Prediction of Gain or Loss of Function in Missense Variants

Victor Maricato Oliveira[1], **Pedro Nuno de Souza Moura**[2]

[1] Bacharelado em Sistemas de Informação (BSI)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
CCET - Avenida Pasteur, 458 - Urca – 22.290-255 – Rio de Janeiro – RJ – Brazil

[2]Programa de Pós-graduação em Informática (PPGI)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
CCET - Avenida Pasteur, 458 - Urca – 22.290-255 – Rio de Janeiro – RJ – Brazil

victor.maricato@edu.unirio.br, pedro.moura@uniriotec.br

Abstract. *Missense variants alter a single amino acid in proteins and can cause loss-of-function (LOF) or gain-of-function (GOF). Accurately classifying these effects is essential for understanding genetic diseases and tailoring precision medicine approaches. Here, we propose the Gain and Loss of Function Dataset (GLOF), the first dataset of LOF, GOF, and Neutral missense variants annotated by specialists, curated with one of Latin America's largest genetic diagnostics laboratories. Using embeddings from the ESM-1v model, our Random Forest model achieves state-of-the-art results without requiring complex feature engineering or multi-sequence alignment. We hope the GLOF dataset will stimulate further research into LOF/GOF prediction for personalized genomics.*

Resumo. *Variantes missense alteram um único aminoácido em proteínas e podem causar perda de função (LOF) ou ganho de função (GOF). A classificação precisa desses efeitos é fundamental para compreender doenças genéticas e adaptar abordagens da medicina de precisão. Propõe-se então o Gain and Loss of Function Dataset (GLOF), o primeiro conjunto de dados contendo variantes LOF, GOF e Neutras anotado por especialistas, curado junto a um dos maiores laboratórios de diagnóstico genético da América Latina. Utilizando embeddings do modelo ESM-1v, nosso modelo Random Forest obtém resultados do estado da arte sem necessidade de engenharia complexa de atributos ou alinhamento de sequências. Espera-se que o conjunto de dados GLOF estimule mais pesquisas sobre a previsão de LOF/GOF em genômica personalizada.*

1. Introduction

Genetic variation underlies many human traits and diseases [Mardis 2008, Shastry 2009], including missense variants, which replace a single amino acid. These can cause devastating diseases or confer adaptive advantages [Aidoo 2002], depending on whether they lead to loss-of-function (LOF), gain-of-function (GOF), or are Neutral [Karki et al. 2015, Jung et al. 2015]. Identifying these functional impacts is vital for disease pathogenesis studies and drug development [Griffiths et al. 2015]. Recent protein language models like ESM [Rives et al. 2021] offer valuable representations of protein structure and function, though explicit LOF/GOF classification remains relatively sparse [Stein et al. 2023].

In this work, we collaborate with Mendelics Análise Genômica S.A, one of Latin America's largest genetic diagnostics laboratories, to publish a new dataset named GLOF (Gain and Loss of Function). The dataset comprises 112,437 missense variants labeled as LOF, GOF, or Neutral, representing the first comprehensive, expertly-curated resource of its kind. We then develop an end-to-end prediction pipeline that uses modern protein language model embeddings (ESM-1v, ESM-2) without complex feature engineering or multi-sequence alignment (MSA). Our approach processes both reference (REF) and alternative (ALT) sequences independently, concatenating their embeddings to capture mutation-induced changes. Experimental results show that our method achieves state-of-the-art performance in classifying missense variants across all three categories.

The remainder of this paper is structured as follows: Section 2 explains the relevance of the research; Section 3 describes the related work; Section 4 discusses the experimental study carried out; Section 5 describes the results obtained; and, finally, Section 6 presents the conclusion and future work.

2. Research Relevance

Understanding LOF and GOF mutations is critical for personalized medicine, as they require different treatment approaches. For example, LOF mutations in the SPTLC1 gene cause hereditary sensory neuropathy (HSN) [Rotthier et al. 2010], while GOF mutations in the same gene cause juvenile amyotrophic lateral sclerosis (ALS) [Johnson et al. 2021]. Treatments effective for one variant type may harm patients with the other [Eilbeck et al. 2017]. Existing computational tools like SIFT [Ng and Henikoff 2003], PolyPhen-2 [Adzhubei et al. 2010], and CADD [Rentzsch et al. 2018] focus on pathogenic vs. benign classification but don't differentiate LOF from GOF, limiting their clinical utility.

Recent approaches like AlphaMissense [Cheng et al. 2023] achieve unprecedented scale in variant prediction but don't classify functional mechanisms. Our work addresses this gap through GLOF and its associated prediction pipeline, aligning with the vision presented in [Lopes et al. 2016], in the document "Grand Research Challenges in Information Systems in Brazil 2016 - 2026", by addressing a data science challenge that could potentially improve people's quality of life.

3. Related Work

AlphaMissense [Cheng et al. 2023] focuses on large-scale missense variant pathogenicity predictions by combining a protein language model with structural context, achieving state-of-the-art pathogenicity classification. While groundbreaking in its scale and accuracy, it does not directly distinguish between LOF and GOF mechanisms.

EVE [Frazer et al. 2021] derives evolutionarily-inspired embeddings to predict variant pathogenicity, incorporating deep learning on multiple sequence alignments. However, like AlphaMissense, it stops short of specifying functional consequences.

On the other hand, **LoGoFunc** [Stein et al. 2023] is closer to our purpose: it classifies LOF and GOF using features derived from multiple sources, such as protein structure models and molecular descriptors, culminating in an ensemble of LightGBM classifiers. Despite strong results, it requires multiple-sequence alignments and extensive feature engineering, which can be complex or infeasible for poorly studied proteins.

Our approach employs protein embeddings from ESM [Rives et al. 2021, Meier et al. 2021, Lin et al. 2022], specifically ESM-1v and ESM-2, which interpret protein sequences as "languages" to reveal representations correlated with protein structure and function. Unlike previous methods, we eliminate the need for multi-sequence alignment or complex feature engineering. This approach simplifies inference and generalizes better to novel proteins, while also removing the need for domain-specific feature construction. The proposed GLOF dataset also consists in the largest open-access LOF/GOF-labeled missense variant resource, annotated and curated by clinical specialists.

4. Experimental Study

We carried out an experimental study to validate our proposed approach, which is described in the next subsections.

4.1. Protein Language Models

Our approach leverages two state-of-the-art protein language models: ESM-1v and ESM-2. Both models are based on the transformer architecture but differ in their training approaches and capabilities.

ESM-1v [Meier et al. 2021] is specifically optimized for variant effect prediction, using a 33-layer transformer architecture with 650 million parameters and 1280 hidden units. Trained on over 1 billion protein sequences from UniRef90, it processes sequences up to 1024 tokens. The model learns evolutionary constraints by predicting masked amino acids during pre-training, enabling it to capture both local and long-range interactions within proteins. ESM-1v generates per-residue embeddings that encode both sequence and implicit structural information, making it particularly suitable for variant effect prediction.

Subsequently, **ESM-2** [Lin et al. 2022] represents a significant advancement in protein language modeling, scaling up to 15 billion parameters and incorporating architectural improvements that allow it to process sequences up to 2048 tokens. Trained on a larger corpus of 250 million protein sequences from UniRef50, ESM-2 employs a deeper architecture with 48 transformer layers and improved attention mechanisms. While ESM-2 demonstrates superior performance in protein structure prediction and zero-shot prediction tasks, our experiments show that ESM-1v's specialized training for variant effects makes it more effective for our specific classification task.

In our implementation, we extract embeddings from both models for each variant's reference and alternative sequences. The embedding vectors capture the protein context around the mutation site, encoding both sequence conservation patterns and potential structural impacts of amino acid substitutions.

4.2. Dataset and Preprocessing

We combined benign missense variants from GnomAD v3.1 [Chen et al. 2024] with pathogenic variants from ClinVar [Landrum et al. 2014] (July 2023). Expert geneticists at Mendelics Análise Genômica S.A curated each variant to assign it to *Neutral*, *LOF*, or *GOF* categories based on extensive literature review and clinical expertise. This expert curation process ensures high-quality labels that capture the functional impact of each variant.

The GLOF[1] dataset contains $112,437$ variants, of which $83,924$ ($\sim 74.64\%$) are Neutral, $25,376$ ($\sim 22.57\%$) are LOF, and $3,137$ ($\sim 2.79\%$) are GOF. This distribution reflects the natural prevalence of these variant types in human populations. We split the data into 80% training, 10% validation, and 10% testing sets, maintaining the class distribution across splits.

4.3. Context Window Cropping

Some language models (e.g., ESM-1v) have length limitations for protein embedding. When a protein exceeded the maximum token length, we cropped around the mutated amino acid, retaining up to 512 amino acids in each direction (Figure 1). This strategy preserves the local sequence context critical for correct embedding but may omit distant residues that could influence the 3D structure [Fowler and Fields 2014, Branden and Tooze 2012].

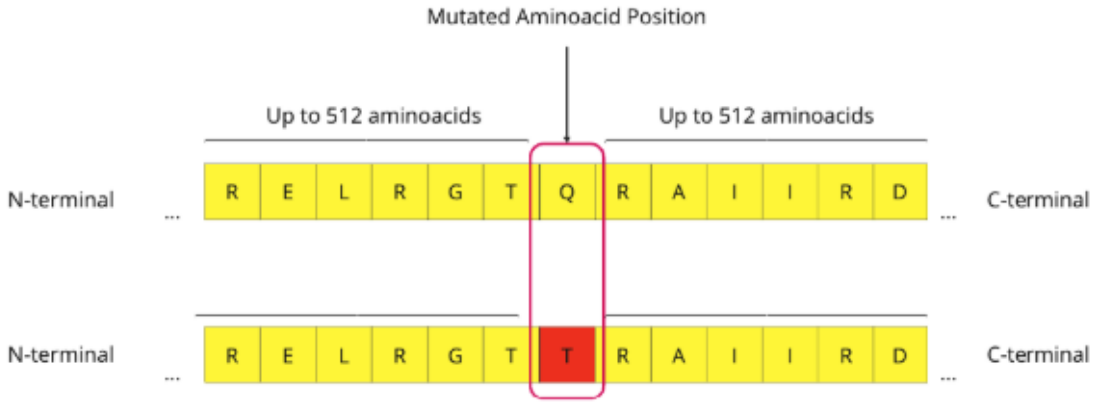

Figure 1. Schematic representation of the context window cropping. Centralizing the mutated amino acid position, up to 512 amino acids are kept towards the N-terminal (beginning of the protein sequence) and C-terminal (end of the protein sequence).

4.4. Embedding Generation

Using fair-esm and HuggingFace libraries [Wolf et al. 2020], we generated embeddings from ESM-1v and ESM-2. Each variant has two sequences: *REF* (reference) and *ALT* (mutated). We independently embed both sequences and concatenate the resulting vectors to capture mutation-induced changes in the protein's language representation.

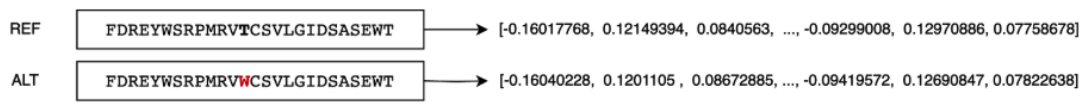

Figure 2. Embedding generation schema. Each protein sequence associated with the variant (*REF* and *ALT*) is embedded distinctly, resulting in two 1280-d vectors that are concatenated before classification.

[1] https://www.kaggle.com/datasets/maricatovictor/loss-and-gain-of-function-variants/data

Figure 2 illustrates the pipeline: the reference (REF) and alternative (ALT) sequences are processed by the language model, typically producing a 1280-dimensional vector (for ESM-1v). These two vectors are merged (concatenation) to form a 2560-dimensional embedding for each variant, capturing both the original protein context and the mutation-induced changes.

4.5. Modeling and Training

In our experiments, we compared five classifiers: Logistic Regression (LogReg), Random Forest (RF), XGBoost, LightGBM, and a small fully-connected neural network (NN). Each model receives the concatenated embeddings as input. The models were trained on the 80% training set and validated on the 10% hold-out validation set. We report results on the remaining 10% test set to ensure unbiased performance evaluation.

4.6. Evaluation Metrics

Due to the significant imbalance across LOF, GOF, and Neutral categories, we measured Precision, Recall, F1-score, and Accuracy for each class. F1-score helps mitigate the misleading effect of Accuracy in imbalanced scenarios, while Precision and Recall clarify how often the model incorrectly labels or misses variants [Aggarwal 2023].

4.7. Computational Environment

The experimental setup consisted of three main components. First, protein embeddings were generated using a dedicated cluster of 4 Tesla T4 GPUs on the Google Cloud Platform (GCP). After generating all embeddings, fine-tuning experiments were conducted using Google Colab with a Pro+ subscription, enabling access to an A100 GPU for significantly faster neural network training. Further exploratory data analysis was performed on an Apple Macbook Pro with an M1 Max chip.

Python 3.10 served as the primary programming language for all the experiments. The following Python libraries were employed: PyTorch 2.2.2 [Paszke et al. 2019] was the main library for deep learning-related tasks; scikit-learn 1.4.2 [Pedregosa et al. 2018] was used for non-neural modeling; and hyperopt 0.2.7 [Bergstra et al. 2015] for hyperparameter tuning of both deep-learning and non-deep-learning models. The code, and instructions for setting up the Python environment, are provided in a repository on GitHub[2].

5. Results

5.1. Classification Performance

We evaluated five different classifiers: Logistic Regression, Random Forest, XGBoost, LightGBM, and a Neural Network, using both ESM-1v and ESM-2 embeddings. Random Forest with ESM-1v embeddings emerged as the best-performing model, achieving F1-scores of 0.93 (Neutral), 0.76 (LOF), and 0.80 (GOF) with default hyperparameters. Table 1 shows the comprehensive performance metrics across all models using ESM-1v embeddings.

We also compared ESM-1v and ESM-2 embeddings to determine their suitability for variant effect classification. While both embedding types showed strong performance,

[2]https://github.com/victormaricato/lof-gof-predictor

Table 1. Performance metrics for different models using ESM-1v embeddings.

Metric	Class	LogReg	RF	XGBoost	LightGBM	NN
Precision	Neutral	0.88	**0.92**	0.90	0.90	0.89
	LOF	0.67	0.78	**0.79**	0.78	0.73
	GOF	0.56	0.84	**0.87**	0.81	0.58
Recall	Neutral	0.92	**0.94**	**0.94**	**0.94**	0.92
	LOF	0.29	**0.74**	0.67	0.67	0.64
	GOF	0.52	**0.76**	0.65	0.73	0.62
F1-score	Neutral	0.88	**0.93**	0.92	0.92	0.91
	LOF	0.58	**0.76**	0.72	0.72	0.68
	GOF	0.38	**0.80**	0.75	0.77	0.60
Accuracy		0.81	**0.89**	0.87	0.87	0.85

ESM-1v consistently outperformed ESM-2 across most metrics. For example, using Random Forest, ESM-2 achieved F1-scores of 0.92 (Neutral), 0.75 (LOF), and 0.71 (GOF), compared to ESM-1v's superior scores of 0.93, 0.76, and 0.80, respectively.

Hyperparameter tuning using Bayesian optimization was performed on the Random Forest model for both embedding types. For ESM-1v, tuning yielded optimal parameters of *max_depth=92*, *n_estimators=168*, *min_samples_leaf=20*, and *min_samples_split=20*. However, the tuned model showed only marginal improvements over the default configuration, with some metrics like GOF recall actually decreasing slightly (0.72 vs 0.76). Given the minimal gains and increased complexity, we retained the default Random Forest configuration with ESM-1v embeddings as our final model.

5.2. Cosine Similarity Analysis

We examined the relationship between reference and alternative protein embeddings using cosine similarity, which measures the angle between two vectors in the embedding space. For vectors R (reference) and A (alternative), the cosine similarity is calculated as described in Equation 1:

$$\text{Sim}(R, A) = \cos(\theta) = \frac{R \cdot A}{||R||_2||A||_2} \tag{1}$$

where the resulting value ranges from -1 to 1. We then define cosine distance as in Equation 2:

$$\text{Dist}(R, A) = 1 - \text{Sim}(R, A) \tag{2}$$

GOF/LOF variants, on average, displayed significantly higher embedding distances than Neutral variants ($p < 0.0005$), suggesting greater structural/functional perturbations in these classes (Figure 3). This aligns with the notion that GOF/LOF arise from more severe disruptions in protein conformation or active site chemistry compared to Neutral substitutions [Teng et al. 2010]. Similar embeddings (low distance) typically indicate conserved function. In contrast, dissimilar embeddings (high distance) suggest functional changes, analogous to how similar words in language models (e.g., "human" and "person") have closer vector representations.

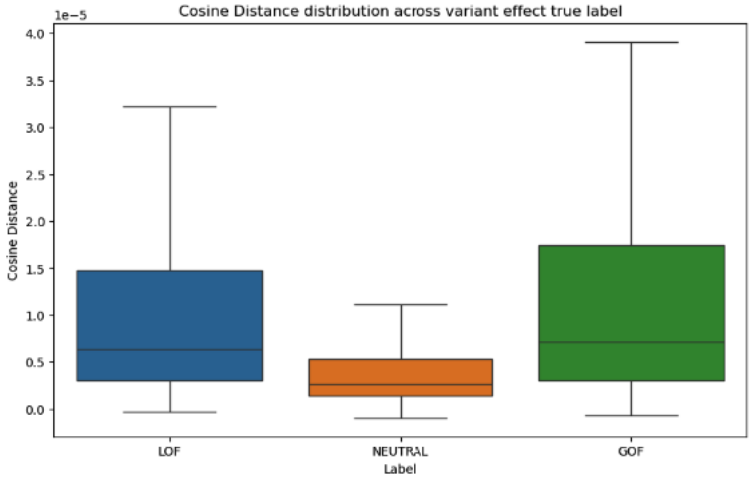

Figure 3. Distribution of cosine distance between REF and ALT embeddings per variant effect GOF and LOF variants show consistently lower similarities, indicating greater predicted structural/functional changes.

5.3. Comparison with LoGoFunc

With regard to comparison, LoGoFunc [Stein et al. 2023] reported F1-scores of 0.56 (GOF), 0.87 (LOF), and 0.89 (Neutral), using complex feature engineering and structural descriptors. Our approach yielded 0.80 (GOF), 0.76 (LOF), and 0.93 (Neutral). Notably, we achieve stronger performance for GOF and Neutral classes without requiring multi-sequence alignment or extensive manual descriptor building, although our LOF score was somewhat lower. This trade-off suggests that while protein language models effectively capture many functional signals, certain LOF mechanisms may benefit from explicit structural information.

5.4. Gene-Level Insights and Protein Length Effects

We observed gene-level performance variations that provide insights into the model's strengths and limitations. Well-studied genes like *FBN1*, associated with Marfan syndrome [Parvizi and Kim 2010], showed near-ideal LOF classification, possibly due to better representation in training data. However, genes with fewer known missense variants had reduced performance.

Analysis across protein lengths revealed performance decline for sequences exceeding 2000 amino acids, particularly for LOF and GOF predictions (Figure 4), while Neutral variant classification remained robust across all lengths.

More specifically, this length-dependent performance decline likely stems from two factors: the context window limitation in ESM-1v and the inherent difficulty of capturing long-range protein interactions in larger sequences. The model must rely on localized sequence context for extremely long proteins like Titin (35,992 amino acids), potentially missing important structural relationships that span distant regions. This limitation suggests that integrating additional structural information or developing specialized architectures for long proteins could improve prediction accuracy for these challenging cases.

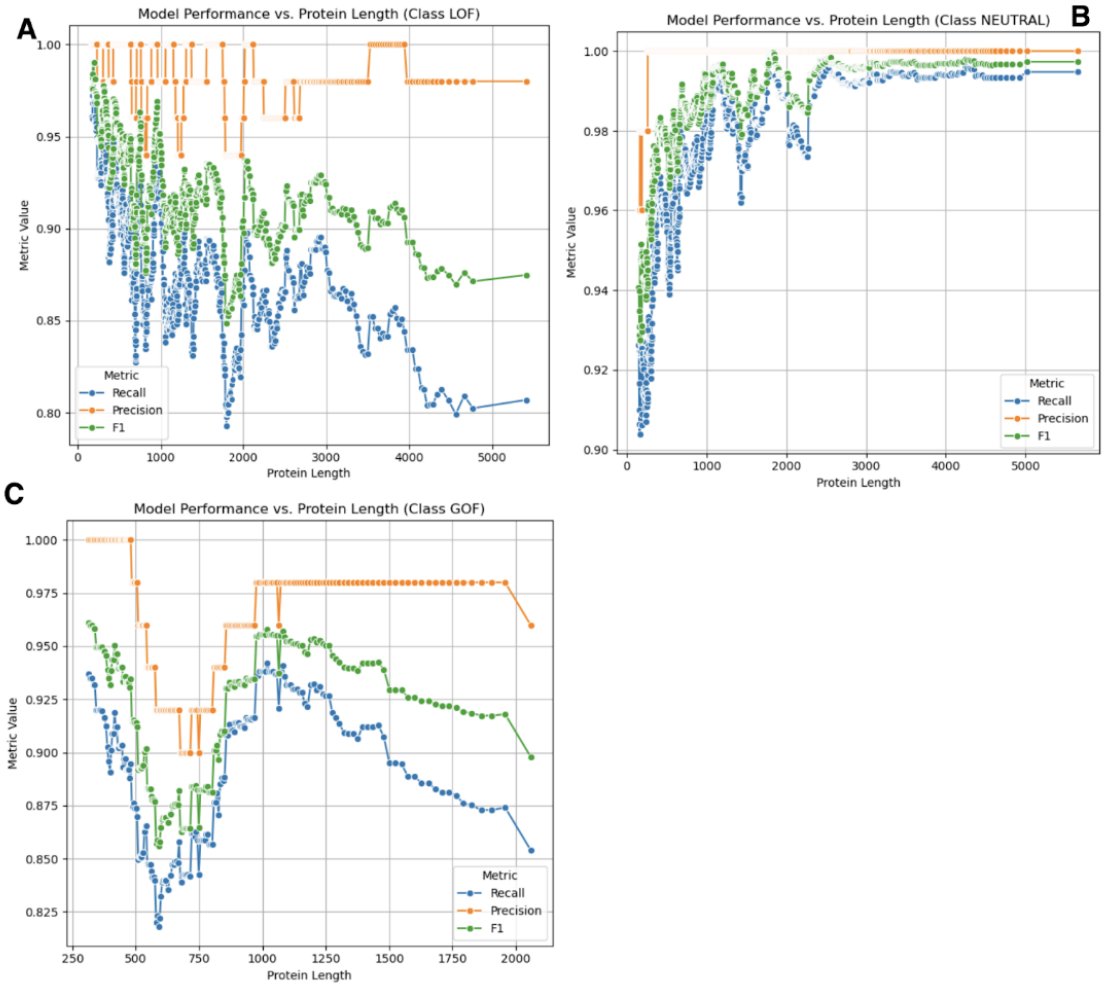

Figure 4. Model performance across protein lengths for different variant classes. Performance metrics plotted against protein length for (A) LOF, (B) Neutral, and (C) GOF variants.

6. Conclusions

In this work, we present the GLOF dataset composed of $112,437$ missense variants, each labeled as Neutral, LOF, or GOF by expert curators. Our end-to-end prediction pipeline, which uses ESM-based embeddings (without multi-sequence alignment) and a Random Forest classifier, achieves robust F1-scores for all classes. These results support the importance of advanced protein language models in capturing functional disruption signals directly from sequence. Compared to existing methods, our approach is simpler, handles less-studied proteins better, and performs competitively in distinguishing GOF, LOF, and Neutral variants.

Future directions include (1) integrating 3D structural data from tools like AlphaFold [Jumper et al. 2021] to handle highly complex or large proteins; (2) applying interpretability techniques [Montavon et al. 2018, Ribeiro et al. 2016] for greater clinical transparency; (3) experimentally validating model predictions in the wet lab. We release GLOF to the public to foster continued improvements in variant effect prediction and precision genomics.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249.

Aggarwal, C. C. (2023). *Neural Networks and Deep Learning: A Textbook*. Springer Publishing Company, Incorporated, 2nd edition.

Aidoo, M. e. a. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet*, 359(9314):1311–1312.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015). Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008.

Branden, C.-I. and Tooze, J. (2012). *Introduction to Protein Structure*. Garland Science, 2nd edition.

Chen, S. et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):492–503.

Cheng, J. et al. (2023). Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 382(6664):eadg7492.

Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and mendelian disease. *Nature Reviews Genetics*, 18(10):599–612. ISSN: 1471-0064.

Fowler, D. M. and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95.

Griffiths, A. J. F., Wessler, S. R., Carroll, D. S. B., and Doebley, J. (2015). *Introduction to Genetic Analysis*. W.H. Freeman, 11th edition.

Johnson, J. O. et al. (2021). Association of variants in the sptlc1 gene with juvenile amyotrophic lateral sclerosis. *JAMA Neurology*, 78(10):1236–1248.

Jumper, J. M. et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589.

Jung, S., Lee, S., Kim, S., and Nam, H. (2015). Identification of genomic features in the classification of loss- and gain-of-function mutation. *BMC Medical Informatics and Decision Making*, 15(Suppl 1):S6.

Karki, R., Pandya, D., Elston, R. C., and Ferlini, C. (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Medical Genomics*, 8(1):37.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2014). Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985.

Lin, Z. et al. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*.

Lopes, F., Silva, L., and Breternitz, V. (2016). *Research and Education in Data Science: Challenges for the Area of Information Systems*, chapter 14, pages 176–184. Sociedade Brasileira de Computação.

Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annual Review of Genomics and Human Genetics*, 9:387–402.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Ng, P. C. and Henikoff, S. (2003). Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814.

Parvizi, J. and Kim, G. K. (2010). *High Yield Orthopaedics*. Saunders/Elsevier, Philadelphia, PA.

Paszke, A. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Pedregosa, F. et al. (2018). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2018). Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Rives, A. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

Rotthier, A. et al. (2010). Mutations in the sptlc2 subunit of serine palmitoyltransferase cause hereditary sensory and autonomic neuropathy type i. *American Journal of Human Genetics*, 87(4):513–522. ISSN: 0002-9297.

Shastry, B. S. (2009). *SNPs: Impact on Gene Function and Phenotype*, pages 3–22. Humana Press, Totowa, NJ.

Stein, D., Liang, J., Abrusán, G., and Itan, Y. (2023). Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. *Genome Medicine*, 15(1):1–19.

Teng, S., Srivastava, A. K., and Wang, L. (2010). Structural assessment of the effects of amino acid substitutions on protein stability and protein protein interaction. *International Journal of Computational Biology and Drug Design*, 3(4):334–349.

Wolf, T. et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.