Federal University of the State of Rio de Janeiro

Center of Exact Sciences and Technology

School of Applied Informatics

PREDICTION OF LOSS OR GAIN OF FUNCTION IN MISSENSE VARIANTS

VICTOR MARICATO OLIVEIRA

Supervisor

PEDRO NUNO DE SOUZA MOURA

Rio de Janeiro, RJ – Brazil

May 2024

PREDICTION OF LOSS OR GAIN OF FUNCTION IN MISSENSE VARIANTS

VICTOR MARICATO OLIVEIRA

Undergraduate project presented at the School of Applied Informatics at Federal University of the State of Rio de Janeiro (UNIRIO) in order to obtain the title of bachelor's in information systems.

Approved by

_____

PEDRO NUNO DE SOUZA MOURA (UNIRIO)

_____

LAURA DE OLIVEIRA FERNANDES MORAES (UNIRIO)

_____

JULIO CESAR DUARTE (IME)

RIO DE JANEIRO, RJ –BRAZIL.

MAY 2024

Oliveira, Victor Maricato
O048p PREDICTION OF LOSS OR GAIN OF FUNCTION IN MISSENSE VARIANTS / Victor Maricato Oliveira. -- Rio de Janeiro, 2024.
100

Orientador: PEDRO NUNO DE SOUZA MOURA.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Sistemas de Informação, 2024.

1. Deep Learning. 2. Molecular Modeling. 3. Variant Effect Prediction. I. NUNO DE SOUZA MOURA, PEDRO, orient. II. Título.

Acknowledgments

First, I would like to thank my brother, Vinicius Maricato, and my mother, Renata Maricato, for all the love, support, education, and joy throughout my life that combined allowed and inspired me to do great things in my personal, professional and academic life.

To my dog, Maria Aleatória (known as: "Tóia" or "Random"), for being the most faithful companion I have had throughout all these years.

To my family, who has incentivized me to ask questions, be curious, and be kind since I was a kid. These are the core of the motivation for this research. Thanks for always prioritizing my education, teaching, and supporting me in whatever I wanted to do and learn. “Live as if you were to die tomorrow. Learn as if you were to live forever.”—*Mahatma Gandhi*.

To my friends, Andre Pimentel and Lucas Tunger, for pushing me to follow the correct path in every aspect of my life since our years at Instituto Federal de Ciência e Tecnologia do Rio de Janeiro (IFRJ). Also, they play a key role in all the fun necessary to go through all the years of hard work. Additionally, Tacio Monteiro, Lara Dorileo, Raira Lima, Nathalia Cardozo, and Kevin Botelho, whom UNIRIO made me meet or get closer to. Also, my friend Mauricio Pedro Vieira, who was present since my early university days and was key in advising all professional and personal decisions, supported me when I decided to do the opposite. To Mariana Batalha, Camila Souza, and Joana Oliveira, who were essential in making the last years bliss. Finally, a special thanks to all the friends not mentioned here but who were a key part of the path to this work.

To the Mendelics team: David Schlesinger, Igor Correa, Flavia Rius, Danilo Imparato, Diego Coelho, Cristiane Moreno, and Lucas Taniguti, who, without the hard trust and guidance, this project would never have been possible.

To my advisor, Pedro Nuno de Souza Moura, for all the guidance, attention, patience, and valuable discussions not only on this work but in the field and throughout the course. Special thanks to him and Jefferson Simões, the program coordinator, for

making an exception during a professor's strike in Brazilian federal universities to allow this study to be concluded and the pursuit of my new future endeavors.

To Regina Barzilay, Hannes Stärk, Alex Rives, Bowen Jing, Gabriele Corso, and Marinalva Silva from Massachusetts Institute of Technology (MIT), who made this research possible by providing guidance and invaluable discussions since 2022.

To my colleagues in the industry, Raul Renteria, Frederico Israel, Frederico Caroli, Augusto Acioli, Alexandre Dedavid, Yuri Zoel, Scott Brownlie, Lucas Rolim, Renata Gotler, Sergio Barreto, David Spechter, Yanyi He, Le Gu, David Bickley, Jenna Liu, Ted Liu, Shyam Jayasankar and Peter Fennel, who play a huge role in how I developed my relationship with computer science and developing software engineering projects in general.

To all the UNIRIO professors and professionals who make Brazilian public education excel at doing great things in the chaotic political and financial environment.

Finally, thanks to the whole BLACKPINK group, especially Park "Rosé" Chae-young, for the thousands of hours of listening to K-POP during this project development.

"People fear what they cannot understand."

- Syndra, League of Legends

ABSTRACT

Genetic variants are necessary for evolution to happen in nature. Through DNA mutations, one can gain or lose the ability to adapt to the environment and compete with other species. In humans, these variants often may cause diseases like Amyotrophic Lateral Sclerosis (ALS) or modify the individual metabolism for certain nutrients and medicines. Missense variants are a subtype of genetic variants that cause an amino acid change in the resulting protein sequence. These variants may incur a protein gain-of-function (GOF), where it may perform its original function in an augmented way or even gain new capabilities. Alternatively, loss-of-function (LOF) variants may cause the protein to be incapable of performing its original role or, for example, lose the capability of being regulated.

This work is based on a dataset containing LOF, GOF, and Neutral variants curated by Mendelics Análise Genômica S.A, one of Latin America's largest genetic diagnostics laboratories. This Gain and Loss of Function Dataset (GLOF) is made publicly available[1] and corresponds to the first dataset of LOF and GOF variants annotated by specialists in the literature. With this dataset, we propose a method that addresses the problem of variant effect prediction, providing an end-to-end approach that reaches state-of-the-art performance, surpassing alternatives that require complex feature engineering and multi-sequence alignment, which is more complex and slower and may not be possible for less well-studied proteins. Moreover, this work uses this dataset to carry out several computational experiments, including the creation of a Random Forest model[2] that uses original and mutated protein sequences embeddings to predict LOF, GOF, and Neutral variant effect, where an F1-score of 0.76, 0.78, and 0.93, respectively, is reported on the test set, placing a potential new state-of-the-art model for variant effect prediction.

Keywords: deep learning, molecular modeling, protein language models, gain-of-function variant prediction, loss-of-function variant prediction

---

[1] https://www.kaggle.com/datasets/maricatovictor/loss-and-gain-of-function-variants
[2] https://github.com/victormaricato/lof-gof-predictor

RESUMO

Variantes genéticas são necessárias para que a evolução aconteça na natureza. Através de mutações no DNA, um indivíduo pode ganhar ou perder a capacidade de se adaptar ao ambiente e competir com outras espécies. Em humanos, essas variantes frequentemente podem causar doenças como a Esclerose Lateral Amiotrófica (ELA) ou modificar o metabolismo individual para certos nutrientes e medicamentos. Variantes *missense* são um subtipo de variante genética que causam uma mudança de aminoácido na sequência protéica resultante. Essas variantes podem ocasionar em um ganho de função (*gain-of-function,* GOF) da proteína, em que ela pode realizar sua função original de forma aumentada ou até mesmo fazê-la ganhar novas capacidades. Alternativamente, variantes de perda de função (*loss-of -function,* LOF) podem fazer com que a proteína seja incapaz de desempenhar seu papel original ou, por exemplo, perder a capacidade de ser regulada.

Este trabalho é baseado em um *dataset* contendo variantes LOF, GOF e Neutras curadas pela Mendelics Análise Genômica S.A, um dos maiores laboratórios de diagnóstico genético da América Latina. Este *dataset*, nomeado *Gain and Loss of Function Dataset* (GLOF) foi disponibilizado publicamente[1], sendo o primeiro conjunto de variantes LOF e GOF anotadas disponível na literatura. Com este *dataset*, é proposto um método que aborda o problema de predição de efeito de variantes, fornecendo uma abordagem de ponta a ponta que alcança um desempenho correspondente ao estado da arte, superando alternativas que requerem engenharia de características complexas e alinhamento de múltiplas sequências, ainda mais complexo, mais lento e que pode não ser possível para proteínas menos estudadas. Além disso, este trabalho utiliza este conjunto de dados para realizar vários experimentos computacionais, incluindo a criação de um modelo[2] *Random Forest* que usa *embeddings* de sequências proteicas originais e mutadas para prever o efeito de variantes LOF, GOF e Neutras, em que um *F1-score* de 0,76, 0,78 e 0,93, respectivamente, é relatado no conjunto de teste, estabelecendo um potencial novo modelo de estado da arte para predição de efeito de variantes.

---

[1] https://www.kaggle.com/datasets/maricatovictor/loss-and-gain-of-function-variants
[2] https://github.com/victormaricato/lof-gof-predictor

Palavras-chave: aprendizado profundo, modelagem molecular, modelos de linguagem de proteínas, previsão de variantes de ganho de função, previsão de variantes de perda de função.

Table of Contents

Abbreviations Index

ALS - Amyotrophic Lateral Sclerosis

GOF - Gain-of-function

LOF - Loss-of-function

GLOF - Gain and Loss of Function Dataset

ELA - Esclerose Lateral Amiotrófica

CNN - Convolutional Neural Networks

NLP - Natural Language Processing

AI - Artificial Intelligence

LLM - Large Language Models

ML - Machine Learning

HSN - Hereditary Sensory Neuropathy

A - Adenine

T - Thymine

G - Guanine

C - Cytosine

mRNA - messenger RNA

SNV - Single Nucleotide Variants

SNP - Single Nucleotide Polymorphisms

indel - Insertion and Deletion

CNV - Copy Number Variations

GWAS - Genome-Wide Association Studies

nsSNPs - non-synonymous Single Nucleotide Polymorphisms

SIFT - Sorting Intolerant From Tolerant

PolyPhen-2 - Polymorphism Phenotyping v2

CADD - Combined Annotation Dependent Depletion

ExAC - Exome Aggregation Consortium

UKBB - UK Biobank

GBMI - Global Biobank Meta-analysis Initiative

CASP - Critical Assessment of Protein Structure Prediction

GDT - Global Distance Test

DMD - Duchenne muscular dystrophy

MPRAs - Massively Parallel Reporter Assays

DMS - Deep Mutational Scanning

RNN - Recurrent Neural Networks

LSTM - Long Short-Term Memory

GRU - Gated Recurrent Units

GNNs - Graph Neural Networks

GCNs - Graph Convolutional Networks

GATs - Graph Attention Networks

Q - Query

K - Key

V - Value

ESM - Evolutionary Scaled Modeling

SVM - Support Vector Machine

RF - Random Forest

DT - Decision Tree

GBM - Gradient Boosting Machines

k-NN - k-Nearest Neighbors

HMM - Hidden Markov Models

MSA - Multiple Sequence Alignment

PLM - Protein Language Model

UniProtKB - UniProt Knowledgebase

GCP - Google Cloud Platform

EDA - Exploratory Data Analysis

REF - Original protein

ALT - Mutated protein

TPE - Tree of Parzen

Cos - Cosine

Sim - Similarity

ANOVA - Analysis of Variance

H0 - Null hypothesis

H1 - Alternative hypothesis

R - Arginine

K - Lysine

D - Aspartic Acid

E - Glutamic Acid

V - Valine

Table Index

Figure Index

1 Introduction

1.1 Motivation

Genome sequencing has revolutionized our understanding of human genetic variation and its impact on health and disease, allowing a more comprehensive understanding of evolution and improving the diagnosis and treatment of genetic diseases [1]. Genetic mutations, while essential for generating diversity and adaptability in species, can have beneficial and harmful effects on individuals [2]; for example, while the sickle cell trait mutation protects against malaria in regions where the disease is endemic [3], mutations in BRCA1 and BRCA2 genes significantly increase the risk of developing breast, ovarian and other types of cancer [4].

The deep learning field has witnessed remarkable progress since 2012, when AlexNet [5], a deep convolutional network (CNN), achieved unprecedented performance in the ImageNet Large Scale Visual Recognition Challenge [6]. In the following years, various deep learning architectures, such as VGGNet [7], GoogLeNet [8], and ResNet [9], were developed, pushing the boundaries of image classification and object recognition. Beyond computer vision, significant advancements also thrived in natural language processing (NLP) and sequence modeling. More recently, the introduction of attention mechanisms [10] and transformer architectures [11] further revolutionized NLP, which is behind artificial intelligence (AI) technologies such as Large Language Models (LLM) and the massively known ChatGPT.

Recent advancements in AI and Machine Learning (ML) have shown promising results in understanding the language of proteins, with biological structure and function emerging from the unsupervised learning of protein sequences [12, 13]. In 2021, Google's DeepMind released its version of AlphaFold 2 [14], a deep learning model capable of predicting protein structures on par with X-ray crystallography, the state-of-the-art tool for the task that requires an enormous amount of time and expertise to run [15].

A particular case of interest corresponds to missense variants, a type of genetic variant that alters the amino acid sequence of the final protein and can cause either a loss-of-function (LOF) or gain-of-function (GOF) effect [16]. LOF variants can lead to partial or complete knockdown of the protein, while GOF variants may cause increased or novel protein activity [17].

Distinguishing between LOF and GOF variants is crucial for understanding the underlying mechanisms of genetic diseases and determining appropriate treatment strategies [18]. For example, LOF mutations in the SPTLC1 gene are associated with hereditary sensory neuropathy (HSN) [19], while GOF mutations in the same gene may cause juvenile amyotrophic lateral sclerosis (ALS) [20]. Serine supplementation, a treatment for HSN, can worsen symptoms in patients with SPTLC1 GOF-related ALS, highlighting the importance of accurate variant effect prediction for personalized medicine [21].

1.2 Objective

In collaboration with Mendelics Análise Genômica S.A, a leading genetic sequencing company in Latin America, this study aims to enhance the understanding of missense variants and develop a predictive model for classifying unseen variants as LOF, Neutral, or GOF. The primary objectives of this work are:

1. To create a publicly available annotated dataset of LOF, Neutral, and GOF variants, establishing the first benchmark of its kind;
2. To develop an AI-based model for predicting missense variants' effects using state-of-the-art representation and transfer learning techniques;
3. To provide the model and code used in this work for public and open use, promoting transparency and reproducibility in the field of genomics research.

1.3 Contributions

This work's main contribution lies in creating a curated benchmark for LOF and GOF variant effect prediction and establishing a novel model for predicting these effects on variants that do not require feature engineering. In this sense, to the best of our knowledge, this is the first end-to-end approach proposed for the variant effect prediction task. It achieves reasonable metrics such as an F1-score of 0.76 for GOF, 0.78 for LOF, and 0.93 for Neutral variants, marking a new state-of-the-art approach for this task. Furthermore, this work also delves into comparing existing pre-trained models for this fine-tuning task, effectively comparing the performance impacts of using ESM or ESM2 with different downstream models.

1.4 Structure

This thesis is organized into five chapters, each focusing on a specific aspect of the study:

- Chapter I: Introduction
  - Presents the motivation behind the study, the main objectives, and the overall structure of the thesis.
- Chapter II: Background Knowledge
  - Provides an overview of the relevant concepts and techniques needed to understand this work, including genetics, protein structure and function, deep learning, representation learning, and related works in the field.
- Chapter III: Experimental Study
  - Describes the experimental setup, dataset annotation process, data preprocessing steps, embedding generation, model fine-tuning, hyperparameter optimization, and evaluation metrics used in the study.
- Chapter IV: Results
  - Presents the main findings of the study, including the distribution of variants across the dataset, the effectiveness of representation learning for protein folding and variant prediction, the relationship between cosine similarity and variant effect, the emergence of biological reasoning from sequences, the impact of biological complexity on model quality, model interpretability, and a comparison with existing methods.
- Chapter V: Conclusions
  - Summarizes the key contributions of the study, discusses its limitations, highlights potential applications in clinical settings and drug discovery, and suggests future research directions.

2 Background Knowledge

This chapter presents the relevant concepts, such as the biochemical context behind protein variants, genomic datasets, machine learning context behind sequence models, more specific concepts on how to represent and learn protein sequences, and, finally, a review of related works. Those acquainted with these concepts may skip this chapter.

2.1 Genetics

Genetics is the study of heredity and variation in inherited traits [18]. It involves passing traits from parents to offspring through genes [22]. DNA, or deoxyribonucleic acid, is the hereditary material found in all living organisms and consists of four nucleotide bases: adenine (A), thymine (T), guanine (G), and cytosine (C) [23]. The sequence of these bases determines the genetic information for an organism's development, functioning, and reproduction [24].

The central dogma of molecular biology describes the flow of genetic information from DNA to RNA to proteins [25]. During transcription, the DNA sequence is used as a template to produce messenger RNA (mRNA) molecules [26]. The mRNA is then translated into a sequence of amino acids, which fold into a functional protein [27]. Proteins are essential macromolecules that perform various functions in living organisms, including catalyzing metabolic reactions (Figure 1), providing structural support, and regulating gene expression [28].

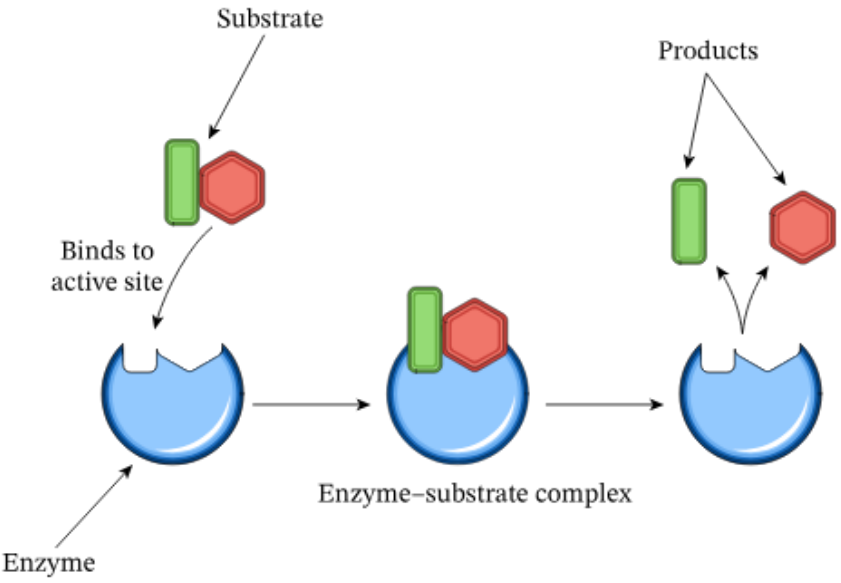

Figure 1: Illustration of how enzymes (proteins) interact with the substrate to produce products. Source: https://www.nagwa.com/en/explainers/726129032520/

Genetic variation arises from mutations, which are changes in the DNA sequence [18]. Mutations can occur spontaneously or be induced by environmental factors like radiation or chemicals [29]. Depending on how the sequence is altered, they can be classified into different types, such as point mutations, called single nucleotide variants (SNV), insertions, deletions, and chromosomal rearrangements [30]. While some mutations have no observable effect on the organism, others can lead to altered protein function, which may be beneficial, neutral, or deleterious [31].

2.2 Amino acids

Amino acids are the building blocks of proteins, essential macromolecules in all living organisms. The genetic code encodes 20 standard amino acids and is used in protein synthesis [28]. The resulting linear chain of amino acids, a polypeptide, can fold into a specific three-dimensional structure determined by the sequence of amino acids (Figure 2) [32].

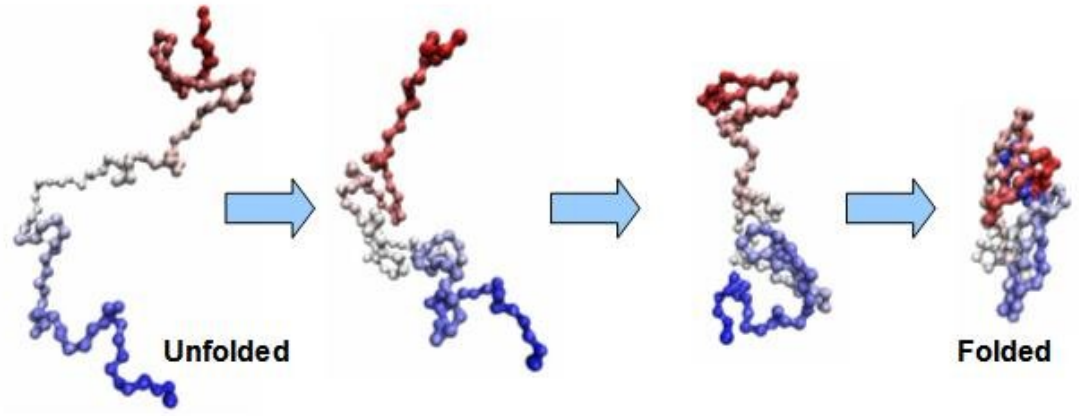

Figure 2: Schematic representation of protein folding. Source: https://www.researchgate.net/publication/293317366_Structurally_Resolved_Coarse-Grained_Modeling_of_Motor_Protein_Dynamics

The side chains of amino acids have diverse chemical properties, such as polarity, charge, and hydrophobicity, which contribute to proteins' folding, stability, and function [33]. For example, amino acids with hydrophobic side chains like leucine and valine tend to be buried within the protein core. In contrast, those with hydrophilic side chains, such as serine and threonine, are often found on the protein surface [26].

Mutations in the genetic code can lead to changes in the amino acid sequence of proteins, which may alter their folding, stability, and activity [31]. These changes can be classified as conservative or non-conservative substitutions, depending on the similarity of the chemical properties between the original and the substituted amino acid [34]. Non-conservative substitutions are more likely to impact protein function significantly and may be associated with genetic disorders [35].

In addition to the 20 standard amino acids, non-standard amino acids can be incorporated into proteins through post-translational modifications or expanded genetic codes [36]. These non-standard amino acids can provide additional chemical functionality and have applications in protein engineering and drug discovery [37].

2.3 Genetic Variants

Genetic variants are differences in DNA sequence between individuals or populations. They can range from single nucleotide changes to large-scale structural variations and play a crucial role in genetic diversity, evolution, and disease susceptibility. Genetic

variants can be classified into several types, including single nucleotide polymorphisms (SNPs), insertions and deletions (indels), copy number variations (CNVs), and chromosomal rearrangements [38].

SNPs are the most common type of genetic variation, occurring when a single nucleotide base is substituted with another. They can be found in coding and non-coding regions of the genome and may have different effects on gene function depending on their location and the specific base change [2]. SNPs in coding regions can be synonymous, causing no change in the amino acid sequence, or non-synonymous, resulting in an amino acid substitution, truncation, or extension of the protein [31].

Genetic variants can arise through various mechanisms, including DNA replication errors, exposure to mutagens, and recombination events during meiosis [39]. Most genetic variants are neutral and have no observable effect on the organism, while others may be beneficial or deleterious [40]. The functional impact of a genetic variant depends on factors such as its location, the specific base change, and the surrounding genomic context [31].

Studying genetic variants has important implications for understanding human health and disease. Through the application of genome-wide association studies (GWAS), numerous genetic variants have been identified to be associated with complex traits and diseases, such as diabetes, cardiovascular disease, and cancer. These findings have contributed to developing personalized medicine approaches, where genetic information guides disease prevention, diagnosis, and treatment strategies [41].

2.4 Missense Variants

Missense variants are a type of genetic variation that occurs when a single nucleotide change in a gene's coding region results in substituting one amino acid for another in the encoded protein. These variants are the most common non-synonymous single nucleotide polymorphisms (nsSNPs) and can have diverse effects on protein structure, stability, and function [31, 42].

The consequences of a missense variant depend on the specific amino acid substitution and its location within the protein [43]. Some amino acid changes are conservative, involving substituting an amino acid with similar chemical properties (like Glutamic Acid and Aspartic Acid, which are both negatively charged), and may have minimal impact on protein function. In contrast, in non-conservative substitutions,

where the new amino acid has distinctly different properties, for example, when a negatively charged amino acid (e.g., Glutamic Acid) is replaced by a positively-charged amino acid (e.g., Arginine), this variant might be more likely to disrupt protein structure and function [34].

Missense variants can affect protein function through various mechanisms, such as altering binding sites, catalytic sites, or post-translational modification sites [44]. They can also disrupt protein folding, making misfolded or unstable proteins prone to degradation or aggregation [45]. Sometimes, missense variants confer novel protein functions, leading to gain-of-function effects [46].

The functional impact of missense variants can range from neutral to pathogenic, depending on the specific protein and the biological context. Neutral missense variants do not significantly affect protein function and are often tolerated, while pathogenic variants can lead to disease states by disrupting essential cellular processes. Predicting the functional impact of missense variants is a major challenge in genomics and essential for interpreting genetic variations' clinical significance [47].

Several computational tools have been developed to predict the functional impact of missense variants, such as SIFT (Sorting Intolerant From Tolerant) [31], PolyPhen-2 (Polymorphism Phenotyping v2) [48], and CADD (Combined Annotation Dependent Depletion) [49]. These tools use various features, including sequence conservation, structural information, and physicochemical properties of amino acids, to estimate the likelihood of a missense variant being deleterious [50].

Missense variants have been implicated in numerous genetic disorders, such as sickle cell anemia [51], cystic fibrosis [52], and certain cancers [53]. Identifying and characterizing pathogenic missense variants is crucial for understanding disease mechanisms, developing targeted therapies, and implementing personalized medicine approaches [41].

In recent years, high-throughput sequencing technologies have greatly expanded our ability to identify missense variants across populations. Large-scale projects, such as the 1000 Genomes Project [54] and the Exome Aggregation Consortium (ExAC) [55], have provided extensive catalogs of human genetic variation, facilitating the study of missense variants and their role in health and disease.

2.5 Biobanks and Genomic Datasets

Biobanks and genomic datasets are essential for advancing our understanding of human health and disease. These repositories store biological samples, such as blood, tissue, and DNA, along with associated clinical and demographic data from large numbers of individuals. Providing access to these resources, biobanks, and genomic datasets enable researchers to conduct large-scale studies, identify genetic risk factors, and develop personalized medicine approaches [56].

The samples are typically collected from volunteers who consent to share their health information [57]. Biobanks can be population-based, disease-specific, or focused on specific ethnic groups or age ranges [58]. Some notable examples of large-scale biobanks include the UK Biobank (UKBB) [59], the All of Us Research Program in the United States [60], and the China Kadoorie Biobank [61].

Genomic datasets, on the other hand, are collections of genetic and genomic data generated through various high-throughput technologies, such as DNA sequencing, genotyping, and gene expression profiling. These datasets can be derived from biobank samples or generated independently by research groups or consortia [62]. One of the most notable examples of a large-scale genomic dataset is the Genome Aggregation Database (gnomAD), which contains genetic data from over 125,000 exomes and 15,000 whole genomes from diverse populations [63]. GnomAD provides a valuable resource for studying genetic variation, identifying rare variants, and filtering candidate disease-causing mutations [64].

One of the key advantages of biobanks and genomic datasets is their ability to provide large sample sizes, which are essential for detecting genetic associations with complex diseases. By combining data from multiple biobanks and genomic datasets, researchers can achieve the statistical power needed to identify rare genetic variants and study gene-environment interactions [65]. For example, the Global Biobank Meta-analysis Initiative (GBMI) has recently examined over 2 million individuals from 19 biobanks, identifying over 10,000 genetic *loci* associated with more than 60 diseases and traits [66].

ML and AI techniques are increasingly important in leveraging biobanks and genomic datasets for research and clinical applications [67]. Deep learning methods, such as convolutional neural networks and recurrent neural networks, have been applied

to analyze genomic sequences, predict the impact of genetic variants, and identify disease subtypes based on multi-omics data [68].

2.6 Protein Structure and Function

Proteins are essential macromolecules that play a vital role in virtually all biological processes, including catalysis, signaling, transport, and structural support [27]. The function of a protein is determined by its unique three-dimensional conformation and chemical characteristics, ultimately determined by the original amino acid sequence.

The three-dimensional structure of a protein can be described at four hierarchical levels: primary, secondary, tertiary, and quaternary [69]. The primary structure refers to the linear sequence of amino acids derived from the genetic code. The secondary structure refers to the local conformations of the amino acid sequence, such as α-helices and β-sheets, which are stabilized by hydrogen bonds between the amino acid residues [70]. The tertiary structure describes the overall three-dimensional arrangement of the secondary structure elements and the spatial relationships between the side chains of the amino acids. Some proteins consist of multiple polypeptide chains, and the quaternary structure refers to the arrangement of these subunits into a functional complex [28].

The relationship between protein structure and function is complex and multifaceted. Proteins can perform a wide range of functions, such as catalyzing biochemical reactions (enzymes), transporting molecules across membranes (channels and transporters), recognizing and binding specific ligands (receptors), and providing structural support (structural proteins). The specific function of a protein is determined by its unique three-dimensional structure, which creates binding sites, catalytic sites, and other functional and structural sequence patterns called motifs [71].

Predicting the three-dimensional structure of a protein from its amino acid sequence, known as the protein folding problem, has been a grand challenge in computational biology for decades [72]. Recently, significant progress has been made in this area with the development of AlphaFold [13], an artificial intelligence system created by DeepMind. AlphaFold uses deep learning models, or more specifically, Transformers [11], trained on vast amounts of protein structure data to predict the three-dimensional structure of proteins with unprecedented accuracy. In the Critical Assessment of Protein Structure Prediction (CASP) competition [73], AlphaFold achieved a median global distance test (GDT) score of 92.4%, indicating that its

predictions are comparable to experimentally determined structures through X-ray crystallography.

The success of AlphaFold and other protein structure prediction methods has opened up new opportunities for understanding the relationship between protein structure and function. By providing accurate models of protein structures, these tools can facilitate the discovery of new drug targets, the design of novel enzymes, and the understanding of disease-causing mutations [74]. Moreover, integrating protein structure predictions with other omics data, such as gene expression and protein-protein interaction networks, can provide a more comprehensive view of cellular processes and enable the development of personalized medicine approaches [75].

2.7 Loss-of-function and Gain-of-functions Variants

Genetic variants can have diverse effects on protein function, ranging from neutral to deleterious or beneficial. Two important classes of functional variants are LOF and GOF variants, which can significantly affect biological processes and are often associated with disease states [69].

LOF variants are genetic alterations that reduce or eliminate a protein's normal function [76]. These variants can occur through various mechanisms, such as nonsense mutations (introducing a premature stop codon), frameshift mutations (altering the reading frame), splice-site mutations (disrupting normal splicing), or deletions of essential coding regions [77]. LOF variants can produce truncated, unstable, or non-functional proteins, which may be rapidly degraded by cellular quality control mechanisms [78]. LOF variants have been implicated in a wide range of genetic diseases, such as cystic fibrosis (CFTR gene) [52], Duchenne muscular dystrophy (DMD gene) [79], and familial hypercholesterolemia (LDLR gene) [80].

In contrast, gain-of-function (GOF) variants are genetic alterations that confer new or enhanced activity to a protein [81]. GOF variants can produce proteins with increased stability, altered substrate specificity, or constitutive activation, which may disrupt normal cellular processes and lead to disease states [46]. Examples of GOF variants include activating mutations in oncogenes, such as BRAF (associated with various cancers) [82] and FGFR3 (associated with achondroplasia) [83], as well as mutations in ion channel genes, such as SCN9A (associated with pain disorders) [84].

LOF and GOF variants are not essentially disease-causing; rather, they just represent changes in protein function. Yet, distinguishing between LOF and GOF variants is crucial for understanding the molecular basis of genetic diseases and developing targeted therapies [29]. However, predicting the functional impact of variants remains a significant challenge, as the effects may depend on the specific protein context, genetic background, and environmental factors [85]. Computational tools, such as SIFT [31], PolyPhen-2 [48], and CADD [49], are commonly capable of scoring variant pathogenicity but lack the capability of identifying the functional effect such as LOF or GOF classification.

Recent advances in high-throughput sequencing technologies and genome-editing tools, such as CRISPR-Cas9, have enabled the systematic characterization of LOF and GOF variants at a genome-wide scale [86]. Functional screens, such as massively parallel reporter assays (MPRAs) [87] and deep mutational scanning (DMS) [88], have been used to assess the effects of thousands of variants on protein function simultaneously. These approaches have provided valuable insights into the distribution and functional consequences of LOF and GOF variants across the human genome.

2.8 Deep Learning

Deep learning is a subfield of machine learning that has revolutionized various domains, including computer vision, natural language processing, and bioinformatics. It involves the use of artificial neural networks with multiple layers (hence "deep") to learn hierarchical representations of data [89]. Deep learning algorithms have achieved state-of-the-art performance on various tasks, often surpassing human-level performance [90].

The core building block of deep learning is the artificial neural network, loosely inspired by biological neurons' structure and function. A neural network consists of interconnected nodes (neurons) organized into layers, with each neuron receiving input from the previous layer, applying a non-linear transformation, and passing the output to the next layer. The input layer receives the raw data, while the output layer produces the final predictions. Between the input and output layers, one or more hidden layers learn increasingly abstract data representations [91].

The relevance and utility of deep learning lie in its ability to automatically learn relevant features from raw data without manual feature engineering [92]. This is achieved through training, where the neural network is exposed to a large dataset and adjusts its internal parameters (weights and biases) to minimize a loss function that quantifies the difference between the predicted and true outputs. The most common training algorithm is backpropagation [93], which uses the chain rule based on dynamic programming [94] to efficiently compute gradients and update the network's parameters paired with optimization techniques such as stochastic gradient descent [94].

Deep learning has been successfully applied to a wide range of problems in bioinformatics, including protein structure prediction [95], functional protein design [96], variant effect prediction [97], and gene expression analysis [56]. For example, DeepVariant [97], a deep learning-based variant caller developed by Google, has accurately identified genetic variations from sequencing data. Deep learning has also been used to predict the effects of non-coding variants on gene expression and disease risk [98].

One of the key advantages of deep learning is its ability to learn from large, complex datasets, especially from non-tabular, high-dimensional data [99]. As biological data grows exponentially, deep learning provides a powerful tool for extracting meaningful insights and making accurate predictions [17]. However, deep learning also has challenges, such as the need for large amounts of labeled training data, the risk of overfitting [100], and the difficulty interpreting the learned models.

To address these challenges, various techniques have been developed or used, such as data augmentation [101], regularization [102, 103], and transfer learning [104]. Data augmentation generates additional training examples by applying transformations to the existing data. Meanwhile, classic regularization adds penalty terms to the loss function to prevent overfitting. Deep learning-specific regularization uses the specific neural network structure and training procedures, such as Dropout and Early Stopping, to apply ways to prevent overfitting. Transfer learning leverages pre-trained models on large datasets to improve performance on related tasks with limited labeled data [105].

In recent years, there has been a growing interest in developing interpretable deep-learning models [106]. Techniques such as attention mechanisms [10], saliency maps [107], and model-agnostic methods [108] have been used to identify the input features that contribute most to the model's predictions. These approaches can provide

valuable insights into the underlying biological mechanisms and guide further experimental validation [109].

2.9 Neural Network Architectures

Neural network architectures are the fundamental building blocks of deep learning models, defining the structure and organization of the network's layers and connections. The choice of architecture greatly influences the model's performance, efficiency, and ability to learn meaningful representations from data [89-91]. Over the years, various neural network architectures have been proposed, each with strengths and applications in different domains, including computer vision, natural language processing, and bioinformatics [110].

One of the most influential and widely used architectures is the CNN [9]. CNNs are particularly well-suited for processing grid-like data, such as images and time series, due to their ability to capture local and translation-invariant features. The key components of CNNs are convolutional layers, which apply learned filters to the input data, and pooling layers, which downsample the feature maps to reduce spatial dimensions and provide invariance to small translations. CNNs have achieved state-of-the-art performance on various computer vision tasks, such as image classification, object detection, and semantic segmentation [90, 111-113].

Another important architecture is the Recurrent Neural Network (RNN), designed to process sequential data, such as text, speech, and time series. RNNs have recurrent connections that allow information to persist across time steps, enabling the model to capture long-term dependencies and context. However, traditional RNNs suffer from the vanishing gradient problem, making learning long-range dependencies difficult [114]. To address this issue, variants of RNNs, such as Long Short-Term Memory networks (LSTM) [115] and Gated Recurrent Units (GRU) [116], have been proposed, which introduce gating mechanisms to control the flow of information and maintain long-term memory.

Autoencoders are a class of neural network architectures used for unsupervised representation learning. They consist of an encoder network that maps the input data to a lower-dimensional latent space and a decoder network that reconstructs the original data from the latent representation. By minimizing the reconstruction error, autoencoders learn to capture the data's most salient features and structures [117].

Variants of autoencoders, such as Denoising Autoencoders [118] and Variational Autoencoders [119], have been proposed to improve these models' robustness and generative capabilities.

Graph Neural Networks (GNNs) are a family of architectures designed to process graph-structured data, where nodes represent entities and edges represent their relationships. GNNs learn node embeddings by iteratively aggregating information from neighboring nodes and updating the node representations based on the aggregated features [120]. The most popular GNN architectures include Graph Convolutional Networks (GCNs) [121], which generalize the convolution operation to graph-structured data, and Graph Attention Networks (GATs) [122], which introduce an attention mechanism to weigh the importance of different neighbors during the aggregation process.

In bioinformatics, neural network architectures have been adapted and extended to address biological data's unique challenges and characteristics. For example, CNNs have been used to learn motif representations from DNA and protein sequences, RNNs have been employed to predict protein secondary structure and solvent accessibility, and GNNs have been applied to explore molecular graphs and predict drug-target interactions [110, 123-126]. Yet, Transformers excel in sequence modeling tasks such as natural language processing and generation [127], time-series forecasting [128], and biological tasks such as biological molecule structure prediction [129], protein folding [6], and variant prediction [130-131].

2.10 Representation Learning

Representation learning (RL) is a fundamental concept in machine learning, particularly deep learning, that focuses on automatically learning meaningful and useful data representations. In contrast to traditional machine learning approaches that rely on handcrafted features, representation learning aims to automatically discover and learn the underlying structure and relevant features from raw data. RL allows machine learning models to capture complex patterns and relationships in the data, improving performance on various tasks [91].

RL aims to transform raw data into a more abstract and compressed form that captures the essential information while discarding the irrelevant noise. The learned representations should be informative, discriminative, and invariant to certain

transformations or variations in the input data [91]. For example, in image recognition tasks, a good representation should be invariant to changes in illumination, scale, or orientation while preserving the distinctive features that distinguish different object classes [132].

Deep learning models, such as CNNs and autoencoders, have been particularly successful in learning hierarchical representations from complex data. CNNs learn local and translation-invariant features by applying convolutional filters and pooling operations to the input data [132]. As the network depth increases, the learned features become increasingly abstract and capture higher-level concepts [133]. Autoencoders, conversely, learn compressed representations by encoding the input data into a lower-dimensional latent space and then reconstructing the original data from the latent representation [117].

One popular bioinformatics RL approach is using word embeddings, which originated in NLP. In this approach, biological sequences are treated as "sentences" composed of "words" (e.g., k-mers or amino acid residues), and neural networks are used to learn dense vector representations (embeddings) for each word in a multidimensional space. The learned embeddings capture the semantic and functional relationships between the words, allowing for efficient downstream analysis and prediction tasks [134, 135].

Despite the success of representation learning, there are still challenges and open questions in this field. One challenge is the interpretability of the learned representations, as deep learning models often produce highly abstract and non-linear transformations of the input data [136]. Developing methods to visualize and interpret the learned representations is an active area of research. Another challenge is the limited availability of labeled data in many bioinformatics tasks, which can hinder learning effective representations [110]. Transfer learning, self-supervised learning, and unsupervised pre-training approaches have been proposed to address this issue, leveraging large unlabeled datasets to learn general-purpose representations that can be fine-tuned for specific tasks [105].

2.11 Transfer Learning

Transfer learning (Figure 3) is a machine learning approach that leverages knowledge gained from solving one problem to improve the performance of a different but related problem [137]. The main idea behind transfer learning is to use pre-trained models,

which have been trained on large datasets for a specific task, typically demanding plenty of computational power, as a starting point for training on a new task with limited labeled data. By transferring the learned features and representations from the source domain to the target domain, transfer learning can significantly reduce the amount of time and data required to train a model from scratch [138].

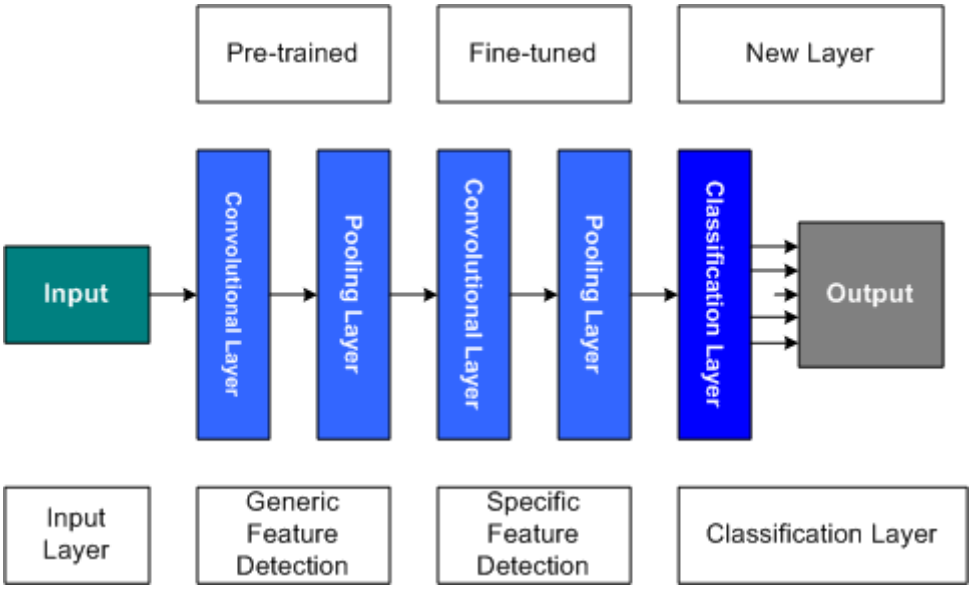

Figure 3: Schematic representation of transfer learning steps using a CNN. First, the input is fed into the pre-trained CNN. Then, resulting weights serve as input to deeper layers of the network that will be responsible for fine-tuning the model and learning task-specific underlying structures in the data. Source: https://developer.ibm.com/articles/transfer-learning-for-deep-learning/

Transfer learning has been successfully applied to various problems in bioinformatics, such as drug-target interaction prediction [139] and genomic sequence analysis [140]. For example, pre-trained language models, such as BioBERT [141], have been fine-tuned on various biomedical text-mining tasks.

2.12 Transformers

Transformers [11] are a type of deep learning architecture that has revolutionized the field of NLP and has been increasingly applied to various tasks in bioinformatics. The key innovation of Transformers is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input sequence when making predictions [11]. Unlike RNNs and CNNs, which process the input sequentially or with fixed-size windows with a limited range. Transformers can observe all positions in the

sequence simultaneously, enabling them to capture long-range dependencies and context [142], but also making the computational cost of this computation $O(n^2)$. As an example (Figure 4), when the input is the sentence "The animal didn't cross the street because it was too tired," the self-attention mechanism is effectively able to capture the dependency between the words "animal" and "it," as seen by the relative value attention scores represented by color intensity.

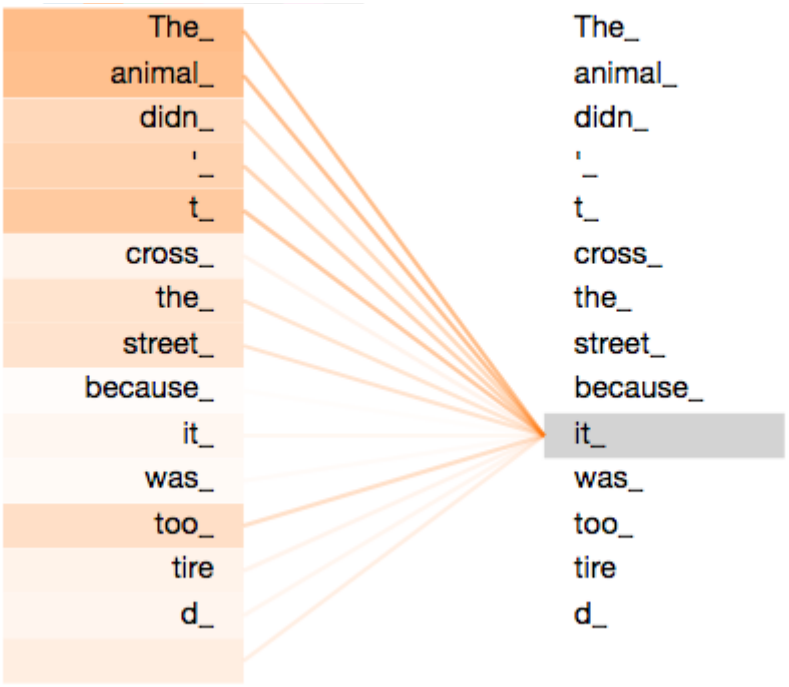

Figure 4: Self-attention scores in a given sentence. The stronger the color, the higher the attention score between the word on the left and "it." Source: https://jalammar.github.io/illustrated-transformer/

In the self-attention mechanism, the input sequence is first transformed into three matrices: the query (Q), key (K), and value (V) matrices (Figure 5). These matrices are obtained by applying linear transformations to the input embeddings. The attention scores are computed as the scaled dot product between the query and key matrices, followed by a softmax function to obtain the attention weights. The output of the self-attention layer is the weighted sum of the value matrix [11].

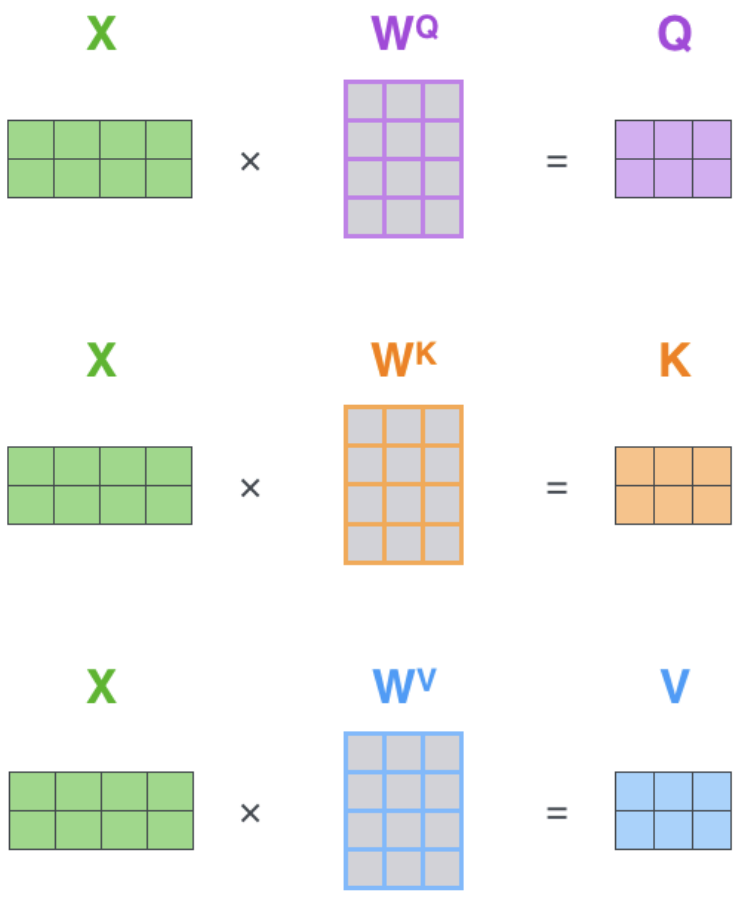

Figure 5: Query, Key, and Value matrix multiplication in the self-attention mechanism. Multiplying a vector x_1 by the W^Q weight matrix produces q_1, while $W^K * x_1 = k_1$, $W^V * x_1 = v_1$, etc. Computing q_n for each n-th token in X results in the Q matrix; the same goes for K and V. Source: https://jalammar.github.io/illustrated-transformer/

Transformers consist of an encoder and a decoder, each containing multiple layers of self-attention and feedforward neural networks (Figure 6) [11]. The architecture is usually applied in sequence-to-sequence tasks, where the encoder takes the input sequence and generates a hidden representation, while the decoder takes the hidden representation and generates the output sequence. In some tasks, such as text classification or protein function prediction, only the encoder is used, and the hidden representation is fed into a classifier or regressor [143].

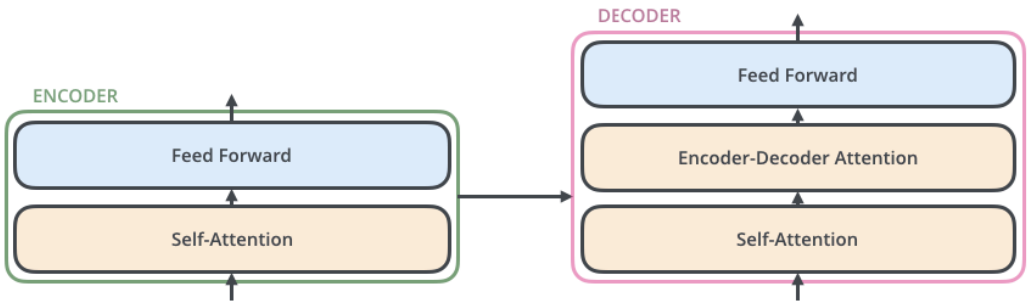

Figure 6: Encoder-Decoder architecture implemented on Transformers. Source: https://jalammar.github.io/illustrated-transformer/

In bioinformatics, Transformers have been applied to a wide range of problems, including protein function prediction [12], protein-protein interaction prediction [144], and genomic sequence analysis [145]. For example, the Evolutionary Scaled Modeling (ESM) Transformer, a 650 million parameters model pre-trained on millions of protein sequences, has achieved state-of-the-art performance on protein function prediction and remote homology detection [12]. Similarly, the DNA Transformer, pre-trained on genomic sequences, has been used to predict the effects of non-coding variants on gene expression [146].

Transformers have shown remarkable promise in advancing state-of-the-art bioinformatics, achieving unparalleled performance in various tasks, from protein function prediction to genomic sequence analysis. By leveraging the self-attention mechanism and pre-training on vast amounts of unlabeled data, Transformers have demonstrated their ability to capture complex biological patterns and unlock the potential of biological data. As bioinformatics evolves, more pre-trained Transformer models and task-specific fine-tuning strategies become available, yet datasets and models are still highly private apart from initiatives such as biobanks and efforts like ESM [12] and AlphaFold2 [13], limiting transparency and collaboration.

2.13 Foundational Models

Foundational models are a new class of AI models that have gained significant attention in recent years due to their impressive performance across various tasks and domains [147]. These models are trained on vast amounts of diverse data using self-supervised learning and usually employ an enormous amount of model parameters. Combined, model size and data enable foundational models to learn general-purpose

representations that can be adapted to various downstream tasks with minimal fine-tuning [127].

The concept of foundational models is inspired by the success of pre-trained language models like BERT [142] and GPT [127] in NLP. These models are trained on a large corpora of unlabeled text data using objectives such as masked language modeling and next-sentence prediction, which allow the models to learn contextual representations of words and sentences. The pre-trained models can then be fine-tuned on specific tasks, such as text classification or question answering, achieving state-of-the-art performance with minimal task-specific training data [143]. By training on large-scale, diverse datasets using self-supervised objectives, foundational models aim to learn universal representations that capture the underlying structure and patterns in the data. These representations can then be transferred to various downstream tasks, enabling efficient learning and improved generalization [148].

One of the key advantages of foundational models is their ability to learn from unlabeled data, which is particularly valuable in bioinformatics, where labeled data is often scarce and expensive to obtain [149]. By leveraging self-supervised learning objectives, foundational models can capture the intrinsic structure and patterns in biological sequences, enabling them to generate informative representations without the need for explicit annotations [150]. These representations can then be used as input features for downstream tasks, reducing the need for task-specific feature engineering and improving the efficiency of model training [151].

2.14 Non-neural Machine Learning Algorithms

While deep learning has achieved remarkable success in various bioinformatics tasks, non-neural machine learning algorithms still play a significant role in the field [152]. These algorithms, which do not rely on neural network architectures, have been widely used for protein function prediction, disease diagnosis, and drug discovery [153]. Non-neural machine learning algorithms offer several advantages, including interpretability, computational efficiency, and the ability to work with smaller datasets [154].

One of the most popular non-neural machine learning algorithms is the Support Vector Machine (SVM) [155]. SVMs are supervised learning models that aim to find the optimal hyperplane that separates different classes in a high-dimensional feature space

[156]. SVMs have been successfully applied to various bioinformatics tasks, such as protein-protein interaction prediction [157], protein fold recognition [158], and cancer classification based on gene expression data [159]. By applying the Kernel Trick, SVMs can handle high-dimensional data and are effective when the number of features is larger than the number of samples [160, 161].

Another widely used non-neural machine learning algorithm is the Random Forest (RF). RF is an ensemble learning method that combines multiple decision trees (DT) to make predictions, introducing randomness in the training process to create a diverse set of trees. This randomness is achieved through feature randomness, where each decision tree node considers a random subset of input features for splitting, and sample randomness, where each tree is trained on a random subset of the training data obtained through bootstrap sampling [162]. During prediction, each tree independently makes a prediction, and the final prediction is obtained by aggregating the predictions of all trees using majority vote for classification tasks or average prediction for regression tasks. The randomness introduced in RF helps to reduce overfitting and improve generalization performance while also enabling the handling of high-dimensional data through automatic feature selection during training. Additionally, RFs provide a measure of feature importance, which can be used for feature selection and interpretation [163].

DTs and RFs are especially good at learning tabular datasets, often beating multilayer perceptron approaches [164]. RFs have been applied to various bioinformatics problems, including protein function prediction [165], DNA methylation pattern analysis [166], and biomarker discovery [167].

Gradient Boosting Machines (GBM) is another ensemble learning method that combines multiple weak learners, typically decision trees, to make predictions. Unlike RF, which trains decision trees independently, GBM trains weak learners sequentially, with each learner trying to correct the mistakes of the previous ones [186]. The training process of GBM involves initializing the model with a constant value that minimizes the loss function, and then iteratively computing the negative gradient of the loss function with respect to the current model's predictions, training a weak learner on the residuals, and adding the predictions of the weak learner, multiplied by a learning rate, to the current model's predictions. The final prediction is obtained by combining the predictions of all weak learners. GBM can use various loss functions, such as squared

error for regression tasks and logarithmic loss for classification tasks, and the learning rate is a hyperparameter that controls the contribution of each weak learner to the final model, helping to prevent overfitting. GBM is known for its high predictive performance, as it can capture complex interactions and non-linearities in the data. Like RF, it also measures feature importance [169]. GBMs have been applied to tasks such as protein-ligand binding affinity prediction [170], gene expression analysis [171], and drug response prediction [172].

Other non-neural machine learning algorithms applied in bioinformatics include k-nearest Neighbors (k-NN), Naïve Bayes, and Hidden Markov Models (HMMs). k-NN is a non-parametric method that makes predictions based on the majority class of the k-closest training examples in the feature space. Naïve Bayes is a probabilistic classifier that assumes the independence of features given the class label. HMMs are probabilistic models that represent the temporal dependencies among a sequence of observations and have been widely used for modeling DNA and protein sequences [173-175].

Despite the success of non-neural machine learning algorithms in bioinformatics, they also have limitations. One challenge is the need for feature engineering, which requires domain expertise to represent the problem properly and can be time-consuming [111]. To address these limitations, researchers have explored hybrid approaches that combine non-neural machine learning algorithms with deep learning, leveraging the strengths of both approaches [176].

2.15 Multiple Sequence Alignment (MSA)

Understanding the relationships between biological sequences is a central theme in bioinformatics. Multiple sequence alignment (MSA) is a fundamental tool that allows researchers to analyze and interpret the similarities and differences between protein or DNA sequences. By aligning three or more sequences, MSAs unveil crucial information about evolutionary relationships, conserved functional domains, and motifs, offering valuable insights into protein function and structure, as well as the impact of genetic variation [177].

MSAs, initially proposed in 1987 by Russell Doolittle and Da-Fei Feng [178], are essential for constructing phylogenetic trees, depicting the evolutionary history and relationships between species. By identifying conserved and variable regions within aligned protein sequences, one can infer evolutionary distances and common ancestry,

providing a deeper understanding of the evolutionary processes that have shaped biological diversity. Additionally, MSAs play a crucial role in protein structure prediction by highlighting conserved residues and motifs within protein families [13]. These conserved regions often correspond to functionally important residues or structural elements, serving as anchors for predicting the three-dimensional structure of proteins.

Furthermore, MSAs contribute significantly to function prediction and annotation by enabling comparisons of protein sequences with unknown functions against databases of known sequences. It allows researchers to infer potential functional roles based on conserved domains or motifs, expanding our understanding of protein functions and their involvement in biological processes. Moreover, MSAs facilitate the identification of functionally important residues by analyzing patterns of conservation and variation across aligned sequences. This knowledge is critical for understanding how proteins function and how mutations may affect their activity [179].

However, introducing MSAs to the data processing significantly increases the complexity of the implementation. To leverage MSAs, it is required first to align the sequence of interest against multiple evolutionary sequences of the same protein [180]. This alignment is computationally expensive due to the employment of pattern-matching algorithms and is also dependent on the availability of similar protein sequences in a reference dataset [181]. Therefore, while increasing the context of conserved functional domains and structural motifs, any approach based on only one sequence is preferred, especially due to the simplicity of implementation. In this sense, this work prioritized using a single sequence approach, which can be done using pretrained models such as Evolutionary Scaled Model, to be described in the following.

2.16 Evolutionary Scaled Model

The Evolutionary Scaled Model (ESM) [12] is a powerful protein language model (PLM) that has gained significant attention in computational biology and bioinformatics. Developed by Rives et al. (2020), ESM is a 650 million-parameter transformer-based model that learns to predict the next amino acid in a protein sequence. Many possible residues occurred in a protein sequence through randomness, yet evolution naturally selected adequate amino acids in each position, either due to

function or structural relevance. Therefore, each amino acid residue carries an indirect evolutionary context of multiple selection steps. By training on a massive dataset, UR50/S, a high-diversity sparse dataset based on the UniRef50 representative sequences of over 250 million protein sequences from various species, ESM has demonstrated remarkable capabilities in capturing the complex patterns and dependencies in protein sequences, thus enabling a wide range of downstream applications [12]. UniRef provides clustered sets of sequences from UniProt Knowledgebase (UniProtKB), and UniRef50 is specifically built by clustering UniRef100 at 50% identity level [182]. The goal behind UniRef50 is to reduce redundancy and increase the detection of distant relationships between sequences.

One of the key features of ESM is its ability to learn from unlabeled protein sequences in a self-supervised manner. By leveraging the vast amount of evolutionary information in protein sequences across diverse species, ESM can capture the intricate relationships between amino acids and the evolutionary constraints that shape protein structure and function. This self-supervised learning approach allows ESM to learn rich and biologically meaningful representations of proteins without the need for explicit annotations or labeled data [12].

The ESM architecture is based on the transformer model, explained in Section 2.12. The transformer architecture enables ESM to capture long-range dependencies and interactions between amino acids in a protein sequence, which is crucial for understanding proteins' complex folding and functional properties. By focusing on different positions in the sequence and learning to weigh their importance, ESM can effectively model the contextual information and dependencies that govern protein behavior [12].

One notable feature of ESM is its ability to generate accurate predictions for a wide range of protein-related tasks, even without task-specific training data. This zero-shot prediction capability has been demonstrated in various applications, such as predicting the effects of mutations on protein function [130], identifying functional sites in proteins [183], and designing novel protein sequences with desired properties [184]. By utilizing the knowledge learned from the vast evolutionary landscape of proteins, ESM can generalize to new and unseen proteins, as well as provide valuable insights into their structure and function.

A significant advantage of ESM is that it does not require MSAs as input, which greatly simplifies its implementation and makes it more accessible to researchers and practitioners. Traditional protein function prediction methods often rely on MSAs to capture evolutionary information and identify conserved regions in proteins. However, constructing accurate and reliable MSAs can be computationally expensive and time-consuming, especially for large and diverse protein families. ESM circumvents this requirement by directly learning from individual protein sequences, making it more efficient and scalable for large-scale analyses.

The authors have made ESM publicly available[1], allowing researchers and practitioners to easily access and utilize this powerful PLM for their studies and applications. The availability of ESM as an open-source resource has facilitated its widespread adoption and spurred further research and development in protein bioinformatics.

Since its introduction, ESM has been applied to various tasks and has shown impressive performance compared to traditional methods. For example, in the task of predicting the effects of single amino acid variants on protein function, ESM has outperformed existing methods, such as PolyPhen-2 [48] and SIFT [31], by leveraging its ability to capture the contextual information and evolutionary constraints in protein sequences.

This specific work utilized a variation of ESM, also released in 2021 by the same research group, ESM-1v [130]. ESM-1v is an advanced version specifically designed for improved protein variant prediction task performance. Building upon the success of the original ESM model, ESM-1v introduces several key enhancements to the training pipeline and model architecture, making it more suitable for accurately predicting the effects of single amino acid variants on protein function.

One of the main improvements in ESM-1v is using a larger and more diverse training dataset. While the original ESM model was trained on a dataset of 250 million protein sequences, ESM-1v leverages an even larger dataset of over 1 billion protein sequences from the UniRef90 database [182]. This expanded training dataset allows ESM-1v to capture a wider range of evolutionary information and to learn more robust and generalizable representations of protein sequences

[1] https://github.com/facebookresearch/esm

ESM-1v also employs the more sophisticated model architecture from ESM-1b, proposed in the original ESM paper. It also utilizes a 650 million parameters transformer model with 33 layers and 1280 hidden units, which allows for a more expressive and nuanced representation of protein sequences. The increased model capacity enables ESM-1v to capture more complex patterns and dependencies in the proteins' structures, leading to approximately 10% improvement in downstream task performance, such as variant prediction [130].

2.17 Related Works

The prediction of LOF and GOF effects in missense variants is a crucial area of research in genomics. Various computational tools and approaches have been developed to address this challenge. In the following, we discuss some notable works related to our study.

AlphaMissense [131] model combines structural context and evolutionary conservation to predict pathogenic effects in missense variants and achieves state-of-the-art results across various genetic and experimental benchmarks. It has been fine-tuned to human and primate variant population frequency data and does not train on clinically curated labels. These datasets can be limited in representing the population's full spectrum of genetic variation. This means that training data may be biased towards variants more likely to have a known functional impact. AlphaMissense provides predictions for all possible single amino acid substitutions in the human proteome, the entire set of proteins expressed by an organism, and classifies 89% of missense variants as either likely benign or likely pathogenic. AlphaMissense expands the number of confidently classified variants and has shown improved performance compared to other models on clinical benchmarks. As a limitation, AlphaMissense uses MSA representation, which, as mentioned, limits the applications of this model and increases its complexity, requiring building the MSA sequences before applying the model [131].

LoGoFunc [185] specifically targets GOF and LOF prediction. The study uses a pre-built dataset by Bayrak et al. [186] comprising 11,370 labeled pathogenic GOF and LOF variants. This work used NLP approaches to identify the classes for the specific variants, which might be a limitation on the biological accuracy of these variants due to

potential language processing ambiguity and errors. They augment this dataset by adding 65,075 variants deposited in the Human Gene Mutation Database (HGMD). The proposed model uses feature engineering to train an ensemble of 27 LightGBM classifiers [185]. The features passed to the model consist of structural representations by AlphaFold 2 (which requires MSA) and molecule descriptors, a feature engineering approach for representing molecule structure (for example, the count of Nitrogen atoms in the molecule).

VariPred [187] compares different approaches to variant prediction, a problem different from the one addressed in this work. Variant prediction, as opposed to variant effect (LOF or GOF) prediction, is the task of distinguishing if a variant will be deleterious, that is, cause a disruption in the original protein role in a biological pathway or benign. This paper effectively compares different versions of ESM models (ESM-1b, ESM-2, ESM-1v) and different approaches to extract the model's learned representations. This study shows that ESM-1v is a better-suited model for variant prediction than ESM-2 [188] when fine-tuned on top of the learned representations. This is potentially due to ESM-1v being specifically tailored for variant prediction tasks, employing a training paradigm and dataset focused on this specific subset of tasks. However, considering a binary zero-shot classification setup, ESM-1b, using simply the log-likelihood on the output layer, outperforms both mentioned models. The study also suggests that ESM-2 might be more suited for protein folding problems, while ESM1 variants (ESM-1v and ESM-1b) perform better in variant function, pathogenicity, and effect prediction. The study hypothesizes that this is potentially due to the closeness of ESM-1 datasets to the human-only proteins included in ClinVar used to train these models. The experimental setup for modeling from VariPred (Figure 7) is similar to what we applied in this work. The most significant change is substituting a fixed shallow network for fine-tuning by trying out different models, including but not only neural networks [187].

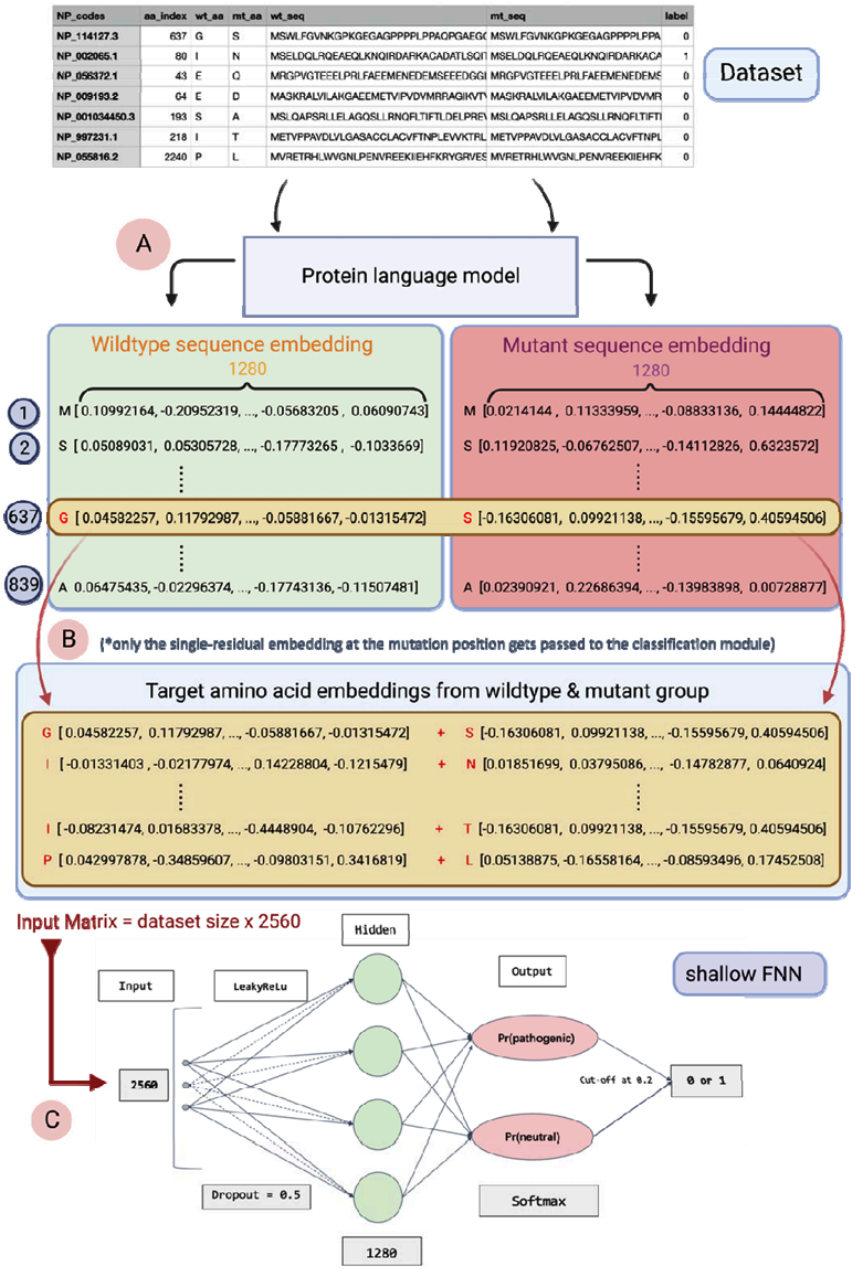

Figure 7: Schematic representation of the steps for embedding generation and downstream classification fine-tuning. A. Protein language model (e.g., ESM) generates wildtype (non-mutated) and mutant sequence embeddings. B and C. Embeddings are fed into a shallow feed-forward neural network model for specific variant prediction tasks (extracted from [187]).

Our work differs from the others mentioned due to the following: increasing the granularity of variant prediction works like AlphaMissense while reducing the inference pipeline complexity by removing the MSA construction step. We also utilize a more refined dataset than LoGoFunc, with manually curated labels on coding genes, while employing a deep-learning-based approach that can effectively remove the feature

engineering used in the mentioned work and potentially improve the quality of predictions. Finally, we explored the VariPred work comparing ESM1 and ESM2 models for variant prediction, finding similar results when fine-tuning both LOF and GOF prediction.

3 Experimental Study

3.1 Computational Environment

This study's 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 (EDA) 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 [189] was the main library for deep learning-related tasks; scikit-learn 1.4.2 [190] was used for non-neural modeling; and hyperopt 0.2.7 [191] for hyperparameter tuning of both deep-learning and non-deep-learning models. The code, and instructions for setting up the Python environment, are provided on GitHub[1].

3.2 Dataset Annotation

A significant contribution of this study is making the dataset provided by Mendelics Análise Genômica S.A. publicly available[2]. The original dataset combines two public variant datasets: benign variants from GnomAD v3.1 [63] and pathogenic variants from ClinVar (July 2023 version) [192]. The team of physicians at Mendelics with PhD in Genetics or Neurology, led by Prof. Dr. Fernando Kok and Dr. David Schlesinger, annotated the dataset, targeting genes with known functions. In the annotation process, each variant was analyzed only once by a specialist whereas in case of uncertainty, a committee of specialists would vote on that specific variant. In case of disagreement between the annotators, the variant was discarded from the final set. The final curated dataset, named Gain and Loss of Function Dataset (GLOF), is available on Kaggle[2].

3.3 Dataset

The experimental dataset contains 112,437 variants, with 3,137 (~2.79%) labeled as GOF, 25,376 (~22.57%) as LOF, and 83,924 (~74.64%) as Neutral. Due to this imbalance in the dataset, specific metrics like F1-score, Recall, and Precision were

---

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

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

chosen for evaluation, further discussed in Section 3.6. The dataset was split into training, validation and test sets using randomized sampling for cross-validation (holdout) [163], respecting the original label distribution in the dataset. From the original dataset, 80% of variants were assigned for training, 10% for validation and hyperparameter tuning, and 10% for tests. Final metrics in this study are reported using the holdout test set. Model selection was conducted using only the validation set to ensure no optimization or decision-making based on the test set.

The only preprocessing required in this study was an integer encoding applied to the labels, where Neutral = 0, LOF = 1, and GOF = 2. This encoding does not introduce ordering bias as the models treat each integer as a distinct class and do not implement ordinal classification objectives [193].

3.4 Embeddings Generation

For generating the embeddings in the Google Cloud Platform, the *fair-esm* library[1] was used for ESM-1v [130], and *HuggingFace* [194] was used for ESM2 [188].

All proteins analyzed were within the length of ESM2 maximum amino acids (tokens). Yet, ESM-1 [12] has a maximum length of 1024 amino acids. While ESM2 did not require any preprocessing, in ESM1 a context window was implemented, as shown in Figure 8. The context window is centralized at the SNP and extends the protein sequence to 512 amino acids to the left and 512 to the right of the SNP. This is a significant limitation on ESM-1v as proteins are 3D-shaped, meaning amino acids further located in the sequence after the 512 amino acids window could still be very close in nature's 3D space as the protein folds, as depicted in Figure 9.

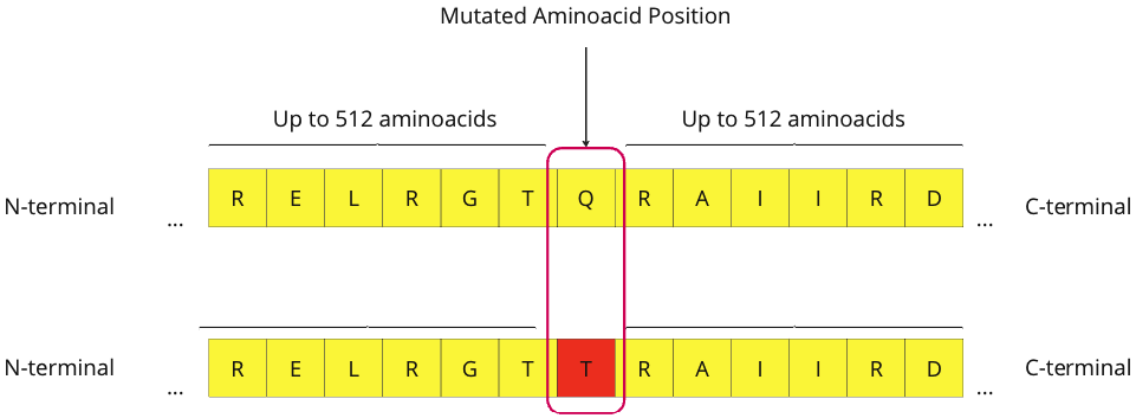

Figure 8: Schematic representation of the context window cropping. Centralizing the mutated amino acid position, up to 512 amino acids are kept towards the N-terminal (beginning of the protein sequence) and C-terminal (end of the protein sequence).

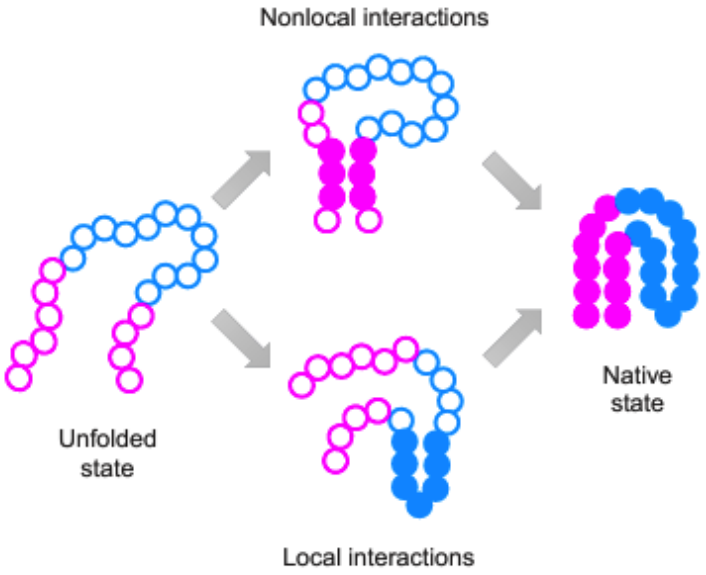

Figure 9: Schematic representation of simple protein folding. Protein is translated into its unfolded state (left), and post-translational chemical interactions and bonds are formed in the tridimensional space (middle), ending in the protein result (right). Amino acids that were far in the unfolded state (linear sequence representation) are then directly interacting with each other. Source: https://www.nature.com/articles/s41467-023-41664-1

The final embedding for both models consists of a 1280-sized vector that ideally is capable of extracting relevant features from the protein sequence [12, 195-196]. Each

---

1 - https://github.com/facebookresearch/esm

protein form has an associated embedding with the protein sequence, as illustrated by Figure 10. It is usually the case where the original and mutated protein will have relatively similar embeddings. Therefore, the wild-type protein has the reference embedding, representing the protein in its natural form. Figure 8 describes the architecture adopted, where the reference protein embedding is concatenated with the mutated protein embedding for each mutation, generating a 2560 input vector for subsequent modeling tasks (Figure 10).

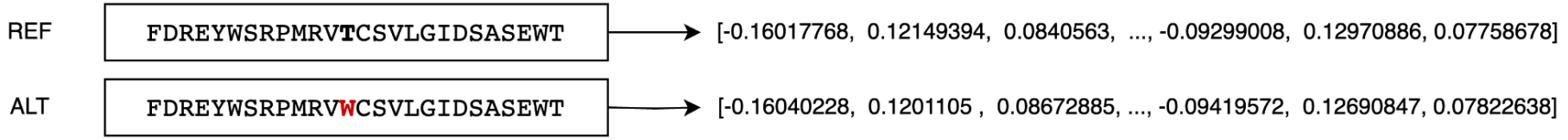

Figure 10: Embedding generation schema. Each protein sequence associated with the variant (REF and ALT) is embedded distinctly.

3.5 Fine-tuning

Five different models were used for fine-tuning the LOF/GOF/Neutral classification task (Figure 11). After generating the embeddings separately for REF (original) and ALT (mutated) proteins, the embeddings are concatenated and fed into an independent classifier (Logistic Regression [163], Random Forest [162], XGBoost [197], LightGBM [169] and Fully Connected Network [90]) for fine-tuning on variant effect. All models were tested with default values for hyperparameters [191], described in Table 1, and the best model was then selected for hyperparameter tuning using the tree of Parzen estimators algorithm.

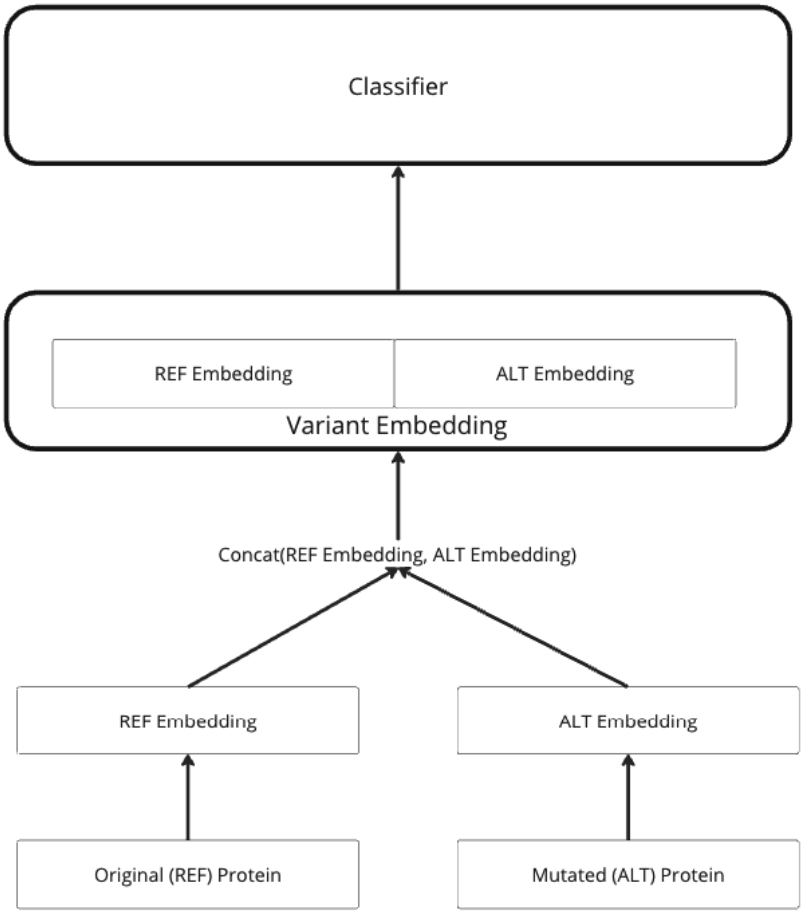

Figure 11: Fine-tuning architecture implementation.

Table 1: Default hyperparameters in tested models.

Model	Parameter list
Logistic Regression	penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None
Random Forest	n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None, monotonic_cst=None
XgBoost Classifier	base_score=0.5, colsample_bylevel=1, colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=10, min_child_weight=1, missing=None, n_estimators=100, nthread=-1, objective='binary:logistic', reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=0, silent=True, subsample=1
LightGBM	boosting_type='gbdt', num_leaves=31, max_depth=-1, learning_rate=0.1, n_estimators=100, subsample_for_bin=200000, objective=None,

	class_weight=None, min_split_gain=0.0, min_child_weight=0.001, min_child_samples=20, subsample=1.0, subsample_freq=0, colsample_bytree=1.0, reg_alpha=0.0, reg_lambda=0.0, random_state=None, n_jobs=None, importanane_type='split'
Fully Connected Network	hidden_layers = [256, 128], epochs=100, batch_size=32, learning_rate=0.001, weight_decay=0.001, optimizer="AdamW"

The Tree of Parzen (TPE) algorithm [191] is a Bayesian optimization method for hyperparameter tuning. It aims to minimize the objective function by intelligently searching the hyperparameter space based on past evaluation results on the holdout set. TPE builds models to approximate the performance of hyperparameters based on historical measurements and then subsequently chooses new hyperparameters to test based on these models.

The TPE algorithm models the objective function as a Parzen estimator, a non-parametric approach to density estimation. It uses a tree-structured scheme to recursively partition the hyperparameter space into subregions based on the previously evaluated hyperparameter configurations and their corresponding objective function values. By modeling the densities, TPE can identify promising regions in the hyperparameter space where good objective function values are more likely to be found. It balances exploration and exploitation by sampling hyperparameter configurations from different density estimations to avoid getting stuck in local optima.

The TPE algorithm is effective and widely used for hyperparameter optimization in various machine-learning tasks. It is particularly useful when the objective function is expensive to evaluate, as it aims to minimize the number of evaluations required to find good hyperparameter configurations. In this study, the RF algorithm, which obtained the best results using default values, was optimized under the TPE algorithm for the hyperparameters: *number of estimators*, *max depth, minimum samples in a leaf, and minimum samples for split*.

3.6 Metrics

For evaluating the models, precision (Equation 1), recall (Equation 2), F1-score (Equation 3), and accuracy (Equation 4) were measured. The first three metrics allow an independent understanding of model behavior across all three classes, which is important due to label imbalance. Precision metric measures the amount of false positive cases in a class, while recall measures false negatives. Good precision means the model classification is not too permissive and labels irrelevant instances, whereas a good recall means the model can correctly identify the relevant cases in a class. F1-score computes a harmonic mean of precision and recall. Accuracy (Equation 4) was also calculated in this study, but due to the imbalance of the labels, it can be misleading if not interpreted alongside auxiliary metrics.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Equation 1: Precision metric calculation.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Equation 2: Recall metric calculation.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Equation 3: F1-score metric calculation.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$$

Equation 4: Accuracy metric calculation.

3.7 Non-Conservative Substitutions

To further understand the model behavior and draw interpretations to the model, a series of investigations were conducted to leverage the biological context of this study.

As mentioned in Section 2.2, non-conservative substitutions in amino acids are more likely to impair protein function [198]. This happens when an amino acid with

different chemical properties replaces another (e.g., a basic amino acid is replaced by an acidic).

In this sense, a chi-squared test was performed to test if the model predictions significantly vary between conservative and non-conservative substitutions [199]. The chi-square test is a statistical test used to determine whether a significant difference exists between the expected occurrence frequency and the observed frequency of the variant effect (LOF, GOF, or Neutral). The chi-square test is performed by calculating the chi-square statistic, which measures the discrepancy between the expected and observed frequency of occurrence. The chi-square statistic is calculated as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Equation 5: Chi-square statistic calculation.

where *O* is the observed frequency of occurrence, *E* is the expected frequency, and the summation is taken over all categories. The chi-square statistic is then compared to a critical value from a chi-square distribution with *k - 1* degrees of freedom, where *k* is the number of categories. If the chi-square statistic is greater than the critical value, then the null hypothesis that the expected and observed frequencies of occurrence are equal is rejected. The chi-square test is a powerful tool for testing the significance of differences between observed and expected frequencies of occurrence. It is a relatively simple test to perform and can be used to test various hypotheses. In this case, it tests whether the frequency of LOF, GOF, or Neutra variants changes significantly given a non-conservative amino acid change.

3.8 Cosine Similarity

It is expected that similar instances are closely represented in the embedding space. In language models, this essentially means that "human" and "person" vectors would be somewhat similar [200]. This similarity can be measured through cosine similarity and Euclidean distance metrics.

This study performed an ANOVA [163] test to check for a statistically significant difference in cosine similarity measure between different labels in ESM-1v embedding space.

In this case, cosine similarity (Equation 6) is calculated between the reference and mutated protein as follows:

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

Equation 6: Calculating the cosine similarity between the non-mutated and mutated proteins in the variant.

where R is the embedding of the reference protein, and A is the embedding of the mutated (alternative) protein. The resulting value is the cosine of the angle between the vectors R and A and can take on values from -1 to 1. Moreover, we define the cosine distance (Equation 7) as the opposite of the cosine similarity:

$$Dist(R, A) = 1 - Sim(R, A)$$

Equation 7: The cosine distance between non-mutated and mutated protein embeddings.

4 Results

This chapter presents computational experiments exploring non-neural and neural models using ESM-1v and ESM2 embeddings as inputs. We investigate the impacts of hyperparameter tuning and examine the connection between model behavior and biological context. Finally, we compare the results with existing methods.

4.1 Classification Metrics

As mentioned, five different model types were evaluated on a validation set extracted from the training set. Evaluation metrics were extracted and reported in Table 2 specifically for the models built using ESM-1v embeddings as input, and Table 3 for ESM-2 results, where the highest value for each metric in each variant effect label is highlighted in bold.

Table 2: Metrics from models trained with default parameters using ESM-1v embeddings.

Metric	Variant Label	Logistic Regression	Random Forest	XGBoost	LightGBM	Neural Network
Precision	Neutral	0.88	**0.92**	0.90	0.90	0.89
	Loss-of-Function	0.67	0.78	**0.79**	0.78	0.73
	Gain-of-Function	0.56	0.84	**0.87**	0.81	0.58
Recall	Neutral	0.92	**0.94**	**0.94**	**0.94**	0.92
	Loss-of-Function	0.29	**0.74**	0.67	0.67	0.64
	Gain-of-Function	0.52	**0.76**	0.65	0.73	0.62
F1-score	Neutral	0.88	**0.93**	0.92	0.92	0.91
	Loss-of-Function	0.58	**0.76**	0.72	0.72	0.68
	Gain-of-Function	0.38	**0.80**	0.75	0.77	0.60
Accuracy		0.81	**0.89**	0.87	0.87	0.85

Table 3: Metrics from models trained with default parameters using ESM-2 embeddings.

Metric	Variant Label	Logistic Regression	Random Forest	XGBoost	LightGBM	Neural Network
Precision	Neutral	0.85	**0.91**	0.89	0.89	0.81
	Loss-of-Function	0.67	**0.78**	0.76	0.76	0.49
	Gain-of-Function	0.58	**0.77**	**0.77**	0.76	0.00
Recall	Neutral	0.92	**0.93**	**0.93**	**0.93**	0.89
	Loss-of-Function	0.53	**0.71**	0.67	0.67	0.38
	Gain-of-Function	0.22	**0.67**	0.58	0.63	0.00
F1-score	Neutral	0.88	**0.92**	0.91	0.91	0.85
	Loss-of-Function	0.59	**0.74**	0.71	0.71	0.43
	Gain-of-Function	0.31	**0.72**	0.67	0.69	0.00
Accuracy		0.80	**0.88**	0.86	0.87	0.75

4.2 Hyperparameter Tuning

Using the initial default hyperparameter values shown in Tables 2 and 3, the winning model is RF for both EMS-1v and EMS-2 embeddings, outperforming the other models tested in most metrics. As a next step, the best model with default hyperparameter values was submitted for hyperparameter tuning using Bayesian optimization. In Table 4, the hyperparameter tuning results of the Random Forest model using ESM-1v are shown, and the highest value for each metric in each variant effect label is highlighted in bold.

Table 4: ESM-1v best model metrics after hyperparameter tuning.

Metric	Variant Label	Default Random Forest	Random Forest with Hyperparameter Tuning
Precision	Neutral	**0.92**	0.91
	Loss-of-function	0.78	**0.79**
	Gain-of-function	0.84	**0.86**
Recall	Neutral	**0.94**	**0.94**
	Loss-of-function	**0.74**	**0.74**
	Gain-of-function	**0.76**	0.72
F1-score	Neutral	**0.93**	**0.93**
	Loss-of-function	**0.76**	**0.76**
	Gain-of-function	**0.80**	0.78
Accuracy		**0.89**	**0.89**

When specifically predicting GOF variants, the RF model, using ESM-1v embeddings as inputs with default hyperparameter values, performed better than the hyperparameter-tuned one, as outlined by the Recall measure. Table 5 depicts the hyperparameter values estimated through Bayesian optimization for the RF model with ESM-1v embeddings.

Table 5: ESM-1v model best hyperparameter values obtained from the hyperparameter tuning step.

Parameter	Optimized Value	Default value
max_depth	92	None
n_estimators	168	100
min_samples_leaf	20	1
min_samples_split	20	2

With respect to ESM-2 embedding, slight improvements were observed in the model after hyperparameter tuning, as shown in Table 6. Although the enhancements were marginal, the hyperparameter-tuned RF was retained for subsequent comparisons with the ESM-1v based model. Table 7 presents the hyperparameter values used in the RF model trained with ESM-2 embeddings after tuning.

Table 6: ESM-2 best model metrics after hyperparameter tuning.

Metric	Variant Label	Default Random Forest	Random Forest with Hyperparameter Tuning
Precision	Neutral	**0.91**	**0.91**
	Loss-of-function	**0.78**	**0.78**
	Gain-of-function	0.77	**0.78**
Recall	Neutral	**0.93**	**0.93**
	Loss-of-function	0.71	**0.72**
	Gain-of-function	**0.67**	0.66
F1-score	Neutral	**0.92**	**0.92**
	Loss-of-function	0.74	**0.75**
	Gain-of-function	**0.72**	0.71
Accuracy		**0.88**	**0.88**

Table 7: ESM-2 model best hyperparameter values obtained from the hyperparameter tuning step.

Parameter	Optimized Value	Default value
max_depth	100	None
n_estimators	209	100
min_samples_leaf	2	1
min_samples_split	6	2

4.3 Comparing ESM1-v and ESM2

At this point, we compared ESM-1v and ESM2 embeddings to determine which is better suited for representing protein molecules for a variant effect classification task. For this, we followed the same methodology for the experiment described in Section 4.1, such that all five model types were trained using ESM2 embedding. The best model was selected, and a hyperparameter tuning step was applied. For ESM2, hyperparameter tuning does enhance model performance, although not very significantly. For the sake of simplicity, we only show the results obtained for the best model obtained, exhibited in Table 5. This table also compares the best models considering both embeddings ESM1-v and ESM2. From the results, it is clear that ESM-1v is the embedding best suited for this downstream task since it is the one that leads to the best-performing model (Table 6).

Table 8: Comparison between best models using ESM-2 and ESM-1v embeddings.

Metric	Variant Label	ESM-2 Best Model (with hyperparameter tuning)	ESM-1v Best Model (without hyperparameter tuning)
Precision	Neutral	0.91	**0.92**
	Loss-of-function	**0.78**	**0.78**
	Gain-of-function	0.78	**0.84**
Recall	Neutral	0.93	**0.94**
	Loss-of-function	0.72	**0.74**
	Gain-of-function	0.66	**0.76**
F1-score	Neutral	0.92	**0.93**
	Loss-of-function	0.75	**0.76**
	Gain-of-function	0.71	**0.80**
Accuracy		0.88	**0.89**

4.4 Cosine Similarity and Variant Effect

From the last experiments, it is clear that the embedding representation derived from ESM models is well suited for subsequent variant effect prediction, as seen by the considerably high metrics obtained, with ESM-1v combined with Random Forest being the best model for this task. However, ESM models can do zero-shot classification for deleteriousness or benign labels [130]. Following this finding, we calculate the cosine distance between the embeddings of reference and alternative proteins within variants of the same type. Then, it investigated how the embeddings correlate with variant effects, as shown in Figure 12.

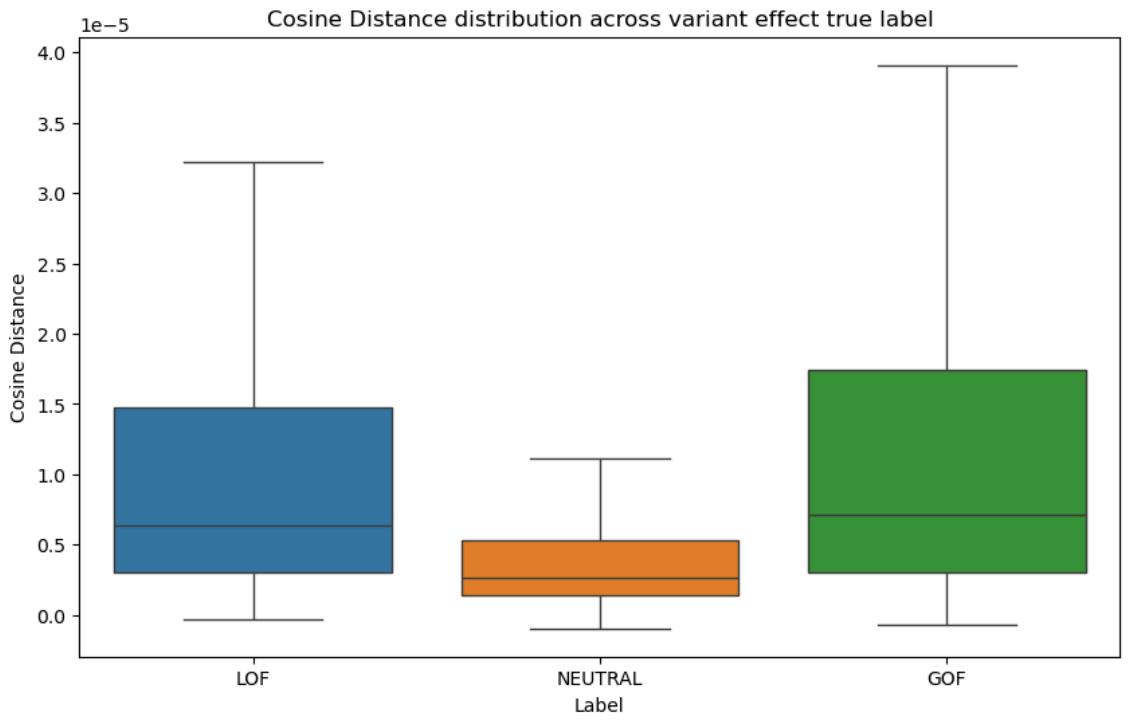

Figure 12: Comparing cosine distance between proteins within variants of the same label. The higher the cosine distance, the more distinct the embedding model represents the proteins in the embedding space.

Overall, LOF and GOF variants have a higher cosine distance (lower cosine similarity) than Neutral variants. A one-way analysis of variance (ANOVA) test was further performed to evaluate the statistical significance of this finding with the resulting $p < 0.0005$. In this statistical test, the null hypothesis (H0) was that the mean cosine distance was the same across all three labels (Neutral, LOF, and GOF). In other words, there is no statistically significant difference among the labels' average cosine distance values. Meanwhile, the alternative hypothesis (H1) was that at least one of the means of the cosine distance values differs from the others. In other words, there is a statistically significant difference in the average cosine distance values among the different labels.

Deleterious variants (LOF or GOF) being more dissimilar (higher cosine distance) corroborate with the biological concept that proteins that remain structurally and chemically similar should keep their original function, while mutated proteins that suffer from structural changes or have non-conservative substitutions will tend to deviate from its original function, potentially gaining or losing it [33].

It is also found that GOF variants have higher dissimilarity in terms of cosine distance ($p < 0.05$). Further biological investigation on this is necessary to understand if there is a biological process behind it. It could be hypothesized that GOF variants that do not have an augmented function but gain a whole new one would need a more distant structure from their original form [201]. For example, in the case of enzymes, losing a function is as simple as not being able to interact with a substrate anymore. However, gaining a function would mean that the protein can interact with different substrates or at least interact with the same substrate in a completely new way [202].

4.5 Biological Reasoning Emerges from Sequences

Amino acids are the representation unit of a protein. As mentioned, roughly 20 amino acids are needed to build a protein sequence. These amino acids share similar chemical properties. For example, Arginine (R) and Lysine (K) are positively charged (basic) amino acids, while Aspartic Acid (D) and Glutamic Acid (E) are negatively charged (acidic) amino acids. A non-conservative substitution from a hydrophobic amino acid (aliphatic) like Valine (V) to a hydrophilic amino acid, such as Glutamic Acid, consisting of an Aliphatic-Acid substitution, could significantly alter the protein structure and function. Therefore, this study examines whether non-conservative substitution relates to a LOF, GOF, or Neutral prediction. To investigate this, the amount of non-conservative and conservative changes in each variant effect class was counted (Figure 13).

A chi-squared test confirmed this finding ($p < 0.0005$), where H0 represented no significant association between the non-conservative changes and the label itself, while H1 was that there is an association between non-conservative changes and the label. Therefore, it is indeed the case where mutations that change the class of amino acids increase the chance of causing a GOF or LOF variant.

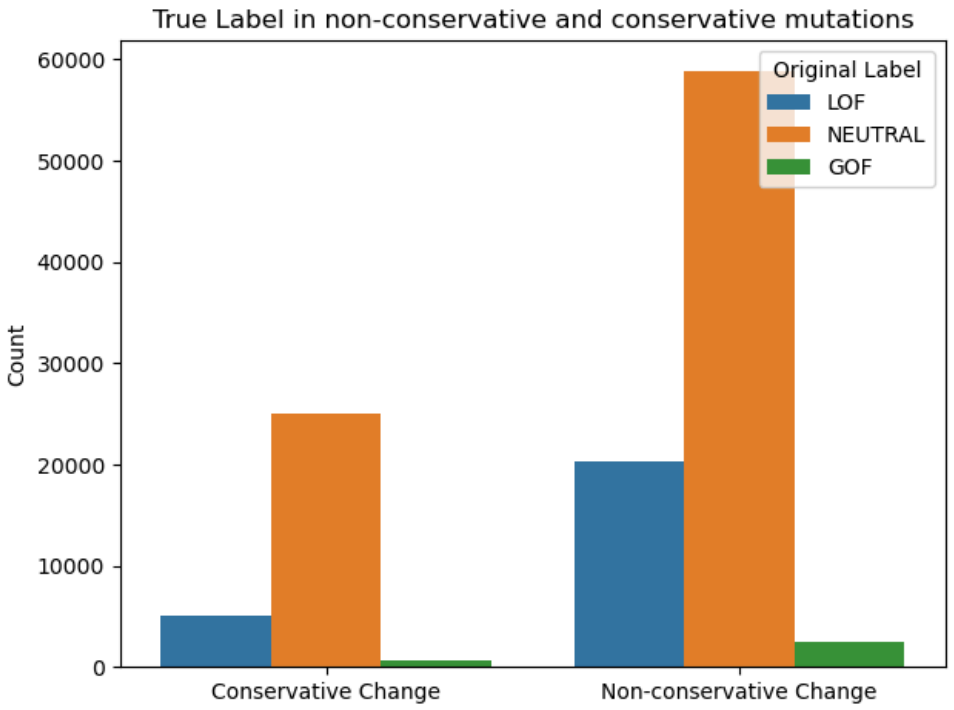

Figure 13: Counting the true labels of distinct variant effect instances regarding non-conservative and conservative changes.

However, this finding does not imply that the model also captures this biological context of non-conservative mutations. The same test was then evaluated using the fine-tuned model's predicted labels, shown in Figure 14. In this case, the model distribution for LOF and GOF variants with amino acid class changes is similar to that seen in the original labels ($p < 0.0005$). Therefore, the model does capture the biological reasoning behind non-conservative mutations and is more prone to predict LOF and GOF classes in non-conservative mutations.

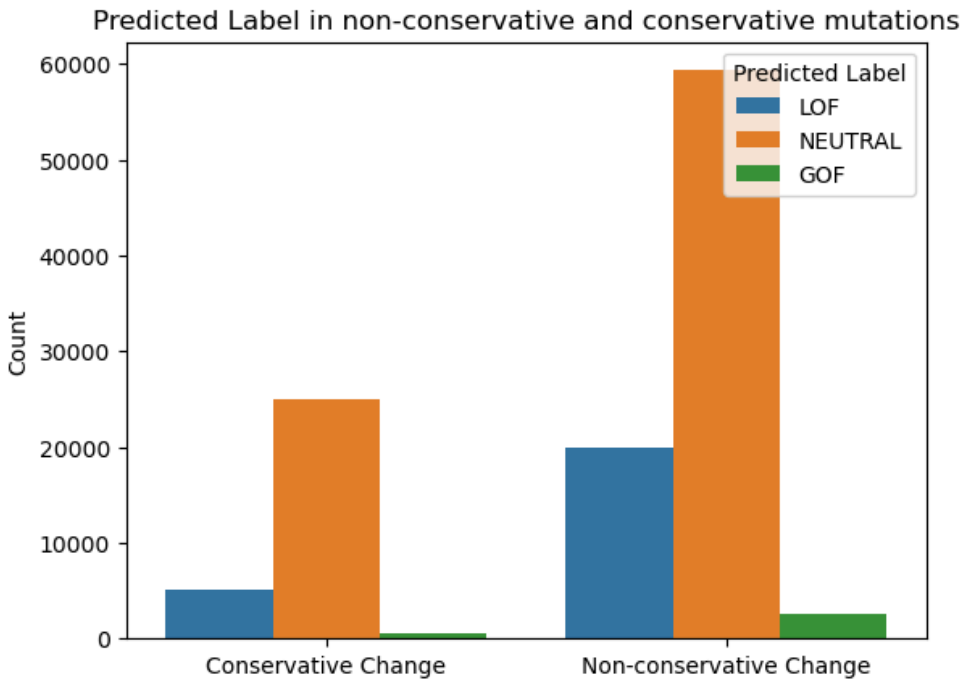

Figure 14: Counting the predicted labels of distinct variant effect instances regarding non-conservative and conservative changes.

It was further evaluated how these variants are distributed across different non-conservative and conservative mutations, which is exhibited in Figure 15.

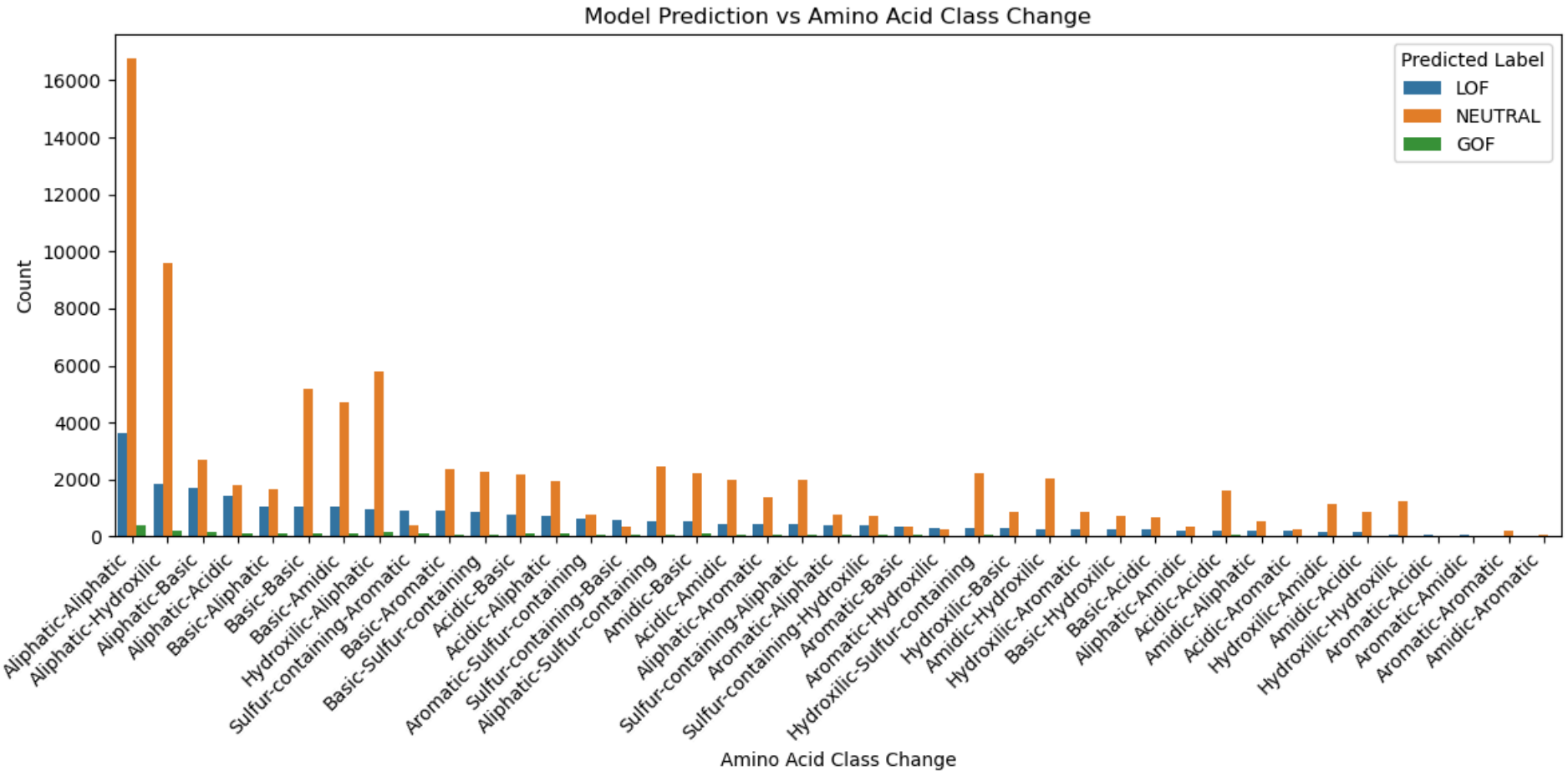

Figure 15: Counting variants per predicted label in different mutations concerning the amino acid class.

In this case, the most relevant finding is that conservative mutations (for example, Aliphatic-Aliphatic) tend to cause neutral variants. The higher prevalence of aliphatic-aliphatic variants is mostly due to most of the amino acids being Aliphatic (6/20). In contrast,

Aromatic and Basic amino acids are the second most frequent amino acid class (3/20, summing 6/20). The remaining classes (Acidic, Hydroxylic, Sulfur-containing, and Amidic) are the least represented (2/20 each, summing 8/20). Conservative mutations are more likely to be identified because they have a lower impact on survival during early developmental stages [203]. This is because conservative mutations tend to be less detrimental to the subject [204]. Yet, this analysis shows that even conservative mutations can generate LOF, especially GOF variants, with GOF being mostly Aliphatic-Aliphatic mutations. This requires further biological investigation.

4.6 Biological Complexity Impacts Model Quality

Genes and proteins serve a biological purpose. Significant changes in the charge of a protein region may affect how a protein regulates itself [205], as in CHMP2B, which is associated with ALS [206]. Some genes are more relevant to core biological functions than others that may play a more supportive role. In this experiment, we investigate if a given protein's biological role importance is implicitly represented by the sequence and, thus, perceived by the model.

In particular, the FBN1 gene, which has been extensively researched in the literature, appears to be highly susceptible to mutations leading to the loss of its function [207]. This gene is associated with microfibrillar bundles, which play a role in elastogenesis. Missense mutations in this gene are known to interfere with the assembly of microfibrils through a loss-of-function and dominant-negative mechanism [208]. The model predicted ~ 93% of the variants for this specific gene to be LOF, with an F1-score of 99.2% observed. Therefore, the result of highly prevalent LOF predictions by the model in this context is on par with what's known about the disruption of biological capabilities in this gene [209]. This gene is associated with Marfan syndrome and lower height in Peruvian populations [210]. The lack of life-threatening phenotypes associated with these variants also affects how the mutations can spread through the population. In this sense, it is reasonable that many LOF variants are present in the FBN1 gene. However, this gene is also responsible for coding a large protein further proteolytically cleaved near its C-terminal [211]. In other words, this could mean that the more relevant sections of the protein are more concentrated in this specific region, meaning that potentially, LOF variants in less important sections are less detrimental to the individual. In this experiment, the context window restriction (1024 amino acids) might also play a role in the full capability of the model to understand the variant effect for long proteins.

Following this, it's reasonable to assume the model will better predict some genes and worse in others. To investigate this hypothesis, we measured the percentage of genes with F1-score higher than multiple thresholds, as shown in Figure 16. The model was calculated to yield an F1-score > 0.75 for 89.62% of genes with GOF

variants, 87.43% with LOF variants, and 95.90% with Neutral variants on the hold-out set.

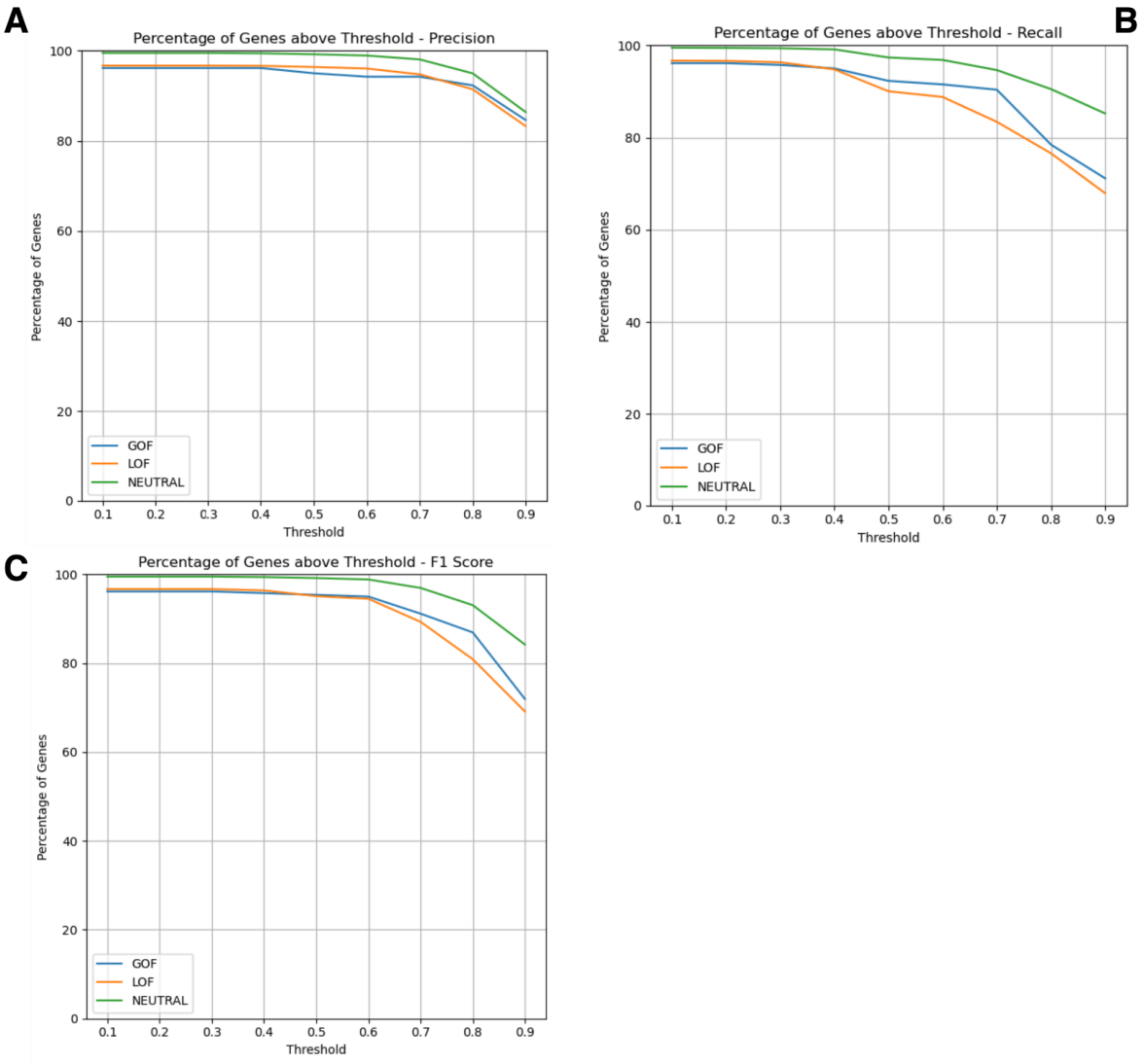

Figure 16: Percentage of unique genes with Precision (A), Recall (B), and F1 Score (C) at different cutoffs.

Genes with rare variants (therefore, underrepresented in the training set) yield worse performance. Therefore, the model can generalize for unseen variants in the same genes but seems to fail to generalize to poorly represented genes in the training set.

4.7 Protein Length Impacts the Model Performance

As mentioned above, it is valid to hypothesize that the larger the protein, the harder it will be for the model to make accurate predictions. This is not only due to context window limitations on the ESM-1v model but also due to Transformers and language models' capabilities to retrieve information from a large context length ("needle-in-a-haystack" problem) [212-213]. In the dataset, the largest observed protein in length is Titin, encoded by the TTN gene, corresponding to 35,992 amino acids. For this, we investigated model performance in terms of F1-score across different protein length inputs and found that the performance degrades slightly as protein length increases (Figure 17). Instead, the phenomenon is observed for LOF and GOF labels.

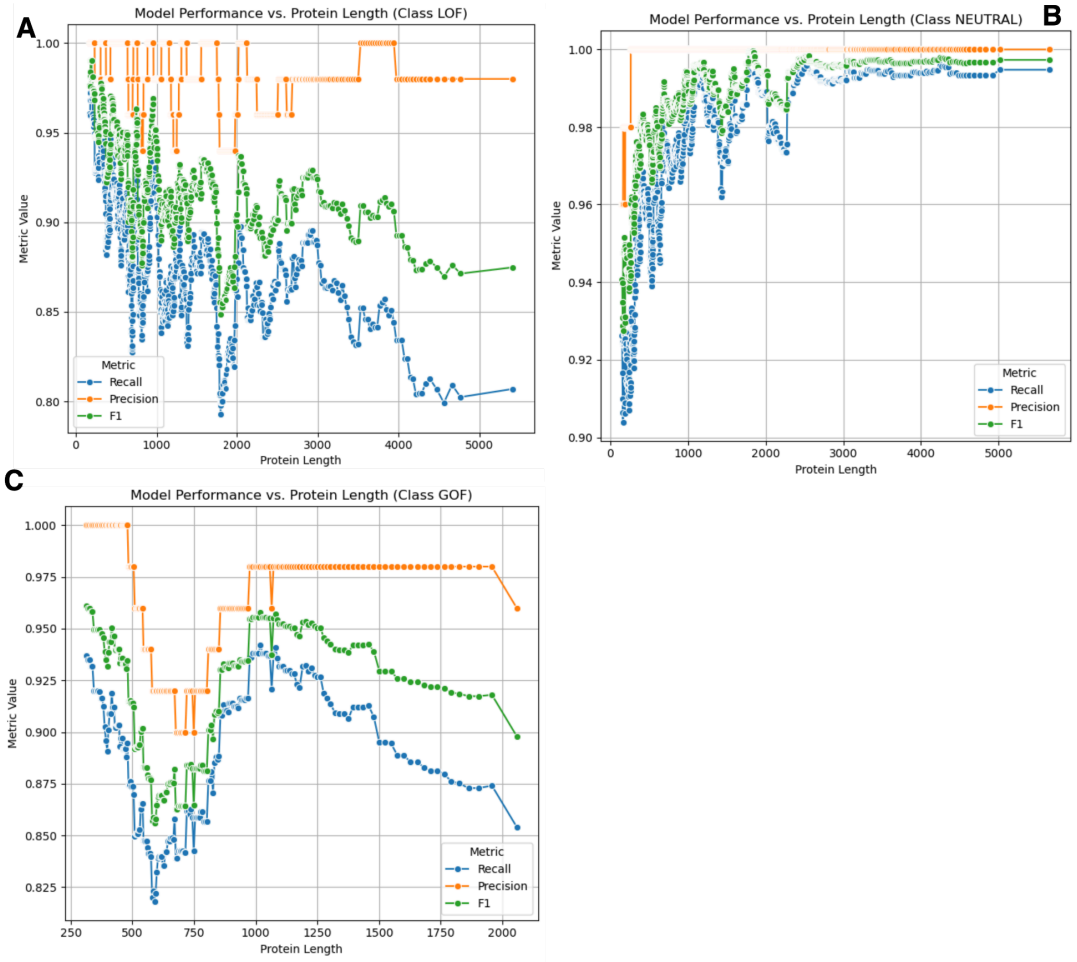

Figure 17: Model performance (blue: Recall, green: F1-score, orange: Precision) at different protein lengths and classes (A: LOF, B: Neutral, C: GOF). Moving average with window size = 50 was used for better visualization.

4.8 Comparison with Existing Methods

Common missense predictors consider facts like the conservation of amino acid sites across species and populations and protein folding effects. Some examples of predictors are EVE [214], AlphaMissense [131], PrimateAI [215], PROVEAN [216], MetaRNN [217], MetaSVM [218], REVEL [219], PERCH [220]. However, none of these predictors are trained to infer LOF and GOF effects. Instead, Stein et al. [185] proposed an ensemble method, LoGoFunc, for genome-wide LOF and GOF prediction based on LightGBM with an initial step of structured feature engineering. In their study, they resorted to NLP-based techniques to extract the variant labels from public datasets. Alternatively, our study relies on data annotated by specialists. LoGoFunc reports F1-scores of 0.56, 0.87, and 0.89 for GOF, LOF, and Neutral variants, respectively on a fivefold cross-validation setup. In comparison, this work does not require structured feature engineering, directly using the protein sequence, reporting F1-scores of 0.80, 0.76, and 0.93 for GOF, LOF, and Neutral variants, respectively, on the hold-out set. However, this comparison should be done carefully since the datasets employed in both works are different.

This work addresses the problem of the lack of a common benchmark dataset with human and biologically-backed rules annotations that is reliable for downstream tasks and model evaluation, as well as the usage of direct protein sequence for predicting LOF, GOF, and Neutral variant effects in missense variants. It represents a good contribution to the area and a building block toward a more comprehensive understanding of pathophysiological mechanisms in disease-causing variants.

5 Conclusions

5.1 Final Considerations

This research's primary contribution is to establish a benchmark for variant effect prediction, meticulously curated by geneticists and physicians at Mendelics Análise Genômica S.A. The benchmark dataset, named Gain and Loss of Function Dataset (GLOF) provides a solid foundation for future research and model development in variant effect prediction, particularly in LOF and GOF variants.

The comparative analysis between ESM-2 and ESM-1v embeddings for LOF, GOF, and Neutral variant effect prediction reveals that ESM-1v outperforms its more recent counterpart, corroborating the findings of previous works. This insight highlights the importance of selecting the most suitable protein embedding model for the specific downstream task, as newer models may not always yield superior performance.

Moreover, this work presented a state-of-the-art end-to-end model for variant effect prediction, as it does not require any intervention from the user apart of providing the reference and alternative sequences. More important, this is done at the granularity of GOF and LOF, which is crucial for understanding pathophysiological mechanisms in genetic diseases. The proposed model eliminates the need for handcrafted feature engineering by leveraging the comparison between wild-type (reference) and mutated (alternative) proteins, streamlining the prediction process and potentially improving its accuracy.

Besides, this study also delved into the model's behavior for distinct biological interpretations, effectively demonstrating that the model captures some of the biological reasoning behind variant effects, such as the impact of non-conservative variants and gene function. Understanding the model performance and behavior in different scenarios is a key contribution to the current demands of deep learning model interpretability. It is essential for building trust in the model's predictions, facilitating its adoption in clinical and research settings.

5.2 Study Limitations

The main limitation of this research is the lack of a diverse population for conducting genetic validation across different populations. The claims made in this study are only valid for populations represented in gnomAD, as the provided dataset is derived from

this resource. Consequently, this research inherits the population bias present in gnomAD. Yet, gnomAD is quite representative, including data from 140,000 individuals from various populations, such as African, Latino, East Asian, South Asian, and non-Finnish European populations. However, there is still an overrepresentation of European ancestry, limited representation of indigenous populations, bias towards disease-specific studies or reference populations, and may not adequately represent populations with a high rate of consanguinity, as the database primarily includes outbred (non-consanguineous) populations. Although it is foreseen that this study is highly expected to be generalizable to general populations due to the underlying dataset and self-supervised learning-based modeling, future research should incorporate a more diverse set of genetic variants from various world populations to increase the reliability and generalizability of the findings.

Regarding the modeling approaches, the study oversimplified the best model selection by using default hyperparameters before applying hyperparameter optimization. The claim that Random Forest is the best model for this dataset would be stronger if all models were compared after hyperparameter optimization. It would also be interesting to validate these findings using wet lab experiments for variant effects and model behavior, augmenting the study's ability to validate *de novo* variant effects. These wet lab and clinical validations are also required to effectively consider these variants annotations in clinical diagnosis per the ACMG/AMP guidelines [221].

5.3 Potential Applications

The proposed model has clinical and research applications. From a clinical perspective, the model can support the diagnosis of genetic diseases in patients. Understanding the effects of LOF or GOF variants is crucial for designing appropriate treatment plans and ensuring that the treatment does not aggravate the patient's condition.

In terms of research, the proposed dataset constitutes a benchmark that can stimulate new research in this area, and that can be used to develop more efficient methods for classifying LOF and GOF variants, as well as to deepen our understanding of disease mechanisms.

5.4 Future Work

As future work, it would be ideal to compare more versions of ESM (like ESM-1b, ESM-1a) and other protein embeddings, potentially leveraging AlphaFold 3 efficient latent representation of protein folding. This comparison would provide a more comprehensive understanding of the performance and limitations of various embedding models in the context of the variant effect prediction task. Furthermore, AlphaFold 3's ability to foster new studies on protein-protein, protein-DNA, protein-RNA, and protein-antibodies interactions will be essential for further understanding underlying biological mechanisms that may explain the LOF or GOF effect on variants.

Further research is necessary to investigate whether the findings regarding the predominance of conservative Aliphatic-Aliphatic amino acid changes in GOF variants hold in a wider population study. This investigation could uncover underlying biological mechanisms or determine if the observed pattern is a consequence of variant frequency in the studied dataset.

A proper comparison with existing approaches, such as LoGoFunc, using the proposed common benchmark is essential for further understanding the limitations and strengths of each model. As the benchmark is publicly available, this research encourages the development of new approaches for modeling variant effects that can push the boundaries of human understanding of genetic variants beyond the scope of this work.

Additionally, implementing hierarchical classification techniques in modeling approaches may prove beneficial in managing class imbalance and the underrepresentation of GOF variants. Moreover, statistical methods, like SVM, particularly focused on anomaly detection, might assist in alleviating the intricate nature of modeling infrequently occurring labels.

Alternatively, as LLMs are typically trained using an enormous corpus that might eventually include biological sequences, they could be capable of predicting proteins biological characteristics such as function. In this case, information retrieval techniques may have a key role in assisting these LLMs to have the relevant information (such as similar protein sequences) directly from the prompt.

References

[1] MARDIS, Elaine R.. **Next-Generation DNA Sequencing Methods. Annual Review Of Genomics And Human Genetics,** [S.L.], v. 9, n. 1, p. 387-402, 1 set. 2008. Annual Reviews. http://dx.doi.org/10.1146/annurev.genom.9.081307.164359.

[2] SHASTRY, Barkur S.. **SNPs: impact on gene function and phenotype. Methods In Molecular Biology**, [S.L.], p. 3-22, 2009. Humana Press. http://dx.doi.org/10.1007/978-1-60327-411-1_1.

[3] AIDOO, Michael; TERLOUW, Dianne J; KOLCZAK, Margarette s; MCELROY, Peter D; KUILE, Feiko O Ter; KARIUKI, Simon; NAHLEN, Bernard L; A LAL, Altaf; UDHAYAKUMAR, Venkatachalam. **Protective effects of the sickle cell gene against malaria morbidity and mortality**. The Lancet, [S.L.], v. 359, n. 9314, p. 1311-1312, abr. 2002. Elsevier BV. http://dx.doi.org/10.1016/s0140-6736(02)08273-9.

[4] KUCHENBAECKER, Karoline B.; HOPPER, John L.; BARNES, Daniel R.; PHILLIPS, Kelly-Anne; MOOIJ, Thea M.; ROOS-BLOM, Marie-José; JERVIS, Sarah; VAN LEEUWEN, Flora E.; MILNE, Roger L.; ANDRIEU, Nadine. **Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers**. Jama, [S.L.], v. 317, n. 23, p. 2402, 20 jun. 2017. American Medical Association (AMA). http://dx.doi.org/10.1001/jama.2017.7112.

[5] KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E.. **ImageNet classification with deep convolutional neural networks**. Communications Of The Acm, [S.L.], v. 60, n. 6, p. 84-90, 24 maio 2017. Association for Computing Machinery (ACM). http://dx.doi.org/10.1145/3065386.

[6] DENG, J. et al. **ImageNet: A large-scale hierarchical image database**. 2009 IEEE Conference on Computer Vision and Pattern Recognition, [s.l.], 2009. ISBN: 9781424439928, DOI: https://doi.org/10.1109/cvpr.2009.5206848.

[7] SIMONYAN, K.; ZISSERMAN, A. **Very Deep Convolutional Networks for Large-Scale Image Recognition.** Computer Science, [s.l.], 2014. DOI: https://doi.org/10.48550/arXiv.1409.1556.

[8] SZEGEDY, C. et al. **Going Deeper with Convolutions**. arXiv (Cornell University), [s.l.], 2014. DOI: https://doi.org/10.48550/arxiv.1409.4842.

[9] HE, K. et al. **Deep Residual Learning for Image Recognition**. arXiv (Cornell University), [s.l.], 2015. DOI: https://doi.org/10.48550/arxiv.1512.03385.

[10] DZMITRY BAHDANAU; CHO, K.; YOSHUA BENGIO. **Neural Machine Translation by Jointly Learning to Align and Translate**. arXiv (Cornell University), [s.l.], 2014. DOI: https://doi.org/10.48550/arxiv.1409.0473.

[11] VASWANI, A. et al. **Attention Is All You Need**. arXiv.org. 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. DOI: https://doi.org/10.48550/arXiv.1706.03762.

[12] RIVES, A. et al. **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. Proceedings of the National Academy of Sciences, [s.l.], no 15, 2021. DOI: https://doi.org/10.1073/pnas.2016239118.

[13] JUMPER, J. et al. Highly accurate protein structure prediction with alphafold. Nature, [s.l.], no 7873, 2021. DOI: https://doi.org/10.1038/s41586-021-03819-2.

[14] **AlphaFold: a solution to a 50-year-old grand challenge in biology - Google DeepMind**. Available at: <https://deepmind.google/discover/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology/>. Access at: Apr 27th, 2024.

[15] BLOW, D. **Protein crystallography**. **Nature 1977 265:5592**, v. 265, n. 5592, p. 390–390, jan. 1977.

[16] KARKI, R. et al. **Defining “mutation” and “polymorphism” in the era of personal genomics**. BMC Medical Genomics, [s.l.], no 1, 2015. DOI: https://doi.org/10.1186/s12920-015-0115-z.

[17] JUNG, S. et al. **Identification of genomic features in the classification of loss- and gain-of-function mutation**. BMC Medical Informatics and Decision Making, [s.l.], no S1, 2015. DOI: https://doi.org/10.1186/1472-6947-15-s1-s6.

[18] GRIFFITHS, A. J. F. et al. **Introduction to genetic analysis**, Tenth edition. **Livre**, p. 862, 2012.

[19] EILBECK, K.; QUINLAN, A.; YANDELL, M. **Settling the score: variant prioritization and Mendelian disease**. Nature Reviews Genetics, [s.l.], no 10, 2017. DOI: https://doi.org/10.1038/nrg.2017.52.

[20] ROTTHIER, A. et al. **Mutations in the SPTLC2 Subunit of Serine Palmitoyltransferase Cause Hereditary Sensory and Autonomic Neuropathy Type I**. The American Journal of Human Genetics, [s.l.], no 4, 2010. DOI: https://doi.org/10.1016/j.ajhg.2010.09.010.

[21] JOHNSON, J. O. et al. **Association of Variants in the SPTLC1 Gene With Juvenile Amyotrophic Lateral Sclerosis**. JAMA neurology, [s.l.], no 10, 2021. DOI: https://doi.org/10.1001/jamaneurol.2021.2598.

[22] ALBERTS, B. et al. **Molecular biology of the cell**. p. 70, [s.d.].

[23] WATSON, J. D.; CRICK, F. H. C. **Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid**. Nature, [s.l.], no 4356, 1953. DOI: https://doi.org/10.1038/171737a0.

[24] CRICK FH. **On protein synthesis**. Symp Soc Exp Biol. 1958;12:138-63. PMID: 13580867.

[25] CRICK, F. **Central Dogma of Molecular Biology**. Nature, [s.l.], no 5258, 1970. DOI: https://doi.org/10.1038/227561a0.

[26] LODISH, H. et al. Section 20.3. G Protein-Coupled Receptors and Their Effectors. in Molecular Cell Biology. **Molecular Cell Biology**, p. 862–871, 2000.

[27] BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. Exploring Proteins. **Biochemistry**, p. 137–194, 2002.

[28] NELSON, D. L.; COX, M. M. Biosignaling. In: Lehninger Principles of Biochemistry. **Lehninger Principles of Biochemistry**, p. 1129–1264, 2017.

[29] CHATTERJEE, N.; WALKER, G. C. **Mechanisms of DNA damage, repair, and mutagenesis. Environmental and molecular mutagenesis**, [s.l.], no 5, 2017. DOI: https://doi.org/10.1002/em.22087.

[30] STRACHAN, T.; READ, A. P. Human Molecular Genetics. **Human Molecular Genetics**, 20 dez. 2018.

[31] NG, P. C. **SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research**, [s.l.], no 13, 2003. DOI: https://doi.org/10.1093/nar/gkg509.

[32] ANFINSEN, C. B. **Principles that Govern the Folding of Protein Chains**. Science, [s.l.], no 4096, 1973. DOI: https://doi.org/10.1126/science.181.4096.223.

[33] BETTS, M. J.; RUSSELL, R. B. Amino Acid Properties and Consequences of Substitutions. **Bioinformatics for Geneticists**, [s.l.], [s.d.]. ISBN: 0470843934, DOI: https://doi.org/10.1002/0470867302.ch14.

[34] ZUCKERKANDL, E.; PAULING, L. Evolutionary Divergence and Convergence in Proteins. **Evolving Genes and Proteins**, p. 97–166, 1 jan. 1965.

[35] WANG, Z.; MOULT, J. **SNPs, protein structure, and disease. Human Mutation**, [s.l.], no 4, 2001. DOI: https://doi.org/10.1002/humu.22.

[36] LIU, C. C.; SCHULTZ, P. G. **Adding New Chemistries to the Genetic Code**. Annual Review of Biochemistry, [s.l.], no 1, 2010. DOI: https://doi.org/10.1146/annurev.biochem.052308.105824.

[37] JOHNSON, J. A. et al. **Residue-specific incorporation of non-canonical amino acids into proteins: recent developments and applications**. Current opinion in chemical biology, [s.l.], no 6, 2010. DOI: https://doi.org/10.1016/j.cbpa.2010.09.013.

[38] ALKAN, C.; COE, B. P.; EICHLER, E. E. **Genome structural variation discovery and genotyping.** Nature Reviews Genetics, [s.l.], no 5, 2011. DOI: https://doi.org/10.1038/nrg2958.

[39] FRIEDBERG, E. C. et al. DNA Repair and Mutagenesis. **DNA Repair and Mutagenesis**, 22 nov. 2005.

[40] KIMURA, M. **The Neutral Theory of Molecular Evolution**. Scientific American, [s.l.], no 5, 1979. DOI: https://doi.org/10.1038/scientificamerican1179-98.

[41] ASHLEY, E. A. **Towards precision medicine**. Nature Reviews Genetics, [s.l.], no 9, 2016. DOI: https://doi.org/10.1038/nrg.2016.86.

[42] STENSON, P. D. et al. **The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics,**

diagnostic testing and personalized genomic medicine. Human Genetics, [s.l.], no 1, 2013. DOI: https://doi.org/10.1007/s00439-013-1358-4.

[43] STEFL, S. et al. **Molecular Mechanisms of Disease-Causing Missense Mutations**. Journal of Molecular Biology, [s.l.], no 21, 2013. DOI: https://doi.org/10.1016/j.jmb.2013.07.014.

[44] SAHNI, N. et al. **Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders**. Cell, [s.l.], no 3, 2015. DOI: https://doi.org/10.1016/j.cell.2015.04.013.

[45] YUE, P.; LI, Z.; MOULT, J. **Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease**. Journal of Molecular Biology, [s.l.], no 2, 2005. DOI: https://doi.org/10.1016/j.jmb.2005.08.020.

[46] YAMPOLSKY, L. Y.; STOLTZFUS, A. **The Exchangeability of Amino Acids in Proteins**. Genetics, [s.l.], no 4, 2005. DOI: https://doi.org/10.1534/genetics.104.039107.

[47] CHOI, Y. et al. **Predicting the Functional Effect of Amino Acid Substitutions and Indels**. *PLoS ONE*, [s.l.], nº 10, 2012. DOI: https://doi.org/10.1371/journal.pone.0046688.

[48] ADZHUBEI, I. A. et al. **A method and server for predicting damaging missense mutations**. *Nature Methods*, [s.l.], nº4, 2010. DOI: https://doi.org/10.1038/nmeth0410-248.

[49] RENTZSCH, P. et al. **CADD: predicting the deleteriousness of variants throughout the human genome**. *Nucleic Acids Research*, [s.l.], nº D1, 2018. DOI: https://doi.org/10.1093/nar/gky1016.

[50] THUSBERG, J.; OLATUBOSUN, A.; VIHINEN, M. **Performance of mutation pathogenicity prediction methods on missense variants**. *Human Mutation*, [s.l.], nº 4, 2011. DOI: https://doi.org/10.1002/humu.21445.

[51] REES, D. C.; WILLIAMS, T. N.; GLADWIN, M. T. **Sickle-cell disease**. *The Lancet*, [s.l.], nº 9757, 2010. DOI: https://doi.org/10.1016/s0140-6736(10)61029-x.

[52] RIORDAN et al. **Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA**. *Science*, [s.l.], nº 4922, 1989. DOI: https://doi.org/10.1126/science.2475911.

[53] CARVALHO, M. A. et al. **Determination of Cancer Risk Associated with Germ Line BRCA1 Missense Variants by Functional Analysis**. [s.l.], nº 4, 2007. DOI: https://doi.org/10.1158/0008-5472.can-06-3297.

[54] THE 1000 GENOMES PROJECT CONSORTIUM. **A Global Reference for Human Genetic Variation**. *Nature*, [s.l.], nº 7571, 2015. DOI: https://doi.org/10.1038/nature15393.

[55] LEK, M. et al. **Analysis of protein-coding genetic variation in 60,706 humans**. *Nature*, [s.l.], nº 7616, 2016. DOI: https://doi.org/10.1038/nature19057.

[56] ZATLOUKAL, K.; HAINAUT, P. **Human tissue biobanks as instruments for drug discovery and development: impact on personalized medicine**. *Biomarkers in Medicine*, [s.l.], nº 6, 2010. DOI: https://doi.org/10.2217/bmm.10.104.

[57] HANSSON, M. G. **Ethics and biobanks**. *British Journal of Cancer*, [s.l.], nº 1, 2008. DOI: https://doi.org/10.1038/sj.bjc.6604795.

[58] GOTTWEIS, H.; ZATLOUKAL, K. **Biobank Governance: Trends and Perspectives**. *Pathobiology*, [s.l.], nº 4, 2007. DOI: https://doi.org/10.1159/000104446.

[59] SUDLOW, C. et al. **UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age**. *PLOS Medicine*, [s.l.], nº 3, 2015. DOI: https://doi.org/10.1371/journal.pmed.1001779.

[60] **The "All of Us" Research Program**. *New England Journal of Medicine*, [s.l.], nº 7, 2019. DOI: https://doi.org/10.1056/nejmsr1809937.

[61] CHEN, Z. et al. **China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up**. *International Journal of Epidemiology*, [s.l.], nº 6, 2011. DOI: https://doi.org/10.1093/ije/dyr120.

[62] SCHATZ, M. C. **Biological data sciences in genome research**. *Genome Research*, [s.l.], nº 10, 2015. DOI: https://doi.org/10.1101/gr.191684.115.

[63] CHEN, S. et al. **A genomic mutational constraint map using variation in 76,156 human genomes**. *Nature*, [s.l.], nº7993, 2024. DOI: https://doi.org/10.1038/s41586-023-06045-0.

[64] WANG, Q. et al. **Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes**. *Nature Communications*, [s.l.], nº 1, 2020. DOI: https://doi.org/10.1038/s41467-019-12438-5.

[65] EVANGELOU, E.; IOANNIDIS, J. P. A. **Meta-analysis methods for genome-wide association studies and beyond**. *Nature Reviews Genetics*, [s.l.], nº 6, 2013. DOI: https://doi.org/10.1038/nrg3472.

[66] ZHOU, W. et al. **Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies**. *Nature Genetics*, [s.l.], nº 9, 2018. DOI: https://doi.org/10.1038/s41588-018-0184-y.

[67] ZOU, J. et al. **A primer on deep learning in genomics**. *Nature Genetics*, [s.l.], nº 1, 2018. DOI: https://doi.org/10.1038/s41588-018-0295-5.

[68] ERASLAN, G. et al. **Deep learning: new computational modelling techniques for genomics**. *Nature Reviews Genetics*, [s.l.], nº 7, 2019. DOI: https://doi.org/10.1038/s41576-019-0122-6.

[69] BRANDEN, C. I.; TOOZE, J. Introduction to Protein Structure. **Introduction to Protein Structure**, 26 mar. 2012.

[70] PAULING, L.; COREY, R. B.; BRANSON, H. R. **The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain**. *Proceedings of the National Academy of Sciences*, [s.l.], nº 4, 1951. DOI: https://doi.org/10.1073/pnas.37.4.205.

[71] BARTLETT, G. J. et al. **Analysis of Catalytic Residues in Enzyme Active Sites**. *Journal of Molecular Biology*, [s.l.], nº1, 2002. DOI: https://doi.org/10.1016/s0022-2836(02)01036-7.

[72] DILL, K. A. et al. **The Protein Folding Problem**. *Annual Review of Biophysics*, [s.l.], nº 1, 2008. DOI: https://doi.org/10.1146/annurev.biophys.37.092707.153558.

[73] ANDRIY KRYSHTAFOVYCH et al. **Critical assessment of methods of protein structure prediction (CASP)—Round XV**. *Proteins: Structure, Function, and Bioinformatics*, [s.l.], nº 12, 2023. DOI: https://doi.org/10.1002/prot.26617.

[74] TUNYASUVUNAKOOL, K. et al. **Highly accurate protein structure prediction for the human proteome**. *Nature*, [s.l.], 2021. DOI: https://doi.org/10.1038/s41586-021-03828-1.

[75] **Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities**. *Information Fusion*, [s.l.], 2019a. DOI: https://doi.org/10.1016/j.inffus.2018.09.012.

[76] MACARTHUR, D. G. et al. **A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes**. *Science*, [s.l.], nº 6070, 2012. DOI: https://doi.org/10.1126/science.1215040.

[77] RIVAS, M. A. et al. **Effect of predicted protein-truncating genetic variants on the human transcriptome**. *Science*, [s.l.], nº 6235, 2015. DOI: https://doi.org/10.1126/science.1261877.

[78] GUDMUNDSSON, S. et al. **Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans**. *Nature*, [s.l.], nº 7874, 2021. DOI: https://doi.org/10.1038/s41586-021-03758-y.

[79] MONACO, A. P. et al. **Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene**. *Nature*, [s.l.], nº 6089, 1986. DOI: https://doi.org/10.1038/323646a0.

[80] HOBBS, H. H.; BROWN, M. S.; GOLDSTEIN, J. L. **Molecular genetics of the LDL receptor gene in familial hypercholesterolemia**. *Human Mutation*, [s.l.], nº 6, 1992. DOI: https://doi.org/10.1002/humu.1380010602.

[81] VEITIA, R. A. **Exploring the Molecular Etiology of Dominant-Negative Mutations**. *The Plant Cell*, [s.l.], nº 12, 2007. DOI: https://doi.org/10.1105/tpc.107.055053.

[82] MILLINGTON, G. W. M. **Mutations of theBRAFgene in human cancer, by Davieset al. (Nature2002; 417: 949-54)**. *Clinical and Experimental Dermatology*, [s.l.], nº 2, 2013. DOI: https://doi.org/10.1111/ced.12015.

[83] ROUSSEAU, F. et al. **Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia**. *Nature*, [s.l.], nº 6494, 1994. DOI: https://doi.org/10.1038/371252a0.

[84] YANG, Y. **Mutations in SCN9A, encoding a sodium channel alpha subunit, in patients with primary erythermalgia**. *Journal of Medical Genetics*, [s.l.], nº 3, 2004. DOI: https://doi.org/10.1136/jmg.2003.012153.

[85] KATSONIS, P.; LICHTARGE, O. **A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness**. *Genome Research*, [s.l.], nº 12, 2014. DOI: https://doi.org/10.1101/gr.176214.114.

[86] FINDLAY, G. M. et al. **Accurate classification of BRCA1 variants with saturation genome editing**. *Nature*, [s.l.], nº7726, 2018. DOI: https://doi.org/10.1038/s41586-018-0461-z.

[87] MELNIKOV, A. et al. **Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay**. *Nature Biotechnology*, [s.l.], nº 3, 2012. DOI: https://doi.org/10.1038/nbt.2137.

[88] FOWLER, D. M.; FIELDS, S. **Deep mutational scanning: a new style of protein science**. *Nature Methods*, [s.l.], nº 8, 2014. DOI: https://doi.org/10.1038/nmeth.3027.

[89] LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep Learning**. *Nature*, [s.l.], nº 7553, 2015. DOI: https://doi.org/10.1038/nature14539.

[90] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. [s.l.]: MIT Press, 2016.

[91] BENGIO, Y.; COURVILLE, A.; VINCENT, P. **Representation Learning: A Review and New Perspectives**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [s.l.], nº 8, 2013. DOI: https://doi.org/10.1109/tpami.2013.50.

[92] RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. **Learning representations by back-propagating errors**. *Nature*, [s.l.], nº 6088, 1986. DOI: https://doi.org/10.1038/323533a0.

[93] BOTTOU, L. Large-Scale Machine Learning with Stochastic Gradient Descent. LECHEVALLIER, Y.; SAPORTA, G. (Orgs.). In: COMPSTAT. [s.l.]: Physica-Verlag, 2010. Disponível em: <http://dblp.uni-trier.de/db/conf/compstat/compstat2010.html#Bottou10>. ISBN: 978-3-7908-2604-3.

[94] SENIOR, A. W. et al. **Improved Protein Structure Prediction Using Potentials from Deep Learning**. *Nature*, [s.l.], nº7792, 2020. DOI: https://doi.org/10.1038/s41586-019-1923-7.

[95] PASCAL NOTIN et al. **Machine learning for functional protein design**. *Nature biotechnology*, [s.l.], nº 2, 2024. DOI: https://doi.org/10.1038/s41587-024-02127-0.

[96] SUNDARAM, L. et al. **Predicting the clinical impact of human mutation with deep neural networks**. *Nature Genetics*, [s.l.], nº 8, 2018. DOI: https://doi.org/10.1038/s41588-018-0167-z.

[97] POPLIN, R. et al. **A universal SNP and small-indel variant caller using deep neural networks**. *Nature Biotechnology*, [s.l.], nº 10, 2018. DOI: https://doi.org/10.1038/nbt.4235.

[98] ZHOU, J.; TROYANSKAYA, O. G. **Predicting effects of noncoding variants with deep learning–based sequence model**. *Nature Methods*, [s.l.], nº 10, 2015. DOI: https://doi.org/10.1038/nmeth.3547.

[99] HALEVY, A.; NORVIG, P.; PEREIRA, F. **The Unreasonable Effectiveness of Data**. *IEEE Intelligent Systems*, [s.l.], nº2, 2009. DOI: https://doi.org/10.1109/mis.2009.36.

[100] STEPHENS, Z. D. et al. **Big Data: Astronomical or Genomical?** *PLOS Biology*, [s.l.], nº 7, 2015. DOI: https://doi.org/10.1371/journal.pbio.1002195.

[101] SHORTEN, C.; KHOSHGOFTAAR, T. M. **A survey on Image Data Augmentation for Deep Learning**. *Journal of Big Data*, [s.l.], nº 1, 2019. DOI: https://doi.org/10.1186/s40537-019-0197-0.

[102] SRIVASTAVA, N. et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. **Journal of Machine Learning Research**, v. 15, n. 56, p. 1929–1958, 2014.

[103] KUKAČKA, J.; VLADIMIR GOLKOV; CREMERS, D. **Regularization for Deep Learning: A Taxonomy**. *arXiv (Cornell University)*, [s.l.], 2017. DOI: https://doi.org/10.48550/arxiv.1710.10686.

[104] **Database and Expert Systems Applications**. *Lecture notes in computer science*. [s.l.]: Springer Science+Business Media, 2013. DOI: https://doi.org/10.1007/978-3-642-40173-2.

[105] ZHUANG, F. et al. **A Comprehensive Survey on Transfer Learning**. [s.l.], 2019. DOI: https://doi.org/10.48550/arxiv.1911.02685.

[106] HOLZINGER, A. et al. **Causability and explainability of artificial intelligence in medicine**. *WIREs Data Mining and Knowledge Discovery*, [s.l.], nº 4, 2019. DOI: https://doi.org/10.1002/widm.1312.

[107] SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. **Deep Inside Convolutional Networks: Visualising Image Classification Models and**

Saliency Maps. *arXiv.org*. 2014. Disponível em: <https://arxiv.org/abs/1312.6034v2>. DOI: https://doi.org/10.48550/arXiv.1312.6034.

[108] RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"Why should I trust you?"** *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, [s.l.], 2016. ISBN: 9781450342322, DOI: https://doi.org/10.1145/2939672.2939778.

[109] JIMÉNEZ-LUNA, J.; GRISONI, F.; SCHNEIDER, G. **Drug discovery with explainable artificial intelligence**. *Nature Machine Intelligence*, [s.l.], nº 10, 2020. DOI: https://doi.org/10.1038/s42256-020-00236-4.

[110] CHING, T. et al. **Opportunities and obstacles for deep learning in biology and medicine**. *Journal of The Royal Society Interface*, [s.l.], nº 141, 2018. DOI: https://doi.org/10.1098/rsif.2017.0387.

[111] LECUN, Y. et al. **Gradient-based learning applied to document recognition**. *Proceedings of the IEEE*, [s.l.], nº 11, 1998. DOI: https://doi.org/10.1109/5.726791.

[112] REN, S. et al. **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [s.l.], nº 6, 2017. DOI: https://doi.org/10.1109/tpami.2016.2577031.

[113] LONG, J.; SHELHAMER, E.; DARRELL, T. **Fully convolutional networks for semantic segmentation**. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, [s.l.], 2015. ISBN: 9781467369640, DOI: https://doi.org/10.1109/cvpr.2015.7298965.

[114] BENGIO, Y.; SIMARD, P.; FRASCONI, P. **Learning long-term dependencies with gradient descent is difficult**. *IEEE Transactions on Neural Networks*, [s.l.], nº 2, 1994. DOI: https://doi.org/10.1109/72.279181.

[115] HOCHREITER, S.; SCHMIDHUBER, J. **Long Short-Term Memory**. *Neural Computation*, [s.l.], nº 8, 1997. DOI: https://doi.org/10.1162/neco.1997.9.8.1735.

[116] CHUNG, J. et al. **Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling**. *arXiv.org*. 2014. Disponível em: <http://arxiv.org/abs/1412.3555>. DOI: https://doi.org/10.48550/arXiv.1412.3555.

[117] PENDRY, J. B. et al. References and Notes Supporting Online Material Reducing the Dimensionality of Data with Neural Networks. **IEEE Trans. Microw. Theory Tech**, v. 47, n. 9, p. 653, 1999.

[118] VINCENT, P. et al. **Extracting and composing robust features with denoising autoencoders**. *Proceedings of the 25th international conference on Machine learning - ICML '08*, [s.l.], 2008. DOI: https://doi.org/10.1145/1390156.1390294.

[119] KINGMA, D. P.; WELLING, M. Auto-Encoding Variational Bayes. **2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings**, 20 dez. 2013.

[120] SCARSELLI, F. et al. **The Graph Neural Network Model**. *IEEE Transactions on Neural Networks*, [s.l.], nº 1, 2009. DOI: https://doi.org/10.1109/tnn.2008.2005605.

[121] KIPF, T. N.; WELLING, M. **Semi-Supervised Classification with Graph Convolutional Networks**. *arXiv (Cornell University)*, [s.l.], 2016. DOI: https://doi.org/10.48550/arxiv.1609.02907.

[122] PETAR VELIČKOVIĆ et al. **Graph Attention Networks**. *arXiv (Cornell University)*, [s.l.], 2017. DOI: https://doi.org/10.48550/arxiv.1710.10903.

[123] ALIPANAHI, B. et al. **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning**. *Nature Biotechnology*, [s.l.], nº 8, 2015. DOI: https://doi.org/10.1038/nbt.3300.

[124] WANG, S. et al. **Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields**. *Scientific Reports*, [s.l.], nº 1, 2016. DOI: https://doi.org/10.1038/srep18962.

[125] ZITNIK, M.; AGRAWAL, M.; LESKOVEC, J. **Modeling polypharmacy side effects with graph convolutional networks**. *Bioinformatics*, [s.l.], nº 13, 2018. DOI: https://doi.org/10.1093/bioinformatics/bty294.

[126] GAO, W. et al. **Deep Learning in Protein Structural Modeling and Design**. *Patterns*, [s.l.], nº 9, 2020. DOI: https://doi.org/10.1016/j.patter.2020.100142.

[127] BROWN, T. et al. **Language Models are Few-Shot Learners**. *arXiv (Cornell University)*, [s.l.], 2020. DOI: https://doi.org/10.48550/arxiv.2005.14165.

[128] RASUL, K. et al. **Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting**. *arXiv.org*. 2024. Disponível em: <https://arxiv.org/abs/2310.08278>. Acesso em: 15/abr./24. DOI: https://doi.org/10.48550/arXiv.2310.08278.

[129] ABRAMSON, J. et al. **Accurate structure prediction of biomolecular interactions with AlphaFold 3**. *Nature*, [s.l.], 2024. DOI: https://doi.org/10.1038/s41586-024-07487-w.

[130] MEIER, J. et al. **Language models enable zero-shot prediction of the effects of mutations on protein function**. [s.l.], 2021. DOI: https://doi.org/10.1101/2021.07.09.450648.

[131] CHENG, J. et al. **Accurate proteome-wide missense variant effect prediction with AlphaMissense**. *Science*, [s.l.], nº6664, 2023. DOI: https://doi.org/10.1126/science.adg7492.

[132] LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. **Convolutional networks and applications in vision**. *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, [s.l.], 2010. ISBN: 9781424453085, DOI: https://doi.org/10.1109/iscas.2010.5537907.

[133] ZEILER, M. D.; FERGUS, R. **Visualizing and Understanding Convolutional Networks**. [s.l.], 2013. DOI: https://doi.org/10.48550/arxiv.1311.2901.

[134] NG, P. **dna2vec: Consistent vector representations of variable-length k-mers**. *arXiv (Cornell University)*, [s.l.], 2017. DOI: https://doi.org/10.48550/arxiv.1701.06279.

[135] ASGARI, E.; MOFRAD, M. R. K. **Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics**. *PLOS ONE*, [s.l.], nº 11, 2015. DOI: https://doi.org/10.1371/journal.pone.0141287.

[136] MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. **Methods for interpreting and understanding deep neural networks**. *Digital Signal Processing*, [s.l.], 2018. DOI: https://doi.org/10.1016/j.dsp.2017.10.011.

[137] AGGARWAL, C. C. Neural Networks and Deep Learning: A Textbook. **Neural Networks and Deep Learning: A Textbook**, p. 1–529, 1 jan. 2023.

[138] PAN, S. J.; YANG, Q. **A Survey on Transfer Learning**. *IEEE Transactions on Knowledge and Data Engineering*, [s.l.], nº 10, 2010. DOI: https://doi.org/10.1109/tkde.2009.191.

[139] ÖZTÜRK, H.; ÖZGÜR, A.; OZKIRIMLI, E. **DeepDTA: deep drug–target binding affinity prediction**. *Bioinformatics*, [s.l.], nº 17, 2018. DOI: https://doi.org/10.1093/bioinformatics/bty593.

[140] KEREPESI, C. et al. **Prediction and characterization of human ageing-related proteins by using machine learning**. *Scientific Reports*, [s.l.], nº 1, 2018. DOI: https://doi.org/10.1038/s41598-018-22240-w.

[141] LEE, J. et al. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**. *Bioinformatics*, [s.l.], nº 4, 2019. DOI: https://doi.org/10.1093/bioinformatics/btz682.

[142] DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. *Proceedings of the 2019 Conference of the North*, [s.l.], 2019. DOI: https://doi.org/10.18653/v1/n19-1423.

[143] QIU, X. et al. **Pre-trained models for natural language processing: A survey**. *Science China Technological Sciences*, [s.l.], nº 10, 2020. DOI: https://doi.org/10.1007/s11431-020-1647-3.

[144] ANTEGHINI, M.; SANTOS; EDOARDO SACCENTI. **P-PPI: accurate prediction of peroxisomal protein-protein interactions (P-PPI) using deep learning-based protein sequence embeddings**. *bioRxiv (Cold Spring Harbor Laboratory)*, [s.l.], 2023. DOI: https://doi.org/10.1101/2023.06.30.547177.

[145] JI, Y. et al. **DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome**. *Bioinformatics*, [s.l.], 2021. DOI: https://doi.org/10.1093/bioinformatics/btab083.

[146] AVSEC, Ž. et al. **Effective gene expression prediction from sequence by integrating long-range interactions**. *Nature Methods*, [s.l.], nº 10, 2021. DOI: https://doi.org/10.1038/s41592-021-01252-x.

[147] RISHI BOMMASANI et al. **On the Opportunities and Risks of Foundation Models**. *arXiv (Cornell University)*, [s.l.], 2021. DOI: https://doi.org/10.48550/arxiv.2108.07258.

[148] YOSINSKI, J. et al. **How transferable are features in deep neural networks?** [s.l.], 2014. DOI: https://doi.org/10.48550/arxiv.1411.1792.

[149] HEINZINGER, M. et al. **Modeling aspects of the language of life through transfer-learning protein sequences**. *BMC Bioinformatics*, [s.l.], nº 1, 2019. DOI: https://doi.org/10.1186/s12859-019-3220-8.

[150] AX, L. et al. **Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization**. *bioRxiv (Cold Spring Harbor Laboratory)*, [s.l.], 2020. DOI: https://doi.org/10.1101/2020.09.04.283929.

[151] RAO, R. et al. **Transformer protein language models are unsupervised structure learners**. [s.l.], 2020. DOI: https://doi.org/10.1101/2020.12.15.422761.

[152] LARRAÑAGA, P. et al. **Machine learning in bioinformatics**. *Briefings in Bioinformatics*, [s.l.], nº 1, 2006. DOI: https://doi.org/10.1093/bib/bbk007.

[153] LIBBRECHT, M. W.; NOBLE, W. S. **Machine learning applications in genetics and genomics**. *Nature Reviews Genetics*, [s.l.], nº 6, 2015. DOI: https://doi.org/10.1038/nrg3920.

[154] TARCA, A. L. et al. **Machine Learning and Its Applications to Biology**. *PLoS Computational Biology*, [s.l.], nº 6, 2007. DOI: https://doi.org/10.1371/journal.pcbi.0030116.

[155] VAPNIK, V. N. **An overview of statistical learning theory**. *IEEE Transactions on Neural Networks*, [s.l.], nº 5, 1999. DOI: https://doi.org/10.1109/72.788640.

[156] NOBLE, W. S. **What is a support vector machine?** *Nature Biotechnology*, [s.l.], nº 12, 2006. DOI: https://doi.org/10.1038/nbt1206-1565.

[157] SHEN, J. et al. **Predicting protein-protein interactions based only on sequences information**. *Proceedings of the National Academy of Sciences of the United States of America*, [s.l.], nº 11, 2007. DOI: https://doi.org/10.1073/pnas.0607879104.

[158] DING, C. H. Q.; DUBCHAK, I. **Multi-class protein fold recognition using support vector machines and neural networks**. *Bioinformatics*, [s.l.], nº 4, 2001. DOI: https://doi.org/10.1093/bioinformatics/17.4.349.

[159] BROWN, M. P. S. et al. **Knowledge-based analysis of microarray gene expression data by using support vector machines**. *Proceedings of the*

National Academy of Sciences, [s.l.], nº 1, 2000. DOI: https://doi.org/10.1073/pnas.97.1.262.

[160] BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. **A training algorithm for optimal margin classifiers**. *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, [s.l.], 1992. ISBN: 089791497X, DOI: https://doi.org/10.1145/130385.130401.

[161] BEN-HUR, A. et al. **Support Vector Machines and Kernels for Computational Biology**. *PLoS Computational Biology*, [s.l.], nº 10, 2008. DOI: https://doi.org/10.1371/journal.pcbi.1000173.

[162] BREIMAN, L. **Random Forests**. *Machine Learning*, [s.l.], nº 1, 2001. DOI: https://doi.org/10.1023/a:1010933404324.

[163] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. New York: Springer, 2004. ISBN: 9780387952840.

[164] BORISOV, V. et al. **Deep Neural Networks and Tabular Data: A Survey**. *arXiv (Cornell University)*, [s.l.], 2021. DOI: https://doi.org/10.48550/arxiv.2110.01889.

[165] BAI, S.; DU, T.; KHOSRAVI, E. **Applying internal coordinate mechanics to model the interactions between 8R-lipoxygenase and its substrate**. *BMC bioinformatics*, [s.l.], nº S6, 2010. DOI: https://doi.org/10.1186/1471-2105-11-s6-s2.

[166] CHOPRA, P. et al. **Microarray data mining using landmark gene-guided clustering**. *BMC Bioinformatics*, [s.l.], nº 1, 2008. DOI: https://doi.org/10.1186/1471-2105-9-92.

[167] DÍAZ-URIARTE, R.; ALVAREZ DE ANDRÉS, S. **Gene selection and classification of microarray data using random forest**. *BMC Bioinformatics*, [s.l.], nº 1, 2006. DOI: https://doi.org/10.1186/1471-2105-7-3.

[168] HE, Z. et al. **Gradient Boosting Machine: A Survey**. *arxiv.org*, [s.l.], 2019. DOI: https://doi.org/10.48550/arXiv.1908.06951.

[169] MACHADO, M. R.; KARRAY, S.; SOUSA, I. T. De. **LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry**. *IEEE Xplore*. 2019. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8845529?casa_token=-enj6arGcq

8AAAAA:TWOf-ILDSHvvvBTrKBX7eqVL9ZMcUosr_UowqTflWhjx1VmZ8 8a1CZqGs5YMs4PcdvliZFiPVQ>. DOI: https://doi.org/10.1109/ICCSE.2019.8845529.

[170] XIA, Z. et al. **Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces**. *BMC Systems Biology*, [s.l.], nº S2, 2010. DOI: https://doi.org/10.1186/1752-0509-4-s2-s6.

[171] THEDINGA, K.; HERWIG, R. **A gradient tree boosting and network propagation derived pan-cancer survival network of the tumor microenvironment**. *iScience*, [s.l.], 2021. DOI: https://doi.org/10.1016/j.isci.2021.103617.

[172] PARK, A.; LEE, Y.; NAM, S. **A performance evaluation of drug response prediction models for individual drugs**. *Scientific Reports*, [s.l.], nº 1, 2023. DOI: https://doi.org/10.1038/s41598-023-39179-2.

[173] ALTMAN, N. S. **An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression**. *The American Statistician*, [s.l.], nº 3, 1992. DOI: https://doi.org/10.1080/00031305.1992.10475879.

[174] Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3.

[175] KROGH, A. et al. **Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen**. *Journal of Molecular Biology*, [s.l.], nº 3, 2001. DOI: https://doi.org/10.1006/jmbi.2000.4315.

[176] MA, J. et al. **Using deep learning to model the hierarchical structure and function of a cell**. *Nature Methods*, [s.l.], nº4, 2018. DOI: https://doi.org/10.1038/nmeth.4627.

[177] CHENNA, R. **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Research*, [s.l.], nº 13, 2003. DOI: https://doi.org/10.1093/nar/gkg500.

[178] FENG, D.-F.; DOOLITTLE, R. F. **Progressive sequence alignment as a prerequisitetto correct phylogenetic trees**. *Journal of Molecular Evolution*, [s.l.], nº 4, 1987. DOI: https://doi.org/10.1007/bf02603120.

[179] CHATZOU, M. et al. **Multiple sequence alignment modeling: methods and applications**. *Briefings in Bioinformatics*, [s.l.], nº 6, 2015. DOI: https://doi.org/10.1093/bib/bbv099.

[180] CHOWDHURY, B.; GARAI, G. **A review on multiple sequence alignment from the perspective of genetic algorithm**. *Genomics*, [s.l.], nº 5-6, 2017. DOI: https://doi.org/10.1016/j.ygeno.2017.06.007.

[181] STEINEGGER, M.; SÖDING, J. **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**. *Nature Biotechnology*, [s.l.], nº 11, 2017. DOI: https://doi.org/10.1038/nbt.3988.

[182] SUZEK, B. E. et al. **UniRef: comprehensive and non-redundant UniProt reference clusters**. *Bioinformatics*, [s.l.], nº10, 2007. DOI: https://doi.org/10.1093/bioinformatics/btm098.

[183] LITTMANN, M. et al. **Embeddings from deep learning transfer GO annotations beyond homology**. *Scientific Reports*, [s.l.], nº 1, 2021. DOI: https://doi.org/10.1038/s41598-020-80786-0.

[184] MADANI, A. et al. **ProGen: Language Modeling for Protein Generation**. *arXiv (Cornell University)*, [s.l.], 2020. DOI: https://doi.org/10.48550/arxiv.2004.03497.

[185] STEIN, D. et al. **Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set**. *Genome Medicine*, [s.l.], nº 1, 2023. DOI: https://doi.org/10.1186/s13073-023-01261-9.

[186] SEVIM BAYRAK, C. et al. **Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants**. *The American Journal of Human Genetics*, [s.l.], nº 12, 2021. DOI: https://doi.org/10.1016/j.ajhg.2021.10.007.

[187] LIN, W. et al. VariPred: Enhancing Pathogenicity Prediction of Missense Variants Using Protein Language Models. bioRxiv (Cold Spring Harbor Laboratory), [s.l.], 2023. DOI: https://doi.org/10.1101/2023.03.16.532942.

[188] LIN, Z. et al. **Evolutionary-scale prediction of atomic level protein structure with a language model**. [s.l.], 2022a. DOI: https://doi.org/10.1101/2022.07.20.500902.

[189] PASZKE, A. et al. **PyTorch: An Imperative Style, High-Performance Deep Learning Library**. *arXiv (Cornell University)*, [s.l.], 2019. DOI: https://doi.org/10.48550/arxiv.1912.01703.

[190] PEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python**. *arXiv.org*. 2018. Disponível em: <http://arxiv.org/abs/1201.0490>. DOI: https://doi.org/10.48550/arXiv.1201.0490.

[191] BERGSTRA, J. et al. **Hyperopt: a Python library for model selection and hyperparameter optimization**. *Computational Science & Discovery*, [s.l.], nº 1, 2015. DOI: https://doi.org/10.1088/1749-4699/8/1/014008.

[192] LANDRUM, M. J. et al. **ClinVar: public archive of relationships among sequence variation and human phenotype**. *Nucleic Acids Research*, [s.l.], nº Database issue, 2014. DOI: https://doi.org/10.1093/nar/gkt1113.

[193] FRANK, E.; HALL, M. **A Simple Approach to Ordinal Classification**. *Machine Learning: ECML 2001*, [s.l.], 2001. ISBN: 9783540425366, DOI: https://doi.org/10.1007/3-540-44795-4_13.

[194] WOLF, T. et al. **HuggingFace's Transformers: State-of-the-art Natural Language Processing**. *arXiv.org*. 2020. Disponível em: <https://arxiv.org/abs/1910.03771v5>. Acesso em: 29/maio/23. DOI: https://doi.org/10.48550/arXiv.1910.03771.

[195] REIMERS, N.; IRYNA GUREVYCH. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. [s.l.], 2019. DOI: https://doi.org/10.48550/arxiv.1908.10084.

[196] HOWARD, J.; RUDER, S. **Universal Language Model Fine-tuning for Text Classification**. *arXiv (Cornell University)*, [s.l.], 2018. DOI: https://doi.org/10.48550/arxiv.1801.06146.

[197] CHEN, T.; GUESTRIN, C. **XGBoost: a Scalable Tree Boosting System**. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, [s.l.], 2016. ISBN: 9781450342322, DOI: https://doi.org/10.1145/2939672.2939785.

[198] TENG, S. et al. **Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein protein interaction**. *International Journal of Computational Biology and Drug Design*, [s.l.], nº 4, 2010. DOI: https://doi.org/10.1504/ijcbdd.2010.038396.

[199] **Ch8 Chi-Square, Pt 1**. Available at: <https://web.archive.org/web/20171022032306/http://vassarstats.net:80/textbook/ch8pt1.html>. Accessed at: 27 abr. 2024.

[200] MIKOLOV, T. et al. **Efficient Estimation of Word Representations in Vector Space**. *arXiv.org*. 2013. Disponível em: <http://arxiv.org/abs/1301.3781>. DOI: https://doi.org/10.48550/arXiv.1301.3781.

[201] GERASIMAVICIUS, L.; LIVESEY, B. J.; MARSH, J. A. **Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure**. *Nature Communications*, [s.l.], nº 1, 2022. DOI: https://doi.org/10.1038/s41467-022-31686-6.

[202] SÉGALAT, L. **Loss-of-function genetic diseases and the concept of pharmaceutical targets**. *Orphanet Journal of Rare Diseases*, [s.l.], nº 1, 2007. DOI: https://doi.org/10.1186/1750-1172-2-30.

[203] GRAUR, D. **Amino acid composition and the evolutionary rates of protein-coding genes**. *Journal of Molecular Evolution*, [s.l.], nº 1, 1985. DOI: https://doi.org/10.1007/bf02105805.

[204] GRANTHAM, R. **Amino acid difference formula to help explain protein evolution**. *Science (New York, N.Y.)*, United States, nº 4154, 1974. DOI: https://doi.org/10.1126/science.185.4154.862.

[205] CLARKE, B. **Selective Constraints on Amino-acid Substitutions during the Evolution of Proteins**. *Nature*, [s.l.], nº5267, 1970. DOI: https://doi.org/10.1038/228159a0.

[206] FONTANA, F.; SIVA, K.; DENTI, M. A. **A network of RNA and protein interactions in Fronto Temporal Dementia**. *Frontiers in Molecular Neuroscience*, [s.l.], 2015. DOI: https://doi.org/10.3389/fnmol.2015.00009.

[207] PARVIZI, J.; KIM, G. K. Marfan Syndrome. **High Yield Orthopaedics**, p. 287–288, 1 jan. 2010.

[208] CHARBONNEAU, N. L. et al. **In VivoStudies of Mutant Fibrillin-1 Microfibrils**. *Journal of Biological Chemistry*, [s.l.], nº 32, 2010. DOI: https://doi.org/10.1074/jbc.m110.130021.

[209] MARTÍNEZ-QUINTANA, E. et al. **A Novel Fibrillin 1 Gene Mutation Leading to Marfan Syndrome with Minimal Cardiac Features**. *Molecular Syndromology*, [s.l.], 2014. DOI: https://doi.org/10.1159/000358846.

[210] ASGARI, S. et al. **A positively selected FBN1 missense variant reduces height in Peruvian individuals**. *Nature*, [s.l.], nº 7811, 2020. DOI: https://doi.org/10.1038/s41586-020-2302-0.

[211] JACQUINET, A. et al. **Neonatal progeroid variant of Marfan syndrome with congenital lipodystrophy results from mutations at the 3' end of FBN1 gene**. *European Journal of Medical Genetics*, [s.l.], nº 5, 2014. DOI: https://doi.org/10.1016/j.ejmg.2014.02.012.

[212] LEVY, M.; JACOBY, A.; GOLDBERG, Y. **Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models**. *arXiv (Cornell University)*, [s.l.], 2024. DOI: https://doi.org/10.48550/arxiv.2402.14848.

[213] ELEFTHERIA BRIAKOU; CHERRY, C.; FOSTER, G. **Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability**. *arXiv (Cornell University)*, [s.l.], 2023. DOI: https://doi.org/10.18653/v1/2023.acl-long.524.

[214] FRAZER, J. et al. **Disease variant prediction with deep generative models of evolutionary data**. *Nature*, [s.l.], nº7883, 2021. DOI: https://doi.org/10.1038/s41586-021-04043-8.

[215] SUNDARAM, L. et al. Predicting the clinical impact of human mutation with deep neural networks. **Nature genetics**, v. 50, n. 8, p. 1161, 1 ago. 2018.

[216] CHOI, Y.; CHAN, A. P. **PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels**. *Bioinformatics*, [s.l.], nº 16, 2015. DOI: https://doi.org/10.1093/bioinformatics/btv195.

[217] LI, C. et al. **MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning**. [s.l.], nº 1, 2022. DOI: https://doi.org/10.1186/s13073-022-01120-z.

[218] KIM, S. et al. **Meta-analytic support vector machine for integrating multiple omics data**. *BioData Mining*, [s.l.], 2017. DOI: https://doi.org/10.1186/s13040-017-0126-8.

[219] IOANNIDIS, N. M. et al. **REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants**. *American Journal of Human Genetics*, [s.l.], nº 4, 2016. DOI: https://doi.org/10.1016/j.ajhg.2016.08.016.

[220] FENG, B.-J. **PERCH: A Unified Framework for Disease Gene Prioritization**. *Human Mutation*, [s.l.], nº 3, 2017. DOI: https://doi.org/10.1002/humu.23158.

[221] RICHARDS, S. et al. **Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology**. *Genetics in medicine : official journal of the American College of Medical Genetics*, [s.l.], nº 5, 2015. DOI: https://doi.org/10.1038/gim.2015.30.