# Protein Language Models for Predicting Mutational Effects in Plants

Eva Pardo-Otero, Verónica Bolón-Canedo, and Rosalía Piñeiro

Evolutionary Biology Research Group (GIBE), Center of Information and Communication Technologies (CITIC), Universidade da Coruña, 15071 A Coruña, Spain
Laboratory for Research and Development in Artificial Intelligence (LIDIA), Center of Information and Communication Technologies (CITIC), Universidade da Coruña, 15071 A Coruña, Spain
Evolutionary Biology Research Group (GIBE), Interdisciplinary Centre of Chemistry and Biology (CICA), Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: eva.pardo@udc.es

DOI: *https://doi.org/10.17979/spu.23.c44*

*Abstract*: Predicting the effect of mutations is key for estimating genetic load—the accumulation of deleterious mutations that may reduce fitness of organisms. Traditional methods rely on predefined features, such as evolutionary conservation or the predicted impact on the protein sequence and structure. Protein language models (PLMs) offer a new data-driven alternative by learning functional constraints directly from protein sequences. In this study, we apply the PLM ESM-1v for predicting mutational effects, using 12,865 functionally annotated mutations from 433 plant species, a group with limited predictive tools. We also explore integrating ESM-1v embeddings as features in supervised machine learning models.

## 1 Introduction

Today, more than 253 million protein sequences representing 1.33 million species are available in public databases such as UniProtKb. This scenario, enhanced by increased computing power and recent advances in artificial intelligence, has stimulated research aimed at discovering how large-scale evolutionary data modeling can reveal relevant aspects of protein biology. One of the areas receiving greater attention is the prediction of mutational effects, which is highly relevant for multiple fields, including clinical diagnosis, crop breeding (Sendrowski et al., 2025) or genetic load estimations (Bertorelle et al., 2022). Mutations can be neutral, beneficial, or deleterious, the latter accumulating as a genetic load that may reduce the fitness of organisms. Assessing the impact of mutations is essential for understanding the dynamics and evolutionary consequences of this load. However, experimentally characterizing the full mutational landscape is extremely costly and time-consuming. For this reason, computational prediction has become a crucial strategy to preselect mutations for further investigation. A number of methods have been developed to predict mutational effects; however, they lack standardization and differ in the input data used, the methodologies, as well as the criteria for defining deleteriousness (Liu et al., 2022).

Biological sequences exhibit variations that reflect evolutionary constraints on protein structure and function. Based on this premise, protein language models (PLMs) have recently emerged as the top performing approaches in benchmarks for the prediction of mutation effects (Livesey and Marsh, 2023). PLMs are trained on massive sets of protein sequences and learn to model the implicit 'grammar' that has been shaped by evolution (Brandes et al., 2023).

In doing so, they can capture structural, functional, and evolutionary constraints in individual aminoacid residues and protein domains, the fundamental and functional units of protein sequences, respectively. This learning is achieved through a self-supervised approach, where instead of having an external label for each example, a "pseudo-label" is defined from the input itself to form a predictive task, like, for example, predicting an aminoacid from the surrounding context.

PLMs are typically implemented using the Transformer architecture, where self-attention layers explicitly model residue-residue interactions throughout the sequence (Rives et al., 2021). In addition, they are trained with the masked language modeling objective, which consists of randomly masking a fraction of aminoacids in each sequence and training the model to recover them from the surrounding context. After training, the model can assign probabilities to any aminoacid at a given position. Deleteriousness is then measured by comparing the probabilities that the model assigns to the wild-type and mutant residues expressed as a log-likelihood ratio (Brandes et al., 2023; Meier et al., 2021). This approach ultimately relies on the same fundamental principle as conservation-based classical methods (Kumar et al., 2009): mutations that deviate from patterns preserved by evolution are more likely to be deleterious. However, they extend this principle further by learning from millions of protein sequences, thereby modeling higher-order residue–context dependencies that shape protein structure and function. In particular, the PLM ESM-1v (Meier et al., 2021) has been shown to be one of the most effective models for the prediction of mutation effects (Livesey and Marsh, 2023). ESM-1v is a 650 million parameter transformer model, trained on approximately 98 million diverse sequences from the UniRef90 database.

In addition to direct probability-based scoring, PLMs can generate high-dimensional embeddings, i.e., latent vector representations of protein sequences and aminoacids (tokens) learned during pre-training on large sequence databases. As shown in Elnaggar et al. (2022), the embeddings obtained from large pre-trained models capture the significant biophysical properties of proteins. When embeddings are used as input features for supervised models, they have proven competitive in tasks such as secondary structure prediction, protein subcellular localization, and the classification of membrane versus soluble proteins (Elnaggar et al., 2022). Recently, this strategy has been extended to mutational effect prediction (Glaser and Brägelmann, 2025).

In this preliminary work, we explore the potential of protein language models to predict the effects of plant protein mutations, which have received much less attention than human proteins. Existing approaches in plants have mainly focused on model species or crop species. Predicting mutational effects in plants poses additional challenges due to the complexity of their genomes, including their large and variable sizes and frequent gene and whole-genome duplications. To address this, we employ two complementary strategies with protein language models: probability-based scoring with log-likelihood ratios, and supervised prediction using embeddings directly derived from pretrained PLM.

## 2 Materials and Methods

For the probability-based scoring approach the ensemble of five models of ESM-1v was used (Meier et al., 2021). Each model shares the same architecture as ESM-1b (Brandes et al., 2023), consisting of a 33-layer Transformer with 670 millions of parameters. Each aminoacid is represented as a 1,280-dimensional embedding. The effects of mutations were quantified as the log-likelihood ratio between the wild-type and mutant sequences. The final score was obtained by averaging the outputs of the five models in the ensemble. All calculations were performed using the official implementation available at *https://github.com/facebookresearch/esm*.

For the supervised embedding-based approach, we extracted 1,280-dimensional embeddings from the last hidden layer of ESM-1v. Different embedding types were considered to capture different contextual levels: (i) the mean embedding of the full mutated sequence, (ii) the

embedding at the mutated residue, (iii) the mean embedding of a 13-residue window centered on the mutation. These embeddings were used as input features for three machine learning models: Support Vector Machines (SVM), Random Forests (RF), and XGBoost. Hyperparameters were optimized via grid search with 5-fold cross-validation. We additionally tested the algorithm ReliefF for feature selection retaining the top 150 features, and applied UMAP to explore the separability of embeddings for classification of mutation effects.

To validate both approaches, we used the test set curated by (Gou et al., 2022) with plant mutations annotated as neutral or functional (deleterious effect proxy). For training the supervised strategy, we relied on the corresponding train set (Gou et al., 2022). Together, both sets comprise 12,865 protein mutations from 6,172 sequences across 433 plant species.

Table 1: Performance of the supervised embedding-based strategies and ESM-1v probability-based scoring

| Method | Accuracy | F1-Score | Precision | Recall | Best model |
|---|---|---|---|---|---|
| Mut site emb | 0.819 | 0.819 | 0.820 | 0.819 | SVM |
| Window emb | 0.810 | 0.810 | 0.813 | 0.810 | SVM |
| Mean emb | 0.796 | 0.795 | 0.799 | 0.796 | XGBoost |
| Window emb + Relief | 0.803 | 0.803 | 0.804 | 0.803 | SVM |
| Mean emb + Relief | 0.800 | 0.800 | 0.803 | 0.800 | XGBoost |
| Mut site emb + Relief | 0.792 | 0.792 | 0.792 | 0.792 | SVM |
| ESM1v $\log(P_{mut}/P_{wt})$ | 0.734 | 0.730 | 0.747 | 0.733 | |

## 3 Results

The prediction of the neutral or functional effect based on the log-likelihood ratio scoring with ESM-1v resulted in lower performance (f1-score = 0.730) compared to supervised embeddings-based method (Table 1). The embedding at the mutated site provided the best results in separating neutral versus functional mutations, reaching an f1-score of 0.819 with balanced precision of 0.820 and recall of 0.819. Averaging embeddings over the full sequence or over a 13-residue window led to lower performance. The superior performance of the mutated site embedding can also be suggested by the UMAP projection (Fig.1), where a tendency towards better class separation is observed regarding the other approaches. Many examples still remain difficult to separate in the three strategies.

Applying ReliefF for feature selection to retain the 150 most informative features did not substantially improve model performance (Table 1).

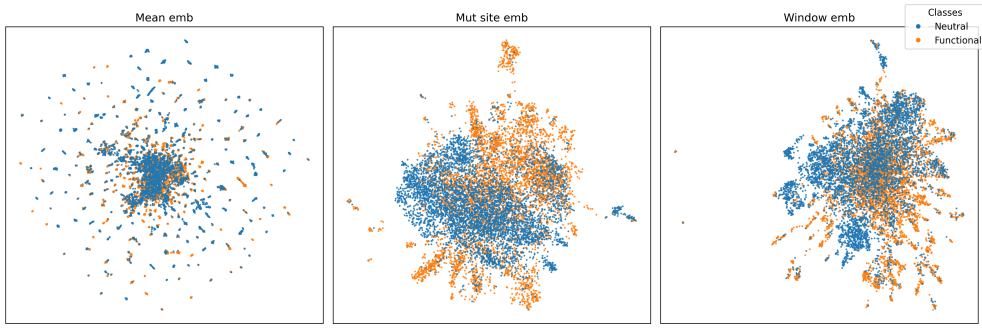


Figure 1: UMAP projection of 1,280-dimensional embeddings for the training dataset.

## 4 Discussion and Future Work

Firstly, the log-likelihood ratio scoring approach with ESM-1v did not achieve the competitive performance expected from recent benchmarks (Livesey and Marsh, 2023). This aligns with other works that pointed to the inconsistency of performance across different datasets or domains (Fawzy and Marsh, 2024; Kono et al., 2018). We plan to compare differences in performance across species within our dataset, which is largely dominated by model organisms. It would also be worth investigating potential biases in the ESM-1v pre-training data, particularly the underrepresentation of plant sequences, which could reduce its generalization ability in this context. Secondly, the supervised embedding-based approach demonstrated that embeddings can capture information that reflects mutational impact, producing better results than log-likelihood scoring, in agreement with previous findings (Glaser and Brägelmann, 2025; Yamaguchi and Saito, 2021). This improvement can be partly explained by the fact that supervised models are explicitly trained and optimized for the classification of the two effects, in contrast to log-likelihood ratio scores that are not task-specific. On the other hand, the superior performance of the mutated site embedding could be explained by the fact that only one mutation per sequence is considered. Averaging across the full sequence might dilute the signal of the mutated residue, as was also observed by Glaser and Brägelmann (2025).

This preliminary work marks a step toward protein language models for predicting the effects of plant mutations. Future work will include comparing their effectiveness with classical methods. Furthermore, given the presence of some mutations that are difficult to classify, we will investigate which biological characteristics may influence performance.

## Bibliography

G. Bertorelle, F. Raffini, M. Bosse, C. Bortoluzzi, A. Iannucci, E. Trucchi, H. E. Morales, and C. van Oosterhout. Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, 23:492–503, 8 2022. ISSN 14710064. doi: 10.1038/s41576-022-00448-x.

N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55:1512–1522, 9 2023. ISSN 15461718. doi: 10.1038/S41588-023-01465-0;SUBJMETA. URL *https://www.nature.com/articles/s41588-023-01465-0*.

A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7112–7127, 10 2022. ISSN 19393539. doi: 10.1109/TPAMI.2021.3095381.

M. Fawzy and J. A. Marsh. Understanding the heterogeneous performance of variant effect predictors across human protein-coding genes. *Scientific Reports*, 14, 12 2024. ISSN 20452322. doi: 10.1038/s41598-024-76202-6.

M. Glaser and J. Brägelmann. Esm-effect: An effective and efficient fine-tuning framework towards accurate prediction of mutation's functional effect. *bioRxiv*, Feb. 2025. doi: 10.1101/2025.02.03.635741. URL *http://biorxiv.org/lookup/doi/10.1101/2025.02.03.635741*.

X. Gou, X. Feng, H. Shi, T. Guo, R. Xie, Y. Liu, Q. Wang, H. Li, B. Yang, L. Chen, and Y. Lu. Ppved: A machine learning tool for predicting the effect of single amino acid substitution on protein function in plants. *Plant Biotechnology Journal*, 20:1417–1431, 7 2022. ISSN 14677652. doi: 10.1111/pbi.13823.

T. J. Kono, L. Lei, C. H. Shih, P. J. Hoffman, P. L. Morrell, and J. C. Fay. Comparative genomics approaches accurately predict deleterious variants in plants. *G3: Genes, Genomes, Genetics*, 8: 3321–3329, 10 2018. ISSN 21601836. doi: 10.1534/g3.118.200563.

P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature Protocols*, 4:1073–1082, 2009. ISSN 17542189. doi: 10.1038/nprot.2009.86.

Y. Liu, W. S. Yeung, P. C. Chiu, and D. Cao. Computational approaches for predicting variant impact: An overview from resources, principles to applications. *Frontiers in Genetics*, 13, 9 2022. ISSN 16648021. doi: 10.3389/fgene.2022.981005.

B. J. Livesey and J. A. Marsh. Updated benchmarking of variant effect predictors using deep mutational scanning. *Molecular Systems Biology*, 19, 8 2023. ISSN 1744-4292. doi: 10.15252/msb.202211474.

J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 12 2021. URL *https://github.*

A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL *https://www.pnas.org/doi/abs/10.1073/pnas.2016239118*.

J. Sendrowski, T. Bataillon, and G. P. Ramstein. In silico prediction of variant effects: promises and limitations for precision plant breeding. *Theoretical and Applied Genetics*, 138, 8 2025. ISSN 14322242. doi: 10.1007/s00122-025-04973-1.

H. Yamaguchi and Y. Saito. Evotuning protocols for transformer-based variant effect prediction on multi-domain proteins. *Briefings in Bioinformatics*, 22, 11 2021. ISSN 14774054. doi: 10.1093/bib/bbab234.