GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Medicine and Life Sciences

Article

EGNN-CMutPred: Predicting Protein Mutational Effects by Integrating Primary and Tertiary Protein Structures

Yuchen Wang [1,*]

[1] Saint Mark's School, Massachusetts, United States
* Correspondence: Yuchen Wang, Saint Mark's School, Massachusetts, United States

Abstract: The study proposed the EGNN-CMutPred, Equivariant Graph Neural Network based Comprehensive Protein Mutational Effect Predictor, a novel approach for predicting the effects of protein mutations by integrating both primary and tertiary protein structures. By combining Evolutionary Scale Modeling 2 (ESM-2) for semantic embedding with Equivariant Graph Neural Network (EGNN) for structural encoding, the model improves its accuracy in predicting how mutations impact protein function and stability. The study aims to address the limitations of traditional sequence and structure-based prediction methods by incorporating both semantic and topological embeddings of proteins, allowing the model to capture a comprehensive understanding of each protein. EGNN-CMutPred was trained on non-redundant protein sequences from the CATH v4.3.0 database and evaluated against benchmarks, including ProteinGym (DMS) and the ProThermDB database, which measure changes in melting temperature (△Tm) and change in the variation of Gibbs free energy (△△G). The model effectively simulates mutations and predicts changes in protein stability, as demonstrated by strong performance in metrics such as Spearman's Correlation and True Positive Rate. These results suggest that EGNN-CMutPred is a valuable tool for precision medicine and protein engineering, offering enhanced prediction capabilities over existing methods. Future research will refine the model's computational techniques and expand its applicability to larger, more diverse datasets, furthering its potential in understanding protein mutations and their implications for disease and therapeutic development.

Keywords: protein mutational effect prediction; EGNN; ESM-2; primary and tertiary protein structures; semantic embedding; structural encoding

1. Introduction

Proteins are responsible for providing structural support, catalyzing biochemical reactions as enzymes, and participating in signal transduction pathways. However, mutations in protein-coding genes can lead to the production of dysfunctional proteins, which may have severe consequences for cellular function and organism health. Figure 1 illustrates how a normal gene differs from a mutant gene, where the former generates a working protein, while the latter may produce either no protein or a dysfunctional one. Therefore, predicting the effects of genetic mutations on protein function and stability is of paramount importance for advancing our understanding of disease mechanisms and facilitating the development of precision medicine and protein engineering.

Received: 03 July 2025
Revised: 15 July 2025
Accepted: 05 August 2025
Published: 09 August 2025

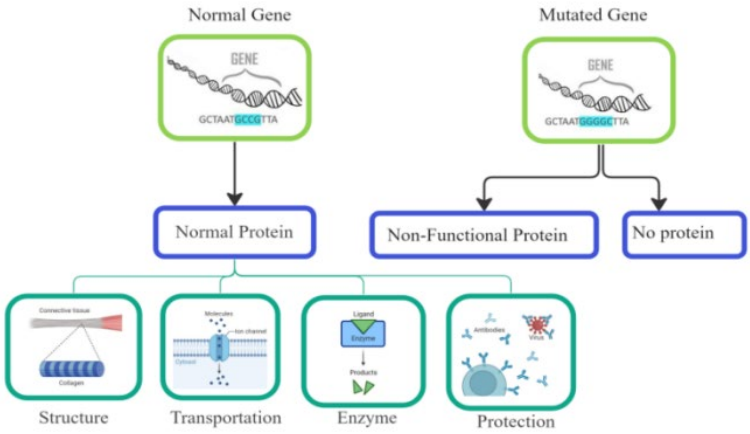

Figure 1. Comparison between a normal and mutated gene [1,2].

Initially, the effects of protein mutations were primarily studied through experimental methods such as microarray analysis and affinity purification mass spectrometry (AP-MS) [3,4]. However, these traditional approaches are often time-consuming and labor-intensive. The advent of bioinformatics and next-generation sequencing technologies has led to an explosion of available protein sequence data, with databases like UniProt now containing over 60 million protein sequences [5]. Concurrently, the Protein Data Bank (PDB) has amassed over 180,000 protein structures [6]. These vast repositories of data have enabled the development of computational models for predicting protein function and the effects of mutations.

Existing protein function prediction models can be broadly categorized into sequence-based and structure-based approaches [7]. Sequence-based methods, such as homology-based models (e.g., BLAST) and deep learning models utilizing 1-dimensional convolutional neural networks (CNNs) or recurrent neural networks (RNNs), primarily focus on identifying evolutionary relationships and sequence patterns [8,9]. However, they often fail to capture the 3-dimensional interactions between amino acids [10,11]. On the other hand, structure-based methods, including Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and Equivariant Graph Neural Networks (EGNNs), leverage the spatial arrangements of molecular components to predict protein function [12]. While these methods provide insights into protein structure, they may overlook important sequence-based features [13].

To address the limitations of existing methods, we developed the Equivariant Graph Neural Network based Comprehensive Protein Mutational Effect Predictor, EGNN-CMutPred, that integrates two powerful approaches—Evolutionary Scale Modeling 2 (ESM-2) and Equivariant Graph Neural Network (EGNN)—to capture both the primary and tertiary structures of proteins, overcoming the limitations of using sequence or structure data alone [13,14]. ESM-2, a transformer-based language model, learns the "grammar" of amino acid sequences by identifying patterns that dictate protein function, while EGNN encodes the 3D structure of proteins, focusing on the geometric relationships between amino acids and their surrounding environment [15]. By combining these components, EGNN-CMutPred can simulate mutations and accurately evaluate the fitness, stability, and functional impact of mutated proteins.

2. Materials and Methods

2.1. Dataset

The model is trained based on the non-redundant Protein Data Bank (PDB) format files in CATH v4.3.0. Protein domains, or the portions of a protein that fold independently from the remainder of the protein, are organized systematically in the CATH database

[16]. There are 5841 superfamilies among the 151 million protein domains that are currently in the database. The database is divided into four levels (listed from top to bottom): class (C), architecture (A), topology (T), and homologous superfamily (H). The C-level categorizes the proteins based on the arrangements of their locally folded structures. The A-level groups the proteins based on the three-dimensional organization of secondary structures within a protein domain. The T-level arranges the domains that have similar topological secondary structures (folds), irrespective of evolutionary relationships. The H-level indicates an evolutionary relationship with domains that share a common ancestor [17]. Figure 2 is a depiction of the CATH Database classification.

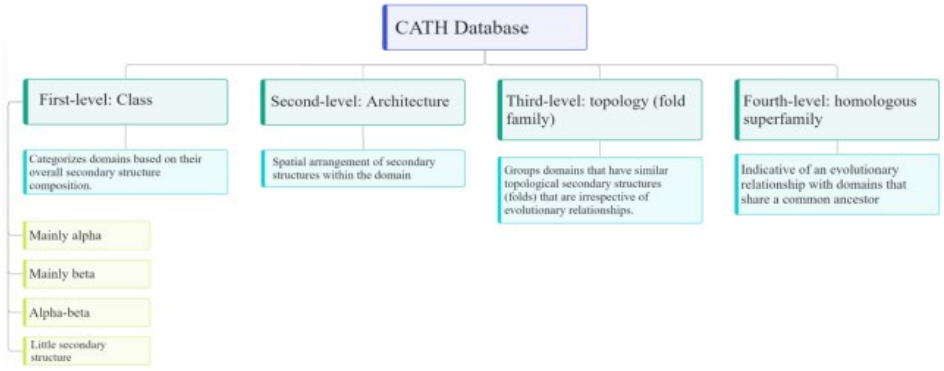

Figure 2. The classification of the CATH Database.

The PDB format is a standard format for files that contain the description and annotation of structural information of proteins in the form of atomic coordinates. This study incorporates the ATOM record that consists of x, y, and z orthogonal Å coordinates of the atoms of proteins or nucleic acids [18]. Table 1 demonstrates the information presented in the ATOM record. The most crucial information is the residue name (18–20) and 3D coordinates (31–54). Additionally, additional data also refine the model's understanding of each atom: atom name (13-16) identifies the element's chemical properties and ability to form bonds, chain and segment identifier (22) (73-76) prevent confusion between different polypeptide chains and protein chain regions, and occupancy (55-60) and temperature factor (61-66) allow the model to predict possible structural conformations. We used ProThermDB, also referred to as the Thermodynamic Database for Proteins and Mutants, to evaluate the performance of our trained model. The database contains more than 32,000 data on protein stability, and it contains data for both wild-type proteins and mutants with point mutations. The two parameters we chose were the change in melting temperature and the change in the variation of Gibbs free energy, with the datasets named $\triangle$Tm and $\triangle\triangle$G respectively [19]. The other benchmark that was used was the ProteinGym benchmark, which includes over 80 proteins of varying taxa [20]. It is used for testing deep mutational effects.

Table 1. The ATOM record of the PDB layout [14].

Columns	Data
1-4	"ATOM"
7-11	Atom serial number
13-16	Atom name
17	Alternative location indicator
18-20	Residue name
22	Chain identifier
23-26	Residue sequence number

27	Code for insertions of residues
31-38	X orthogonal Å coordinate
39-46	Y orthogonal Å coordinate
47-54	Z orthogonal Å coordinate
55-60	Occupancy
61-66	Temperature factor
73-76	Segment identifier
77-78	Element symbol
79-80	Charge

To gain insights into testing data, we conducted a preliminary analysis of the ProThermDB dataset by creating graphs that compare the original and mutated protein sequences. We made the comparisons based on several factors, including the distribution of atom and amino acid types, as well as 3-dimensional structural changes resulting from the mutations. See Appendix A for analyzed protein data files and codes for the graphs. This analysis is essential for understanding the mutant's degree of change. Figure 3 illustrates the atomic composition of three protein structures: 1w7s (the original wild-type sequence), 1w7t, and 1w7u (both are mutants of 1w7s), depicted through 3D scatter plots. Each plot represents the spatial distribution of key atoms within the proteins, with carbon atoms shown in red, oxygen in yellow, nitrogen in blue, and sulfur in green. Despite the subtle differences in atomic positioning, the scatter plots highlight the overall structural consistency between the proteins. The visualization underscores the minimal structural changes resulting from mutations, providing insight into the stability and geometric arrangement of these proteins.

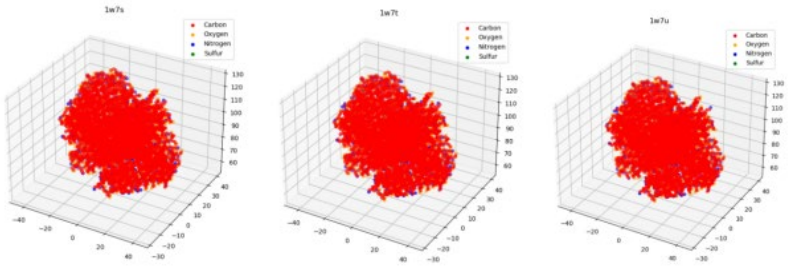

Figure 3. 3D scatter plot representation of each protein.

Even though there are no significant structural changes by appearance as displayed in the scatter plot, subtle alterations in the protein sequence are highlighted in Figures 4 and 5. Figure 4 compares the amino acid distribution of the original and mutant sequences, and Figure 5 compares the atom type distribution. This illustrates the effect of point mutations on amino acid and atom distributions. Notably, the amino acid composition of the two mutants is the same, suggesting that they differ structurally due to variations in atom positioning. Exposing the model to proteins with minor mutations allows it to gain a better comprehension of such mutations and improve the ability to predict their effects.

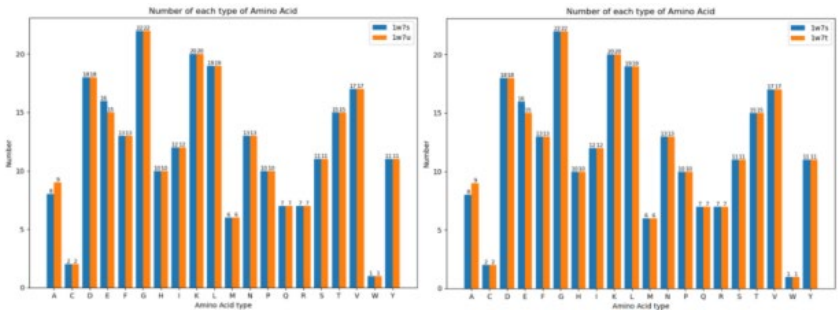

Figure 4. Bar graph presentation of amino acid distribution.

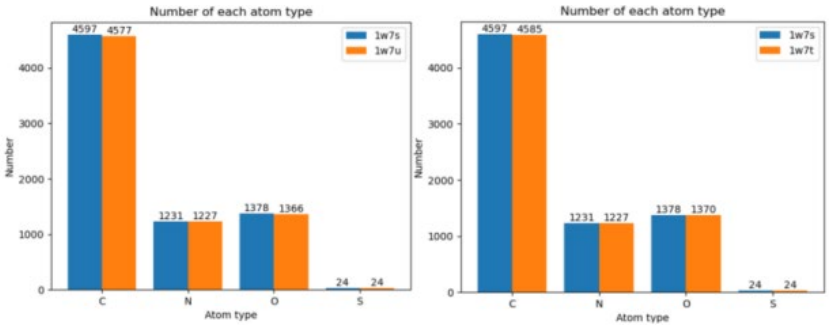

Figure 5. Bar graph presentation of atom type distribution.

2.2. Model Architecture

By synthesizing the semantic embeddings from ESM-2 and topological embeddings from EGNN, the model will generate predictions for mutational effects with higher accuracy than previous models because it will learn about the protein from both local and global perspectives. This allows the model to understand relationships between amino acids and the overall conformation patterns of the protein. The effect of mutations on protein sequence and structure is minuscule, as observed in our analysis of the ProThermDB benchmark data. Due to this reason, the model may potentially struggle to forecast the impact of mutations.

The approach that resolves this potential issue is the introduction of observation perturbations. They are applied to the primary structures when the protein sequence v is inputted during training, where noise is added to the model to prevent it from overfitting and to mimic blind mutations that occur in nature. This enables the model to detect underlying patterns in protein sequences despite the subtle effects of mutations. The perturbed protein sequences then become the input for the ESM-2 language model. The model learns the grammatical rules and patterns of the protein sequences through extracting evolutionary information. The final output is the last hidden state W_v of the protein sequence, W_v = ESM-2 ($\tilde{v}$). Next, we represent the output, along with nodes, node and edge attributes, and amino acid coordinates, using the kNN graph. We embed this representation into the topological encoding, W_v^l = EGNN(G) that contains L layers. The output layer, which represents a joint distribution, predicts the potential mutations at each position of the protein sequence.

To generate a query for the blind mutational prediction task, the model samples a set of queries that are possible point mutations through nucleotide substitution. Once the queries have been generated, the model computes their semantic embeddings by passing them through the ESM-2 once more. The model then compares the embeddings of the queries and the reference sequences (original and unmutated) to determine the impact of mutations on their robustness and to calculate a fitness score. The scores obtain probabilities that represent the likelihood that a particular mutation will lead to a stable or unstable protein. The fitness score is not solely based on the difference and similarity between the reference and query sequence. Rather, it considers the entire context of the protein sequence and structure, as well as the interactions between amino acids within the protein. The fitness score of a mutated protein with the mutated sites T ($|T| \geq 1$) is calculated using the log-odds-ratio:

$$\sum_{t \in T} \log p(y_t) - \sum_{t \in T} \log p(v_t) \quad (1)$$

y_t is the mutant amino acid and v_t is the original wild-type amino acid at site t [12]. Positive values signify that the mutation is beneficial and more stable than the original while negative values imply the opposite outcome. Figure 6 portrays the workflow of the EGNN-CMutPred model.

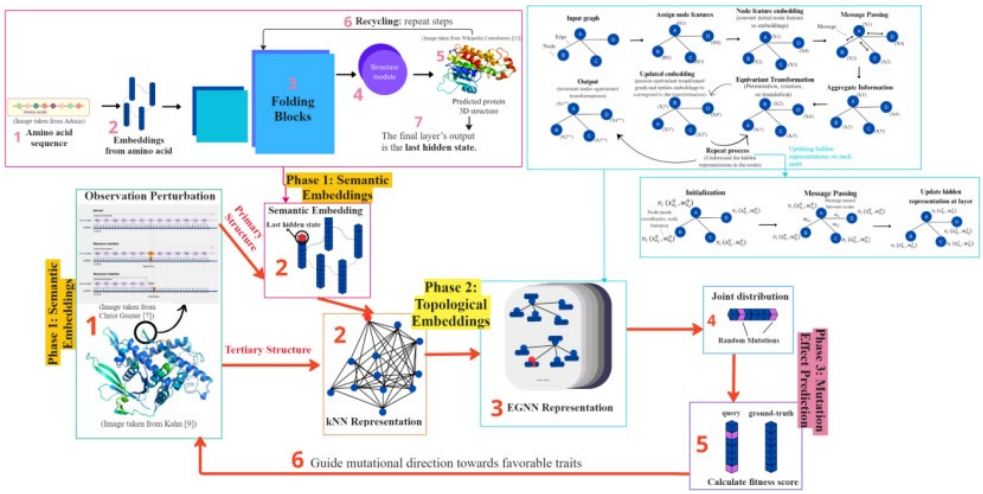

Figure 6. A pipeline of the EGNN-CMutPred's architecture.

2.3. Observation Perturbations

The observed amino acid (AA) type $\tilde{v}$ from the input protein files is initially subjected to random observation perturbations during training. The perturbations allow the model to learn a more robust representation of protein sequences and to account for uncertainty in the observations by simulating blind point mutations. The observation perturbations introduce a small amount of noise into the protein sequences to help avoid overfitting, which occurs when the model performs well on training data but fails to generalize to unseen data. When training data contains trivial or erroneous information, the model may mistakenly learn patterns from this noise. The observation perturbation addresses this issue by introducing noise, which reduces the model's overconfidence. This, in turn, encourages the model to explore multiple possible solutions and refine its predictions. Encouraging the model to rely less on specific details helps it recognize underlying protein sequence patterns and generalize predictions to unseen data [21]. The probability of amino acid variation is governed by a tunable parameter p, which determines the mutation rate. The observation perturbation, based on the Bernoulli distribution, is represented by equation (2):

There are three possible outcomes from the equation. The variable $\delta(\tilde{v} - v)$ represents the Dirac delta function. If $\tilde{v}$ equals v, the function equals 1; if else, it equals 0. The $\Theta(\cdot)$ is a replacement distribution, in which a random amino acid or a masked token will replace the original amino acid at the site. It means that the replacement distribution is not activated when the perturbed amino acid type $\tilde{v}$ is equal to the original amino acid type v with the probability (1-p). In contrast, it means that the replacement is activated when $\tilde{v}$ is a new amino acid type or a masked token that replaces the original amino acid type. Therefore, depending on the mutation rate p, the outcome is either the original amino acid or a replacement drawn from the 20 amino acids or a masked token.

2.4. Capture of Hidden Representations

To fully understand the protein, the model must capture the hidden representations of both its semantic and topological embeddings.

2.5. Semantic Embedding

The semantic encoding analyzes protein sequences $\tilde{v}$, extracts evolutionary information about the protein sequence, and embeds them in hidden representations W_v (a vector of numbers that capture important features and properties) through the Transformer model ESM-2. ESM-2's process of learning sequential information is depicted in Figure 7. After the protein sequences are inputted into the model, the model begins to

create embeddings that identify unique patterns in the protein sequences. The embeddings are then passed to the folding block mechanism, which updates predictions of protein structure. This mechanism creates two representations, sequential and pairwise.

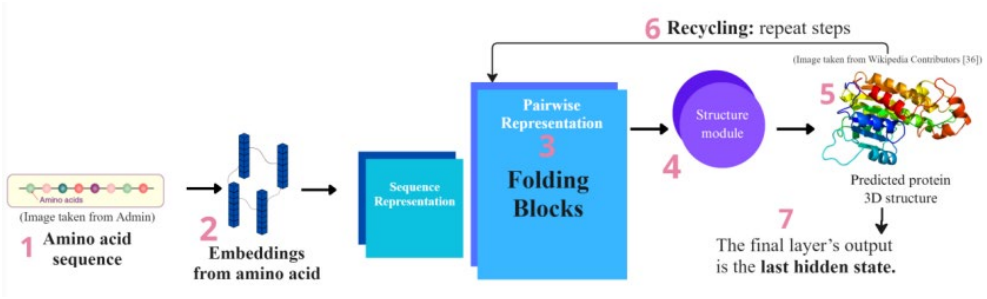

Figure 7. A representation of ESM-2's architecture.

$$\pi(\tilde{v} \mid v) = p\,\Theta(\pi 1, \pi 2, \ldots, \pi 20) + (1 - p)\,\delta(\tilde{v} - v) \quad (2)$$

The sequence representation updates information on sequential information by learning the function of each amino acid in each location. On the other hand, the structural links between pairs of amino acids are updated through pairwise representation. This encoding includes the evolutionary constraints (or selective pressure) on each position. The evolutionary constraint is the level of variability of amino acids in each position based on their functional importance [22]. Positions important for stability or activity undergo fewer substitutions over time since they face higher selective measures while less important positions are more variable since they tend to be evolutionary tolerant to mutations and substitutions. More conserved positions, therefore, have more distinct semantic representations compared to variable ones. This information then goes through the structure module, where the model gains a basic understanding of how mutations create structural change in the protein. Through the recycling process, the model updates embedding information to improve the accuracy of the prediction [14]. The semantic encoding captures the long-range dependencies and global sequence patterns beyond local graph neighborhoods. After updating every piece of sequential information, the final layer's output is the last hidden state representation, which highlights the significance and purpose of each amino acid and its inter-dependencies. This information enables the model to speculate the influence of alterations in protein sequences on their resilience.

2.6. Tertiary Structure Representation

The k-nearest neighbor (kNN) algorithm constructs the protein's geometric configuration by representing spatial configuration and interactions. It is characterized by the equation of $G = (V, E, W_v, W_E, X_v)$. Each node V represents an amino acid. Every node is linked with k neighboring nodes in a Euclidean distance E of 30Å. Node attributes W_v are the hidden representations encoded by the semantic embeddings. Edge attributes W_E are feature relationships of connected nodes within the contact region, local N-C positions (the relative positions or spatial relationships of nitrogen and carbon atoms that influence the protein structure and function), and sequential position encoding. X_v stores the 3D x, y, and z orthogonal Å coordinates of the protein sequence, which is key to the topological embedding of EGNN [15].

2.7. Topological Encoding

We use the Equivariant Graph Neural Networks (EGNN) to encode the geometric structure of the proteins. The neural network captures the hidden representation to node properties $W_{v_i}^{l+1} = \{W_{v_1}^{l+1}, \ldots, W_{v_n}^{l+1}\}$ and node coordinates $X_{pos}^{l+1} = \{x_{v_1}^{l+1}, \ldots, x_{v_n}^{l+1}\}$ at the l + 1th layer with equations (3) (4) (5):

$$m_{ij} = \emptyset_e(w_{v_i}^l, w_{v_j}^l, \left\| x_{v_i}^l - x_{v_j}^l \right\|^2, w_{e_{ij}}) \quad (3)$$

$$x_{v_i}^{l+1} = x_{v_i}^l + \frac{1}{n}\sum_{j \neq i}(x_{v_i}^l - x_{v_j}^l)\emptyset_x(m_{ij}) \quad (4)$$

$$W_{v_i}^{l+1} = \emptyset_v(W_i^l, \sum_{j \neq i} m_{ij}) \quad (5)$$

Each layer revises the hidden representation of nodes using the results from the preceding strata. The node refines embeddings grounded on the information from contiguous nodes. The propagation rules, defining how the messages are updated and propagated through the nodes and edges of the graph during each layer of the network, are $\emptyset_e$ (encodes edge information), $\emptyset_x$ (encodes node information), and $\emptyset_y$ (specify how node features are updated at each layer). The ultimate hidden representation encapsulates the immediate surroundings and spatial arrangement of the protein [15]. Figure 8 portrays EGNN's processing of information, and Figure 9 offers a deeper insight into how the model updates each layer.

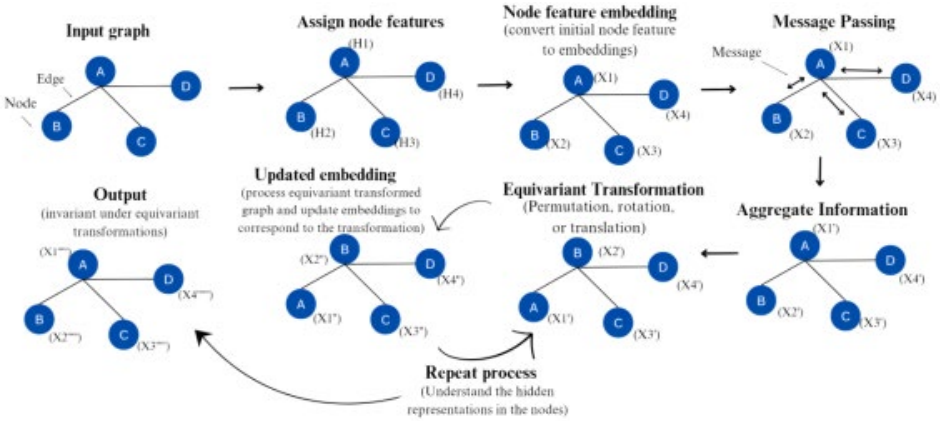

Figure 8. A representation of EGNN's encoding process.

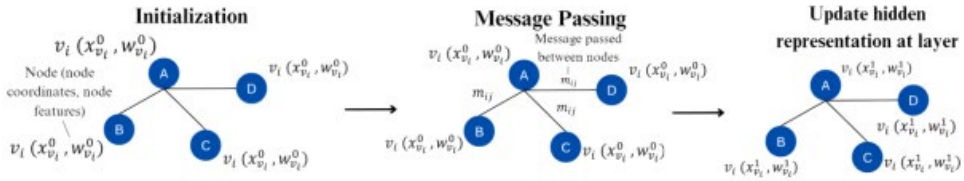

Figure 9. Depiction of EGNN's update on hidden representations.

2.8. Evaluation Metrics

Spearman's correlation: Spearman's correlation determines the extent to which two attributes fluctuate together. It measures the monotonic relationship (nondecreasing or nonincreasing function when the independent variable increases) between two ranked variables. The scale ranges from -1 to 1 [23]. Spearman's correlation in this study assesses the model's ability to rank mutation predictions in accordance with the ranking of the ground truth benchmarks. The model first ranks both the predictions and the ground truth based on the magnitude of their effects on protein stability. Next, it computes the association between the model's predictions and the actual results. This assesses the model's ability to assign higher rankings to predictions with greater effects. Figure 10 illustrates the workings of Spearman's Correlation. A positive Spearman's correlation indicates that the model is capable of distinguishing mutations with larger effects from those with smaller effects, while a negative value indicates the opposite.

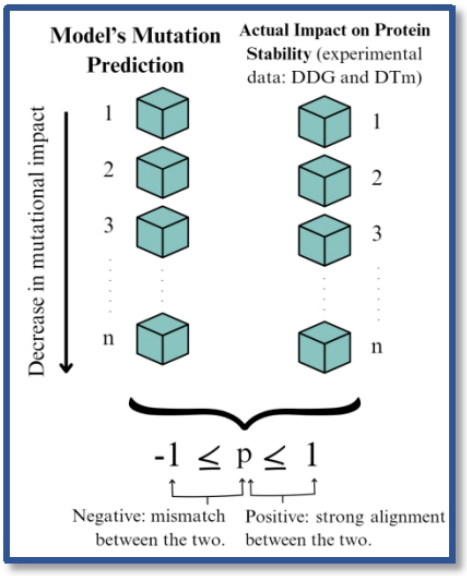

Figure 10. A representation of the Spearman's Correlation.

True Positive Rate: The True Positive Rate (TPR), also known as sensitivity, is the likelihood that genuine positive cases are correctly classified as positive [24]. It evaluates the model's ability to detect mutations that have a substantial impact on protein integrity. For example, when the TPR is set at 50%, the model is expected to correctly identify the top 50% of mutations ranked by their magnitude of effect. The model's TPR will be evaluated at 5%, 25%, and 50% thresholds. Figure 11 presents the operation of TPR.

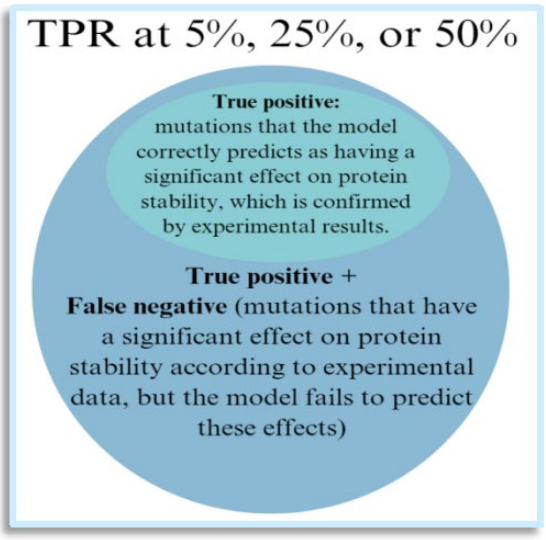

Figure 11. A representation of the True Positive Rate.

3. Result and Discussion

3.1. Biological Analysis

To gain a better insight into the biological characteristics of the protein sequences from the ProteinGym benchmark, we collected data and analyzed the protein sequences in terms of taxonomy, conserved domain, signal peptide, and hydropathicity. We used NCBI tblastn, a tool that compares protein queries to six-frame translations of nucleotide sequences in the database to identify homologous protein-coding regions in uncharacterized sequences [25,26]. As a result, the program predicted which organisms contained the protein sequences and traced their evolutionary lineage.

We then used the NCBI Batch CD-Search tool to identify conserved domains for each protein and recorded their associated superfamilies. Proteins with conserved domains exhibit unaltered sequence patterns that serve distinct purposes [27].

The PSORT II Prediction tool was used to estimate the likelihood of each protein residing in the following subcellular locations: cell wall, cytoplasm, endoplasmic reticulum, extracellular, nuclear, nucleocapsid, periplasmic, and plasma membrane.

We used both versions 4.1 and 5.0 of SignalP to verify result accuracy and predict whether each protein contains a signal peptide. Signal peptides (SP) are short strings of amino acids (peptides), located at the amino-terminal end of proteins that mark the protein secretory pathway and direct protein targeting [28].

We used ProtScale to determine the hydropathicity of the proteins.

3.1.1. Protein Taxonomy and Conserved Domains

We analyzed 86 mutated proteins by collecting data on their taxonomy and conserved domains. We organized the proteins into 11 groups based on taxonomy: virus, pseudomonadota, primate, synthetic, fungi, macroscelidea, terrabacteria, rodentia, thermotogota, archaea, and artiodactyla.

Among the proteins, we identified shared conserved domains, which are explained in detail in the following section.

3.1.2. Subcellular Location

We determined the number of proteins in each intracellular region, as shown in Figure 12. Synthetic proteins were neglected because their domains cannot be determined. We categorized every virus protein into the nucleocapsid area.

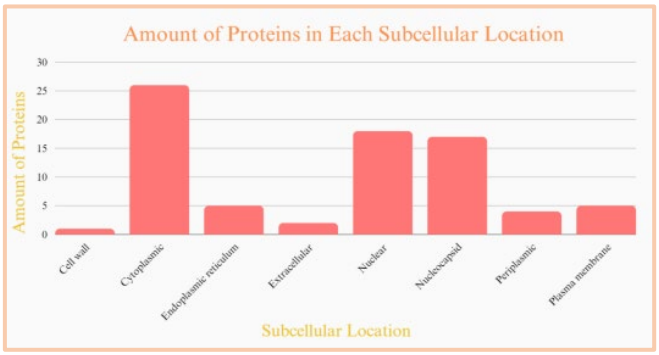

Figure 12. The number of proteins in each subcellular location.

We found correlations between the proteins' intracellular sites and their conserved domains. Two bacterial proteins and one archaea protein all belong to the Triosephosphate isomerase (TIM) superfamily. It is a glycolytic enzyme that is essential for the glycolysis pathway of cellular respiration [29,30]. All three of these prokaryotic proteins are found in the cytoplasm, where this biochemical process occurs. Additionally, two eukaryotic proteins share the ION_Tras superfamily, which contains sodium, potassium, and calcium ion channels [31]. These two eukaryotic proteins are both found in the plasma membrane since these channels are important for transmembrane transport.

3.1.3. Analysis of Signal Peptide Predicted

We generated graphs that show the predicted probability for each protein to contain a signal peptide. We determined the amount of proteins expected to include a signal peptide. As portrayed in Figure 13, there are significantly more proteins without a signal peptide than those that possess one. Proteins with signal peptides are typically those that need to be transported, whereas proteins without signal peptides function within the cell where

they are synthesized. Because fewer proteins require transport, fewer proteins contain signal peptides.

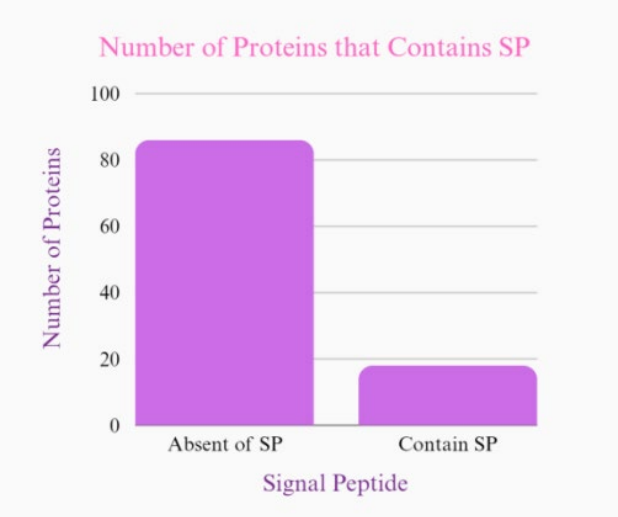

Figure 13. The quantity of proteins that contains SP.

We also calculated the number of proteins with or without signal peptides in each subcellular location, as displayed in Figure 14.

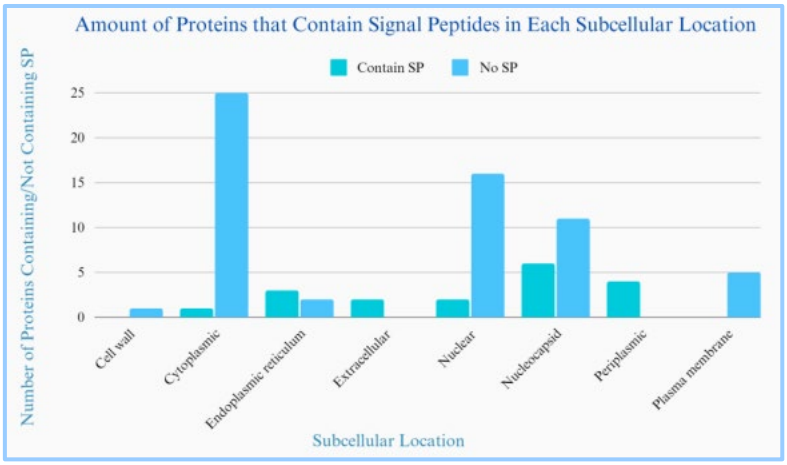

Figure 14. The quantity of proteins that contain signal peptides in each intracellular compartment.

It can be observed that most proteins that contain a signal peptide are located in the endoplasmic reticulum, extracellular regions, or the periplasmic space (the region between the two plasma membranes of bacteria). All eukaryotic proteins belong to the secretory pathway, where some facilitate protein construction and adjustment in the ER and Golgi complex while others are delivered to other locations [32]. Therefore, the ER and extracellular regions are two of the most prevalent areas for proteins containing signal peptides. On the other hand, the Sec translocase system, the primary protein targeting mechanism in prokaryotes, involves proteins which include signal peptides. A molecule named Signal Recognition Particle (SRP) first attaches to the signal peptide on the protein. The protein then crosses the plasma membrane to the periplasmic space via a protein called Sec translocase. In this space, the protein may experience further modifications and then carry out its specific function. As a result, a large number of prokaryotic proteins with signal peptides are found in the periplasmic space since it is the location for most secretory proteins [33].

Conversely, most proteins without a signal peptide are usually found in the nucleus, cytoplasm, and plasma membrane. Many proteins without a signal peptide are located in the nucleus because they often play key roles in processes including DNA synthesis, transcription, and repair. Proteins in the cytoplasm generally do not contain signal peptides because they do not need to be translocated anywhere. Other proteins in the plasma membrane do not contain signal peptides because of the post-translational insertion process. In this process, ribosomes synthesize membrane proteins in the cytoplasm, and other molecules insert them into the membrane. Moreover, since parts of the plasma membrane are hydrophobic, it is common to find proteins without hydrophilic signal peptides in these regions.

3.1.4. Analysis of Hydropathicity Prediction

We found the hydropathicity of each protein, as depicted in Figure 15.

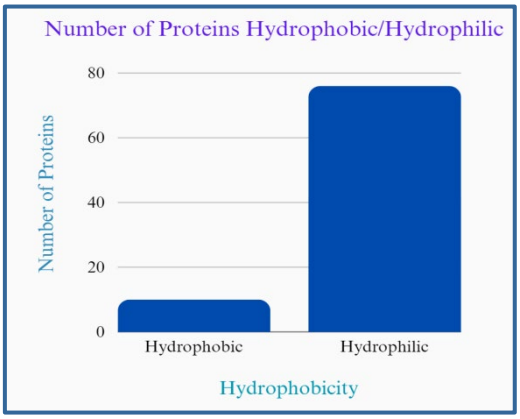

Figure 15. The distribution of proteins that exhibit hydrophobic or hydrophilic properties.

Proteins tend to be more hydrophilic than hydrophobic because they are often found in aqueous environments such as the cytoplasm and nucleus, where they need to interact with other molecules. Many of these molecules are polar or charged. The hydrophilic regions of proteins allow them to interact with these molecules and the surrounding water, which is crucial for their function. Other proteins are hydrophobic as they are embedded in the cell membrane, which is composed primarily of hydrophobic fatty acid chains.

3.1.5. Implications of protein-analysis

The analysis of the ProteinGym protein data is vital as it reveals diverse characteristics and behavior of the studied proteins, allowing us to comprehend the data that the model will analyze. Moreover, the analysis ensures the diversity of proteins. As displayed in previous sections, the proteins come from a wide variety of taxonomic classes, and each serves a different function, resulting from components such as conserved domains and intracellular regions. Therefore, by exposing the model to various proteins with differing degrees of mutations, it can learn to generalize better and improve prediction accuracy for protein mutations.

3.1.6. Limitation of Protein-Level Analysis

While the analysis of protein data offers various insights, it is essential to recognize that it cannot capture the full spectrum of protein characteristics. Proteins function in a system with myriad other molecules, and their interactions with the environment constantly influence their behavior. Therefore, while the results obtained from the analysis of protein data are informative, they represent only a small part of the protein's overall biological landscape. This implies that any conclusions derived from this analysis should be interpreted cautiously.

3.2. AI Results

The biological analysis in Section 3.1 provides crucial context for understanding the data within the ProteinGym benchmark. This information serves as a foundation for validating and interpreting the model's predictions. The following section presents the results of the AI model, and the evaluation values will be compared with current, state-of-art models. By combining our knowledge of the biological data and EGNN-CMutPred's performance, we can assess its effectiveness in predicting the effects of protein mutations and identify areas for future improvement.

Figure 14 characterizes the median evaluation scores from different versions of the model. See Appendix C for the evaluation scores by △△G, △Tm, and ProteinGym. In the graph, k stands for the number of nearest neighbors, and h represents the number of hidden layers. The version with the maximum score is k20_h1280 with a score of 0.640, while the version with the greatest Spearman's Correlation is k20_h512 with a score of 0.587.

From the experiments, it can be observed that setting k = 10 may cause the model to have insufficient ability to interpret the data. In contrast, setting k = 30 may cause the model to overfit to noise. Therefore, among the three k values tested, k = 20 is the optimal value since the versions with the best Spearman's Correlation typically use k set to 20.

There is no clear trend indicated for h. One feasible explanation is that the number of hidden layers does not have a notable influence on the outcomes, or that the effects of h are obscured by the considerable influence of k. Another possibility is that the model is more complex than necessary for practical purposes (Figure 16).

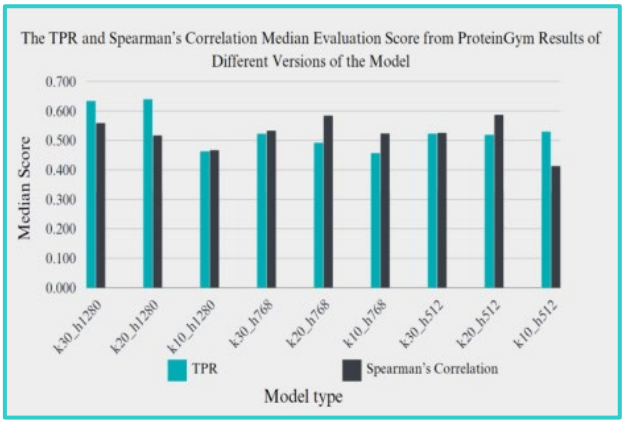

Figure 16. The TPR and Spearman's Correlation Median Evaluation Score from △△G Results of Different Versions of the Model.

The log-likelihood scores of EGNN-CMutPred for mutational predictions are compared to those of an ESM model based on the AlphaFold2 mechanism. This model utilizes inverse folding, an autoregressive encoder-decoder architecture, to predict protein sequences from their structure [34]. The log-likelihood score depicts the likelihood that a given mutated sequence occurs based on wild-type protein sequences. Higher scores indicate that a sequence is more likely to resemble naturally occurring protein sequences, while a low score suggests that the mutation is unfavorable in natural evolution. The log-likelihood score comparison for two proteins, O61594-6.0 and P0A3D9-7.0, is shown below:

As depicted in Tables 2 and 3 and Figure 17, the sequences generated by ESM show a concentrated distribution with a small standard deviation of plausible mutations, and all results fall within a reasonable range. In contrast, EGNN-CMutPred produces a wider scoring range, distinguishing highly favorable mutations (e.g., L100A and V19I) from unfavorable ones (e.g., N391A and A102V). Therefore, the ESM model focuses on predicting sequences with moderate plausibility (scores close to 0), while EGNN-CMutPred identifies more extreme values of beneficial and detrimental mutations. Since the ESM model is

a transformer-based language model trained on large-scale evolutionary sequence data, it captures evolutionary constraints that favor stability. When predicting mutations, ESM is biased towards mutations that are evolutionarily plausible and maintain protein stability. Conversely, EGNN-CMutPred is partially trained on structural data, enabling it to capture mutational disruptions of protein function based on spatial relationships. This structural sensitivity allows it to better distinguish between beneficial and detrimental mutations, leading to a wide range of scores.

Table 2. Comparison for O61594-6.0.

	EGNN-CMutPred	ESM
Average	-0.817	-0.07
Highest Value	-0.764 (sampled_seq_6)	0.52 (V19I)
Lowest Value	-0.922 (sampled_seq_16)	-0.69 (A102V)

Table 3. Comparison for P0A3D9-7.0.

	EGNN-CMutPred	ESM
Average	-0.873	-1.85（-0.59 with the removal of N391A ）
Highest Value	-0.797 (sampled_seq_15)	0.81 (L100A)
Lowest Value	-0.954 (sampled_seq_9)	-23.21 (N391A)

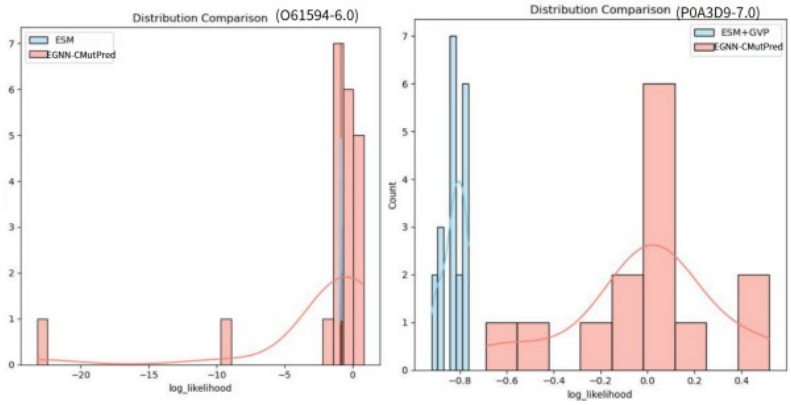

Figure 17. Visualization of data distribution of the comparison between the log_likelihood scores.

Table 4 compares the accuracy of sequence-based, structure-based, and combined sequence and structure-based models based on Spearman's correlation values. Each model is tested on the △△G benchmark. According to the table, our model outperforms most current models. EGNN-CMutPred has the second-highest Spearman's correlation among the models presented, with a score of 0.587, while the highest value is achieved by the structure and sequence-based model SSIPe with a score of 0.62. However, it is important to note that the performance of these models may vary slightly because the Spearman's correlation results come from different studies, and these studies may approach the input data differently. There may also be newer models with enhanced capabilities that we have not yet recognized. The table also reveals that structure-based and combined sequence and structure-based models generally perform better than sequence-based models, as shown by their higher Spearman's correlation scores. Sequence-based models have an average Spearman's correlation of 0.2485, which is significantly lower than the averages of structure-based and combined models, which are 0.49 and 0.47 respectively. One likely explanation is that structural information helps AI models gain a better understanding of protein characteristics than sequential data. This is probably because spatial relationships

and interactions between residues are better represented in three-dimensional space. Additionally, models that analyze structural information can account for epistatic effects, in which a mutation in one gene depends on the presence or absence of mutations in other genes [35].

Table 4. The comparison of Spearman's values among several model types. The 3 highest values in each category are marked as First, Second, and Third.

Category	Model	Version	Spearman's (p)
Sequence-based	ProGen2 [15] (large-scale transformer model with up to billions of parameters [36])	Small (151M params)	0.194
		Medium (764M params)	0.214
		Base (764M params)	0.253
		Large (2700M params)	0.226
		xlarge (6400M params)	0.270
	ESM-2 [15] (transformer-based language model trained via protein sequences for masked language modeling [14])	t12	0.216
		t30	0.317
		t33	0.392
		t36	0.351
		t48	0.252
	RITA [15] (autoregressive, generative model with 1.2B parameters [37])	Small (30M params)	0.143
		Medium (300M params)	0.188
		Large (680M params)	0.236
		xlarge (1200M params)	0.264
	Tranception [15] (transformer model with autoregressive predictions [38])	Small (85M params)	0.169
		Medium (300M params)	0.256
		Large (700M params)	0.284
Structure-based	DSMBind [39] (unsupervised, energy-based model that predicts effects of mutations in protein-protein interactions [40])	-	0.53
	ProteinMPNN [39] (neural network with 128 hidden dimensions that passes messages [41]	-	0.45
Structure and sequence-based	DDMut [42] (deep learning model that integrates both graph-based convolutional layers and transformer encoder [42])	Result of S552 blind test set	0.54
		Result of S2024 blind test set	0.41
	Alphafold [39] (deep learning model that predicts protein structure and estimates mutational effects on protein-protein interactions [39,43])	AF3 ranking_score	0.51
		AF3 iptm	0.50
		AF3 ptm	0.33
		AF3 mean_pae	0.37
		AF2 ranking_score	0.23
		Effective Strain	0.31
		AF2 mean_pae	0.22
	Force Field and Profile-based [39] (Combine structural information of force fields with sequence information of profiles to estimate the	SSIPe	0.62
		FlexddG	0.58
		FoldX	0.54

mutational effects on protein-protein binding affinity [39,44])		
	k30_h1280	0.559
	k20_h1280	0.517
	k10_h1280	0.467
	k30_h768	0.533
EGNN-CMutPred	k20_h768	0.584
	k10_h768	0.524
	k30_h512	0.526
	k20_h512	0.587
	k10_h512	0.413

Our study successfully developed a new method for predicting the effects of protein mutations by integrating both primary (sequence) and tertiary (structure) embeddings from the ESM-2 and EGNN models. By combining semantic and topological embeddings, the model captures both the grammatical rules and geometric structures of proteins, addressing limitations of traditional sequence- and structure-based prediction methods [45-47]. Training was performed on non-redundant protein sequences from the CATH v4.3.0 database, and evaluation was conducted using the ProteinGym (DMS) and ProThermDB (△Tm and △△G) benchmarks [48]. The model demonstrates its effectiveness in forecasting mutational impacts on protein activity and stability [49].

Evaluation metrics, including Spearman's correlation and True Positive Rate (TPR), indicate that the model can reliably rank mutation effects and identify significant mutations, as all evaluation scores are positive. Therefore, the results suggest that the model can effectively forecast the impact of alterations in protein sequences [50,51]. The analysis of the biological input data enhances understanding of the biological characteristics of the protein sequences used as model input. The strengths and limitations of our model are also revealed through analysis of its predictions. Notably, the model's ability to generate a fitness score for mutated proteins provides valuable insights for biomolecular research and the study of disease-related protein mutations [52,53].

4. Conclusion

Future work will focus on advancing computational methodologies, experimenting with more detailed parameters, and testing the model on a wider variety of databases that measure changes in protein stability. Enhancements in the model's architecture and training pipeline could further improve its predictive accuracy and efficiency. Experimenting with more specific parameters of k and h is also important for finding the optimal version of the model. Additionally, we plan to assess the model on an expanded range of benchmarks to gain deeper insights into the wide variety of protein factors affected by mutations. Additional thermodynamic parameters include enthalpy change, which measures the total amount of heat absorbed or emitted during a reaction; heat capacity change, which estimates the energy required to increase the temperature of a protein; and protein-protein interaction assays, which measure changes in a protein's ability to interact with others. Furthermore, the datasets used to train the model are well characterized, potentially limiting its generalizability to less-studied proteins. Testing the model on large-scale databases with more diverse protein data, such as Deep Sequence and GEMME, is therefore essential. Another limitation is that the model is mostly trained on complete protein structural data, which may prevent accurate prediction of mutational effects when data are incomplete. Possible future improvements include incorporating probabilistic models to account for structural uncertainties. Once the model's capabilities are improved comprehensively, future studies could apply it to pharmacological research, aiding drug development by identifying drug targets and predicting mutational effects on mutated drug targets.

References

1. J. A. Lycklama a Nijeholt and A. J. M. Driessen, "The bacterial Sec-translocase: structure and mechanism," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 367, no. 1592, pp. 1016-1028, 2012, doi: 10.1098/rstb.2011.0201.
2. M. Knudsen and C. Wiuf, "The CATH database," *Hum. Genomics*, vol. 4, no. 3, p. 207, 2010, doi: 10.1186/1479-7364-4-3-207.
3. R. Bonetta and G. Valentino, "Machine learning techniques for protein function prediction," *Proteins: Struct., Funct., Bioinf.*, vol. 88, no. 3, pp. 397-413, 2020, doi: 10.1002/prot.25832..
4. W. Lu et al., "AlphaFold3, a secret sauce for predicting mutational effects on protein-protein interactions," bioRxiv, 2024, doi: 10.1101/2024.05.25.595871.
5. The UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 46, no. 5, p. 2699, 2018, doi: 10.1093/nar/gky092.
6. P. W. Rose et al., "The RCSB protein data bank: integrative view of protein, gene and 3D structural information," *Nucleic Acids Res.*, 2016, doi: 10.1093/nar/gkw1000.
7. G. R. Reeck et al., "“Homology” in proteins and nucleic acids: a terminology muddle and a way out of it," *Cell*, vol. 50, no. 5, pp. 667, 1987.
8. S. Sinha, B. Eisenhaber, and A. M. Lynn, "Predicting protein function using homology-based methods," in *Bioinformatics: sequences, structures, phylogeny*, Singapore: Springer Singapore, 2018, pp. 289-305, doi: 10.1007/978-981-13-1562-6_13.
9. J. C. Whisstock and A. M. Lesk, "Prediction of protein function from protein sequence and structure," *Q. Rev. Biophys.*, vol. 36, no. 3, pp. 307-340, 2003, doi: 10.1017/S0033583503003901.
10. M. Camps et al., "Genetic constraints on protein evolution," *Crit. Rev. Biochem. Mol. Biol.*, vol. 42, no. 5, pp. 313-326, 2007, doi: 10.1080/10409230701597642.
11. N. V. Prabhu and K. A. Sharp, "Heat capacity in proteins," *Annu. Rev. Phys. Chem.*, vol. 56, no. 1, pp. 521-548, 2005, doi: 10.1146/annurev.physchem.56.092503.141202.
12. S. Zhang et al., "Graph convolutional networks: a comprehensive review," *Comput. Soc. Netw.*, vol. 6, no. 1, pp. 1-23, 2019, doi: 10.1186/s40649-019-0069-y.
13. O. Handa et al., "Reduction of butyric acid-producing bacteria in the ileal mucosa-associated microbiota is associated with the history of abdominal surgery in patients with Crohn’s disease," *Redox Rep.*, vol. 28, no. 1, p. 2241615, 2023, doi: 10.1080/13510002.2023.2241615.
14. A. R. Katebi and R. L. Jernigan, "The critical role of the loops of triosephosphate isomerase for its oligomerization, dynamics, and functionality," *Protein Sci.*, vol. 23, no. 2, pp. 213-228, 2014, doi: 10.1002/pro.2407.
15. P. Veličković et al., "Graph attention networks," arXiv preprint arXiv:1710.10903, 2017.
16. V. G. Satorras, E. Hoogeboom, and M. Welling, "E (n) equivariant graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021.
17. K. H. Choo and S. Ranganathan, "Flanking signal and mature peptide residues influence signal peptide cleavage," *BMC Bioinformatics*, vol. 9, Suppl 12, p. S15, 2008, doi: 10.1186/1471-2105-9-S12-S15.
18. Y. Tan et al., "Multi-level protein representation learning for blind mutational effect prediction," arXiv preprint arXiv:2306.04899, 2023.
19. S. Yo et al., "Exercise affects mucosa-associated microbiota and colonic tumor formation induced by azoxymethane in high-fat-diet-induced obese mice," *Microorganisms*, vol. 12, no. 5, p. 957, 2024, doi: 10.3390/microorganisms12050957.
20. S. Xue et al., "Comprehensive analysis of signal peptides in Saccharomyces cerevisiae reveals features for efficient secretion," *Adv. Sci.*, vol. 10, no. 2, p. 2203433, 2023, doi: 10.1002/advs.202203433.
21. F. M. G. Pearl et al., "The CATH database: an extended protein family resource for structural and functional genomics," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 452-455, 2003, doi: 10.1093/nar/gkg062.
22. S. Velankar et al., "PDBe: protein data bank in Europe," *Nucleic Acids Res.*, vol. 38, suppl. 1, pp. D308-D317, 2010, doi: 10.1093/nar/gkp916.
23. R. Nikam et al., "ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D420-D424, 2021, doi: 10.1093/nar/gkaa1035.
24. C. Aliferis and G. Simon, "Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI," in *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, 2024, pp. 477-524, doi: 10.1007/978-3-031-39355-6_10.
25. M. Eltehiwy and A. B. Abdul-Motaal, "A new Method for Computing and TestingThe significance of the Spearman Rank Correlation," *Comput. J. Math. Stat. Sci.*, vol. 2, no. 2, pp. 240-250, 2023, doi: 10.21608/cjmss.2023.229746.1015.
26. J. N. Suojanen, "False false positive rates," *N. Engl. J. Med.*, vol. 341, no. 2, p. 131, 1999, doi: 10.1056/NEJM199907083410217.
27. M. Sasahira et al., "The relationship between bacterial flora in saliva and esophageal mucus and endoscopic severity in patients with eosinophilic esophagitis," *Int. J. Mol. Sci.*, vol. 26, no. 7, p. 3026, 2025, doi: 10.3390/ijms26073026.
28. E. M. Gertz et al., "Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST," *BMC Biol.*, vol. 4, no. 1, p. 41, 2006, doi: 10.1186/1741-7007-4-41.

29. W. Deng et al., "ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets," *Bioinformatics*, vol. 23, no. 17, pp. 2334-2336, 2007, doi: 10.1093/bioinformatics/btm331.
30. S. A. Shiryev et al., "Improved BLAST searches using longer words for protein seeding," *Bioinformatics*, vol. 23, no. 21, pp. 2949-2951, 2007, doi: 10.1093/bioinformatics/btm479.
31. IBM, "What are Recurrent Neural Networks?," IBM.com, Oct. 6, 2021. [Online]. Available: https://www.ibm.com/topics/recurrent-neural-networks.
32. H. Owji et al., "A comprehensive review of signal peptides: Structure, roles, and applications," *Eur. J. Cell Biol.*, vol. 97, no. 6, pp. 422-441, 2018, doi: 10.1016/j.ejcb.2018.06.003.
33. S. Grasso et al., "Signal peptide efficiency: from high-throughput data to prediction and explanation," *ACS Synth. Biol.*, vol. 12, no. 2, pp. 390-404, 2023, doi: 10.1021/acssynbio.2c00328.
34. Y. Zhou et al., "DDMut: predicting effects of mutations on protein stability using deep learning," *Nucleic Acids Res.*, vol. 51, no. W1, pp. W122-W128, 2023, doi: 10.1093/nar/gkad472.
35. C. Pancotti et al., "Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset," *Brief. Bioinform.*, vol. 23, no. 2, 2022, doi: 10.1093/bib/bbab555.
36. M. A. Pak et al., "Using AlphaFold to predict the impact of single mutations on protein stability and function," *PLoS One*, vol. 18, no. 3, p. e0282689, 2023, doi: 10.1371/journal.pone.0282689.
37. G. R. Buel and K. J. Walters, "Can AlphaFold2 predict the impact of missense mutations on structure?," *Nat. Struct. Mol. Biol.*, vol. 29, no. 1, pp. 1-2, 2022, doi: 10.1038/s41594-021-00714-2.
38. Y. Peng, E. Alexov, and S. Basu, "Structural perspective on revealing and altering molecular functions of genetic variants linked with diseases," *Int. J. Mol. Sci.*, vol. 20, no. 3, p. 548, 2019, doi: 10.3390/ijms20030548.
39. J. Meier et al., "Language models enable zero-shot prediction of the effects of mutations on protein function," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 29287-29303, 2021.
40. M. H. Høie et al., "Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation," *Cell Rep.*, vol. 38, no. 2, 2022, doi: 10.1016/j.celrep.2021.110207.
41. N. Brandes et al., "Genome-wide prediction of disease variant effects with a deep protein language model," *Nat. Genet.*, vol. 55, no. 9, pp. 1512-1522, 2023, doi: 10.1038/s41588-023-01465-0.
42. X. Liu et al., "Deep geometric representations for modeling effects of mutations on protein-protein binding affinity," *PLoS Comput. Biol.*, vol. 17, no. 8, p. e1009284, 2021, doi: 10.1371/journal.pcbi.1009284.
43. P. Notin et al., "Proteingym: Large-scale benchmarks for protein fitness prediction and design," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 64331-64379, 2023.
44. Z. Lin et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123-1130, 2023, doi: 10.1126/science.ade2574.
45. E. Krieger, S. B. Nabuurs, and G. Vriend, "Homology modeling," in *Struct. Bioinf.*, 2003, pp. 509-523, doi: 10.1002/0471721204.
46. D. J. Diaz et al., "Using machine learning to predict the effects and consequences of mutations in proteins," *Curr. Opin. Struct. Biol.*, vol. 78, p. 102518, 2023, doi: 10.1016/j.sbi.2022.102518.
47. A. Zhou et al., "Proteolytic processing in the secretory pathway," *J. Biol. Chem.*, vol. 274, no. 30, pp. 20745-20748, 1999, doi: 10.1074/jbc.274.30.20745.
48. N. Shah et al., "Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows," *Bioinformatics*, vol. 35, no. 9, pp. 1613-1614, 2019, doi: 10.1093/bioinformatics/bty833.
49. X. Liu, "Deep recurrent neural network for protein function prediction from sequence," arXiv preprint arXiv:1701.08318, 2017.
50. H. Matsumoto et al., "Characteristics of mucosa-associated microbiota in ulcerative colitis patients with 5-aminosalicylic acid intolerance," *Biomedicines*, vol. 12, no. 9, p. 2125, 2024, doi: 10.3390/biomedicines12092125.
51. V. Gligorijević et al., "Structure-based protein function prediction using graph convolutional networks," *Nat. Commun.*, vol. 12, no. 1, p. 3168, 2021, doi: 10.1038/s41467-021-23303-9.
52. S. Aizawa et al., "Adenosine stimulates neuromedin U mRNA expression in the rat pars tuberalis," *Mol. Cell. Endocrinol.*, vol. 496, p. 110518, 2019, doi: 10.1016/j.mce.2019.110518.
53. D. E. V. Pires, D. B. Ascher, and T. L. Blundell, "mCSM: predicting the effects of mutations in proteins using graph-based signatures," *Bioinformatics*, vol. 30, no. 3, pp. 335-342, 2014, doi: 10.1093/bioinformatics/btt691.