Exploring evolution to enhance mutational stability prediction

Pauline Hermans[1,2,◇], Matsvei Tsishyn[1,2,◇], Martin Schwersensky[1,2], Marianne Rooman[1,2,†], Fabrizio Pucci[1,2,†]

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles, 1050 Brussels, Belgium

[2] Interuniversity Institute of Bioinformatics in Brussels, 1050 Brussels, Belgium

◇ Equal contributor, † Equal contributor

May 28, 2024

Abstract

Determining the impact of mutations on the thermodynamic stability of proteins is essential for a wide series of applications such as rational protein design and genetic variant interpretation. Since protein stability is a major driver of evolution, evolutionary data are often used to guide stability predictions. Many state-of-the-art stability predictors extract evolutionary information from multiple sequence alignments (MSA) of proteins homologous to a query protein, and leverage it to predict the effects of mutations on protein stability. To evaluate the power of such methods and their limitations, we used the massive amount of stability data recently obtained by deep mutational scanning to study how best to construct MSAs and optimally extract evolutionary information from them. The parameters considered include the protein sequence dataset used for the homologous search, as well as MSA depth, E-value and curation criteria. We also tested different evolutionary models and unexpectedly found that independent-site models achieve the similar accuracy as more complex epistatic models. Interestingly, by combining any of the evolutionary features with a simple structural feature, the relative solvent accessibility of the mutated residue, we obtained similar prediction accuracy of supervised, machine learning-based, protein stability change predictors. Our results provide new insights into the relationship between protein evolution and stability, and show how evolutionary information can be exploited to improve the performance of mutational stability prediction.

1 Introduction

Understanding how mutations impact protein thermodynamic stability is of fundamental interest for a wide range of applications spanning from protein design [1, 2] to the interpretation of genetic variants [3, 4]. Evolution has largely been used to guide computational or experimental mutagenesis aimed at improving protein stability, the latter being the primary driver of evolution [5, 6, 7].

Directed evolution methods, which aim to replicate evolutionary processes *in vitro* through iterative steps of random mutagenesis, selection and amplification, have played a fundamental role

in the (re)design of new proteins with increased stability [8, 9, 10]. Computational mutagenesis approaches have been extensively improved in the last decade [2, 11, 12, 13, 14] and have often been combined with experimental methods to limit the time-consuming exploration of the vast protein sequence space [1, 15]. Recently, deep learning techniques have also been introduced to the field, but their application has often been constrained by the limited availability of experimental thermodynamic data. New high-throughput stability assays, such as the assay recently developed in [16], are currently changing the field by generating folding stability measurements on an impressively large scale, providing essential information to better understand protein stability and to constitute sufficiently large datasets to train deep learning models.

Some current state-of-the-art models for stability prediction [17, 18, 19] already take advantage of evolution and of the huge amount of sequence data available in metagenomic databases [20]. They extract evolutionary information from multiple sequence alignments (MSAs) of proteins homologous to a query protein, which is used to predict how mutations affect protein stability. Although evolution and protein stability are known to be closely related [21], they are not always directly linked. Indeed, functional regions such as catalytic or binding sites are not at all optimized for stability, whereas they clearly show a very strong conservation [22, 23]. In addition, other open issues are still being debated in the field. For example, it is unclear whether or not the effect of mutations is conserved along evolution [24, 25], and what is the role that epistatic effects play in shaping evolution [25]. From a practical point of view, it is also unclear what is the best way to generate and manage MSAs in order to extract evolutionary signals that can improve protein stability prediction and protein design.

In the last couple of years, much attention has been paid to optimizing or subsampling the input MSAs to improve downstream tasks such as predicting three-dimensional (3D) protein structures and of their complexes [26, 27, 28, 29, 30]. A very recent analysis focused on how MSA construction impacts the outcome of mutational fitness prediction [31]. The authors of this paper presented interesting findings, in particular that prediction accuracy does not always correlate with alignment depth. In fact, they showed that evolutionary information extracted from shallow MSAs can sometimes predict the effect of mutations on protein fitness with high accuracy. Adding homologous sequences that are evolutionarily distant from the query sequence does not necessarily improve predictions [31, 32].

In order to validate and extend these findings, we leveraged the massive amount of protein stability data generated in [16] and analyzed how best to extract evolutionary information from MSAs for mutational stability predictions. One of our goals is to gain insight into the relationship between protein stability and evolution. To this end, we compared the ability of various kinds of evolutionary models to predict protein stability. We also highlighted the relationships between residue solvent accessibility (RSA), which is a simple structural feature, the evolutionary conservation of these residues, and the effect of their substitutions on protein stability. Based on these relationships, we devised a way to modulate evolutionary features with RSA, resulting in simple unsupervised models that yield results similar to those of more complex state-of-the-art protein stability change predictors.

2 Methods

2.1 Multiple Sequence Alignment Construction

For each query sequence, we generated MSAs using the iterative JackHMMER method [33] against a given sequence dataset. Basically, JackHMMER performs a quick sequence similarity search on the considered sequence dataset to build a first MSA, which is used to set up a hidden Markov model (HMM). This HMM is then searched against the sequence dataset to select new sequences and construct the next MSA and HMM, and so on. Each iteration refines both the MSA and the HMM used for its construction.

In order to assess the relevance of the MSAs for stability change predictions and compare different ways of constructing them, we tested various numbers of iterations and E-value thresholds in JackHMMER (see Section 3.1). We also evaluated different sequence datasets. We primarily used the UniRef90 set [34], but also tested searching in two larger datasets: UniRef100 and Metagenomics, a metagenomics dataset constructed by merging UniRef90 [34], the Big Fantastic Database [35], the Gut Phage Database [36], MetaEuk [37], the Metagenomic Gut Virus (MGV) [38], SMAG [39], TOPAZ [40] and MGnify [41].

As MSAs can contain clusters of very closely related proteins, measuring the amount of information contained in them by N_{tot}, the total number of sequences it contains, can sometimes be inaccurate. For that reason, we rather used N_{eff}, the effective number of independent sequences, defined as the sum of the weights w_j assigned to each sequence s_j in the MSA [42]:

$$N_{\text{eff}} = \sum_{j=1}^{N_{\text{tot}}} w_j \tag{1}$$

in order to assess the impact of the "depth" of a MSA on predictions.

The weight w_j of a sequence s_j is defined as the inverse of m_j, the number of sequences in the MSA sharing at least 80% of identity with s_j:

$$w_j = \frac{1}{m_j} = \frac{1}{|\{s \in S \,|\, \text{seqid}(s_j, s) \geq 80\%\}|} \tag{2}$$

where S is the set of sequences in the MSA. In this way, sequences without similar sequences in the dataset have weight one, and sequences with similar sequences are downweighted. These weights were computed as in [42] using the **plmc** software from the EVcouplings suite [43].

2.2 Variant Dataset Construction

We used the recent deep mutational scanning dataset constructed in [16], which contains about one million experimental estimations of protein folding free energy (ΔG) obtained through cDNA display proteolysis applied to 396 "wild-type" sequences. We first curated this dataset by averaging duplicate measured ΔG values for specific sequences (both "wild-type" and "mutant"). We focused on single amino acid substitutions and computed the corresponding changes in folding free energy upon mutations, defined as $\Delta\Delta G = \Delta G_{mt} - \Delta G_{wt}$, where wt and mt stand for "wild-type" and for "mutant" respectively.

To reduce protein-specific imbalances in the dataset, we employed the clustering method of CD-HIT [44] with default parameters (sequence identity threshold of 0.9). In this way, we removed closely related sequences and limited the number of "wild-type" sequences to 308. As many of

these proteins have a very small number of known homologs (among which *de novo* designed proteins), we further refined the dataset by removing proteins whose MSA (obtained by JackHMMER on UniRef90 with one iteration and default parameters) contains less than 100 sequences. The resulting dataset, called $\mathcal{D}$, contains $135,056$ $\Delta\Delta G$ values of single-site mutations in 129 protein domains with lengths ranging from 32 to 72 amino acids. The obtained distribution of the experimental $\Delta\Delta G$ values is presented in Supplementary Section 1. As expected, most of the mutations are destabilizing, with an average destabilizing $\Delta\Delta G$ of 0.82 kcal/mol.

In order to investigate the effect of MSA depth on different evolutionary models, we used the N_{eff} value (Eq. 1) as a criterion to partition the dataset $\mathcal{D}$ into subsets: $\mathcal{D}_{10}$ ($10 \leq N_{\text{eff}} < 100$, 17 proteins), $\mathcal{D}_{100}$ ($100 \leq N_{\text{eff}} < 1000$, 27 proteins), $\mathcal{D}_{1000}$ ($1000 \leq N_{\text{eff}} < 10000$, 41 proteins) and $\mathcal{D}_{10000}$, ($10000 \leq N_{\text{eff}}$, 44 proteins). Values of N_{eff} were derived on MSAs constructed on UniRef90 with two JackHMMER iterations and an E-value threshold of 10^{-7}, as we showed in Section 3.1 that these parameter values are optimal for stability predictions.

The structures of all "wild-type" proteins were modeled using AlphaFold2 [27]. The complete dataset as well as the PDB structures used in this paper are available in our GitHub repository `https://github.com/3BioCompBio/EvoStability`.

2.3 Residue and Mutation Properties

We first define three evolution-based features derived from positional amino acid frequencies in a MSA. To avoid divergent values and manage the lack of information in small MSAs, we defined a way to regularize the observed amino acid frequencies. Considering the positional frequency $f_i(a)$ of amino acid a at position i in the MSA and the frequency $f(a)$ of amino acid a in the full MSA, we define the regularized frequencies $\bar{f}_i(a)$ and $\bar{f}(a)$ as:

$$\bar{f}_i(a) = f_i(a)(1-\theta) + \frac{\theta}{21} \quad \text{and} \quad \bar{f}(a) = f(a)(1-\theta) + \frac{\theta}{21}, \tag{3}$$

where θ is a regularization factor that we set to 0.01, and 21 stands for the number of possible states (20 standard amino acids, and 1 gap). Using these regularized frequencies, the three evolution-based features are:

- **Conservation Index (CI)** [45]. It is a measure of the conservation of residues at each position in a MSA:

$$\text{CI}(i) = \left[\frac{1}{2}\sum_{a\in A}\left(\bar{f}_i(a) - \bar{f}(a)\right)^2\right]^{1/2}, \tag{4}$$

  where A is the set of 20 amino acids.

- **Log-Odd Ratio (LOR)** [46]. By comparing the frequencies of wild-type (wt) and mutant (mt) amino acids in the MSA at a given position, we can statistically estimate their prevalence in their evolutionary context. The log-odd ratio of observing the wt amino acid with respect to the mt amino acid at position i is given by:

$$\text{LOR}(i, wt, mt) = \log\frac{\bar{f}_i(wt)}{1-\bar{f}_i(wt)} - \log\frac{\bar{f}_i(mt)}{1-\bar{f}_i(mt)}. \tag{5}$$

  The sign of LOR is set such that a mutation from a highly represented amino acid wt to a poorly represented amino acid mt has a positive LOR, which is comparable to a destabilizing mutation with $\Delta\Delta G > 0$.

- **Weighted Log-Odd Ratio (WLOR)**. To reduce the impact of clusters of closely related sequences in the MSA, we slightly modified the definition of LOR by rectifying the count for amino acid frequencies using the weights assigned to each sequence in the MSA (computed with **plmc** as for N_{eff}, see Section 2.1). The weighted log-odd ratio is defined as:

$$\text{WLOR}(i, wt, mt) = \log \frac{\bar{f}_i^w(wt)}{1 - \bar{f}_i^w(wt)} - \log \frac{\bar{f}_i^w(mt)}{1 - \bar{f}_i^w(mt)} \tag{6}$$

  where the weighted frequency $f_i^w(a)$ of an amino acid a at position i is defined as,

$$f_i^w(a) = \frac{\sum_j w_j \delta(a, a_{ij})}{\sum_j w_j}, \tag{7}$$

  with w_j is the weight of sequence j, a_{ij} the amino acid at position i in sequence j and $\delta(A, B)$ is the Kronecker symbol, which equals one if the A and B coincide, and zero otherwise.

For comparison, we also used the following mutational evolutionary methods:

- **EVcouplings** [43, 47] suite contains two models for predicting the effects of mutations, the epistatic and the independent-site models, which are referred to as EVcouplings-epi and EVcouplings-ind, respectively. The EVcouplings-epi model accounts for epistatis by modeling interactions between all pairs of residues using a Potts model, with coefficients inferred from the input MSA through a pseudo-likelihood maximization approach [48]. EVcouplings-ind is a site-wise maximum entropy model, which does not take into account interactions between sites.

- **GEMME** [49] combines an epistatic and an independent-site version, which will be considered independently here. The epistatic model, noted GEMME-epi, infers the conservation of combinations of mutations across the sequences in the MSA according to an evolutionary tree. The independent-site model, GEMME-ind, consists of the relative per-site frequencies of occurrence of the wild type and mutant residues.

- **ArDCA** [50] is an epistatic method based on an autoregressive model that employs a maximum-likelihood approach to infer the probability of the target sequence.

- **pycofitness** [51] is a direct coupling analysis model that uses a pseudo-likelihood maximization approach [48] to infer the corresponding Potts model parameters. This method is analogous to the EVcouplings-epi model.

Finally, we used one feature based on the protein 3D structure, and defined a way of combining it with the evolutionary features defined above:

- **Relative Solvent Accessibility (RSA)**. The RSA of a residue at position i, RSA(i), is defined as the ratio (in %) of its solvent-accessible surface area in its 3D structure and in a Gly-X-Gly tripeptide in extended conformation [52]. We calculated it using our in-house software MuSiC [53], which uses an extension of the DSSP algorithm [54] and is available on the `www.dezyme.com` website.

- **RSA and evolution (RSA $\odot$ E).**

  As mutations in the protein core tend to be more destabilizing than mutations at the protein surface (see Section 3.4), RSA is a crucial feature in estimating protein stability. For this reason, we devised a simple way of combining any mutational evolutionary score with the RSA as:

$$\text{RSA} \odot E(i, wt, mt) = \left(\frac{100\% - \text{RSA}(i)}{100\%} \right) \cdot E(i, wt, mt). \tag{8}$$

  where $E(i, wt, mt)$ is the evolutionary score of the substitution from residue wt to residue mt at position i.

3 Results

3.1 Impact of MSA construction and curation

We started by evaluating the impact of MSA construction and curation on the ability of independent-site evolutionary scores (i.e., LOR and WLOR, defined in Section 2.3) to predict changes in protein stability. In particular, we tested different protein sequence datasets, numbers of JackHMMER iterations, E-value thresholds and MSA curation criteria. As evaluation metric, we computed the average per-protein Spearman correlation coefficients between the evolutionary scores and the experimental $\Delta\Delta G$ values from the variant dataset $\mathcal{D}$. We present here the results obtained with WLOR scores (Fig. 1), as they turn out to be slightly more predictive than LOR values; results obtained with LOR scores are given in Supplementary Section 2.

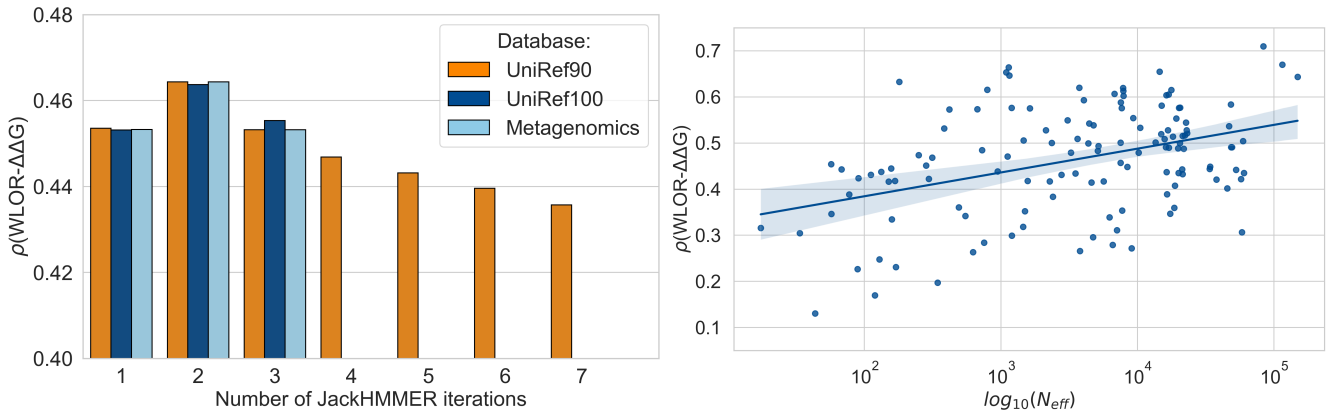
For the numbers of JackHMMER iterations (from 1 to 7), corresponding to successive levels of MSA refinement, we found that MSAs obtained with two iterations yield WLOR scores that best correlate with $\Delta\Delta G$ values (Fig. 1a). This is true regardless of the sequence dataset used. Sequences added during the second iteration contribute to the diversity of the MSAs, which is beneficial for predicting changes in protein stability. Subsequent iterations appear to introduce noise in the MSA rather than adding any useful information.

We also constructed MSAs from three sequence datasets: UniRef90, UniRef100 and Metagenomics (described in Section 2.1). As represented in Fig. 1a, the WLOR-$\Delta\Delta G$ Spearman correlation coefficients are almost identical irrespective of the sequence dataset used. The use of larger sequence datasets enriches the MSAs with homologous sequences. However, this neither increases nor decreases the correlation, which means that the additional sequences probably duplicate information already present in the MSAs constructed with the smaller UniRef90 sequence dataset. Note that this is different to what is happening for protein structure prediction where querying large metagenomics datasets for the input MSA construction seems to boost method performance [26, 27]. In the following, we will use MSAs built from the UniRef90 sequence dataset and two JackHMMER iterations.

In addition to the factors mentioned above, the quality of a MSA also depends on the E-value threshold used to construct it. We tested E-value thresholds between 1 and 10^{-22}, knowing that the default value in JackHMMER is 10^{-3}. As shown in Fig. 2, the WLOR-$\Delta\Delta G$ correlation is relatively stable for E-values between 10^{-2} and 10^{-8}. The highest correlation coefficient (0.467) is reached for an E-value threshold of 10^{-7}. For E-value thresholds below 10^{-13}, the correlation drops

significantly, since the obtained MSAs become significantly smaller and composed of sequences that are very close to the query.

We then evaluated the impact of N_{eff}, the effective number of sequences in a MSA, on the correlation between evolutionary scores and protein stability changes. As shown in Fig. 1b, the WLOR-$\Delta\Delta G$ correlation is better for proteins whose MSA has larger N_{eff} values. This ensures that the MSA better represents the diversity of homologous proteins. However, it is noteworthy that there are performance differences among the proteins in the dataset $\mathcal{D}$, even with similar N_{eff}. Indeed, the WLOR-$\Delta\Delta G$ correlation coefficients vary considerably from one protein to another (see Fig. 1b and more on this phenomenon in Section 3.2).



(a) (b)

Figure 1: (a) Average per-protein Spearman correlation coefficients ρ between WLOR and $\Delta\Delta G$ on the dataset $\mathcal{D}$, as a function of the protein sequence dataset and the number of JackHMMER iterations used to build the MSAs. (b) Per-protein Spearman correlation coefficients ρ between WLOR and $\Delta\Delta G$ on the dataset $\mathcal{D}$, as a function of the effective number of sequences N_{eff} in the MSAs (derived from MSAs built using UniRef90 and two JackHMMER iterations).

To further investigate the effect of MSA depth, we separately measured correlations on subsets of $\mathcal{D}$ split based on the N_{eff} criterion, as defined in Section 2.2. As shown in Fig. 2, MSAs with lower (higher) N_{eff} consistently lead to weaker (stronger) correlations. Indeed, for the best E-value threshold (10^{-7}), the average Spearman correlation coefficients between WLOR and $\Delta\Delta G$ are equal to 0.37 on $\mathcal{D}_{10}$, 0.44 on $\mathcal{D}_{100}$, 0.48 on $\mathcal{D}_{1000}$ and 0.50 on $\mathcal{D}_{10000}$.

Finally, we evaluated the effect of MSA curation by removing sequences based on their gap ratio and sequence identity with the query sequence. We found almost no effect of the curation on the WLOR-$\Delta\Delta G$ Spearman correlation coefficients, regardless of the sequence dataset used (see in Supplementary Section 2). While MSA curation has been reported as an important step for better predictions [49], its effect is marginal on the dataset $\mathcal{D}$. This is probably due to the fact that $\mathcal{D}$ only contains small protein domains, which makes MSA construction simpler. For instance, larger, multi-domain proteins usually have MSAs with sequences that only partially cover the query sequence, on which curation could have a stronger effect.

Following the above analysis, in the rest of the article we will use MSAs constructed from UniRef90 with two JackHMMER iterations and an E-value threshold of 10^{-7}. This choice yields the highest WLOR-$\Delta\Delta G$ correlation, i.e. $\rho = 0.47$.

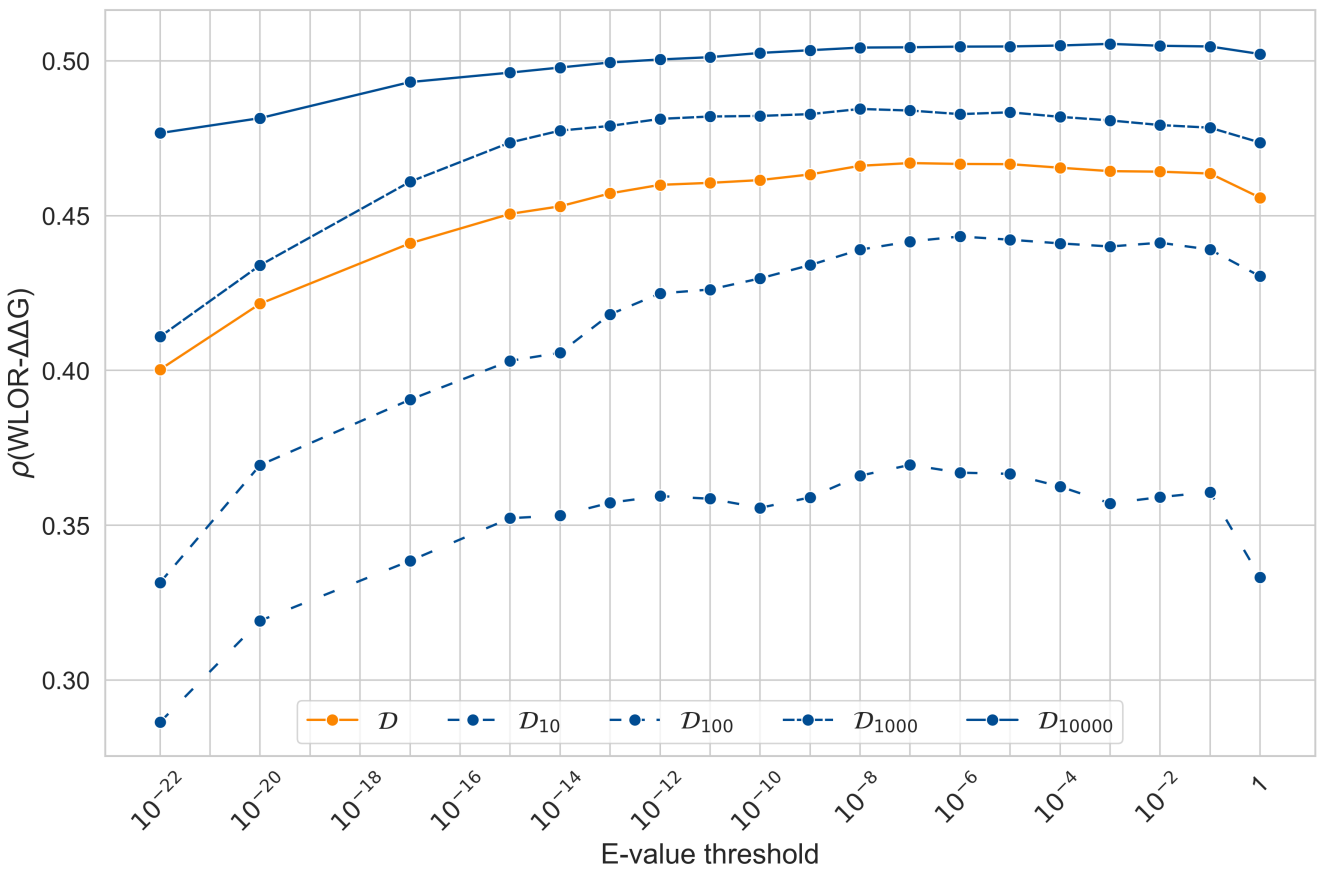


Figure 2: Average per-protein Spearman correlation coefficients ρ between WLOR and $\Delta\Delta G$ on the dataset $\mathcal{D}$ and the subsets $\mathcal{D}_{10}$, $\mathcal{D}_{100}$, $\mathcal{D}_{1000}$ and $\mathcal{D}_{10000}$; WLOR values were derived from MSAs built using UniRef90, two JackHMMER iterations, and E-value thresholds between 10^{-22} and 1.

3.2 Divergent predictions on highly similar proteins

As already discussed in the previous subsection, evolutionary scores tend to become more predictive for $\Delta\Delta G$ as the depth of the MSA increases. However, as shown in Fig. 1b, the performance on different proteins varies (from correlations below 0.2 to correlations above 0.7) even for a similar MSA depth. The strong protein-to-protein variability of the predictions therefore cannot be explained by MSA depth alone. Moreover, note that proteins for which $\Delta\Delta G$ is well predicted by one of the evolutionary models tend to also be well predicted by other models, and vice versa (see Supplementary Section 3).

To explore this performance variability, we focused on two proteins from $\mathcal{D}$ with a high sequence identity but notable differences in stability prediction performance. Both are 57-residue $\beta1$ domains of the streptococcal G protein. The first, $G\beta1_{pH}$, is an engineered pH-sensitive mutant of *Finegoldia magna* whose sequence is identical to that of the PDB structure 2ZW1 [55] with the substitution V54S (referred to in our dataset as `2zw1_A_0-56_V54S`). The second, $G\beta1_{T_m}$, is a redesigned hyperthermophilic heptamutant from *Streptococcus sp. G148*; its sequence is identical to that of the PDB structure 1GB4 [56] with the substitution F53D (referenced as `1gb4_A_1-57_F53D`).

The two proteins share 70% sequence identity and very similar 3D structures, with a root mean square deviation (RMSD) of their C_α atoms of 0.4 Å (Fig. 3a). However, they show very different results in terms of WLOR-$\Delta\Delta G$ Spearman correlation coefficients : $\rho = 0.53$ and 0.17 for $G\beta1_{pH}$

and $G\beta1_{T_m}$, respectively (Tab. 1 and Fig. 3c-3d).

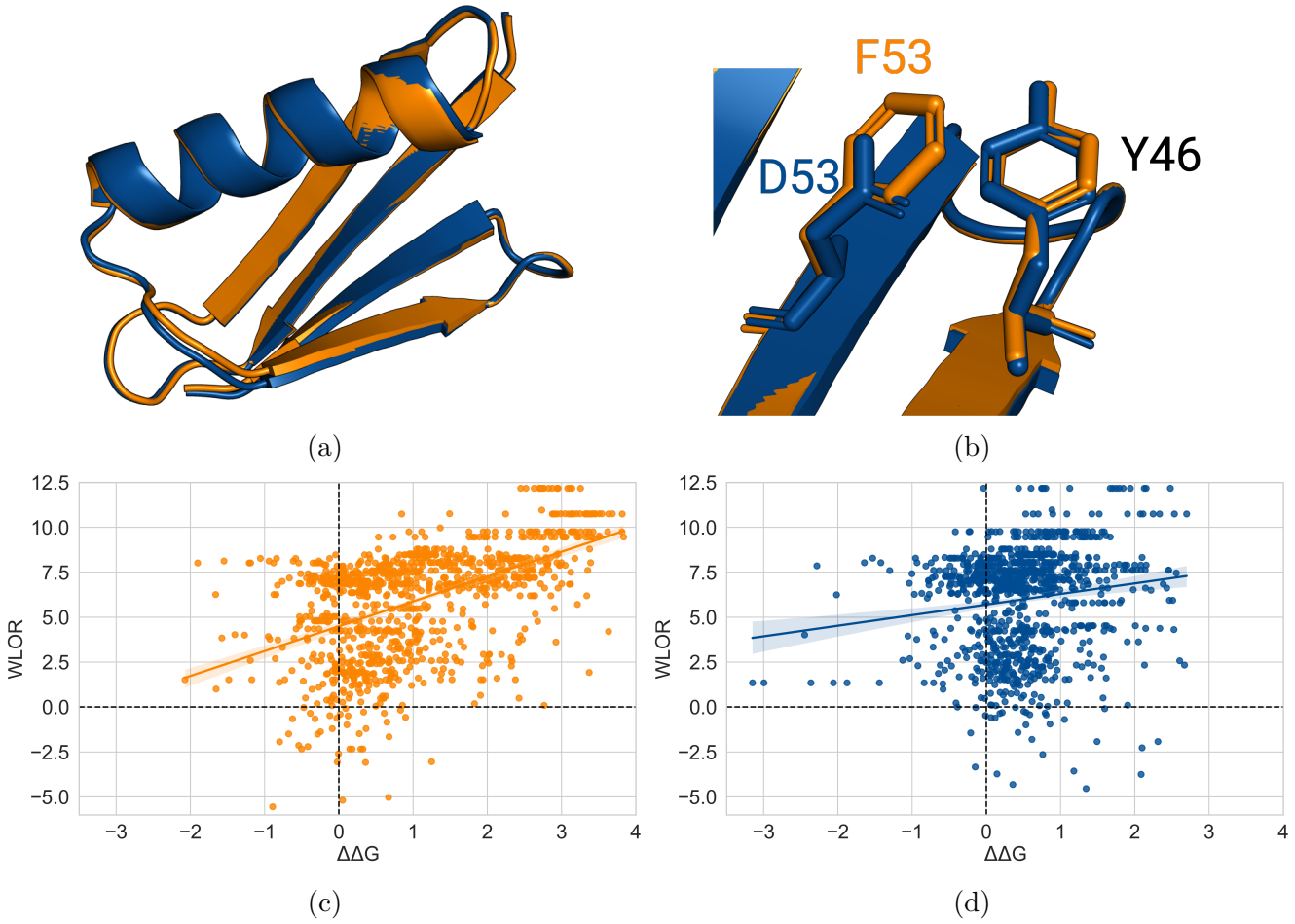


Figure 3: (a-b) Superimposed structures of proteins $G\beta1_{pH}$ (in orange) and $G\beta1_{T_m}$ (in blue). (b) Interactions between residues 46 and 53. (c-d) Comparison between WLOR scores and experimental $\Delta\Delta G$ values for $G\beta1_{pH}$ (c, in orange) and $G\beta1_{T_m}$ (d, blue).

Table 1: Spearman correlation coefficients ρ between experimental $\Delta\Delta G$ values and scores obtained with WLOR, pycofitness (PCF) [51] and PoPMuSiC (PoP) [12] for the two Gβ1 proteins.

Protein	Length (aa)	N_{eff}	ρ(WLOR-$\Delta\Delta G$)	ρ(PCF-$\Delta\Delta G$)	ρ(PoP-$\Delta\Delta G$)
$G\beta1_{pH}$	57	30.3	0.53	0.56	0.63
$G\beta1_{T_m}$	57	30.5	0.17	0.14	0.55

Since the two proteins have a close evolutionary relationship, they have almost identical MSAs and their evolutionary scores are expected to be similar. This is indeed the case: the Spearman correlation between their WLOR values at positions sharing the same amino acid is equal to 0.99. Notably, however, their respective $\Delta\Delta G$ scores at the same positions display a much lower Spearman correlation of 0.75. This suggests that epistatic effects, which are not captured by WLOR scores, play an important role in the stability properties of these proteins.

An example of epistatic effect can be observed by looking at residue 46, which is occupied by a tyrosine in both $G\beta1_{pH}$ and $G\beta1_{T_m}$. This position is highly conserved, with 98% of tyrosine in both

MSAs; it thus has a strongly unfavorable WLOR score for substitutions. However, while mutations at that position are all strongly destabilizing for $G\beta1_{pH}$ (average $\Delta\Delta G$ of 3.0 kcal/mol), their effect is only mildly destabilizing to neutral for $G\beta1_{T_m}$. Analyzing this residue in its 3D structure, we found it to be located in a partially buried region and to interact with residue 53. It forms a stabilizing $\pi - \pi$ (parallel-displaced) interaction with F53 in $G\beta1_{pH}$, and an anion$-\pi$ interaction with D53 in $G\beta1_{T_m}$, known to be barely stabilizing (Fig. 3b). Phenylalanine at position 53 is highly conserved in evolution, whereas D is an engineered mutation that only occurs in the wild-type. The F53D substitution from $G\beta1_{pH}$ to $G\beta1_{T_m}$ has thus a strong epistatic effect on position 46: in short, it changes Y46 mutations from very destabilizing to stabilizing or neutral, an effect that cannot be captured by independent-site WLOR scores. As a consequence, WLOR is less accurate for Y46 mutations in $G\beta1_{T_m}$ than in $G\beta1_{pH}$.

Coevolution-based models that account for epistatic interactions should perform better and capture these effects. Unfortunately, this does not appear to be the case, as the mean pycofitness scores for all substitutions of Y46 are the same for both proteins. In addition, the Spearman correlations of $\Delta\Delta G$ values with pycofitness scores are very similar to those with WLOR scores for both proteins, as shown in Tab. 1, suggesting a difficulty in efficiently learning epistatic patterns. Similar results were observed for epistatic models other than pycofitness. We will return to the ability of such models to inform about protein stability in the next subsection.

Finally, we compared the results of evolutionary models with those of PoPMuSiC [12], a structure-based $\Delta\Delta G$ prediction model that relies on statistically derived physical features. Although a small difference in performances is still observed between $G\beta1_{pH}$ and $G\beta1_{T_m}$, PoPMuSiC is able to better capture the effect of mutations in both proteins, with correlations as high as 0.63 and 0.55, respectively.

These observations show that the ability of evolutionary information to explain stability differs from one protein to another. Although we were able to observe and analyze such differences on a per-protein basis, we were unable to identify biophysical protein characteristics or properties of their MSAs that could consistently explain this variability.

For example, although we interpreted that the low prediction accuracy of Y46 mutations in $G\beta1_{T_m}$ was due to the interacting residue at position 53 being occupied by an unusual amino acid, this can in no way be generalized to the entire sequence. Rather, it appears that a few key positions, such as 53, have complex epistatic effects on the mutational landscape and make the effects of mutations less predictable using only their evolutionary history [57]. Analyses performed on experimental deep mutational scanning data of domain $\beta1$ of protein G reveal that a minority of pairs (about 5%) exhibit significant epistatic effects on structural stability, but that a larger fraction (about 30%) exhibit weaker forms of epistasis [58] that are even more difficult to detect.

In the next section, we therefore test a series of epistatic models that take into account these epistatic effects and compare their performance in predicting $\Delta\Delta G$ with several independent models, such as WLOR.

3.3 Assessing independent-site and epistatic evolutionary models for $\Delta\Delta G$ prediction

Here, we evaluated the ability of various evolution-based methods (introduced in Section 2.3) to detect the impact of mutations on protein stability. More precisely, we tested the independent-site models LOR, WLOR, GEMME-ind [49] and EVcouplings-ind [47], as well as the epistatic models pycofitness [51], ArDCA [50], GEMME-epi [49] and EVcouplings-epi [47]. For that purpose, we computed the average per-protein Spearman correlation coefficient between these evolutionary

scores and experimental $\Delta\Delta G$ on dataset $\mathcal{D}$ and on its subsets (split by MSA depth, see Section 2.2). The performance of the evolutionary models is shown in Fig. 4 and Tab. 2.

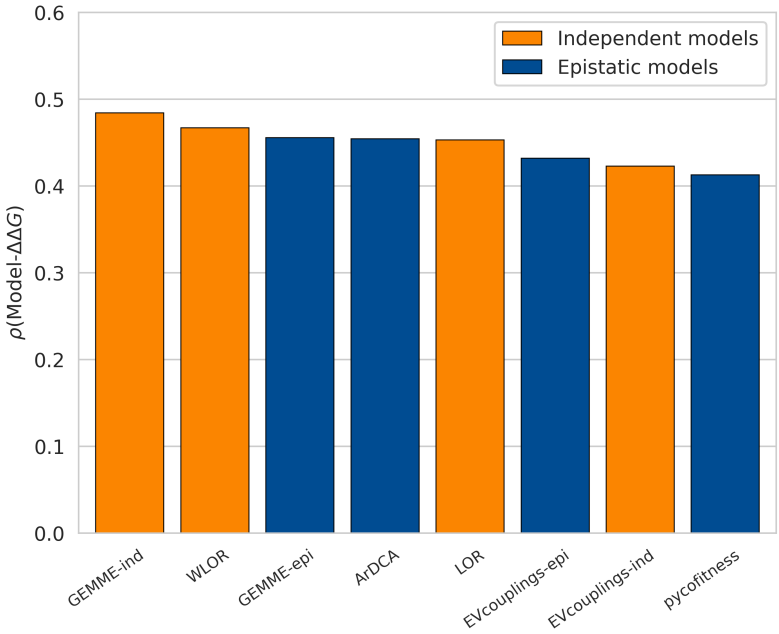


Figure 4: Average per-protein Spearman correlation coefficients ρ between experimental $\Delta\Delta G$ values and evolutionary scores derived from the tested independent-site models (in orange) and epistatic models (in blue), on the dataset $\mathcal{D}$.

Table 2: Average per-protein Spearman correlation coefficients ρ between experimental $\Delta\Delta G$ values and evolutionary scores derived from independent-site and epistatic models, on the dataset $\mathcal{D}$ and the N_{eff}-dependent subsets $\mathcal{D}_M$, where $M \in \{10, 100, 1000, 10000\}$.

Method	Type	$\mathcal{D}$	$\mathcal{D}_{10}$	$\mathcal{D}_{100}$	$\mathcal{D}_{1000}$	$\mathcal{D}_{10000}$
GEMME-ind	Independent	**0.48**	**0.38**	**0.44**	**0.51**	0.53
WLOR	Independent	0.47	0.37	0.44	0.48	0.50
GEMME-epi	Epistatic	0.46	0.37	0.42	0.48	0.49
ArDCA	Epistatic	0.45	0.25	0.39	0.50	**0.53**
LOR	Independent	0.45	0.35	0.43	0.47	0.50
EVcouplings-epi	Epistatic	0.43	0.36	0.42	0.43	0.47
EVcouplings-ind	Independent	0.42	0.36	0.42	0.42	0.45
pycofitness	Epistatic	0.41	0.34	0.40	0.42	0.45

First, we observe that the performance is in general surprisingly good for evolutionary features that are not always directly related to protein stability. Indeed, a low frequency mutant residue in a MSA does not necessarily imply reduced stability, but can be caused by a different effect such as protein function. Regardless of the method tested, we observe that, as expected, the correlations are better, on the average, for proteins that have a large number of homologs. While correlation coefficients close to 0.5 are reached for all tested methods on the dataset $\mathcal{D}_{10000}$ (with $N_{\text{eff}} \geq 10000$), performance is surprisingly still quite good, albeit lower, for shallow MSAs with correlation coefficients in the range of $0.25 - 0.4$ on the dataset $\mathcal{D}_{10}$. How the performance of these

different methods depends on the MSA depth is explored in more detail in Supplementary Section 3 and 4.

Surprisingly, while epistatic models seem to be more complete and have already established their relevance in multiple applications [59], simple single-site independent methods seem to be equally or even slightly more predictive for stability changes. Whether this trend is due to the relatively low impact of epistatic effects or to the poor ability of these models to describe mutational effects remains to be understood. It may be that epistatic models are more sensitive to MSA depth and curation than single-site methods. As we can see in Tab. 2, ArDCA indeed seems to be much more sensitive to MSA depth than other models, showing the lowest correlations of 0.25 for proteins with the smallest MSA ($\mathcal{D}_{10}$) and the highest correlation of 0.53 for proteins with the largest MSA ($\mathcal{D}_{10000}$). However, this high sensitivity to MSA depth is not general to all epistatic models and thus cannot fully explain their slightly lower performance. To further highlight these observations, we performed a MSA subsampling analysis, which indeed confirmed the high sensitivity of ArDCA to MSA depth compared to other methods (see Supplementary Section 4).

In summary, evolutionary-based features are able to accurately estimate stability changes upon mutations. Single-site independent models, which assume that residues have evolved independently of their context, tend to perform as well or even slightly better than more complex epistatic models when correlated to $\Delta\Delta G$, even if the latter take into account epistatic interactions that are known to play an important role in the evolutionary trajectory of proteins [25, 57, 60].

3.4 Residue solvent accessibility and protein stability

It is instructive to analyze how the residue solvent accessibility of the mutated residue is related to both protein stability and evolution. It is well known that RSA plays an essential role in shaping protein stability, as substitutions in the core tend to be more destabilizing than those at the surface. Indeed, residues in the protein core are more constrained by multiple interactions with surrounding residues and therefore cannot be easily substituted without losing structural stability. As a result, most of the state-of-the-art stability change predictors (such as [12, 61, 18, 62]) directly or indirectly use RSA as a feature in their model. The large scale and systematic nature of the deep mutational scanning dataset $\mathcal{D}$ allows us to further explore the relations between RSA and protein stability and evolution.

We first note the very strong anti-correlation of -0.50 between RSA and $\Delta\Delta G$. As shown in Fig. 5a, mutations in low RSA regions display a wide distribution with high destabilizing median values, while mutations in high RSA regions display a narrow distribution with a median only slightly above zero kcal/mol. Mutations of completely buried residues (with RSA = 0%) reach highest destabilizing median values of 2.0 kcal/mol. The median then decreases up to RSA values around 50% and from there the median remains stable around 0.2 kcal/mol.

The RSA dependency is particularly strong for the effect of variants on stability. This can be observed by examining the RSA dependence of single-site mutation scores from ProteinGym [63], a large-scale standard dataset on the effect of protein variants on fitness, which collects many multi-label deep mutational scanning experiments. RSA shows an average per-protein Spearman correlation of -0.50 on mutagenesis data from stability-based assays (essentially composed of values from [16] similar to our dataset $\mathcal{D}$). In contrast, data from other type of assays exhibit still remarkable but much weaker average correlations equal to -0.27, -0.28, -0.32 and -0.33 for binding-, organismal fitness-, expression- and activity-based assay, respectively.

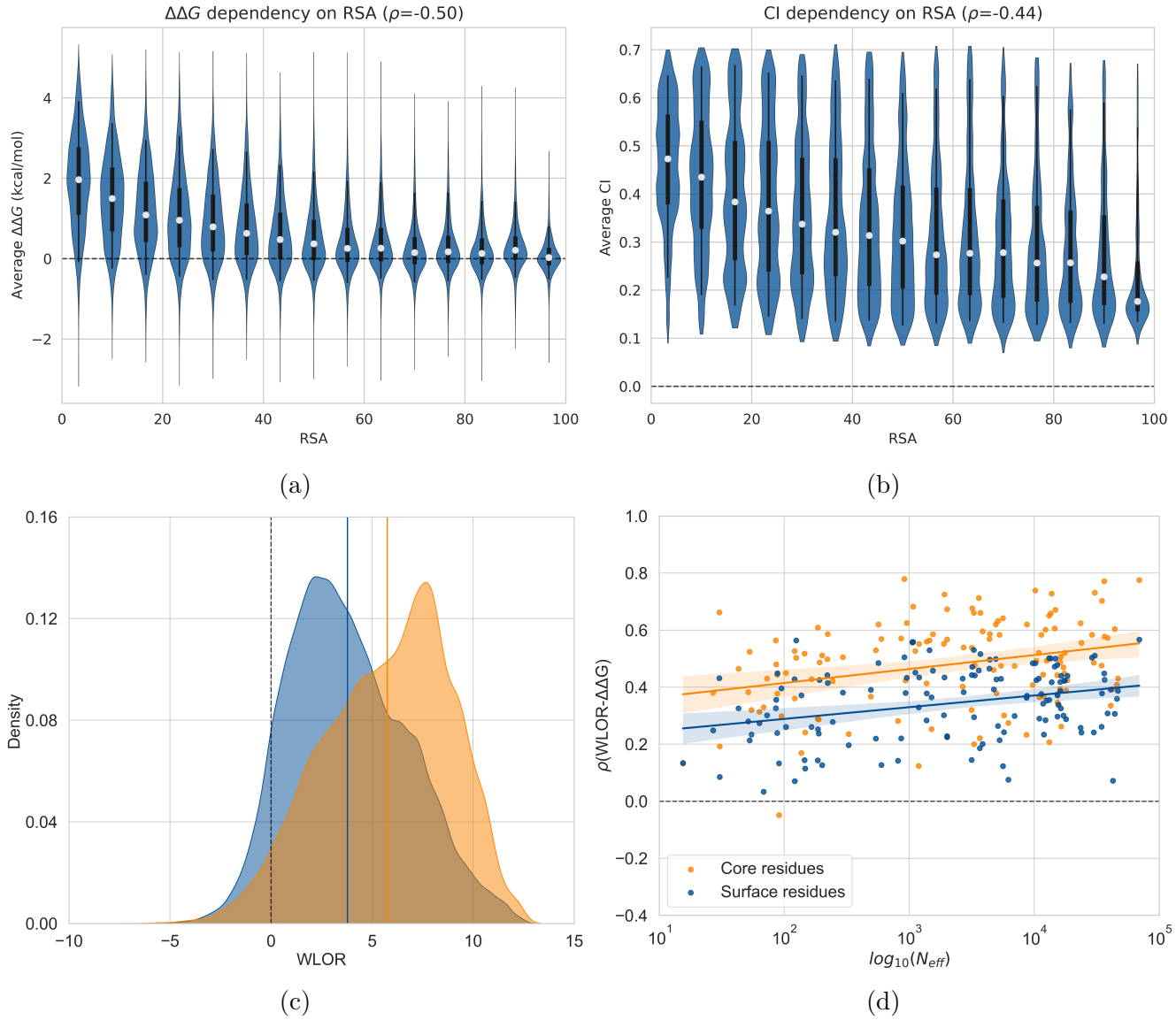


Figure 5: Distributions from the dataset $\mathcal{D}$ of (a) $\Delta\Delta G$ values and (b) CI values as a function of the RSA of the wild-type residues. White dots represent the median, boxes (thick black lines) represent the inter-quartile range from quartile 1 to quartile 3, and the whiskers (thin back lines) represent the inter-percentile range from percentile 5 to percentile 95. (c) Distribution of WLOR for mutations located in the core (in orange) and at the surface (in blue). (d) Per-protein Spearman correlation coefficients ρ between WLOR and $\Delta\Delta G$ for mutations located in the core (in orange) and at the surface (in blue) as a function of N_{eff} in the corresponding MSA.

However, it is important to note that the observed anti-correlation between RSA and $\Delta\Delta G$ may be partially overestimated on the $\mathcal{D}$ dataset, as it contains only small monomeric protein domains. Indeed, large multi-domain protein structures are expected to exhibit more complex $\Delta\Delta G$ distributions as, for example, large proteins are mainly composed of buried regions (with near-zero RSA values), which reduces the RSA signal.

Values of RSA are also strongly related to evolution, as residues at the surface tend to evolve much faster than those in the core [64, 65, 66]. This is confirmed by the anti-correlation of -0.44 between RSA and the conservation index CI (Fig. 5b). Average conservation monotonically

decreases as RSA increases, with higher rates in low RSA regions ($0\% - 20\%$). Unconserved average values for high RSA regions are explained by the fact that such residues are mostly located in variable loop regions or at the C- and N-terminus of the protein where the evolutionary pressure is low compared with higher evolutionary pressure that is exerted on core residues.

From the perspective of mutational evolutionary score WLOR, we observe its average value is notably higher on core residues (RSA $< 20\%$) than on surface residues (RSA $\leq 20\%$) with averages of 5.75 and 3.78, respectively as shown in Fig. 5c. Moreover, WLOR displays a stronger predictive power on $\Delta\Delta G$ for the core residues than for the surface ones ($\rho = 0.48$ and $\rho = 0.34$ respectively, see Fig. 5d). However, it is important to note that other factors may influence this result. First, surface residues are submitted to strong evolutionary pressures from other protein properties such as the optimization of the binding to other proteins, ligands or nucleic acids. Second, since surface residues display smaller stability variations upon mutations, these effects are more subtle and thus more difficult to predict.

3.5 Combining RSA and evolution for $\Delta\Delta G$ prediction

As shown in the previous subsection, RSA is strongly related to both protein stability and conservation. The higher evolutionary pressure for stability in the core of proteins makes RSA a promising indicator to modulate signals extracted from MSAs. We therefore used the RSA as defined in Eq. 8 in order to improve the predictive power of the evolutionary scores (analyzed in Section 3.3) for $\Delta\Delta G$ prediction. Note that we "reversed" the RSA factor: we used $(100\% - \text{RSA})/100\%$ in order to reflect the strong anti-correlation between RSA and the amplitude of $\Delta\Delta G$, which helps to modulate the amplitude of the evolutionary score. Notably, loop regions and C- and N-termini of proteins tend to be poorly conserved and thus exhibit variable MSAs with a high gap ratio. Consequently, while mutations in these flexible regions are characterized by neutral $\Delta\Delta G$ values, evolutionary models tend to wrongly assign them as destabilizing. The RSA factor helps to modulate this effect by lowering the predicted scores.

The performance of all evolutionary methods combined with RSA to predict protein stability changes on the dataset $\mathcal{D}$ is presented in Fig. 6 and Tab. 3. Additionally, we compared them to four state-of-the-art $\Delta\Delta G$ prediction models (PoPMuSiC [12], PremPS [18], DDMut [62], and MAESTRO [61, 67]). PoPMuSiC and MAESTRO are pure structure-based predictors that contain only biophysical and statistical features and use experimental data only to identify the parameters of their respective models. In contrast, PremPS and DDMut also include some evolutionary features. PremPS extracts evolutionary features from the MSA of query proteins, whereas DDMut uses only general mutation matrices. Of these four $\Delta\Delta G$ predictors, only DDMut uses more complex deep learning techniques.

Modulation by RSA significantly improves the performance of all evolutionary models, improving correlations ρ ranging from 0.42 to 0.48 up to ρ values ranging from 0.56 to 0.60. Surprisingly, these scores are comparable or in some cases even better than those of the supervised $\Delta\Delta G$ prediction models. For example, the best evolutionary score of 0.60 achieved by RSA $\odot$ ArDCA outperforms all $\Delta\Delta G$ predictors except PoPMuSiC (which has a slightly higher correlation of 0.61). Overall, when modulated by RSA, independent-site and epistatic models achieve similar performances.

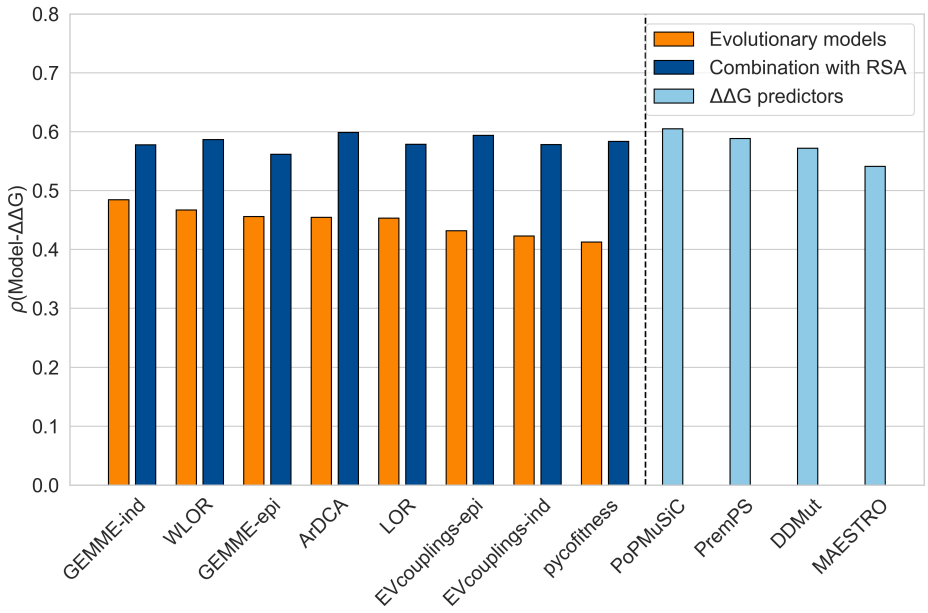


Figure 6: Average per-protein Spearman correlation coefficients ρ between experimental $\Delta\Delta G$ values and evolutionary scores (in orange), evolutionary scores combined with RSA (in dark blue) and predicted $\Delta\Delta G$ values (in light blue), on the variant dataset $\mathcal{D}$.

Table 3: Average per-protein Spearman (ρ) correlation coefficients between experimental $\Delta\Delta G$ values and evolutionary scores or predicted $\Delta\Delta G$ values on the variant dataset $\mathcal{D}$.

Method	Type	ρ(Model-$\Delta\Delta G$)
GEMME-ind	Independent	**0.48**
WLOR	Independent	0.47
GEMME-epi	Epistatic	0.46
ArDCA	Epistatic	0.45
LOR	Independent	0.45
EVcouplings-epi	Epistatic	0.43
EVcouplings-ind	Independent	0.42
pycofitness	Epistatic	0.41
RSA $\odot$ GEMME-ind	Str+Independent	0.58
RSA $\odot$ WLOR	Str+Independent	0.59
RSA $\odot$ GEMME-epi	Str+Epistatic	0.56
RSA $\odot$ ArDCA	Str+Epistatic	**0.60**
RSA $\odot$ LOR	Str+Independent	0.58
RSA $\odot$ EVcouplings-epi	Str+Epistatic	0.59
RSA $\odot$ EVcouplings-ind	Str+Independent	0.58
RSA $\odot$ pycofitness	Str+Epistatic	0.58
PoPMuSiC	$\Delta\Delta G$ predictor	**0.61**
PremPS	$\Delta\Delta G$ predictor	0.59
DDMut	$\Delta\Delta G$ predictor	0.57
MAESTRO	$\Delta\Delta G$ predictor	0.54

Finally, we note that, as expected, the RSA-modulated models exhibit similar behavior to

the evolutionary models alone with respect to MSA depth (section 3.3), showing performances that increase with deeper MSAs (see Supplementary Section 5). Note however, that for shallow MSAs, their performance is only slightly better than the RSA alone, while for deeper MSAs the performance is notably higher.

In conclusion, in the era of deep learning and large language models, it is impressive to see that simple unsupervised models including only amino acid frequencies and their localization in the structure achieve state-of-the-art performance.

4 Conclusion

Evolution and protein stability are two correlated quantities. Indeed, protein stability is known to be one of the major factors driving evolution, and on the other hand, evolution puts constraints on protein stability. In this paper, we take advantage of the huge mutational stability dataset published in [16] to carefully analyze how to leverage evolutionary information to predict the effect of mutations on protein stability. The main findings of the paper are summarized below:

- The quality of the used MSA plays a fundamental role in the ability of evolution-based methods to extract information about protein stability. In particular, proteins with deeper MSAs tend to be largely better predicted than proteins with shallow MSAs. However, having more evolution-derived data does not necessarily lead to better scores, as already highlighted in [31].
- Among the various evolutionary models tested to predict the effect of mutations on protein stability, we found that independent-site models perform as well as or slightly better than coevolutionary models. Thus, although these models achieve very good performance in predicting 3D contacts between residues, they do not seem to have an advantage over independent-site models in predicting changes in protein stability.
- RSA is a key feature strongly anti-correlated with changes in protein stability caused by mutations. Even more impressively, the simple unsupervised combination of RSA with the evolutionary features we analyzed leads to performance similar to that of the state-of-the-art supervised $\Delta\Delta G$ predictors.

In summary, we have demonstrated that evolutionary data can be effectively used in an unsupervised way to predict the impact of mutations on protein stability. Since the performance obtained by modulating evolutionary signals with RSA is similar to that of complex, supervised $\Delta\Delta G$ predictors, an immediate perspective is to combine these models with the aim of further improving $\Delta\Delta G$ prediction.

Acknowledgment

We acknowledge financial support from the Belgian Fund for Scientific Research (F.R.S.-FNRS) through a PDR project. M.T. benefits from a FNRS-FRIA PhD grant. P.H. benefits from a Win4Doc grant from SPW Recherche of the Walloon Region.

References

[1] I. Coluzza, "Computational protein design: a review," *Journal of Physics: Condensed Matter*, vol. 29, no. 14, p. 143001, 2017.

[2] F. Pucci, M. Schwersensky, and M. Rooman, "Artificial intelligence challenges for predicting the impact of mutations on protein stability," *Current opinion in structural biology*, vol. 72, pp. 161–168, 2022.

[3] L. Gerasimavicius, X. Liu, and J. A. Marsh, "Identification of pathogenic missense mutations using protein stability predictors," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.

[4] S. Iqbal, E. Pérez-Palma, J. B. Jespersen, P. May, D. Hoksza, H. O. Heyne, S. S. Ahmed, Z. T. Rifat, M. S. Rahman, K. Lage, *et al.*, "Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants," *Proceedings of the National Academy of Sciences*, vol. 117, no. 45, pp. 28201–28211, 2020.

[5] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, "Protein stability promotes evolvability," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5869–5874, 2006.

[6] K. B. Zeldovich, P. Chen, and E. I. Shakhnovich, "Protein stability imposes limits on organism complexity and speed of molecular evolution," *Proceedings of the National Academy of Sciences*, vol. 104, no. 41, pp. 16152–16157, 2007.

[7] R. Kurahashi, S. Sano, and K. Takano, "Protein evolution is potentially governed by protein stability: directed evolution of an esterase from the hyperthermophilic archaeon sulfolobus tokodaii," *Journal of Molecular Evolution*, vol. 86, no. 5, pp. 283–292, 2018.

[8] M. S. Packer and D. R. Liu, "Methods for the directed evolution of proteins," *Nature Reviews Genetics*, vol. 16, no. 7, pp. 379–394, 2015.

[9] Y. Wang, P. Xue, M. Cao, T. Yu, S. T. Lane, and H. Zhao, "Directed evolution: methodologies and applications," *Chemical reviews*, vol. 121, no. 20, pp. 12384–12444, 2021.

[10] F. H. Arnold, "Directed evolution: bringing new chemistry to life," *Angewandte Chemie International Edition*, vol. 57, no. 16, pp. 4143–4148, 2018.

[11] T. Sanavia, G. Birolo, L. Montanucci, P. Turina, E. Capriotti, and P. Fariselli, "Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine," *Computational and structural biotechnology journal*, vol. 18, pp. 1968–1979, 2020.

[12] Y. Dehouck, J. M. Kwasigroch, D. Gilis, and M. Rooman, "PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–12, 2011.

[13] D. E. Pires, D. B. Ascher, and T. L. Blundell, "mcsm: predicting the effects of mutations in proteins using graph-based signatures," *Bioinformatics*, vol. 30, no. 3, pp. 335–342, 2014.

[14] L. Montanucci, E. Capriotti, Y. Frank, N. Ben-Tal, and P. Fariselli, "DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations," *BMC bioinformatics*, vol. 20, pp. 1–10, 2019.

[15] I. V. Korendovych, "Rational and semirational protein design," *Protein engineering: methods and protocols*, pp. 15–23, 2018.

[16] K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani, J. J. Weinstein, N. M. Mangan, S. Ovchinnikov, and G. J. Rocklin, "Mega-scale experimental analysis of protein folding stability in biology and design," *Nature*, vol. 620, no. 7973, pp. 434–444, 2023.

[17] P. Fariselli, P. L. Martelli, C. Savojardo, and R. Casadio, "Inps: predicting the impact of non-synonymous variations on protein stability from sequence," *Bioinformatics*, vol. 31, no. 17, pp. 2816–2821, 2015.

[18] Y. Chen, H. Lu, N. Zhang, Z. Zhu, S. Wang, and M. Li, "PremPS: Predicting the impact of missense mutations on protein stability," *PLoS computational biology*, vol. 16, no. 12, p. e1008543, 2020.

[19] M. H. Høie, M. Cagiada, A. H. B. Frederiksen, A. Stein, and K. Lindorff-Larsen, "Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation," *Cell reports*, vol. 38, no. 2, p. 110207, 2022.

[20] Q. Hou, F. Pucci, F. Pan, F. Xue, M. Rooman, and Q. Feng, "Using metagenomic data to boost protein structure prediction and discovery," *Computational and Structural Biotechnology Journal*, 2022.

[21] N. Ota, R. Kurahashi, S. Sano, and K. Takano, "The direction of protein evolution is destined by the stability," *Biochimie*, vol. 150, pp. 100–109, 2018.

[22] Q. Hou, M. Rooman, and F. Pucci, "Enzyme stability-activity trade-off: new insights from protein stability weaknesses and evolutionary conservation," *Journal of Chemical Theory and Computation*, 2023.

[23] M. Chen, X. Chen, N. P. Schafer, C. Clementi, E. A. Komives, D. U. Ferreiro, and P. G. Wolynes, "Surveying biomolecular frustration at atomic resolution," *Nature communications*, vol. 11, no. 1, p. 5944, 2020.

[24] O. Ashenberg, L. I. Gong, and J. D. Bloom, "Mutational effects on stability are largely conserved during protein evolution," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 21071–21076, 2013.

[25] T. N. Starr and J. W. Thornton, "Epistasis in protein evolution," *Protein science*, vol. 25, no. 7, pp. 1204–1218, 2016.

[26] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, "Protein structure determination using metagenome sequence data," *Science*, vol. 355, no. 6322, pp. 294–298, 2017.

[27] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[28] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "Colabfold: making protein folding accessible to all," *Nature methods*, vol. 19, no. 6, pp. 679–682, 2022.

[29] S. Petti, N. Bhattacharya, R. Rao, J. Dauparas, N. Thomas, J. Zhou, A. M. Rush, P. Koo, and S. Ovchinnikov, "End-to-end learning of multiple sequence alignments with differentiable smith–waterman," *Bioinformatics*, vol. 39, no. 1, p. btac724, 2023.

[30] B. Wallner, "Afsample: improving multimer prediction with alphafold using massive sampling," *Bioinformatics*, vol. 39, no. 9, p. btad573, 2023.

[31] M. Abakarova, C. Marquet, M. Rera, B. Rost, and E. Laine, "Alignment-based protein mutational landscape prediction: doing more with less," *Genome Biology and Evolution*, p. evad201, 2023.

[32] F. Pucci, M. B. Zerihun, E. K. Peter, and A. Schug, "Evaluating dca-based method performances for rna contact prediction by a well-curated data set," *RNA*, vol. 26, no. 7, pp. 794–802, 2020.

[33] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. 39, no. suppl_2, pp. W29–W37, 2011.

[34] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium, "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, 2015.

[35] M. Steinegger, M. Mirdita, and J. Söding, "Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold," *Nature methods*, vol. 16, no. 7, pp. 603–606, 2019.

[36] L. F. Camarillo-Guerrero, A. Almeida, G. Rangel-Pineros, R. D. Finn, and T. D. Lawley, "Massive expansion of human gut bacteriophage diversity," *Cell*, vol. 184, no. 4, pp. 1098–1109, 2021.

[37] E. Levy Karin, M. Mirdita, and J. Söding, "MetaEuk—Sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics," *Microbiome*, vol. 8, pp. 1–15, 2020.

[38] S. Nayfach, D. Páez-Espino, L. Call, S. J. Low, H. Sberro, N. N. Ivanova, A. D. Proal, M. A. Fischbach, A. S. Bhatt, P. Hugenholtz, *et al.*, "Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome," *Nature Microbiology*, vol. 6, no. 7, pp. 960–970, 2021.

[39] B. Ma, C. Lu, Y. Wang, J. Yu, K. Zhao, R. Xue, H. Ren, X. Lv, R. Pan, J. Zhang, *et al.*, "A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources," *Nature communications*, vol. 14, no. 1, p. 7318, 2023.

[40] H. Alexander, S. K. Hu, A. I. Krinos, M. Pachiadaki, B. J. Tully, C. J. Neely, and T. Reiter, "Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton," *mBio*, pp. e01676–23, 2023.

[41] A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, *et al.*, "MGnify: the microbiome analysis resource in 2020," *Nucleic acids research*, vol. 48, no. D1, pp. D570–D578, 2020.

[42] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[43] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, and D. S. Marks, "Mutation effects predicted from sequence co-variation," *Nature Biotechnology*, vol. 35, no. 2, pp. 128–135, 2017.

[44] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[45] R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, "Functional annotations improve the predictive score of human disease-related mutations in proteins," *Human Mutation*, vol. 30, no. 8, pp. 1237–1244, 2009.

[46] D. Raimondi, A. M. Gazzo, M. Rooman, T. Lenaerts, and W. F. Vranken, "Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects," *Bioinformatics*, vol. 32, no. 12, pp. 1797–1804, 2016.

[47] T. A. Hopf, A. G. Green, B. Schubert, S. Mersmann, C. P. Schärfe, J. B. Ingraham, A. Toth-Petroczy, K. Brock, A. J. Riesselman, P. Palmedo, *et al.*, "The EVcouplings Python framework for coevolutionary sequence analysis," *Bioinformatics*, vol. 35, no. 9, pp. 1582–1584, 2019.

[48] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, "Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models," *Phys. Rev. E - Stat. Nonlin. Soft Mat. Phys.*, vol. 87, no. 1, pp. 1–16, 2013.

[49] E. Laine, Y. Karami, and A. Carbone, "Gemme: a simple and fast global epistatic model predicting mutational effects," *Molecular biology and evolution*, vol. 36, no. 11, pp. 2604–2619, 2019.

[50] J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, and M. Weigt, "Efficient generative modeling of protein sequences using simple autoregressive models," *Nature communications*, vol. 12, no. 1, p. 5800, 2021.

[51] F. Pucci, M. B. Zerihun, M. Rooman, and A. Schug, "pycofitness—Evaluating the fitness landscape of RNA and protein sequences," *Bioinformatics*, p. btae074, 2024.

[52] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, "Hydrophobicity of amino acid residues in globular proteins," *Science*, vol. 229, no. 4716, pp. 834–838, 1985.

[53] G. A. Dalkas, F. Teheux, J. M. Kwasigroch, and M. Rooman, "Cation–π, amino–π, π–π, and H-bond interactions stabilize antigen–antibody interfaces," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. 9, pp. 1734–1746, 2014.

[54] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983.

[55] H. Watanabe, H. Matsumaru, A. Ooishi, Y. Feng, T. Odahara, K. Suto, and S. Honda, "Optimizing ph response of affinity between protein g and igg fc: how electrostatic modulations affect protein-protein interactions," *Journal of biological chemistry*, vol. 284, no. 18, pp. 12373–12383, 2009.

[56] S. M. Malakauskas and S. L. Mayo, "Design, structure and stability of a hyperthermophilic protein variant," *Nature structural biology*, vol. 5, no. 6, pp. 470–475, 1998.

[57] Y. Park, B. P. Metzger, and J. W. Thornton, "Epistatic drift causes gradual decay of predictability in protein evolution," *Science*, vol. 376, no. 6595, pp. 823–830, 2022.

[58] C. A. Olson, N. C. Wu, and R. Sun, "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain," *Current biology*, vol. 24, no. 22, pp. 2643–2651, 2014.

[59] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, "Inverse statistical physics of protein sequences: a key issues review," *Reports on Progress in Physics*, vol. 81, no. 3, p. 032601, 2018.

[60] P. Dasmeh and A. W. Serohijos, "Estimating the contribution of folding stability to nonspecific epistasis in protein evolution," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. 12, pp. 1242–1250, 2018.

[61] J. Laimer, H. Hofer, M. Fritz, S. Wegenkittl, and P. Lackner, "MAESTRO-multi agent stability prediction upon point mutations," *BMC bioinformatics*, vol. 16, pp. 1–13, 2015.

[62] Y. Zhou, Q. Pan, D. E. Pires, C. H. Rodrigues, and D. B. Ascher, "DDMut: predicting effects of mutations on protein stability using deep learning," *Nucleic Acids Research*, p. gkad472, 2023.

[63] P. Notin, A. Kollasch, D. Ritter, L. Van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, *et al.*, "ProteinGym: large-scale benchmarks for protein fitness prediction and design," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[64] M. Schwersensky, M. Rooman, and F. Pucci, "Large-scale in silico mutagenesis experiments reveal optimization of genetic code and codon usage for protein mutational robustness," *BMC biology*, vol. 18, pp. 1–17, 2020.

[65] E. A. Franzosa and Y. Xia, "Independent effects of protein core size and expression on residue-level structure-evolution relationships," *PLoS ONE*, vol. 7, p. e46602, 2012.

[66] S.-W. Yeh, J.-W. Liu, S.-H. Yu, C.-H. Shih, J.-K. Hwang, and J. Echave, "Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure," *Molecular Biology and Evolution*, vol. 31, no. 1, pp. 135–139, 2014.

[67] J. Laimer, J. Hiebl-Flach, D. Lengauer, and P. Lackner, "MAESTROweb: a web server for structure-based protein stability prediction," *Bioinformatics*, vol. 32, no. 9, pp. 1414–1416, 2016.