Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Artificial intelligence challenges for predicting the impact of mutations on protein stability

Fabrizio Pucci[1,2], Martin Schwersensky[1,2] and Marianne Rooman[1,2]

Abstract

Stability is a key ingredient of protein fitness, and its modification through targeted mutations has applications in various fields, such as protein engineering, drug design, and deleterious variant interpretation. Many studies have been devoted over the past decades to build new, more effective methods for predicting the impact of mutations on protein stability based on the latest developments in artificial intelligence. We discuss their features, algorithms, computational efficiency, and accuracy estimated on an independent test set. We focus on a critical analysis of their limitations, the recurrent biases toward the training set, their generalizability, and interpretability. We found that the accuracy of the predictors has stagnated at around 1 kcal/mol for over 15 years. We conclude by discussing the challenges that need to be addressed to reach improved performance.

Addresses

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium

[2] Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium

Corresponding author: Rooman, Marianne (Marianne.Rooman@ulb.be)

Current Opinion in Structural Biology 2022, **72**:161–168

This review comes from a themed issue on **Artificial Intelligence (AI) Methodologies in Structural Biology**

Edited by **Feixiong Cheng** and **Nurcan Tuncbag**

For a complete overview see the Issue and the Editorial

https://doi.org/10.1016/j.sbi.2021.11.001

0959-440X/© 2021 Elsevier Ltd. All rights reserved.

Keywords

Protein stability, Residue mutations, Folding free energy, Machine learning, Prediction biases, Overfitting, Model interpretability.

Introduction

The accurate prediction of mutational effects on protein stability is of utmost importance in many fields ranging from biotechnology to medicine. In rational protein engineering applications, for example, the targeted redesign of proteins makes it possible to optimize the biotechnological and biopharmaceutical processes in which they are involved [1,2]. Stability prediction also plays a key role in interpreting the impact of human genetic variants and may provide a better understanding of how these variants lead to disease conditions [3,4]. Note that stability is all the more important as it is the dominant factor in protein fitness [5].

For these reasons, many studies have been devoted over the last decade to the development of computational tools that aim to predict in a fast and reliable way the change in protein stability upon mutations [6–28]. These methods use information about protein sequence, structure, and evolution, which are combined through a variety of machine learning methods ranging from simple linear regression to more complex models. For more information, we refer to excellent recent reviews [29,30] and comparative tests [31–33].

It has to be noted that, although recent advances in the field of artificial intelligence (AI) and more specifically in deep learning have considerably improved feature selection and combination in multiple bioinformatics problems such as three-dimensional (3D) protein structure prediction [34,35,66,67], so far, they are not often used in predicting the effects of mutations on protein stability. Indeed, most current predictors use shallow algorithms, probably because the amount of experimental training data is too limited to allow for deeper algorithms.

In this review, we concisely present the protein stability prediction methods that are available and functional, and test their performance on an independent set of experimentally characterized point mutations, which are not part of any of the training sets. Our main goal, here, is to take a critical look at the predictors by investigating their algorithms, limitations, and biases, as schematically shown in Figure 1. We also discuss the main challenges the field will have to face in the years to come to strengthen the role of computational approaches in protein design and personalized medicine.

Brief overview and benchmark of the current computational models

We collected existing computational methods predicting the change in protein thermodynamic stability upon point mutations defined by the change in folding free energy $\Delta\Delta G$. We restricted ourselves to predictors that are commonly used and currently available through a working web server or downloadable code. These methods, listed in Table 1, are almost all based on the 3D protein structure and use a series of features, such as the relative solvent accessible surface area of the mutated residue, the change in folding free energy ($\Delta\Delta W$) estimated by various types of energy functions, the change in volume of the mutated residue (ΔVol), and the change in residue hydrophobicity (ΔHyd). They also often use evolutionary information either extracted from multiple sequence alignments of the query protein or from substitution matrices, such as BLOSUM62 [36]. Several machine learning algorithms were used to combine the different features. These are most often algorithms that have become classical, such as artificial neural networks, support vector machines or random forests. Only a few very recent predictors use novel deep learning approaches [18,20,27]. At the other extreme, a predictor published this year uses a very simple model consisting of a linear combination of only three features [37].

It is a difficult task to rigorously evaluate the accuracy of predictors [32,33]. Indeed, performances depend on the training and test sets, as well as on the evaluation metric. Here, we have chosen to benchmark the collected methods by estimating their accuracy in terms of the root mean square error (RMSE) and the Pearson correlation coefficient (r) between experimental and predicted values for 830 mutations inserted in the 56-residue β1 extracellular domain of streptococcal protein G (PDB code 1PGA) [38]. It has to be underlined that this set of mutations is not included in the training sets of the methods tested and is thus a truly independent set.

The RMSE of the predictors varies between 0.9 and 1.4 kcal/mol, and the correlation coefficients vary between 0.3 and 0.7, as shown in Table 1. We observe a low correlation between these two metrics; the method with the worst RMSE (1.42 kcal/mol) has the best r (0.66). This follows from the fact that Pearson correlation coefficients are essentially driven by the points that are far from the mean, in contrast to RMSE, which takes all points equally into account.

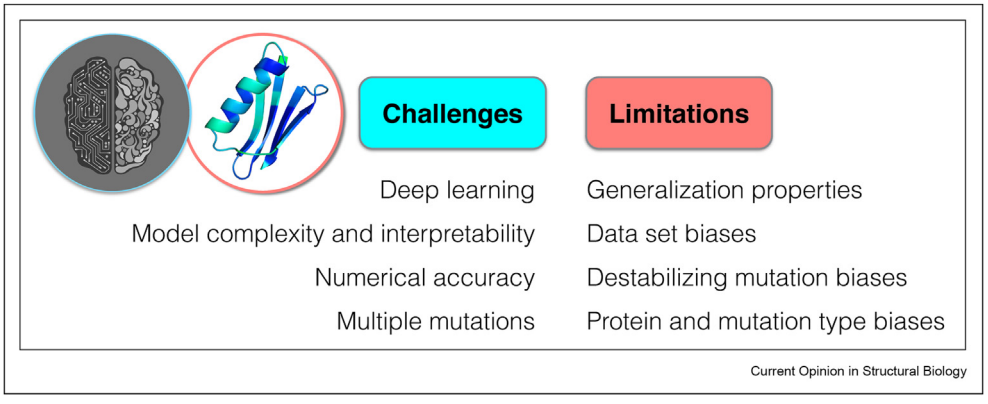
Note that these results must be interpreted with care. Indeed, both RMSE and r values depend on the distribution of experimental $\Delta\Delta G$s and, more specifically, on its variance [39]. The ranking of the prediction methods and their scores thus crucially depend on the metric used and the test $\Delta\Delta G$ distribution.

In addition, we also tested two other widely known stability predictors, FoldX [14] and Rosetta [15], which are physics-based rather than AI-based and use full-atom representations rather than simplified descriptions of protein structures. These two methods reach reasonable correlations with r values of 0.36 and 0.44, respectively, slightly lower than AI-based methods ($\langle r \rangle = 0.48$). In contrast, their RMSE values are above 3 kcal/mol, which is much worse than the average RMSE of 1.02 kcal/mol of AI-based methods. The lesser performance of these two methods has already been observed [31] and could be due to the use of detailed atomic representation, which makes them sensitive to resolution defects.

Evolution of predictor performance over time

We have analyzed the average performance of all the methods according to their year of development. We clearly see in Figure 2a that the average accuracy has not improved in the last 15 years, but basically remains constant, despite all efforts and the improved performances claimed by the authors of the newly published

Figure 1



Schematic representation of the challenges and limitations that protein stability prediction methods have to address in the coming years.

Table 1

List of artificial intelligence-based $\Delta\Delta G$ predictors studied.

Method (Year)	3D	Feature type	RMSE (kcal/mol) r	Run time (min)	AI method	Ref.
MUpro (2006)		Neighbors	1.17 $r = 0.26$	< 1	Support vector regression	[16]
I-Mutant 3.0 (2007)	✓	Residue type, RSA, Residue environment	0.92 $r = 0.38$	~ 400	Support vector regression	[10]
PoPMuSiC v2.1 (2011)	✓	Statistical potentials, ΔVol, RSA	0.95 $r = 0.56$	< 1	Artificial neural network	[6]
SDM (2011)	✓	RSA, Environment-specific Substitution frequencies	0.95 $r = 0.46$	~ 250	Linear combination	[43]
mCSM (2014)	✓	Graph-based signatures, Atomic distance patterns	1.10 $r = 0.44$	~ 250	Regression via Gaussian process	[11]
MAESTRO (2014)	✓	Statistical potentials, PSize, ASA, SS, ΔHyd, ΔIP	0.91 $r = 0.58$	< 1	Linear regression, ANN, SVM	[19]
AUTOMUTE 2.0 (2014)	✓	4-Body statistical potential ASA, depth, SS, Vol	1.16 $r = 0.30$	~ 1	Random forest, Tree regression	[21]
INPS-3D (2016)	✓	Contact potential, RSA, EvolInfo, Bl62, ΔHyd, ΔMW, MutI	0.96 $r = 0.52$	~4	Support vector regression	[8]
STRUM (2016)	✓	Energy functions, homology modeling, ΔHyd, ΔVol, ΔIP, ΔMW, EvolInfo	1.05 $r = 0.49$	~200	Gradient boosting regression	[9]
PoPMuSiC^{sym} (2018)	✓	Statistical potentials ΔVol, RSA	0.98 $r = 0.54$	< 1	Artificial neural network	[44]
DDGun3D (2019)	✓	BL62, ΔHyd, RSA, Statistical potentials	0.94 $r = 0.57$	~ 30	Non-linear regression	[26]
DeepDDG (2019)	✓	ASA, SS, H-bonds, EvolInfo, Residue distances/orientations	1.42 $r = 0.66$	~ 5	Shared residue-pair deep neural network	[20]
ThermoNet (2020)	✓	Aromatic, positive, negative, Hyd, H-bond donor/acceptor	1.01 $r = 0.29$	~ 100	3D convolutional neural network	[18]
PremPS (2020)	✓	EvolInfo, RSA, ΔHyd, Hyd, Aromatic, charged, Leu	0.95 $r = 0.57$	~4	Random forest	[17]
SimBa (2021)	✓	RSA, ΔVol ΔHyd	0.99 $r = 0.53$	< 1	Linear regression	[37]
SAAFEC-SEQ (2021)		EvolInfo, neighbors, ΔVol, ΔHyd, ΔFlex, PSize, H-bond	0.91 $r = 0.49$	~ 30	Gradient boosting decision tree	[28]
Mean		⟨ **RMSE** ⟩= ⟨ **r** ⟩ = ***σ*(Exp)**=	1.02 ± 0.13 0.48 ± 0.12 0.98			

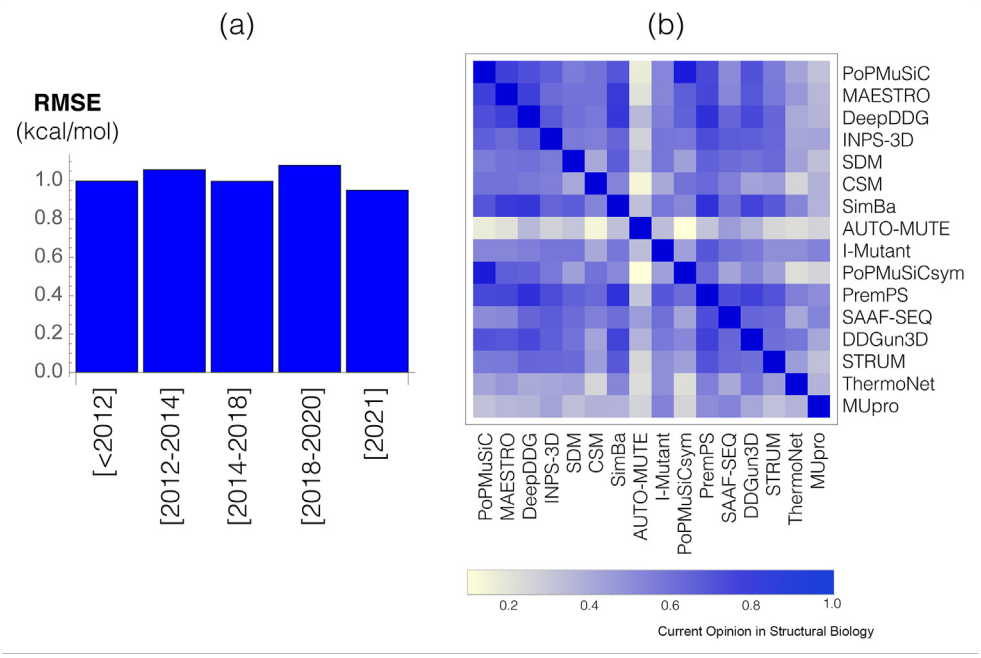
The RMSE (in kcal/mol) and linear correlation coefficient r are computed for the experimentally characterized mutations in the β1-extracellular domain of streptococcal protein G [38]; σ(Exp) is the standard deviation of the experimental $\Delta\Delta G$ distribution (in kcal/mol). Abbreviations used: AI, artificial intelligence; ANN, artificial neural network; ASA, solvent accessible surface area; RSA, relative ASA; Depth, localization on surface, undersurface, or core; PSize, Protein size; Vol, residue volume; ΔVol, change in residue volume upon mutation; ΔMW, change in molecular weight; ΔFlex, change in flexibility; Hyd, residue hydrophobicity; ΔHyd, change in Hyd; SS, secondary structure; MutI, mutability index of the native residue [45]; BL62, BLOSUM62 matrix [36]; Neighbors, type of residues in the neighborhood along the sequence; EvolInfo, evolutionary information from protein families; SVM, supporting vector machine.

methods. This is strikingly different from the situation in the field of protein structure prediction, for example, which has experienced an impressive improvement during the same period [40]. Whether the accuracy limit on predicted $\Delta\Delta G$s is due to the relatively low number of mutations in the training set, more fundamental reasons, or uncontrolled biases in the predictors is currently a topic of debate [39,41,29]. We discuss this issue more extensively in the next sections. It must again be noted that the RMSE threshold and the ranking of the performance of the method can be somewhat different on other test mutations [31–33]. But the lower limit on RMSE is basically always around 1 kcal/mol.

It is instructive to look at the correlations between the predictions of the different methods, shown in Figure 2b. They are all reasonably good, with an average correlation coefficient of 0.5. This reflects that the different methods use roughly the same information but that there is room for improvement and further boosting the prediction accuracy by selecting informative features that have not yet been combined.

Another important characteristic of a prediction method is its speed. Indeed, as many current projects require investigating protein stability properties at a large, proteome, scale [42], the predictors have to be able to run

Figure 2



Evaluation of the $\Delta\Delta G$ prediction methods listed in Table 1 on the basis of the experimentally characterized mutations in the β1-extracellular domain of streptococcal protein G [38]. (**a**) Average root mean square error (RMSE) of the predictors as a function of their development date. (**b**) Correlation coefficients r between the $\Delta\Delta G$s predicted by the different methods.

fast enough to scan the proteome in a reasonable time. All the methods tested are relatively fast, with some extremely fast such as PoPMuSiC, SimBa, MAESTRO, and AUTOMUTE (see Table 1).

Limitations and prediction biases

The generalization property in machine learning is the ability of the algorithm to correctly predict unseen data. The protein stability predictors, such as all machine learning-based methods, tend, however, to be biased toward the data sets on which they are trained. The majority of the methods analyzed here [7,11,43,44,8,19,37,17,28] were trained on the data set known as S2648 [7]. It contains 2648 mutations with experimental $\Delta\Delta G$ values collected from the literature and the ProTherm database [46], which were thoroughly checked and manually curated. Other predictors use subsets of S2648 or a slightly larger data set known as Q3421 [9].

Multiple hidden biases such as feature and hyperparameter selection biases that are difficult to control can affect the generalization properties of the predictors trained on these data sets. These problems are even more severe when complex algorithms are used or when the training sets are small and unbalanced. In the following, we quantitatively analyze a series of biases that often affect stability predictors and are primarily caused by various imbalances in the training data sets and discuss the strategies used to limit their impact.

Cross-validation biases

Often, prediction performance is evaluated using a k-fold cross-validation procedure. This is not always sufficient to estimate the accuracy of the methods and assessments on test sets are usually also provided, even though their sizes are usually small. Going back to cross-validation, there are different ways to perform the random split of the data set into k folds, at the level of the mutation, position, protein, and even protein cluster. Random splitting at the mutation level introduces some distortions because the knowledge of the effect of a mutation at a given position makes the prediction of another substitution at the same position easier. Splitting at the position level can also introduce some biases. To have more reliable estimations, cross-validation at the protein level has to be performed or even at the protein cluster level where all proteins that are similar to the target protein one wants to predict are removed from the training set.

It should be noted that the extent to which the type of data set splitting affects prediction performances is highly dependent on the prediction model. For example, the drop in performance of predictors that do not use complex machine learning, such as PoPMuSiC and

SimBa, is almost negligible when passing from residue level to protein level [47]. In contrast, a substantial decrease in accuracy is undergone, for example, by STRUM, with correlation coefficients and RMSE between experimental and predicted $\Delta\Delta G$s that pass from (0.77, 0.94 kcal/mol) for 5-fold cross validation at mutation level to (0.64, 1.14 kcal/mol) at position level and (0.54, 1.25 kcal/mol) at protein level [9]. A similar drop in performance of about 20–30% when strict cross validation procedures are used has also been observed in [17].

Bias toward destabilizing mutations

At fixed environmental conditions, the change in folding free energy upon mutation is antisymmetric by definition. More precisely, if protein B is a mutant of protein A, we have that $\Delta\Delta G\ (A \rightarrow B) = -\ \Delta\Delta G\ (B \rightarrow A)$. However, most of the stability predictors violate this relation, as shown by a series of studies [48,44,49,41]. This is mainly because training data sets are dominated by destabilizing mutations, which, in turn, results from the vast majority of mutations in a given protein being destabilizing. For example, the ratio between the numbers of destabilizing and stabilizing mutations in the data sets S2648 [7] and Q3421 [9], which are widely used as training sets, are equal to 3.7 and 3.2, respectively, with a mean $\langle\Delta\Delta G\rangle$ of about 1 kcal/mol in both sets.

Some of the recent prediction methods got rid of this bias and satisfy the antisymmetry property by construction [17,27,8]. To check the extent to which it is the case, a balanced data set such as $\mathcal{S}^{\mathrm{sym}}$ [44] must be considered, which contains, for each mutation $A \rightarrow B$, the backward mutation $B \rightarrow A$ and thus an even number of stabilizing and destabilizing mutations. The deviation from antisymmetry $\delta = \Delta\Delta G\ (A \rightarrow B) + \Delta\Delta G\ (B \rightarrow A)$ is an important measure for the evaluation of the lack of bias.

Protein and mutation biases

Another type of bias arises from the fact that training data sets do not provide a good sampling of the types of mutations and proteins, as recently discussed in [41]. Often, mutation data sets are dominated by a few proteins which contain most of the entries and are therefore likely to bias the prediction toward them. For example, the 10 proteins from S2648 and Q3421 that contain the largest number of mutations represent 50% and 40% of the entries, respectively. The types of substitutions are also not well sampled; among the $20 \times 19 = 380$ possible amino acid substitutions, 78 and 38 are not sampled at all in S2648 and Q3421, respectively. The top 10 types are substitutions into alanine, which account for 25% of the entries in the data sets.

The way in which different methods are affected by this bias is extensively evaluated in [41] by introducing an unbiased test set with respect to mutation types. Most of the prediction methods are shown to be biased. They are able to correctly predict the effect of certain types of mutations, while they completely miss others.

Current and future challenges

Deep learning approaches

Deep learning algorithms, such as convolutional neural networks, have provided spectacular improvements in a series of bioinformatics problems, such as protein structure prediction [40]. Such methods are starting to be used in the prediction of the impact of mutations on protein stability [18,20,50,27], but most of the current methods still use standard shallow machine learning approaches. This is due to the fact that deep learning methods require large amounts of input data for training [51], while standard training data sets such as S2648 [6] or Q3421 [9] only include a few thousand entries and are thus too small for these approaches. New mutation data have recently been collected [52–54], which will certainly increase the size of the training data sets after proper curation. However, these sets will probably remain too limited, with the consequence that deep learning is unlikely to outperform standard machine learning approaches without overfitting issues in the near future, even though unsupervised pre-training can help prevent these issues to some extent [51,27].

Prediction model complexity and interpretability

The application of a wide variety of AI algorithms with different complexity to the prediction of protein stability is very informative. These algorithms range from deep learning approaches such as 3D convolutional neural networks [18] to extremely simple models such as linear regression [37]. Complex algorithms can capture the intricate relationships between input features and training data better than simpler models, but they are in general more prone to overfitting. Moreover, most of them act as black boxes, which makes their results more difficult to interpret. Note that both over- and underfitting are serious problems for generalization. Therefore, the development of a prediction model must be a trade-off between these two extremes. We would like to point out that the best current methods are not always those that use the most complex AI techniques (see Table 1).

The interpretability of the model at biophysical and biochemical levels can be another characteristic to be considered in the model design. For example, it has been shown in [37] that just three simple features, that is, the relative solvent accessible surface area of the mutated residues, and the change in residue volume and in hydrophobicity upon mutations, combined using a linear model, can achieve performances similar to state-of-the-art prediction methods that use up to hundred features and complex machine learning. Novel techniques for

interpreting model predictions [55–57], such as SHAP (SHapley Additive exPlanations) [55], have recently been introduced in the AI field. Their application to protein stability predictors could help to better identify the relative importance of features and lead to more accurate prediction models retaining interpretability properties.

Are we stuck with the limit of 1 kcal/mol RMSE ?

Surprisingly enough, all the methods developed over the past fifteen years have an accuracy evaluated in cross validation by an RMSE slightly greater than 1 kcal/mol, while most validations on independent test sets are even worse with RMSEs between 1.5 and 2.5 kcal/mol [32,33]. On the test protein we used here, the situation is somewhat more favorable, with a lower value of 0.9 kcal/mol (Table 1); this is, however, related to the particularly low standard deviation of the experimental $\Delta\Delta G$ distribution in this case (1 kcal/mol). The idea that 1 kcal/mol represents a hard limit for the prediction accuracy has already been suggested in [41].

Several reasons can explain this limit. First, all the predictors are based on a series of approximations, such as the use of the wild type structure but not the mutant structure. They, thus, neglect the possible structural modifications caused by the mutations to the folded structure and, moreover, also overlook perturbations to the unfolded state [41]. In addition, entropy contributions to the folding free energy are largely overlooked, even though the methods based on statistical mean force potentials do not neglect them completely. Another reason comes from the intrinsic errors on experimental $\Delta\Delta G$ values. In particular, both thermal and chemical measurements of $\Delta\Delta G$ generally involve approximations [58]. In addition, all the $\Delta\Delta G$ values in the data sets have not been determined under the same conditions, and the dependence of $\Delta\Delta G$ on, for example, temperature or pH can be important.

Whether the value of 1 kcal/mol is a true limit that cannot be circumvented, as suggested through a theoretical estimation of the experimental $\Delta\Delta G$ distribution and noise [39,59], is an open question. Our observation that the performance of the methods does not increase with time (Figure 2a) supports this view. This question must be further investigated to understand if and how the current state-of-the-art predictors can be significantly improved.

To address these issues, a systematic blinded experiment fully dedicated to the evaluation of protein stability changes upon mutations would be of great benefit, in the same way that CASP (predictioncenter.org) and CAPRI (capri-docking.org) are for structure predictions and CAGI (genomeinterpretation.org), for genome variant interpretation.

Metagenomic data

Metagenomic sequence data are a valuable source of sequence information that started to be used in protein structure prediction since a seminal article [60] and is now also extensively used in enzyme discovery [61]. For example, most methods used such information as input in the last round of the CASP experiment (CASP14) [60]. Indeed, the enrichment of sequence data from metagenomic databases, even though they are often noisy, can improve protein sequence alignments and thus provide a more accurate assessment of how evolution shapes families of homologous proteins.

Metagenomic sequence data are not yet used in the field of protein stability prediction, even not by the methods that have sequence conservation among their features. This could be a way to boost the prediction accuracy.

Multiple mutations versus single-point mutations

Another challenge is to predict the effect of multiple mutations. It is of particular interest in protein design because multiple mutations can clearly lead to a higher degree of protein stabilization or destabilization [62,63]. Yet, the vast majority of computational methods predict only the effect of single-site substitutions [29]. Point mutations can of course be combined to model multiple mutations but this leads to neglect any direct or indirect epistatic interactions between mutated residues [64,65]. The scarcity of experimental data on multiple mutations in a variety of proteins, as well as the degree of complexity compared to point mutations are the current limitations that prevent obtaining satisfactory prediction accuracy.

Conflict of interest statement

Nothing declared.

Acknowledgements

We acknowledge financial support from the FNRS Fund for Scientific Research through a PDR and a PER research project. FP and MR are FNRS postdoctoral researcher and research director, respectively, and MS benefits from a FNRS-FRIA PhD grant.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

1. Korendovych IV, DeGrado WF: **De novo protein design, a retrospective**. *Q Rev Biophys* 2020, **53**:e3.

2. Coluzza I: **Computational protein design: a review**. *J Phys Condens Matter* 2017, **29**:143001.

3. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Aguilera MA, Meyer R, Massouras A: **Varsome: the human genomic variant search engine**. *Bioinformatics* 2019, **35**:1978.

4. Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL, Wright CF: **Assessing performance of pathogenicity predictors using clinically relevant variant datasets**. *J Med Genet* 2020:107003. jmedgenet–2020.

5. Tokuriki N, Tawfik DS: **Stability effects of mutations and protein evolvability**. *Curr Opin Struct Biol* 2009, **19**:596–604.

6. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M: **PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality**. *BMC Bioinf* 2011, **12**:1–12.

7. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M: **Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0**. *Bioinformatics* 2009, **25**:2537–2543.

8. Savojardo C, Fariselli P, Martelli PL, Casadio R: **INPS-MD: a web server to predict stability of protein variants from sequence and structure**. *Bioinformatics* 2016, **32**:2542–2544.

9. Quan L, Lv Q, Zhang Y: **STRUM: structure-based prediction of protein stability changes upon single-point mutation**. *Bioinformatics* 2016, **32**:2936–2946.

10. 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–W310.

11. Pires DE, Ascher DB, Blundell TL: **mCSM: predicting the effects of mutations in proteins using graph-based signatures**. *Bioinformatics* 2014, **30**:335–342.

12. Pires DE, Ascher DB, Blundell TL: **DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach**. *Nucleic Acids Res* 2014, **42**: W314–W319.

13. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: **The FoldX web server: an online force field**. *Nucleic Acids Res* 2005, **33**(suppl 2):W382–W388.

14. Delgado J, Radusky LG, Cianferoni D, Serrano L: **FoldX 5.0: working with RNA, small molecules and a new graphical interface**. *Bioinformatics* 2019, **35**:4168–4169.

15. 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**:830–838.

16. 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**:1125–1132.

17. Chen Y, Lu H, Zhang N, Zhu Z, Wang S, Li M: **PremPS: predicting the impact of missense mutations on protein stability**. *PLoS Comput Biol* 2020, **16**, e1008543.

18. Li B, Yang YT, Capra JA, Gerstein MB: **Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks**. *PLoS Comput Biol* 2020, **16**, e1008291.

19. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P: **Maestro – multi agent stability prediction upon point mutations**. *BMC Bioinf* 2015, **16**:1–13.

20. Cao H, Wang J, He L, Qi Y, Zhang JZ: **DeepDDG: predicting the stability change of protein point mutations using neural networks**. *J Chem Inf Model* 2019, **59**:1508–1514.

21. Masso M, Vaisman II: **AUTO-MUTE 2.0: a portable framework with enhanced capabilities for predicting protein functional consequences upon mutation**. *Advances in Bioinformatics* 2014:278385. 2014.

22. Huang L-T, Gromiha MM, Ho S-Y: **iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations**. *Bioinformatics* 2007, **23**:1292–1293.

23. Witvliet DK, Strokach A, Giraldo-Forero AF, Teyra J, Colak R, Kim PM: **ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity**. *Bioinformatics* 2016, **32**:1589–1591.

24. Giollo M, Martin AJ, Walsh I, Ferrari C, Tosatto SC: **NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation**. *BMC Genom* 2014, **15**:1–11.

25. Chen C-W, Lin M-H, Liao C-C, Chang H-P, Chu Y-W: **: predicting protein thermal stability changes by integrating various characteristic modules**. *Comput Struct Biotechnol J* 2020, **18**: 622–630.

26. * Montanucci L, Capriotti E, Frank Y, Ben-Tal N, Fariselli P: **DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations**. *BMC Bioinf* 2019, **20**:1–10.
Method based on evolutionary information (DDGun) and additional structural information (DDGun3D) which predicts stability changes caused by point mutations but also by multiple mutations.

27. Benevenuta S, Pancotti C, Fariselli P, Birolo G, Sanavia T: **An antisymmetric neural network to predict free energy changes in protein variants**. *J Phys Appl Phys* 2021, **54**:245403.

28. Li G, Panday SK, Alexov E: **SAAFEC-SEQ: a sequence-based method for predicting the effect of single point mutations on protein thermodynamic stability**. *Int J Mol Sci* 2021, **22**:606.

29. Sanavia T, Birolo G, Montanucci L, Turina P, Capriotti E, Fariselli P: **Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine**. *Comput Struct Biotechnol J* 2020, **18**:1968–1979.

30. Marabotti A, Scafuri B, Facchiano A: **Predicting the stability of mutant proteins by computational approaches: an overview**. *Briefings Bioinf* 2021, **22**. bbaa074.

31. Kepp KP: **Towards a "golden standard" for computing globin stability: stability and structure sensitivity of myoglobin mutants**. *Biochim Biophys Acta Protein Proteonomics* 2015, **1854**:1239–1248.

32. Fang J: **A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation**. *Briefings Bioinf* 2020, **21**:1285–1292.

33. Iqbal S, Li F, Akutsu T, Ascher DB, Webb GI, Song J: **Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations**. *Briefings in Bioinf* 2021, **22**:bbab184.

34. Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X: **Deep learning in bioinformatics: introduction, application, and perspective in the big data era**. *Methods* 2019, **166**:4–21.

35. Torrisi M, Pollastri G, Le Q: **Deep learning methods in protein structure prediction**. *Comput Struct Biotechnol J* 2020, **18**: 1301–1310.

36. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci Unit States Am* 1992, **89**: 10915–10919.

37. * Caldararu O, Blundell TL, Kepp KP: **Three simple properties explain protein stability change upon mutation**. *J Chem Inf Model* 2021, **61**:1981–1988.
Simple model based on a linear combination of only three features (RSA, and change in hydrophobicity and volume upon mutation), which reaches stability change prediction scores similar to those of much more complex algorithms using hundreds of features.

38. Nisthal A, Wang CY, Ary ML, Mayo SL: **Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis**. *Proc Natl Acad Sci Unit States Am* 2019, **116**: 16367–16377.

39. Montanucci L, Martelli PL, Ben-Tal N, Fariselli P: **A natural upper bound to the accuracy of predicting protein stability changes upon mutations**. *Bioinformatics* 2019, **35**:1513–1517.

40. * AlQuraishi M: **Machine learning in protein structure prediction**. *Curr Opin Chem Biol* 2021, **65**:1–8.
Up–to–date review on protein structure prediction, which presents new ideas that could also be applied to boost the accuracy of predictors of protein stability changes upon mutation.

41. * Caldararu O, Mehra R, Blundell TL, Kepp KP: **Systematic investigation of the data set dependency of protein stability predictors**. *J Chem Inf Model* 2020, **60**:4772–4784.
Extensive analysis of the impact of training data set properties on the accuracy of the prediction of protein stability changes upon mutations;

the type of mutation, the extent of stabilization, the type of structure and the solvent exposure are carefully analyzed as possible sources of bias.

42. Schwersensky M, Rooman M, Pucci F: **Large-scale in silico mutagenesis experiments reveal optimization of genetic code and codon usage for protein mutational robustness**. *BMC Biol* 2020, **18**:1–17.

43. 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–W222.

44. * Pucci F, Bernaerts KV, Kwasigroch JM, Rooman M: **Quantification of biases in predictions of protein stability changes upon mutations**. *Bioinformatics* 2018, **34**:3659–3665.
In–depth study of the bias toward destabilizing mutations, quantified through the introduction of a new balanced mutation data set; almost all stability predictors are shown to suffer from this bias.

45. Dayhoff M, Schwartz R, Orcutt B. In *Atlas of protein sequence and structure*. Washington DC: National Biomedical Research Foundation; 1978:345–352. Ch. A model of evolutionary change in proteins.

46. Gromiha MM, Sarai A: **Thermodynamic database for proteins: features and applications**. *Methods Mol Biol* 2010, **609**:97–112.

47. Ancien F, Pucci F, Godfroid M, Rooman M: **Prediction and interpretation of deleterious coding variants in terms of protein structural stability**. *Sci Rep* 2018, **8**:4480.

48. Pucci F, Bernaerts K, Teheux F, Gilis D, Rooman M: **Symmetry principles in optimization problems: an application to protein stability prediction**. *IFAC-PapersOnLine* 2015, **48**: 458–463.

49. * Usmanova DR, Bogatyreva NS, Ariño Bernad J, Eremina AA, Gorshkova AA, Kanevskiy GM, Lonishin LR, Meister AV, Yakupova AG, Kondrashov FA, *et al.*: **Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation**. *Bioinformatics* 2018, **34**:3653–3658.
Analysis of the prediction bias toward destabilizing mutations of a series of predictors of protein stability changes upon mutations, which are all shown to be biased.

50. Zhou X, Cheng J: **DNpro: a deep learning network approach to predicting protein stability changes induced by single-site mutations**. *Journal of Bioengineering and Life Sciences* 2016, **10**:1–7.

51. LeCun Y, Bengio Y, Hinton G: *Deep learning, Nature* 2015, **521**: 436–444.

52. * Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM: **ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years**. *Nucleic Acids Res* 2021, **49**: D420–D424.
New manually curated database reporting experimental data on the impact of mutations on the thermal and thermodynamic stability of proteins.

53. * Xavier JS, Nguyen T-B, Karmarkar M, Portelli S, Rezende PM, Velloso JP, Ascher DB, Pires DE: **ThermoMutDB: a thermodynamic database for missense mutations**. *Nucleic Acids Res* 2021, **49**:D475–D479.
New manually curated database reporting experimental data on the impact of mutations on the thermal and thermodynamic stability of proteins.

54. * Stourac J, Dubrava J, Musil M, Horackova J, Damborsky J, Mazurenko S, Bednar D: **FireProtDB: database of manually curated protein stability data**. *Nucleic Acids Res* 2021, **49**: D319–D324.
New manually curated database reporting experimental data on the impact of mutations on the thermal and thermodynamic stability of proteins.

55. Lundberg SM, Lee S-I: **A unified approach to interpreting model predictions**. In *Proceedings of the 31st international conference on neural information processing systems*; 2017: 4768–4777.

56. Shrikumar A, Greenside P, Kundaje A: **Learning important features through propagating activation differences**. In Precup D, Teh YW. *Proceedings of the 34th international conference on machine learning*, vol. 70. PMLR; 2017:3145–3153. of Proceedings of Machine Learning Research.

57. Štrumbelj E, Kononenko I: **Explaining prediction models and individual predictions with feature contributions**. *Knowl Inf Syst* 2014, **41**:647–665.

58. Pucci F, Bourgeas R, Rooman M: **High-quality thermodynamic data on the stability changes of proteins upon single-site mutations**. *J Phys Chem Ref Data* 2016, **45**, 023104.

59. Benevenuta S, Fariselli P: **On the upper bounds of the real-valued predictions**. *Bioinf Biol Insights* 2019, **13**, 1177932219871263.

60. Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D: **Protein structure determination using metagenome sequence data**. *Science* 2017, **355**:294–298.

61. Robinson SL, Piel J, Sunagawa S: **A roadmap for metagenomic enzyme discovery**. *Nat Prod Rep* 2021, https://doi.org/10.1039/D1NP00006C.

62. Campeotto I, Goldenzweig A, Davey J, Barfod L, Marshall JM, Silk SE, Wright KE, Draper SJ, Higgins MK, Fleishman SJ: **One-step design of a stable variant of the malaria invasion protein rh5 for use as a vaccine immunogen**. *Proc Natl Acad Sci Unit States Am* 2017, **114**:998–1002.

63. Musil M, Stourac J, Bendl J, Brezovsky J, Prokop Z, Zendulka J, Martinek T, Bednar D, Damborsky J: **FireProt: web server for automated design of thermostable proteins**. *Nucleic Acids Res* 2017, **45**:W393–W399.

64. Schmiedel JM, Lehner B: **Determining protein structures using deep mutagenesis**. *Nat Genet* 2019, **51**:1177–1186.

65. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, Marks DS: **Inferring protein 3D structure from deep mutation scans**. *Nat Genet* 2019, **51**:1170–1176.

66. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589.

67. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, *et al.*: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science* 2021, **373**:871–876.