Bioinformatics, 2025, **41(6)**, btaf322
https://doi.org/10.1093/bioinformatics/btaf322
Advance Access Publication Date: 28 May 2025
Applications Note

OXFORD

Structural bioinformatics

Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins

Matsvei Tsishyn[1,2,*], **Pauline Hermans**[1,2], **Marianne Rooman**[1,2,†], **Fabrizio Pucci**[1,2,*,†]

[1]Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels 1050, Belgium
[2]Interuniversity Institute of Bioinformatics in Brussels, Bruxelles 1050, Belgium

*Corresponding authors. Matsvei Tsishyn, Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels 1050, Belgium. E-mail: matsvei.tsishyn@ulb.be; Fabrizio Pucci, Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels 1050, Belgium. E-mail: fabrizio.pucci@ulb.be.

[†]These authors contributed equally to this study.

Associate Editor: Arne Elofsson

Abstract

Motivation: Predicting how mutations impact protein biophysical properties remains a significant challenge in computational biology. In recent years, numerous predictors, primarily deep learning models, have been developed to address this problem; however, issues such as their lack of interpretability and limited accuracy persist.

Results: We showed that a simple evolutionary score, based on the log-odd ratio of wild-type and mutated residue frequencies in evolutionary related proteins, when scaled by the residue's relative solvent accessibility, performs on par with or slightly outperforms most of the benchmarked predictors, many of which are considerably more complex. The evaluation is performed on mutations from the ProteinGym deep mutational scanning dataset collection, which measures various properties such as stability, activity or fitness. This raises further questions about what these complex models actually learn and highlights their limitations in addressing prediction of mutational landscape.

Availability and implementation: The RSALOR model is available as a user-friendly Python package that can be installed from the PyPI repository. The code is freely available at https://github.com/3BioCompBio/RSALOR.

1 Introduction

Accurately estimating the fitness of variants is essential both from a biomedical perspective, to deepen our understanding of the mechanisms underlying pathogenesis (Fowler *et al.* 2023), and from a biotechnological perspective, to improve protein engineering approaches (Freschlin *et al.* 2022). As a result, an impressive number of computational tools have been developed over the last decade to predict the effects of variants on different protein biophysical properties (Pucci *et al.* 2022; Livesey and Marsh 2023; Rastogi *et al.* 2024). These tools are characterized by a wide range of architectures, ranging from simple linear models applied to a few features to complex deep learning methods.

In recent years, the field has witnessed an even more remarkable growth, with the emergence of protein language models (pLMs), which have significantly advanced variant effect prediction (Notin *et al.* 2024; Rastogi *et al.* 2024). However, these deep learning techniques also come with notable drawbacks. Their immense number of parameters requires extensive training, making them computationally expensive and more prone to overfitting. While there are methods to mitigate overfitting, it remains challenging to disentangle true biophysical properties from unwanted biases and to achieve good generalizability. Additionally, their inherent complexity often makes it nearly impossible to extract meaningful biophysical insights from their predictions.

An alternative approach is to introduce simple prediction models with biological significance that retain the same accuracy as these more complex approaches. Following this direction (Hermans *et al.* 2024), we recently showed how a simple evolutionary score, when scaled by the relative solvent accessibility (RSA) of the mutated residues, can accurately predict changes in folding free energy upon mutations. The model is extremely simple, interpretable, and performant, and has no free parameters to optimize. Its score reflects the impact of a variant on protein fitness, which is broadly defined as the

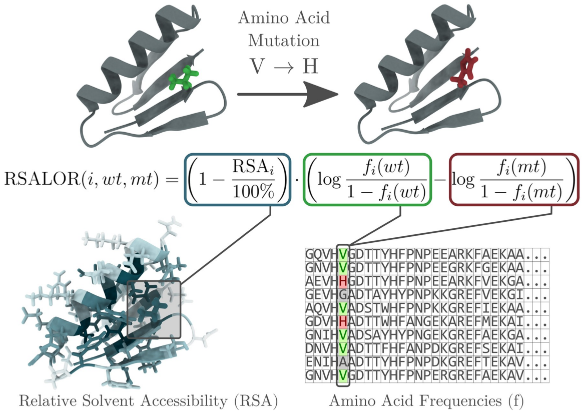
Amino Acid
Mutation
V → H
$\mathrm{RSALOR}(i, wt, mt) = \left(1 - \frac{\mathrm{RSA}_i}{100\%}\right) \cdot \left(\log \frac{f_i(wt)}{1 - f_i(wt)} - \log \frac{f_i(mt)}{1 - f_i(mt)}\right)$
GQVHVGDTTYHFPNPEEARKFAEKAA...
GNVHVGDTTYHFPNPEEGRKFGEKAA...
AEVHHGDTTYHFPNPEEARKFVEKGA...
GEVHGADTAYHYPNPKKGREFVEKGI...
AQVHVADSTWHFPNPKKGREFIEKAA...
GDVHHADTTWHFANGEKAREFMEKAI...
GNIHVADSAYHYPNGEKGREFAEKGA...
DNVHVADTTFHFANPDKGREFSEKAI...
ENIHAADTTYHFPNPDKGREFTEKAV...
GNVHVGDTTYHFPNPERGREFAEKAV...
Relative Solvent Accessibility (RSA)
Amino Acid Frequencies (f)

Figure 1. Graphical representation of the RSALOR model.

Received: 31 January 2025; **Revised:** 23 April 2025; **Editorial Decision:** 23 May 2025; **Accepted:** 27 May 2025

ability of the protein to effectively perform its biological function. As such, it is related not only to stability but also to other protein properties. We thus extended the analysis of this model, called RSALOR, by providing additional evidence of its accuracy across a broader range of protein biophysical properties, including stability, binding affinity, organismal fitness, activity, and expression, using ProteinGym (Notin *et al.* 2024), a widely used benchmark dataset specifically designed for this purpose. A graphical representation of the model is provided in Fig. 1.

2 RSALOR model

We present a very simple, independent-site, unsupervised approach for mutational effects prediction that combines evolutionary and structural information. In this section, we outline the key steps of the RSALOR model. Detailed implementation aspects are described in Supplementary Materials Sections 1 and 2.

The evolutionary information used in the model is derived from amino acid frequencies at the mutated position in a multiple sequence alignment (MSA), using the LOR between wild-type and mutant amino acid frequencies. The MSA is first curated by removing redundant identical sequences and those falling within the 'twilight zone' (Rost 1999) (i.e. sequences too evolutionary distant from the target sequence based on a sequence identity criterion). Indeed, we observed that the presence of very distant sequences adds noise rather than improving RSALOR predictions.

Since MSAs can be dominated by clusters of closely related sequences, we computed the 'weighted' amino acid frequencies by reducing the contribution of sequences from larger clusters, as done in coevolutionary models (e.g. Weigt *et al.* 2009; Morcos *et al.* 2011). We discussed and analyzed the impact of the weighting step in Supplementary Materials Section 3.2, showing that it mildly but consistently improves our model's performance. To prevent LOR values from diverging and to handle the lack of information in small MSAs, we applied regularization to these frequencies. Using the weighted and regularized frequencies $f_i(wt)$ and $f_i(mt)$ for the wild-type and mutant amino acids at position i, the LOR is defined as:

$$\mathrm{LOR}(i, wt, mt) = \log \frac{f_i(wt)}{1 - f_i(wt)} - \log \frac{f_i(mt)}{1 - f_i(mt)}. \qquad (1)$$

The sign of LOR is defined such that the result of mutations from a highly represented amino acid *wt* to a less represented amino acid *mt* is positive, which generally corresponds to a decrease in protein stability or fitness.

As structural information, we used the per-residue RSA, reflecting the observation that mutations in the protein core tend to have a greater impact than those on the surface. The simple product of LOR and the "complement" of RSA defines the RSALOR:

$$\mathrm{RSALOR}(i, wt, mt) = \left(1 - \frac{\mathrm{RSA}_i}{100\%}\right) \cdot \ \mathrm{LOR}(i, wt, mt). \qquad (2)$$

Since the RSA factor in this equation is always positive, the sign of RSALOR is the same as the sign of LOR.

Note that RSALOR nearly perfectly preserves the symmetry property (i.e. the effect of a mutation from *wt* to *mt* is opposite to the effect of the mutation from *mt* to *wt*). Indeed, while the evolutionary component LOR is perfectly symmetric, the fact that RSA is calculated using only the wild-type structure can, in principle, introduce asymmetry into the model. However, as shown in Supplementary Materials Section 3.4, this approximation does not substantially impact the predictions, which remain almost perfectly symmetric. This symmetry, which is often violated by predictors, has been shown to be an important feature in the prediction of changes in protein stability and binding affinity (Pucci *et al.* 2018; Usmanova *et al.* 2018; Tsishyn *et al.* 2025).

3 RSALOR implementation

We provide RSALOR as a freely available, easy-to-install, and user-friendly Python package. It can be installed by cloning our GitHub repository at github.com/3BioCompBio/RSALOR or via the Python Package Index (PyPI) using **pip**. It takes as input the MSA of the target protein and its three-dimensional structure in PDB format.

The package automatically maps RSA values extracted from the structure to the corresponding positions in the MSA, even if the template structure contains missing residues or is homologous but not identical to the target sequence of the MSA. Indeed, RSA values are relatively robust to small structural changes. The model's performance, therefore, remains almost unchanged when using structures with a few amino acid substitutions (see Supplementary Materials Section 3.5 for details).

The RSALOR package outputs or saves to a CSV file the following information for each possible single-site mutation in the target protein: the frequencies of gaps, wild-type and mutant residues in the MSA; the RSA of the mutated residue; and the LOR and RSALOR scores of the mutation.

4 RSALOR performances

In Hermans *et al.* (2024), we evaluated RSALOR on its ability to predict the impact of mutations on protein stability. To further assess the robustness of the model, we tested it on ProteinGym (Notin *et al.* 2024), consisting of 218 standardized deep mutational scanning (DMS) experiments, covering a total of 2.5 million mutations with annotated experimental effects on protein stability, binding affinity, fitness, activity, and expression.

To ensure a fair comparison with the other benchmarked models, we used the MSAs and structures provided by the ProteinGym repository without any modifications. These structures are AlphaFold-generated models (Jumper *et al.* 2021), as the target sequences of most DMS experiments are not, or only partially, covered by experimental structures. Importantly, this means that our model does not rely on the availability of high-quality experimental structures. We additionally assessed the robustness of our predictions using alternative MSAs and structures, and observed essentially the same results (see details in Supplementary Materials Section 3.5).

While we present here a summary of the results, a more comprehensive analysis is provided in Supplementary Materials Section 3.1. It includes both overall performances and per-category performances on single-site and all (single-site and multiple) mutations from ProteinGym, evaluated using various metrics and compared among 27 different predictors (including 19 pLM-based models).

We first focused on the approximately 700,000 single-site mutations in the dataset. In Table 1, we present a comparison of the Spearman correlations, averaged over all DMS experiments or over specific DMS categories. The predictions of RSALOR were compared with some of the top-performing tools included in the ProteinGym benchmark. Our results show that the simple RSALOR model achieves performance in line with the other state-of-the-art methods, with an average Spearman correlation of 0.473 across all 217 datasets. Among the 27 models in the unsupervised category, only ProSST (Li *et al.* 2024), which combines structure and pLM features, achieves better results.

We note that the structural contribution, RSA, has a variable impact on performance depending on the DMS target property. Although it provides great improvements to the LOR score on stability and binding datasets, it enhances accuracy to a lesser extent for expression, activity, and fitness datasets. This is consistent considering that stability and binding affinity are more directly related to protein structure. For instance, the RSA score alone outperforms most benchmarked predictions on stability datasets.

In addition, we have shown that using RSA computed from more appropriate input 3D conformations can further boost predictions (see Supplementary Materials Section 3.6). For example, when studying the impact of mutations on protein–protein binding affinity, using the structure of the protein complex instead of the monomeric structure provided by the ProteinGym dataset leads to substantially improved performance.

It is worth noting, as already highlighted in the ProteinGym benchmark (Notin *et al.* 2024), that predictions also vary greatly between datasets, with some DMSs being exceptionally well predicted (Spearman correlation above 0.7), while in others, all methods essentially fail. In contrast, the number of homologous sequences in the input MSA has only a minor impact on the performance of RSALOR (see details in Supplementary Materials Section 3.3).

To predict the effect of multiple mutations using the RSALOR model, we made the approximation that there are no epistatic effects. Therefore, the effect of a multiple mutation is simply the sum of the effects of its individual single-site mutations. Even with this simplistic assumption, RSALOR achieved a correlation of 0.484 on all ProteinGym mutations, outperformed only by ProSST (Li *et al.* 2024) and PoET (Truong Jr and Bepler, 2023) (with correlations of 0.523 and 0.490, respectively). All values are provided in Supplementary Materials Section 3.1.

We would like to underline that the RSALOR model is clearly a rough approximation for estimating the effects of protein mutations. First, it assumes that mutations at fully exposed residues (with a RSA of 100%) have no effect on the protein. While the link between the RSA of mutated residues and mutational impacts is well known (Wei *et al.* 2013; Ancien *et al.* 2018), the strength of its effect can vary significantly depending on the biophysical property considered. Second, RSALOR completely ignores epistatic effects and potential evolutionary information from other residue positions. While their contribution could be less significant in describing protein stability (Hermans *et al.* 2024; Sternke *et al.* 2025), they seem to play an essential role in protein activity and fitness (Russ *et al.* 2020; Sternke *et al.* 2025). Despite these limitations, the model is on par with or slightly outperforms nearly all benchmarked models, highlighting that predicting mutational landscapes remains a challenge for current state-of-the-art methods.

Remarkably, combining the predictions of other models with RSA values (using Equation (2)) substantially improves the performance for almost all of the 27 benchmarked predictors. This holds true even for models that already incorporate structural information as input (see details in Supplementary Materials Section 4). We thus show that effectively incorporating RSA and other types of structural knowledge into evolutionary- or pLM-based models can lead to improved performance.

Finally, an important feature of RSALOR is its ease of use. It requires no model training or external dependencies, is easily installed with a single **pip** command, and runs in a straightforward manner (see the GitHub repository). From a computational perspective, RSALOR is highly efficient. Its weighting step, being the most computationally intensive, is implemented in C++ and supports multi-threading. For instance, we evaluated the 2.5 million mutations from ProteinGym in less than 20 min on a laptop using 8 CPUs. Results for each individual protein were generated in a time range of 1 s to 1 min.

Table 1. Average per-DMS Spearman correlations across ProteinGym subclasses (categorized by DMS target properties), comparing the RSALOR model with some of the top-performing models from the ProteinGym benchmark.[a]

Method	Model type	Overall	Stability	Binding	Expression	Activity	Fitness
– RSA	STR[b]	0.356	0.480	0.281	0.323	0.330	0.286
LOR	ALI[c]	0.427	0.447	0.375	0.390	0.452	0.414
RSALOR	STR & ALI	0.473	0.551	0.455	0.428	0.472	0.419
ProSST-2048 (Li *et al.* 2024)	STR & pLM	0.522	0.638	0.527	0.527	0.486	0.441
PoET (Truong Jr and Bepler 2023)	ALI & pLM	0.470	0.458	0.440	0.459	0.495	0.474
SaProt (650M) (Su *et al.* 2024)	STR & pLM	0.462	0.565	0.441	0.482	0.459	0.375
VespaG (Marquet *et al.* 2024)	pLM	0.461	0.479	0.415	0.450	0.489	0.440
TranceptEVE (L) (Notin *et al.* 2022)	ALI & pLM	0.450	0.424	0.405	0.447	0.489	0.458
GEMME (Laine *et al.* 2019)	ALI	0.447	0.452	0.367	0.430	0.477	0.444
EVE (ensemble) (Frazer *et al.* 2021)	ALI	0.431	0.410	0.382	0.398	0.466	0.446
ESM2 (650M) (Lin *et al.* 2023)	pLM	0.428	0.496	0.382	0.409	0.431	0.381

[a] Only single-site mutations were considered. Note that ProteinGym's benchmark uses a slightly different method of averaging correlations, so their values and ours do not always perfectly match.
[b] STR, structure-based.
[c] ALI, alignment-based.

Conflict of interest: None declared.

Funding

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

References

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

Fowler DM, Adams DJ, Gloyn AL *et al.* An atlas of variant effects to understand the genome at nucleotide resolution. *Genome Biology* 2023;**24**:147.

Frazer J, Notin P, Dias M *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5.

Freschlin CR, Fahlberg SA, Romero PA. Machine learning to navigate fitness landscapes for protein engineering. *Curr Opin Biotechnol* 2022;**75**:102713.

Hermans P, Tsishyn M, Schwersensky M *et al.* Exploring evolution to uncover insights into protein mutational stability. *Mol Biol Evol* 2024;**42**:msae267.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9.

Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. *Mol Biol Evol* 2019; **36**:2604–19.

Li M, Tan Y, Ma X *et al.* ProSST: Protein language modeling with quantized structure and disentangled attention. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024, Vancouver (CA).

Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023; **379**:1123–30.

Livesey BJ, Marsh JA. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol Syst Biol* 2023; **19**:e11474.

Marquet C, Schlensok J, Abakarova M *et al.* Expert-guided protein language models enable accurate and blazingly fast fitness prediction. *Bioinformatics* 2024;**40**:btae621.

Morcos F, Pagnani A, Lunt B *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011;**108**:E1293–E1301.

Notin P, Kollasch A, Ritter D *et al.* ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Adv Neural Inform Process Syst* 2024;**36**.

Notin P, Van Niekerk L, Kollasch AW *et al.* TranceptEVE: combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. In: *The Thirty-Sixth Annual Conference on Neural Information Processing System,* 2022, New Orleans (USA).

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

Pucci F, Schwersensky M, Rooman M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Curr Opin Struct Biol* 2022;**72**:161–8.

Rastogi R, Chung R, Li S *et al.* Critical assessment of missense variant effect predictors on disease-relevant variant data. *Hum Genet* 2025;**144**:281–93.

Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94.

Russ WP, Figliuzzi M, Stocker C *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* 2020;**369**:440–5.

Sternke M, Tripp KW, Barrick D. Protein stability is determined by single-site bias rather than pairwise covariance. *bioRxiv*, 2025; pages 2025–01. preprint: not peer reviewed.

Su J, Han C, Zhou Y *et al.* SaProt: protein language modeling with structure-aware vocabulary. In: *The Twelfth International Conference on Learning Representations,* 2024, Vienna (AU).

Truong T, Jr, Bepler T. Poet: a generative model of protein families as sequences-of-sequences. *Adv Neural Inform Process Syst* 2023; **36**:77379–415.

Tsishyn M, Pucci F, Rooman M. Quantification of biases in predictions of protein–protein binding affinity changes upon mutations. *Brief Bioinform* 2025;**25**:bbad491.

Usmanova DR, Bogatyreva NS, Ariño Bernad J *et al.* Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics* 2018;**34**:3653–8.

Wei Q, Xu Q, Dunbrack Jr RL. Prediction of phenotypes of missense mutations in human proteins from biological assemblies. *Proteins: Struct Funct Bioinf* 2013;**81**:199–213.

Weigt M, White RA, Szurmant H *et al.* Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009;**106**:67–72.

Bioinformatics, 2025, 41, 1–4
https://doi.org/10.1093/bioinformatics/btaf322
Applications Note