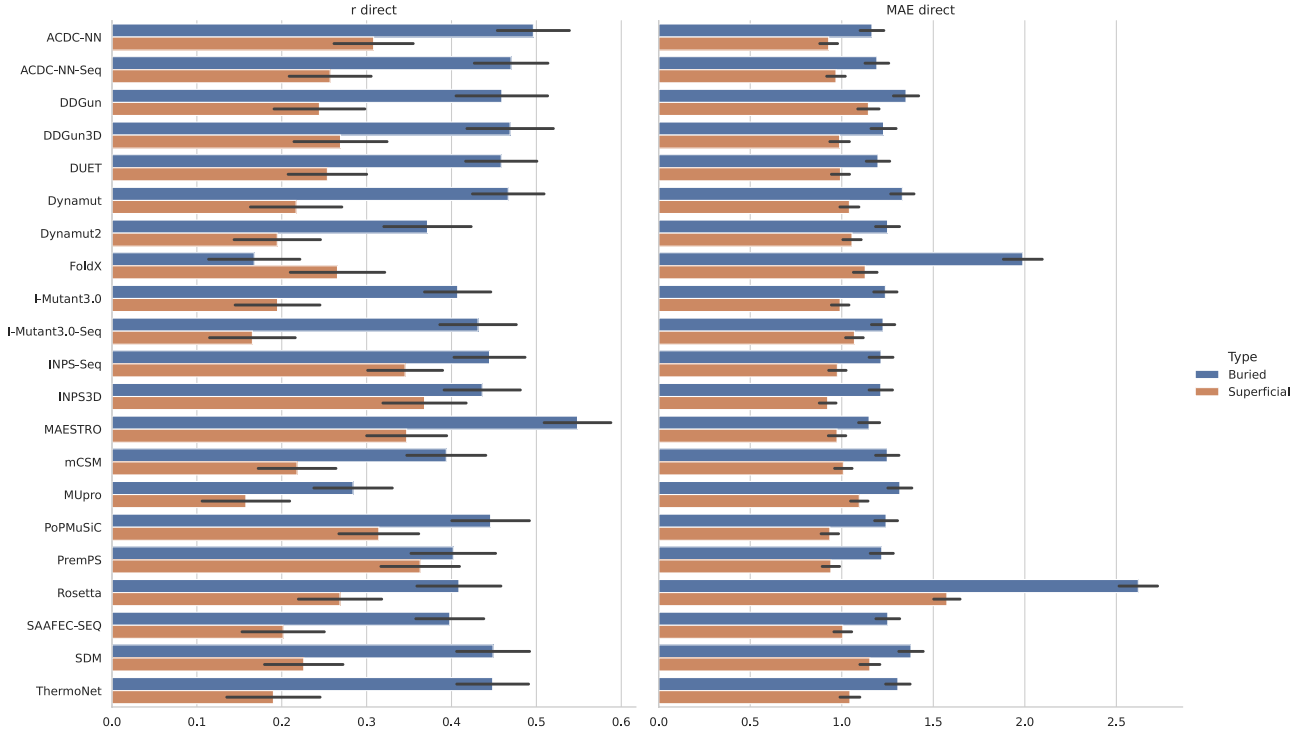

Figure 5. Assessment of the effects of the RA of an amino acid on the prediction of the protein stability. The effects of the RA are estimated by splitting the direct variants in the s669 dataset with respect to the RA median value (24%). Pearson correlation coefficients (r) and mean absolute errors (MAE direct) are displayed in the left and in the right plot, respectively. RA ranges from 0 to 1, with 0 representing a completely buried residue and 1 representing a residue on the surface. The black error bars represent the bootstrap estimated standard error.

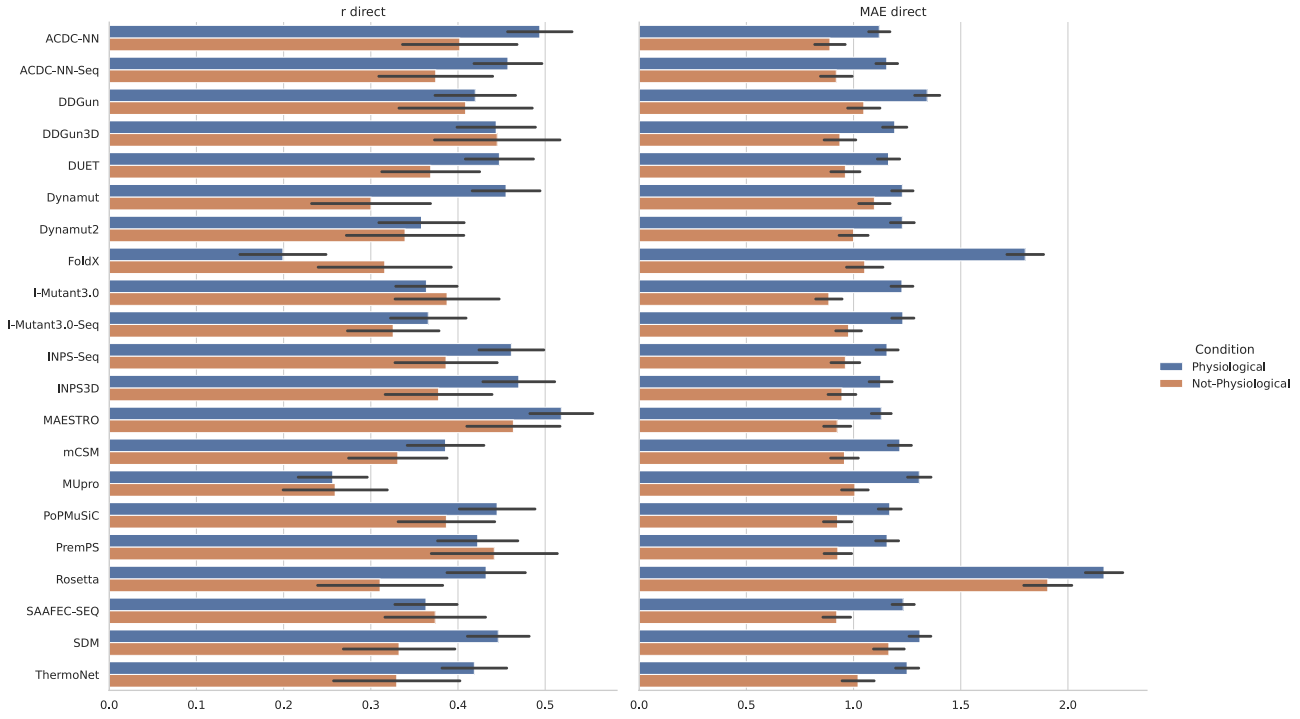

Figure 6. Assessment of the protein stability predictions tools on S669 at different temperature and pH conditions. We compared all the prediction tools at physiological ($T \in [293.15, 313.15]$ K,$pH \in [6.0, 8.0]$, 443 variations) and not-physiological temperature and pH conditions (226 variations). After dividing the S669 dataset accordingly, the effects of different temperature and pH conditions were estimated by calculating the Pearson correlation coefficients (r) and MAE between predicted and real values on the two classes. These two measures are displayed in the left and in the right figure respectively. The black error bars represent the bootstrap estimated standard error.

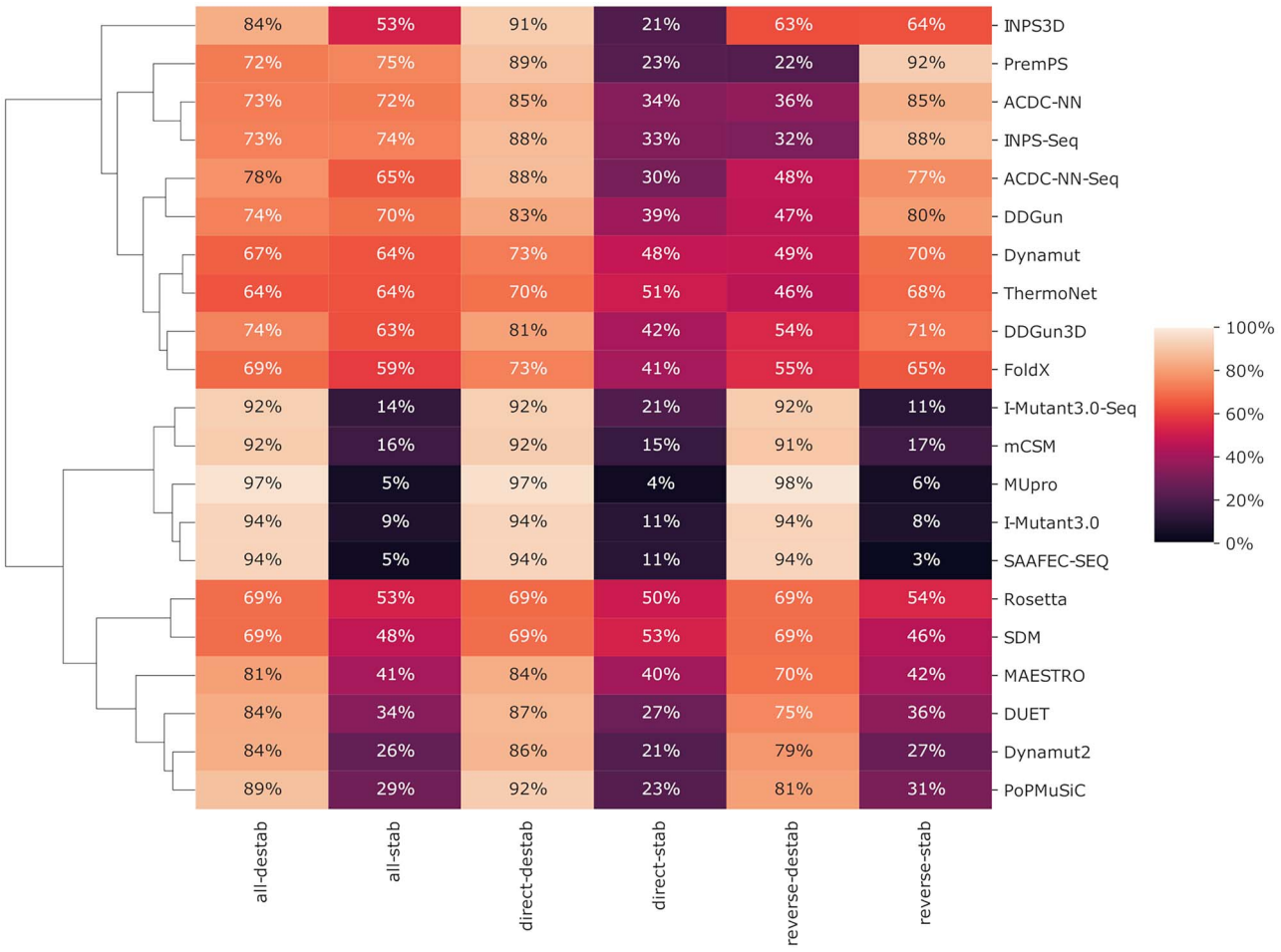

Figure 7. ΔΔG classification accuracy. Here we explore the classification accuracy when predicting the stability change direction. Predicted and experimental ΔΔG values were split in two classes: stabilizing ($\Delta\Delta G \geq 0$) and destabilizing ($\Delta\Delta G < 0$). For each subset (direct, reverse and both direct and reverse together) and for each experimental ΔΔG class (columns), the heatmap shows the ratio of variants predicted to be in the correct ΔΔG class for each predictor (rows).

INPS-Seq, suggesting a trade-off where they 'sacrifice' accuracy in the smaller classes for greater scores in the whole dataset. A second group includes I-Mutant3.0, I-Mutant3.0-Seq, mCSM, MUpro and the newer SAAFEC-SEQ. They are heavily biased toward destabilizing predictions, therefore their accuracies on stabilizing variants are extremely low in all the datasets. The remaining group includes Rosetta, SDM, Maestro, DUET, Dynamut2 and PoPMuSiC, which still showed a bias toward destabilization but to a lower extent, with Rosetta and SDM being quite balanced across different classes. Among the tested sequence-based methods, INPS-Seq, ACDC-NN-Seq and DDGun are the more balanced and the best performing.

An analysis of the prediction distributions of the various methods is reported in Fig. 8, where we split the variants by their experimental ΔΔG into destabilizing ($\Delta\Delta G \leq -0.5$), neutral ($|\Delta\Delta G| < 0.5$) and stabilizing variants ($\Delta\Delta G \geq 0.5$). Compared with the experimental distributions, all the methods tended to compress their predictions toward zero (neutral), generating a significant overlap among the three distributions (Rosetta is the only exception here). However, many of them maintained the relation of order among the three classes except for SAAFEC-SEQ (Fig. 8). The sequence-based DDGun appeared to be the only one that consistently keeps a minimum of difference among the three types of prediction distributions separating the means and the box quantile borders (stabilizing, neutral and destabilizing) for all direct and reverse sets.

Discussion and Conclusions

This paper introduces a new manually curated dataset (S669) containing variants from protein sequences with no homology with the two most widely used training data resources (S2648 and Varibench). We showed that using models based on Rosetta is almost equivalent to using experimental structures for ΔΔG prediction. At the same time, this leads to balance the variant distribution, typically skewed toward destabilizing variations, by adding reverse variants and their corresponding simulated models. The Pearson correlation of the methods tested on S669 is lower than those reported in the original papers. However, this is expected since S669 can be considered as external validation. Nonetheless, the Pearson

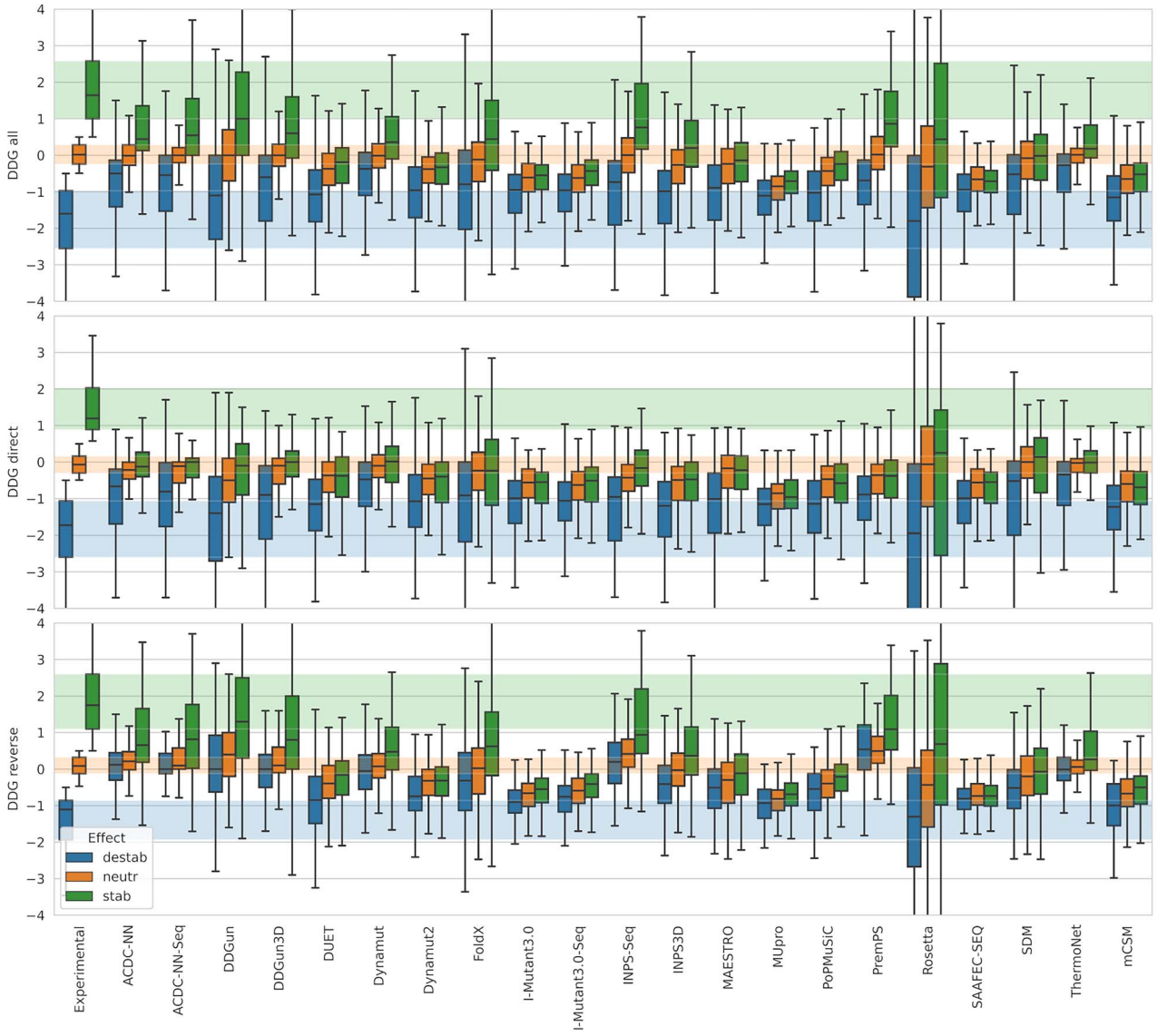

Figure 8. ΔΔG prediction distribution by stability class. These boxplots show the distribution of the predicted ΔΔG values. Variants were split into three classes by their experimental ΔΔG: destabilizing ($\Delta\Delta G \leq 0.5$), neutral ($-0.5 < \Delta\Delta G \leq 0.5$), stabilizing ($0.5 < \Delta\Delta G$). The experimental ΔΔG values were plotted (to the left) and their boxes extended as transparent horizontal bands as a reference. The plot was repeated for all (top), direct (center) and reverse variants only (bottom).

correlations are not too far from those reported for the methods evaluated by limiting sequence identity between training and test sets. The more antisymmetric methods tend to perform better, and those built to be antisymmetric perform better in the regression task (prediction of ΔΔG value), in particular PremPS, ACDC-NN and INPS-Seq. It is also worth noticing that the methods perform equally well on NMR or X-ray structure and are relatively insensitive to pH and temperature outside the physiological conditions, making them useful also when these types of information are not available.

Overall, our assessment highlighted that the predictors satisfying the antisymmetry property can perform better than the other tools in regression or when the test set is balanced. For some of them, as in the case of ACDC-NN and DDGun, their sequence-based version showed similar results compared with their structure-based counterpart. This indicates that both evolutionary information and antisymmetry are important features for narrowing down the gap in the performance between sequence- and structure-based methods. Most methods, especially the non-antisymmetric, show a bias toward the destabilizing class. This makes them unsuitable for variant classification because they tend to predict every variant as destabilizing, misclassifying most stabilizing variants.

When only stabilization/destabilization information is considered, the antisymmetric methods tend to predict better on the whole datasets. However, it appears that the direct stabilizing variants in the datasets are more challenging to assign correctly. In particular, SDM, Rosetta, ThermoNet and Dynamut are the most

balanced. All the tested methods tend to compress the predictions toward neutrality, generating a significant overlap between stabilizing, neutral and destabilizing variations. The compression of the predictions indicates that a possible improvement for future methods is to work on the calibration of the prediction distributions. The destabilizing variants show a stronger signal in terms of $\Delta\Delta G$, which are easier to detect on average. Indeed, the antisymmetric predictors showed to capture very well the reverse variations, as stabilizing. These contrasting results may open a future direction of study, improving our understanding of these types of variants and possibly increasing the method performance.

Key Points

- We performed a thorough benchmark of current predictors of protein stability changes upon single-point mutations, using never-seen-before protein variants.
- We provide a dataset consisting of 669 variants never-seen-before by the current methods.
- We showed the relevance of incorporating the thermodynamic antisymmetric principle to improve prediction robustness.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank the Italian Ministry for Education, University and Research for the PRIN 2017 201744NR8S 'Integrative tools for defining the molecular basis of the diseases' and for the program 'Dipartimenti di Eccellenza 20182022D15D18000410001'.

Author contributions statement

E.C. and P.F. conceived and funded the study. E.C. end P.F. derived the first raw version for the data. C.P., S.B., V.A., T.S. and P.F. manually cleaned the variants and read the original papers. C.P., S.B, G.B., V.R. and P.F. run the predictions. C.P., S.B, G.B. generated the figures and the tables. C.P., S.B, G.B., T.S. and P.F. analyzed the results. C.P., S.B, G.B. and P.F. sketched the draft of the paper. C.P., S.B, G.B., T.S, E.C. and P.F. wrote the manuscript.

References

1. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel* 2009;**22**(9):553–60.
2. Sanavia T, Birolo G, Montanucci L, *et al.* (eds). Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput Struct Biotechnol J* 2020;**18**:1968–79.
3. Marabotti A, Scafuri B, Facchiano A. Predicting the stability of mutant proteins by computational approaches: an overview. *Brief Bioinform* 2021;**22**(3):bbaa074.
4. Caldararu O, Blundell TL, Kepp KP. A base measure of precision for protein stability predictors: structural sensitivity. *BMC bioinformatics* 2021;**22**(1):1–14.
5. Iqbal S, Li F, Akutsu T, *et al.* Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. *Brief Bioinform* 2021;**22**(6):bbab184.
6. Ulrich F, Hartl. Protein misfolding diseases. *Annu Rev Biochem* 2017;**86**:21–6.
7. Martelli PL, Fariselli P, Savojardo C, *et al.* Large scale analysis of protein stability in omim disease related human protein variants. *BMC Genomics* 2016;**17**(2):239–47.
8. Cheng TMK, Yu-En L, Vendruscolo M, *et al.* Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput Biol* 2008;**4**(7):e1000135.
9. Compiani M, Capriotti E. Computational and theoretical methods for protein folding. *Biochemistry* 2013;**52**(48):8601–24.
10. Reza MN, Ferdous N, Emon MTH, *et al.* Pathogenic genetic variants from highly connected cancer susceptibility genes confer the loss of structural stability. *Sci Rep* 2021;**11**(1):19264.
11. Cheng L, Han X, Zhu Z, *et al.* Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *Brief Bioinform* 2021;**22**(2):1442–5003.
12. Ancien F, Pucci F, Rooman M. In Silico analysis of the molecular-level impact of SMPD1 variants on Niemann-pick disease severity. *Int J Mol Sci* 2021;**22**(9):4516.
13. Birolo G, Benevenuta S, Fariselli P, *et al.* Protein stability perturbation contributes to the loss of function in haploinsufficient genes. *Front Mol Biosci* 2021;**8**:10.
14. Pires DE, Chen J, Blundell TL, *et al.* In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep* 2016;**6**:19848.
15. Hou Q, Pucci F, Ancien F, *et al.* SWOTein: a structure-based approach to predict stability strengths and weaknesses of prOTEINs. *Bioinformatics* 2021;**37**(14):1963–71.
16. Fang J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief Bioinform* 2020;**21**(4):1285–92.
17. Usmanova DR, Bogatyreva NS, Bernad JA, *et al.* Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics* 2018;**34**(21):3653–8.
18. 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.
19. Montanucci L, Savojardo C, Martelli PL, *et al.* On the biases in predictions of protein stability changes upon variations: the INPS test case. *Bioinformatics* 2019;**35**(14):2525–707.
20. Capriotti E, Fariselli P, Rossi I, *et al.* A three-state prediction of single point mutations on protein stability changes. *BMC bioinformatics* 2008;**9**(2):1–9.
21. Savojardo C, Martelli PL, Casadio R, *et al.* On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief Bioinform* 2021;**22**(1):601–3.
22. Marabotti A, Del Prete E, Scafuri B, *et al.* Performance of web tools for predicting changes in protein stability caused by mutations. *BMC Bioinformatics* 2021;**22**(Suppl 7):345.

23. 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.
24. Nikam R, Kulandaisamy A, Harini K, *et al.* ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res* 2021;**49**(D1):D420–4.
25. Xavier JS, Nguyen TB, Karmarkar M, *et al.* ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Res* 2021;**49**(D1):D475–9.
26. 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 bioinformatics* 2011; **12**(1):1–12.
27. Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. *Hum Mutat* 2013;**34**(1):42–9.
28. Pires DEV, Rodrigues CHM, Ascher DB. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res* 2020;**48**(W1):W147–53.
29. Kulandaisamy A, Zaucha J, Frishman D, *et al.* MPTherm-pred: analysis and prediction of thermal stability changes upon mutations in transmembrane proteins. *J Mol Biol* 2021;**433**(11): 166646.
30. Song Y, DiMaio F, Wang RY-R, *et al.* High-resolution comparative modeling with rosettacm. *Structure* 2013;**21**(10):1735–42.
31. Benevenuta S, Pancotti C, Fariselli P, *et al.* An antisymmetric neural network to predict free energy changes in protein variants. *J Phys D Appl Phys* 2021;**54**(24):245403.
32. Pancotti C, Benevenuta S, Repetto V, *et al.* A deep-learning sequence-based method to predict protein stability changes upon genetic variations. *Gen* 2021;**12**(6):911.
33. 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 bioinformatics* 2019; **20**(14):335.
34. Bastolla U, Farwer J, Knapp EW, *et al.* How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins: Structure, Function, and Bioinformatics* 2001;**44**(2): 79–96.
35. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 2014;**30**(3):335–42.
36. Worth CL, Preissner R, Blundell TL. Sdm-a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res* 2011;**39**(suppl_2):W215–22.
37. Pires DEV, 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**(W1):W314–9.
38. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* 2018;**46**(W1):W350–5.
39. 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.
40. Schymkowitz J, Borg J, Stricher F, *et al.* The FoldX web server: an online force field. *Nucleic Acids Res* 2005;**33**(suppl_2):W382–8.
41. 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):606.
42. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics* 2006; **62**(4):1125–32.
43. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics* 2011;**79**(3):830–8.
44. 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.
45. Chen Y, Haoyu L, Zhang N, *et al.* PremPS: predicting the impact of missense mutations on protein stability. *PLoS Comput Biol* 2020;**16**(12):e1008543.
46. Laimer J, Hiebl-Flach J, Lengauer D, *et al.* MAESTROweb: a web server for structure-based protein stability prediction. *Bioinformatics* 2016;**32**(9):1414–6.
47. 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.
48. 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**(suppl_2):W306–10.
49. Montanucci L, Martelli PL, Ben-Tal N, *et al.* A natural upper bound to the accuracy of predicting protein stability changes upon mutations. *Bioinformatics* 2019;**35**(9):1513–7.
50. Benevenuta S, Fariselli P. On the upper bounds of the real-valued predictions. *Bioinform Biol Insights* 2019;**13**:1177932219871263.
51. Caldararu O, Mehra R, Blundell TL, *et al.* Systematic investigation of the data set dependency of protein stability predictors. *J Chem Inf Model* 2020;**60**(10):4772–84.