

# Beyond ESM2: Graph-Enhanced Protein Sequence Modeling with Efficient Clustering

Shujian Jiao<sup>1</sup> Bingxuan Li<sup>2</sup> Lei Wang<sup>1</sup>  
Xiaojin Zhang<sup>1</sup> Wei Chen<sup>1✉</sup> Jiajie Peng<sup>3</sup> Zhongyu Wei<sup>1✉</sup>

<sup>1</sup>Huazhong University of Science and Technology

<sup>2</sup>Fudan University <sup>3</sup>Northwestern Polytechnical University

{shujianjiao, wanglei94, xiaojinzhang, lemuria\_chen}@hust.edu.cn

windskileoli@gmail.com jiajiepeng@nwpu.edu.cn zywei@fudan.edu.cn

## Abstract

Proteins are essential to life's processes, underpinning evolution and diversity. Advances in sequencing technology have revealed millions of proteins, underscoring the need for sophisticated pre-trained protein models for biological analysis and AI development. Facebook's ESM2, the most advanced protein language model to date, leverages a masked prediction task for unsupervised learning, crafting amino acid representations with notable biochemical accuracy. Yet, it lacks in delivering functional protein insights, signaling an opportunity for enhancing representation quality. Our study addresses this gap by incorporating protein family classification into ESM2's training. This approach, augmented with Community Propagation-Based Clustering Algorithm, improves global protein representations, while a contextual prediction task fine-tunes local amino acid accuracy. Significantly, our model achieved state-of-the-art results in several downstream experiments, demonstrating the power of combining global and local methodologies to substantially boost protein representation quality.

## 1 Introduction

Proteins are involved in almost all the life activities of organisms, and the study of their sequences, structures, characteristics, and roles is a major area of research in the life sciences in the post-genomic era (Papin et al., 2003). A protein sequence can be thought of as a string of amino acid letters. The residues, structural domains, and families of amino acids that make up a protein resemble words, phrases, and sentences in human language. Therefore, machine learning methods developed for natural language and other sequences are well suited to the task of predicting proteins (Ofer et al., 2021). Most sequence-based language models [e.g.,

BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020)] are designed to process natural language (with a bias towards English). Considering the notable parallels between protein sequences and the structure of natural language, employing natural language processing (Chowdhary and Chowdhary, 2020) (NLP) techniques to analyze protein sequences emerges as a logical approach.

Currently, ESM-2 (Lin et al., 2022), developed by Facebook, is recognized as the most extensive protein sequence language model to date, featuring a sophisticated architecture with 48 layers and over 15 billion parameters. This groundbreaking model is trained on an expansive dataset comprising up to 250 million protein sequences sourced from Uniparc (Leinonen et al., 2004), which encompasses 86 billion amino acids (Ng and Henikoff, 2006). The dataset's vast scale mirrors the extensive text corpora used in developing large-scale neural networks for natural language processing, highlighting the model's unmatched breadth and depth. Leveraging this comprehensive base, ESM-Fold emerges as an innovative 3D protein structure prediction tool, leveraging ESM-2's insights with just a single sequence input to significantly speed up predictions. In cases where it deeply understands sequences, ESMFold achieves atomic-level accuracy, matching or even surpassing leading models like AlphaFold2 (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021). This efficiency in generating swift, precise predictions from single inputs showcases that a more profound grasp of sequences leads to a better understanding of protein structures.

Challenges accompany ESM-2 and its derivatives like ESMFold, primarily in their reliance on statistical analysis of amino acid compositions to generate amino acid representations. This approach, focused on predicting missing amino acids within contexts, proves adequate for classifying

\*Equal contribution. ✉Corresponding authors.

protein families and identifying remote homologies but falls short in capturing the full functional complexity of proteins. Consequently, the unsupervised training method yields protein representations that lack interpretability, underscoring the insufficiency of statistical properties alone to fully elucidate protein functions. Moreover, since the computational complexity is the square of the sequence length and the length of protein sequences is much larger than the text length, and since ESM2 and ESMFold require a large amount of computational resources, the large number of parameters, and the complexity of the computation pose obstacles for researchers who do not have high-performance computing facilities. Additionally, while ESMFold offers promising capabilities in predicting protein structures from single sequences, it struggles with proteins exhibiting complex folding patterns or those necessitating interactions with other molecules for accurate structural and functional modeling, indicating the potential need for integrating additional sequence or molecular information for more precise predictions.

In general, our contributions are of three-folds:

- We have fused graph pre-training with masked language modeling to refine the ESM2 model, achieving unparalleled performance in protein-centric tasks beyond ESM2.
- Our proposed Community Propagation-Based Clustering Algorithm is a novel, resource-efficient training method for graph neural networks.
- We provide a detailed demonstration of the role of the graph network asynchronous information propagation algorithm in the pre-training tasks of protein sequences.

## 2 Related Work

### 2.1 Protein Large Language Models

The technology of language models is currently very popular and being explored for application across a variety of professional domains, including healthcare (Bao et al., 2023), finance (Chen et al., 2023a), industry (Gu et al., 2024), education (Dan et al., 2023), and the judiciary (Yue et al., 2023). In the healthcare sector, there is a particular focus on AI-based intelligent diagnostics (Zhong et al., 2022; Chen et al., 2023a,b; Fan et al., 2024), with primary modalities including text, medical imaging, and medical examinations (Gu et al., 2020). As

research delves deeper, it has become evident that understanding the microscopic information, such as genetics and proteins, is crucial for gaining a more profound comprehension of disease processes.

Proteins, with their sequences of amino acids, exhibit a remarkable resemblance to natural language, where residues and structural domains can be likened to letters and words. This similarity has sparked interest in developing AI methodologies tailored for protein analysis. By leveraging the parallels between the intricate 'language' of proteins and human language, researchers are now turning their attention to AI-driven approaches to better understand and predict protein sequences, which are fundamental to the mechanisms of life. The development of specialized Protein Large Language Models (Prot-LLMs) has significantly advanced with models categorized into encoder-only, decoder-only, and encoder-decoder architectures, each tailored for different protein research applications. Encoder-only models are optimized for predicting protein functions or properties, while decoder-only models focus on protein sequence generation. Central to encoder-only Prot-LLMs is the Transformer encoder architecture, which efficiently translates protein sequences into concise vector representations, crucial for identifying unique protein patterns. Prominent pre-trained protein sequence encoders include ProteinBert (Brandes et al., 2022), ProtTrans (El-naggar et al., 2021), PMLM (He et al., 2021), ProtFlash (Wang et al., 2023a), ProtNPT (Notin et al., 2023), ESM1b (Rives et al., 2021), ESM-1v (Meier et al., 2021), and ESM-2 (Lin et al., 2022). The integration of techniques such as Multiple Sequence Alignment (MSA) with ESM-MSA-1b (MSA Transformer) (Rao et al., 2021) has been instrumental in breakthroughs like AlphaFold2 and AlphaMissense (Cheng et al., 2023). Incorporating 3D structural information has marked a significant progression, exemplified by models like ESM-GearNet (Zhang et al., 2023), SaProt (Su et al., 2023), LM-GVP (Wang et al., 2022c), and PromptProtein (Wang et al., 2022b), indicating a shift towards structure-aware pretraining to refine protein representation. Efforts to enhance encoder architectures and training methodologies include self-supervised learning through masked language modeling (MLM) tasks, effectively reconstructing corrupted tokens from contextual cues, with the ESM series leveraging the Transformer encoder architecture, akin to BERT and RoBERTa (Liu et al.,

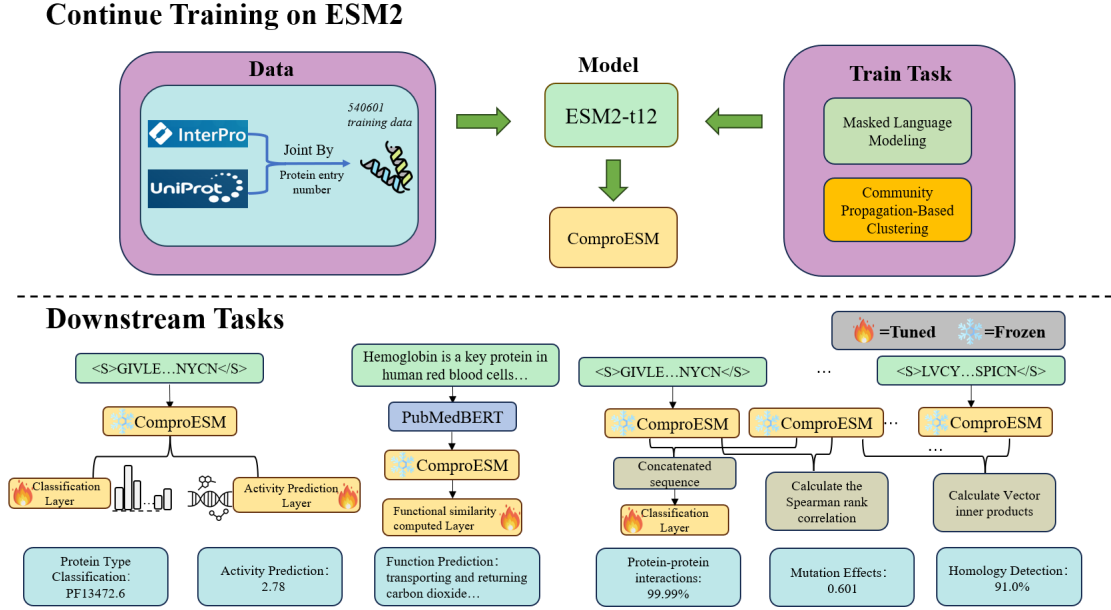


Figure 1: The overview of our work.

2019), to predict protein structure and function.

## 2.2 Advancements in ESM Models

The field of protein prediction has witnessed transformative advancements through the evolution of Evolutionary Scale Modeling (ESM). A significant leap forward was achieved with the adaptation of the ESM1b model, which now provides highly accurate variant effect predictions for an extensive array of over 40,000 protein isoforms, markedly surpassing the capabilities of conventional methodologies. The introduction of the ESM-2 model represents another milestone, featuring an exponential increase in model parameters from 8 million to an impressive 15 billion. This upscaling has been instrumental in achieving unprecedented accuracy in protein structure prediction, demonstrating the critical role of model complexity in computational protein analysis. Furthermore, the application of ESM models in specialized tasks, such as MHC peptide binding prediction (Hashemi et al., 2023) and single-sequence protein structure prediction (Wang et al., 2022a), showcases the models’ adaptability and effectiveness across diverse research applications. The release of cutting-edge pre-trained models and codes (Facebook Research, 2024) specifically designed for protein design, coupled with the recent updates to the ESM Metagenomic Atlas, underscores the continuous efforts toward refining the predictive accuracy and utility of ESM models. Moreover, the integration of

LR-ESM in the NetGO 3.0 (Wang et al., 2023b) framework for advanced large-scale functional prediction exemplifies the broadening scope of ESM models in providing critical insights into the vast expanse of uncharacterized proteins. Collectively, these advancements not only enhance our understanding of protein functions and interactions but also pave the way for novel discoveries in protein science.

## 3 Method

The masking language task is crucial for the model to discern each amino acid’s biochemical traits, while Community Propagation-Based Clustering connects these sequences to the protein’s structure and function based on the statistical nature of amino acids. The masked language model deciphers semantics at the residue level, and the clustering approach interprets the entire protein sequence. Together, they meld local and global insights, greatly enhancing protein representation and capturing the intricate relationship between a protein’s structure, function, and statistics. Community Propagation-Based Clustering Algorithms, inspired by hierarchical clustering, enable nuanced classification of proteins into detailed families or broader superfamilies, establishing hierarchical relationships. This mimics the organization of communities and towns, where synchronizing tasks among members—akin to information exchange within a network—is managed through regular updates to

a central system. This model of synchronization is analogous to the information propagation on a graph, streamlining the integration and coordination of work progress.

### 3.1 Community Propagation-Based Clustering

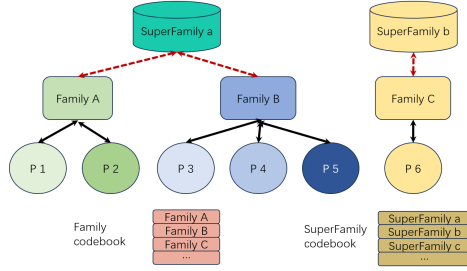


Figure 2: Direction of information flow for Community Propagation-Based Clustering Algorithm

In Figure 2, P1 and P2 are two proteins in family A, and P3 and P4 are two proteins in family B. Together they belong to a superfamily. The black realisation arrows represent the information exchange between the protein nodes and the family nodes they belong to, and this exchange should be executed immediately along with the update of the protein nodes. The red dotted line represents the information exchange between the family node and the superfamily node, which should be time-delayed, i.e., when the update on the family accumulates a certain number of steps before it is performed.

What needs to be considered in the Community Propagation-Based Clustering Algorithm is how to design the initial representation of each node and how information is propagated between nodes. For the initial representation of a protein node, this paper uses the average pooling of the protein representation model for the representation of each amino acid residue output as the initial representation of the protein node:

$$P = \frac{1}{l} \sum_{i=1}^l H_i, \quad (1)$$

where  $l$  represents the input sequence length and  $H_i$  represents the representation vector of the  $i$ th amino acid in the last layer of the protein representation model. In this paper, we use randomly initialised vectors  $F \in \mathbb{R}^d$  and  $SF \in \mathbb{R}^d$  as the initial representations of family and superfamily nodes, respectively.

In this paper, link prediction is used as the information propagation algorithm. Specifically, for the input protein representation  $\{P_1, P_2, \dots, P_N\}$ ,  $N$  being the batch size, and the corresponding family node representation  $\{F_1, F_2, \dots, F_N\}$  and the superfamily node representation  $\{SF_1, SF_2, \dots, SF_N\}$ . The score function  $S_F(P_i, F_j)$  is used to evaluate the likelihood of establishing a link between protein node  $i$  and protein family node  $j$ . The score function  $S_{SF}(P_i, SF_j)$  is used to evaluate the possibility of establishing links between protein nodes and superfamily nodes, and in practice this paper adopts the vector inner product as the score function. It should be ensured that the distribution of the output of the score function conforms to the distribution of the actual graph.

In this paper, Softmax is used to measure the distribution of the score function. Specific:

$$\mathcal{L}_{family} = \frac{1}{N} \sum_{i=1}^N -\log \text{Sm}(S_F(P_i, F_i)), \quad (2)$$

$$\mathcal{L}_{superfamily} = \frac{1}{N} \sum_{i=1}^N -\log \text{Sm}(S_{SF}(P_i, SF_i)), \quad (3)$$

where  $\text{Softmax}()$  measures the distribution of samples in a batch. The loss representation of the final Community Propagation-Based Clustering Algorithm is:

$$\mathcal{L}_{commu} = w\mathcal{L}_{family} + (1-w)\mathcal{L}_{superfamily}, \quad (4)$$

where  $w$  represents the weight magnitude of the two losses. The Community Propagation-Based Clustering Algorithm proposed in this paper addresses how data with multiple category labels can be incorporated into a pre-trained language model by means of graph pre-training. Although the external manifestation of the algorithm is the propagation of information on a graph, it essentially addresses the problem of hierarchical clustering of data from different categories.

Clustering algorithms, such as hierarchical clustering (Murtagh and Contreras, 2012) and k-means clustering (Hartigan and Wong, 1979), segment data into classes based on feature similarity, aiming to maximize intra-class similarity and minimize inter-class similarity. In the realm of protein families, this translates to grouping proteins into superfamilies and families, where each family represents a fine-grained cluster and each superfamily



a coarse-grained one. This research treats each family as a precise clustering focus, reflecting the need for protein representations to encapsulate the familial information. Hence, we introduce the Community Propagation-Based Clustering Algorithm for a supervised approach to clustering protein sequences, aligning the sequence representation with its familial heritage. A detailed exposition on the algorithm’s interpretability is provided in the appendix A.

### 3.2 Masked Language Modeling

Intertwined with clustering, the masked language task equips the model to discern individual amino acid biochemical properties. The Community Propagation-Based Clustering Algorithm then takes this a step further, forging connections between amino acid sequences and the protein’s structural and functional attributes. These attributes are deeply linked to the statistical properties of amino acids. Together, these twin objectives—masking language tasks and community propagation—create a synergy, reinforcing each other to refine the protein representation quality. In our methodology, we harness the BERT model’s architecture, renowned for its multiple stacked Transformer Blocks, as the encoder for protein sequences. To achieve our dual goals, we deploy two distinct fully connected neural networks to project the encoded protein vectors into separate representation spaces for the masked prediction task and the Community Propagation-Based Clustering Algorithm. Each Transformer Block in the BERT framework features a multi-head self-attention mechanism, recalculating token representations by considering contextual token relationships. Multi-heads permit multiple attention computations, capturing diverse sequence characteristics, followed by concatenation and mapping back to the original input size through a perceptron. The forward propagation network—a classic two-layer fully connected network—serves as an intermediary, transforming representations between layers.

Each token vector encoded by the BERT model will have its loss computed by the mask prediction task and the Community Propagation-Based Clustering Algorithm, respectively, in order to run the stochastic gradient descent algorithm to update all network parameters of the model. Specifically, for the mask prediction task the loss function is

computed as:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_M \sum_{i \in M} -\log p(x_i | x_{/M}), \quad (5)$$

For each sequence  $x = \{x_1, x_2, \dots, x_n\}$ , a set of tokens  $M$  to be masked is sampled, and the actual amino acid at each position  $i$  is replaced with the mask symbol  $\langle \text{mask} \rangle$ . For each masked position, the loss function is set to the negative logarithmic value of the probability of predicting the correct amino acid  $x_i$ , taking the sequence  $x_{/M}$  as the context of the protein sequence with the masked portion removed. Intuitively, in order to predict the masked position, the model must identify the correlation between the masked site and the unmasked portion of the sequence. The probability is calculated as:

$$p(x_i | x_{/M}) = \text{Softmax}(W_{Mj} \cdot z_{li}), \quad (6)$$

where  $W_M \in \mathbb{R}^{d \times V}$  is the vocabulary prediction matrix used to map the size of the representation vector of each token to the size of the vocabulary list,  $V$  is the size of the vocabulary list, and  $W_{Mj}$  stands for the vector of prediction matrices corresponding to the  $j$  word (the correct word corresponding to  $x_i$ ) in the vocabulary list.

For the Community Propagation-Based Clustering Algorithm this paper uses the link prediction task as the goal, using the average of the token representations at each location as the final representation vector of the protein sequence i.e:

$$P = \frac{1}{l} \sum_{i=1}^l z_{li}, \quad (7)$$

where  $l$  represents the input sequence length and  $z_{li}$  represents the representation vector of the  $i$ th amino acid output from the last layer of BERT. In this paper, we use the randomly initialised vectors  $F \in \mathbb{R}^d$  and  $SF \in \mathbb{R}^d$  as the initial representations of family and superfamily nodes, respectively. Assume that all protein sequences involved in a single batch calculation are denoted as  $\{P_1, P_2, \dots, P_N\}$ ,  $N$  being the batch size, and the corresponding family node representation for each protein  $\{F_1, F_2, \dots, F_N\}$  and the superfamily node representation  $\{SF_1, SF_2, \dots, SF_N\}$ . The score function  $S_F(P_i, F_j)$  is used to evaluate the likelihood of establishing a link between protein node  $i$  and protein family node  $j$ . The score function  $S_{SF}(P_i, F_j)$  is used to evaluate the possibility

of establishing links between protein family nodes and superfamily nodes, and in practice this paper adopts the vector inner product as the score function. It should be ensured that the distribution of the score function output matches the distribution of the actual graph.

The final model loss is the sum of the two optimisation objective losses:

$$\mathcal{L}_{\text{model}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{commu}}. \quad (8)$$

The specifics concerning the masked language model and the Community Propagation-Based Clustering are elaborated in the appendix B.

## 4 Continue Training on ESM2

### 4.1 Pre-training Dataset

We obtained the classification data and corresponding amino acid sequences of proteins from two databases, InterPro (<https://www.ebi.ac.uk/interpro/>) (Paysan-Lafosse et al., 2023) and UniprotKB (<https://www.uniprot.org/>) (Boutet et al., 2016) are two databases to obtain the classification data and corresponding amino acid sequences of the proteins, respectively. InterPro classifies each protein into a family and provides predictions of protein structural domains and sites of importance in the protein sequence to analyse the protein function. To classify proteins in this way, InterPro uses prediction models provided by several different databases that make up the InterPro consortium (called member databases). By combining protein features from these member databases into one searchable resource, their individual strengths are exploited to generate a powerful integrated database and diagnostic tool.

UniProtKB contains the functional information of the proteins and preserves the amino acid sequences of the proteins. UniProtKB consists of two parts:

- The information in this part of the UniProtKB/Swiss-Prot-Database is manually annotated and reviewed, and is therefore of high quality and non-redundant.
- UniProtKB/TrEMBL-The information in this part of the database is computationally annotated by algorithms and is not reviewed, thus providing a high level of coverage of the proteome.

The protein classification data from InterPro and the protein sequence data from UniprotKB are finally joined by protein entry numbers to obtain all the training data.

Considering the limitations of local computing resources and the quality of the data, all the proteins from UniProtKB/Swiss-Prot database are finally extracted in this paper, which are all These proteins are all manually annotated and thus representative and highly reliable.

A total of 568744 proteins were obtained from UniProtKB /SwissProt, and all the obtained proteins were queried for the corresponding protein interfaces in InterPro, and finally 540601 training data were obtained. For each protein, the complete amino acid sequences of all families and superfamilies were recorded. It should be noted that some proteins may lack the corresponding superfamily or family classification information, but they must have one of the two categories, in this case, we will randomly sample to fill in the missing items, for example, if a protein lacks a family category, the model will randomly select one of the superfamily categories as the family category each time it reads the sample. Detailed statistics of the data are given in the table 2.

### 4.2 Training Details

A 12-layer ESM2 model was used as the protein representation model in the pre-training experiments, which was subjected to knowledge distillation using Tinybert to generate the student model. For the mask prediction task, 15% of the amino acid residues in the protein sequence are randomly selected for replacement, of which 80% are replaced with <mask>, 10% are replaced with a random representation, and the remaining 10% are kept unchanged. For Community Propagation-Based Clustering Algorithm, the weights of family score loss and superfamily score loss were 0.2 and 0.8, respectively. an AdamW was used as the optimiser with a learning rate of  $5e-5$ .

### 4.3 Validation of Protein Representation

The visualisation method demonstrates whether the stability and functional properties possessed by a protein can be deduced from the representation of the model. The stability of a protein is mainly characterised by the biochemical properties of the amino acid residues that make up the protein, whereas the function of a protein is mainly reflected in the proximity of representations of pro-

Table 1: Experimental validation of the model on downstream tasks.

Model	Family Class.	Mut. Effects	Act. Pred.	Interact.	Func. Pred.	Homol. Det.
ESM2-12layer	27.51%	0.514	0.040	99.68%	90.38%	87.8%
ESM2-6layer	20.55%	0.423	0.043	99.58%	88.39%	81.3%
TinyBert	13.30%	0.178	0.075	99.17%	88.09%	76.5%
<b>OurModel</b>	<b>31.08%</b>	<b>0.601</b>	<b>0.038</b>	<b>99.99%</b>	<b>91.47%</b>	<b>91.0%</b>
Ablation Experiment						
only MLM	27.93%	0.601	0.042	99.75%	90.27%	88.1%
only Graph	30.05%	0.132	0.052	99.69%	91.0%	89.2%

Table 2: Statistics of the dataset.

Statistics	Values
Protein samples count	540,601
Family categories count	17,132
Superfamily categories count	3,189
Family memberships per protein	1.23
Family memberships per protein	1.43
Average length of amino acid sequences	367.01

teins with similar structure and function. In order to investigate whether the model has learnt to encode physicochemical properties in the representation of amino acid residues, this paper uses TSNE to downscale the parameters of the final prediction layer of the network so that they can be plotted onto 2D coordinates. In Fig. 3, it is found that the space can clearly distinguish the boundaries between amino acids with different biochemical properties. These biochemical properties include hydrophobicity, polarity, and aromaticity, as well as molecular size and electrode properties, all of which affect the interchangeability of amino acid residues at a site in a protein, i.e., the stability of the protein structure.

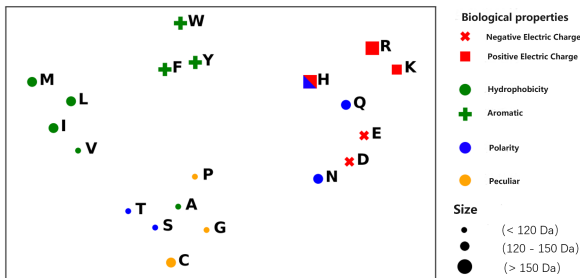


Figure 3: Model prediction layer representation of the biochemical properties of the embedded amino acids.

A representation of each amino acid residue can be obtained through the final layer of the model, and the final protein can be obtained by averaging

these representations. These protein representations project a sequence to a point in space, thus different protein sequences can be distinguished. In this paper, the distribution of homologous genes in space can be studied intuitively in two-dimensional coordinates by TSNE dimensionality reduction. In addition, Principal Component Analysis (PCA) was applied to recover the main direction of change in the representation, and four homologous genes in four species were selected to find the direction of change. In order to exclude the statistical properties of the number of amino acids itself in the amino acid sequence, this paper also introduces the protein representation before untrained and the protein representation using the unitary bag-of-words model. Figure 4 shows that the modelled protein representations encode structural and functional information about the proteins thereby allowing for a significant aggregation effect of protein representations within the same protein family.



Figure 4: Protein representation after TSNE reduction

## 5 Downstream task experiments

### 5.1 Downstream task introduction

To evaluate the effectiveness of the protein representation model post-pretraining, we established the following downstream tasks based on six different protein aspects: Protein Classification, Protein Activity Prediction, Protein-Protein Interaction, Protein Function Prediction, Mutation Effects, and Homology Testing. For task validation,

datasets specific to each task were procured, which were not involved in the training of the protein representation model. The model developed in this study solely generates a numerical vector replete with structural-functional information from the input protein sequences, which is then utilized for the respective downstream tasks. For tasks requiring multiple protein sequences (Protein-Protein Interaction, Mutation Effects, Homology Testing), we adopted an approach of training multiple protein sequences over several iterations. Tasks such as Protein Classification, Protein-Protein Interaction, and Protein Function Prediction necessitate the construction of a multi-layer fully connected neural network and subsequent fine-tuning on a training set. It is important to note that the parameters of the protein representation model proposed in this paper remain constant during the experiments of these downstream tasks. Detailed experimental procedures are provided in the appendix (from appendixD to appendixI).

## 5.2 Enhanced Performance in Protein Representation Tasks

The table highlights optimal performances in bold, underscoring that ComproESM—leveraging both mask prediction tasks and the graph network asynchronous information propagation algorithm—surpasses the ESM2 model in a range of downstream tasks. Given ComproESM’s integration of protein family data, which encapsulates structural and functional protein information, enhancements in family classification, function prediction, and homology detection are anticipated. Moreover, ComproESM outperforms the ESM2 model in tasks requiring a comprehensive understanding of the biochemical properties of amino acid residues, such as mutation effects, activity prediction, and protein-protein interactions. Ablation studies focusing individually on these tasks further suggest that while targeting specific tasks can yield improvements, a combined approach garners optimal results across all downstream tasks. Considering both ESM2 and our proposed model are trained on mask prediction tasks, we partitioned data into a training set and test set at a 9:1 ratio. Performance on the mask prediction task is evaluated using the Exponential Cross Entropy Score (ECE), where a perfect model scores 1, and a uniform random prediction across 25 amino acids scores 25. Lower ECE scores indicate superior mask prediction performance. As shown in Table 3, our model slightly

underperforms the sole MLM model trained on the same dataset, attributed to the incorporation of graph link prediction tasks altering the original protein representation space by integrating protein family data. Further validation of the changes induced by the protein representation is detailed in the appendixC.

Table 3: Mask prediction task.

Model	Parameter Count	ECE
Ideal value		1
Uniform Distribution		25
4-gram		19.21
LSTM-small	28.4M	15.13
LSTM-Large	113.4M	13.97
ESM2-12layer	35.8M	5.80
OurModel	35.8M	5.78
only MLM	35.8M	5.62

## 5.3 Protein Classification Model Analysis

In order to analyse the advantages of the model in this paper on the protein classification task even further, the 10 most common protein families were selected from the dataset PF13649.6, PF00560.33, PF13508.7, PF06580.13, PF02397.16, PF00677.17, PF01035.20, PF02417.15, PF13472.6, PF00684.19.60 samples of proteins were sampled for each category, of which 50 samples were used for training and 10 samples for testing. The results after 20 epochs of iteration under the same experimental setup are shown in Table 4, and it can be found that the accuracy of the model proposed in this paper is much larger than that of the ESM2 model in the training set, which indicates that the protein representation given by the model proposed in this paper has a better decision boundary between the different protein family classes, and so the model can be in training to find this boundary quickly. The confusion matrix of the prediction results is shown in Figure 9, which shows that the ESM2 model has a lower prediction accuracy for the proteins in the protein family categories PF13649.6, PF13508.7, and PF13472.6, whereas there is no such bias in the model proposed in this paper, which indicates that the protein representations given in the proposed model are more consistent with the protein family categories than those given in the ESM2 model. protein representation has a more direct correlation with the protein family categories.



Table 4: Comparison of results on the protein classification task

Models	Training Sets	Test Sets
ESM2-12layer	38.46%	80.00%
OurModel	78.24%	94.00%

## 6 Conclusion

In this paper, we propose ComproESM, and demonstrate the efficacy of integrating protein family classification information with the ESM2 model’s training process, thereby enhancing the model’s capability to accurately represent global protein structures and predict local amino acid residues. Our innovative Community Propagation-Based Clustering Algorithm, when combined with the masking algorithm, has significantly advanced the model’s performance across a spectrum of downstream tasks, far surpassing the capabilities of the ESM2 model with a comparable number of parameters. These findings not only underscore the potential of tailored algorithmic strategies in understanding complex biological data but also pave the way for future research in protein sequence analysis and functional prediction. Through this work, we contribute to the broader field of computational biology by providing a robust framework for protein representation that is sensitive to both the structural and functional nuances of proteins.

## Limitation

Although this paper achieves optimal performance on various protein downstream tasks, due to the limitation of experimental resources, we do not use the mask prediction task and the Community Propagation-Based Clustering Algorithm to train directly on the randomly initialised Transformer model, but uses the 12-layer model of ESM2 for initialisation. The representation space of the final model trained by this method is highly offset, which means that the model only converges to a locally optimal solution. Meanwhile, the training data taken in this paper only has 500,000 proteins, which is still too small compared to the hundreds of millions of proteins stored in protein databases. In the future, this paper will train the model in this paper from scratch on larger protein sequence data, so as to obtain a higher quality protein representation model.

## Ethics Consideration

Given the profound potential impact of this research on the life sciences and medical fields, several ethical considerations must be taken into account. Firstly, ensuring data privacy is critical, particularly if any human-derived protein sequences are included in the dataset; researchers must handle such data responsibly and with consideration for consent and confidentiality. Additionally, the accessibility of the developed tools and methods should be addressed to prevent exacerbating inequalities in research capabilities across different institutions or countries, especially considering the computational resources required for high-performance models like ESM-2. Moreover, it is imperative to transparently communicate both the capabilities and limitations of the model to avoid misinterpretation or overreliance on its predictions, which could have significant consequences for downstream applications such as drug discovery or personalized medicine. Finally, there should be a commitment to open science by making the methodologies, source code, and potentially the trained model accessible to other researchers, fostering collaboration and accelerating advancements while respecting intellectual property rights.

## References

- Monica Agrawal, Marinka Zitnik, and Jure Leskovec. 2018. Large-scale analysis of disease pathways in the human interactome. In *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 111–122. World Scientific.
- Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876.
- Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.
- Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. 2016. Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: how to use the entry view. *Plant bioinformatics: methods and protocols*, pages 23–54.

- Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rapoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.
- Wei Chen, Zhiwei Li, Hongyi Fang, Qianyu Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023a. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.
- Wei Chen, Cheng Zhong, Jiajie Peng, and Zhongyu Wei. 2023b. Dxfomer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations. *Bioinformatics*, 39(1):btac744.
- Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. 2023. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492.
- KR1442 Chowdhary and KR Chowdhary. 2020. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ahmed Elnaggar, Michael Heinzinger, Christian Dalago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127.
- Facebook Research. 2024. [ESM: Evolutionary Scale Modeling](#). Accessed: 2024-02-14.
- Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.
- Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. 2014. Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309.
- Vanessa E Gray, Ronald J Hause, Jens Luebeck, Jay Shendure, and Douglas M Fowler. 2018. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell systems*, 6(1):116–124.
- Qizheng Gu, Cong Nie, Ruixiang Zou, Wei Chen, Chaojun Zheng, Dongqing Zhu, Xiaojun Mao, Zhongyu Wei, and Dong Tian. 2020. Automatic generation of electromyogram diagnosis report. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1645–1650. IEEE.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2024. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1932–1940.
- John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.
- Nasser Hashemi, Boran Hao, Mikhail Ignatov, Ioannis Ch Paschalidis, Pirooz Vakili, Sandor Vajda, and Dima Kozakov. 2023. Improved prediction of mhc-peptide binding using protein language models. *Frontiers in Bioinformatics*, 3.
- Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, et al. 2021. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.
- Rasko Leinonen, Federico Garcia Diez, David Binns, Wolfgang Fleischmann, Rodrigo Lopez, and Rolf Apweiler. 2004. Uniprot archive. *Bioinformatics*, 20(17):3236–3237.
- Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

- Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303.
- Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.
- Pauline C Ng and Steven Henikoff. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, 7:61–80.
- Pascal Notin, Ruben Weitzman, Debora S Marks, and Yarin Gal. 2023. Proteinntp: Improving protein property prediction and design with non-parametric transformers. *bioRxiv*, pages 2023–12.
- Dan Ofer, Nadav Brandes, and Michal Linial. 2021. The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758.
- Jason A Papin, Nathan D Price, Sharon J Wiback, David A Fell, and Bernhard O Palsson. 2003. Metabolic pathways in the post-genome era. *Trends in biochemical sciences*, 28(5):250–258.
- Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. 2023. Interpro in 2022. *Nucleic Acids Research*, 51(D1):D418–D427.
- Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR.
- Adam J Riesselman, John B Ingraham, and Debora S Marks. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822.
- Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.
- Johannes Söding and Michael Remmert. 2011. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Current opinion in structural biology*, 21(3):404–411.
- Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. 2023. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pages 2023–10.
- Lei Wang, Hui Zhang, Wei Xu, Zhidong Xue, and Yan Wang. 2023a. Deciphering the protein landscape with protflash, a lightweight language model. *Cell Reports Physical Science*, 4(10).
- Shaojun Wang, Ronghui You, Yunjia Liu, Yi Xiong, and Shanfeng Zhu. 2023b. Netgo 3.0: Protein language model improves large-scale functional annotations. *Genomics, Proteomics & Bioinformatics*.
- Wenkai Wang, Zhenling Peng, and Jianyi Yang. 2022a. Single-sequence protein structure prediction using supervised transformer protein language models. *Nature Computational Science*, 2(12):804–814.
- Zeyuan Wang, Qiang Zhang, HU Shuang-Wei, Hao-ran Yu, Xurui Jin, Zhichen Gong, and Huajun Chen. 2022b. Multi-level protein structure pre-training via prompt learning. In *The Eleventh International Conference on Learning Representations*.
- Zichen Wang, Steven A Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O Salawu, Colby J Wise, Sri Priya Ponnappalli, et al. 2022c. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific reports*, 12(1):6832.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.
- Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.
- Zuobai Zhang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. 2023. Enhancing protein language models with structure-based encoder and pre-training. *arXiv preprint arXiv:2303.06275*.
- Cheng Zhong, Kanganbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, 38(16):3995–4001.

## A Interpretability of Community Propagation-Based Clustering Algorithms

In the Community Propagation-Based Clustering Algorithm, the data input to the model

is a set of samples  $\{P_1, P_2, \dots, P_N\}$ , where  $N$  is the batch size, and each sample corresponds to the family category representation  $\{F_1, F_2, \dots, F_N\}$  and the superfamily category representation  $\{SF_1, SF_2, \dots, SF_N\}$ . For the  $i$ th sample  $P_i$  of these,  $\text{Softmax}(Z_i)$  is required to be equal to 1. That is, the score  $Z_i$  is required to be as large as possible and the other scores as small as possible. Since  $Z = wZ_F + (1 - w)Z_{SF}$ , i.e., the scores  $Z_F$  and  $Z_{SF}$  are required to be as large as possible. Take the calculation of  $Z_F$  as an example:

$$Z_F = P \cdot F. \quad (9)$$

where  $P = \{p_1, p_2, \dots, p_d\}$ ,  $F = \{f_1, f_2, \dots, f_d\}$ , then  $Z_F$  can be expressed as:

$$Z_F = \sum_{j=1}^d p_j \times f_j, \quad (10)$$

And if you use the Euclidean distance measure between  $P, F$ , the distance distance is given by:

$$S = \frac{\sum_{j=1}^d (p_j - f_j)^2}{|P||F|}, \quad (11)$$

$$\frac{|P|^2 + |F|^2 - 2Z_F}{|P||F|}, \quad (12)$$

Assuming that the stretching of the vectors does not affect the properties of the vectors themselves, it can be assumed that  $|P| = |F| = 1$ , and thus the distance  $S$  can be expressed as:

$$S = 2 - 2Z_F. \quad (13)$$

It follows that  $\max(Z_F)$  is equivalent to  $\min(S)$ , in other words the optimisation objective treats the category representation as a clustering centre, and the purpose of the loss function is to bring in closer the distance between nodes within the class and the centre of the class, and to bring out further the distance between nodes outside the class and the centre of the class. The parameters  $w$  and size are used to control the coverage of different granularity categories in the representation space. In order to demonstrate the clustering effect of Community Propagation-Based Clustering Algorithm in a more graphical way, in this paper, 120 random 2-dimensional vectors are initialised using a normal distribution, and each vector corresponds to a fine-grained category and a coarse-grained category, with a total of 30 small categories and 10 large categories. The backpropagation algorithm was used

on this randomly generated dataset using Community Propagation-Based Clustering Algorithm until convergence, and finally the representations of all the 2-dimensional vectors were presented on the axes (as shown in Fig. 5).

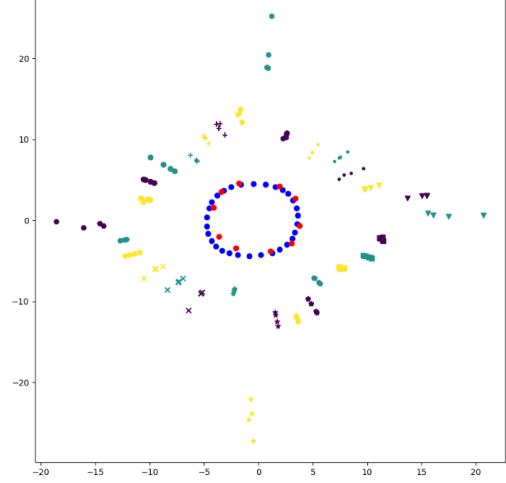


Figure 5: Community Propagation-Based Clustering Algorithm on random initialisation vectors

In the graph 5, all red nodes represent coarse-grained cluster centres and all blue nodes represent fine-grained cluster centres. The density of the red nodes is less than that of the blue nodes due to the weight parameters  $w_f$  and  $w_s$ . The rest of the nodes are all nodes in the dataset, nodes belonging to the same coarse-grained level are represented using the same shape, and nodes of the same colour under the same shape represent their same fine-grained category. From the figure, it can be seen that Community Propagation-Based Clustering Algorithm achieves a significant clustering effect and the sample distribution is in accordance with the assumption above, i.e., the stretching of the vectors does not affect the nature of the vectors themselves.

## B Mask Language Model Setting Details

The input used by the model is a protein sequence, which is a string of letters where each letter represents a particular amino acid. Also for each protein sequence there is a family class and superfamily class to which the protein belongs, and these labels will be used for subsequent loss function calculations. The model will first obtain a token sequence from the input protein sequences (each letter, i.e., an amino acid, is a token) and add special identifiers  $\langle \text{bos} \rangle$  and  $\langle \text{eos} \rangle$  at the beginning and the end of the sequence, respectively, and use a masking technique to blur the input so as to force the model



to derive the missing token based on the context. 15% of the tokens will be randomly selected and used with the identifier <mask>. A random 15% token is selected and replaced with the identifier <mask>. Note that in order to improve the model’s performance, the strategy of replacing the original token with <mask>, replacing it with a random token, and keeping the original token unchanged is implemented in a ratio of 8:1:1 for all the replaced words. The inputs to the model are  $\{x_1, x_2, \dots, x_n\}$  where  $x_i \in \mathbb{R}$ , the first layer of the BERT model is the word encoder, which is to convert the token input after word splitting into a vector input, and here the approach is to use the shaping index corresponding to each token to obtain the corresponding column of vectors in the word encoding matrix. The  $W_e \in \mathbb{R}^{d \times V}$  and  $W_p \in \mathbb{R}^{d \times S}$  are the word encoding matrix and position encoding matrix respectively,  $V$  is the size of the vocabulary list,  $S$  is the maximum length of the input that can be accepted by the model, and  $d$  is the size of the representation vector dimension. The vector encoding for each token is the sum of word encoding and position encoding. After word encoding, we obtain  $\{z_{01}, z_{02}, \dots, z_{0n}\}$  where  $z_{0i} \in \mathbb{R}^d$ . Next these word vectors will be passed through multiple Transformer Blocks, and it should be noted that the inputs and outputs of the Transformer Blocks passing through any of the layers will be of the exact same form, so that the final representation of each token after capturing the contextual information is obtained  $\{z_{l1}, z_{l2}, \dots, z_{ln}\}$ , where  $z_{li} \in \mathbb{R}^d$ .

### C Verify that the protein representation

In order to further validate the changes produced by the protein representation, in this paper, 1010 pairs of proteins belonging to a family were randomly sampled as positive samples and an equal number of pairs of proteins that do not belong to a protein family were randomly sampled as negative samples, and the similarity between the pairs of proteins was calculated using cosine similarity. The similarity takes values between 0 and 1. The results of the two models are plotted in Figure 6 and Figure 7.

From the figure, it can be seen that the model proposed in this paper can clearly distinguish between positive and negative samples, and there is still a large part of overlapping space in the distribution of positive and negative samples for ESM2. Therefore, the biggest advantage of the proposed model is that it incorporates family information into the

protein representation with little or no loss of the quality of the ESM2 representation, which itself carries the structural and functional information of the protein. Therefore, the protein representations generated using the model proposed in this paper can outperform the original ESM2-generated protein representations for a variety of protein downstream prediction tasks.

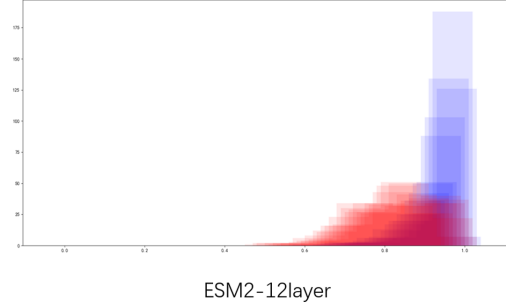


Figure 6: ESM2 model cosine similarity distribution (positive samples in blue, negative samples in red)

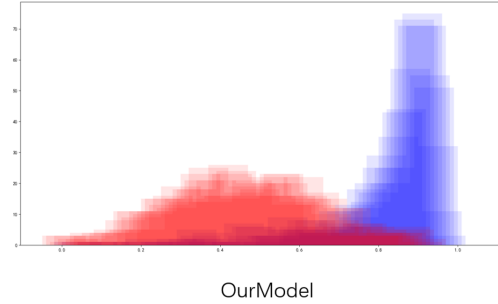


Figure 7: ComproESM cosine similarity distribution (positive samples in blue, negative samples in red)

### D Protein Classification

Search engines compute target sequences by sequence comparison algorithms to compare them with known sequences in databases to achieve classification, such as BLAST, and the performance of this method is limited by the comparison algorithm. Therefore, the use of machine learning methods to classify protein families based on protein sequences has become a hot research topic in recent years.

In this paper, we obtain the protein classification dataset on Kaggle, and organise the dataset, each sample is composed of protein sequences and their corresponding family names, and count the number of proteins corresponding to each family type (see table 5 and figure 8), and it can be seen from figure 8, it can be seen that the number of

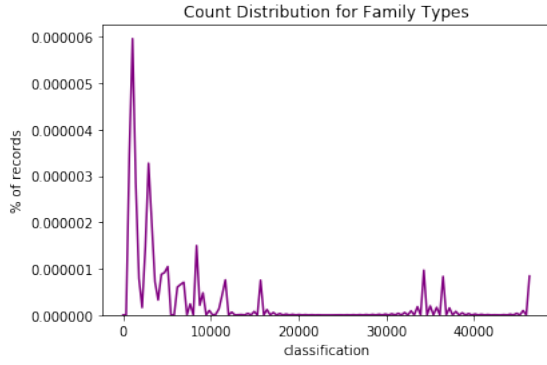


Figure 8: Distribution of protein number in the protein classification

Table 5: Statistics on the number of proteins in each category in the Protein Classification Dataset

Family	Number
HYDROLASE	29,726
TRANSFERASE	24,338
OXIDOREDUCTASE	22,578
IMMUNE SYSTEM	11,087
HYDROLASE/HYDROLASE INHIBITOR	9569
LYASE	8445
...	...
ENDOCYTOSIS, PROTEIN BINDING	1
Transferase, Cell Cycle	1
HORMONE/GENE REGULATION	1

protein family distributions in this dataset shows a long-tailed distribution, so the top 250 classes of families with the highest number of proteins are taken as the predicted labels, and the final data obtained is 278,866. Considering the time of the test, this paper uniformly samples each class to one percent of the original one, and divides the training set and the test set according to the ratio of 1:1. Finally, 1394 training samples and test samples were obtained, and the model uses the simplest single-layer fully connected neural network to classify the input protein representations and use the classification accuracy for representation quality assessment. The test results are shown in Table 1.

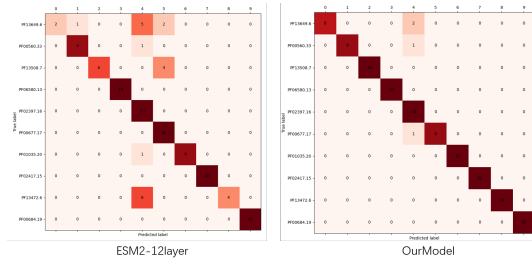


Figure 9: Confusion matrix for protein classification tasks

## E Mutation Effects

**Masked Language Modelling Objective** The training objective of a trained protein language model is to output the probability of an amino acid occurring at a position in a protein given its surroundings. In this paper, we use this property to score mutations in amino acid sequences. For a given mutation, the amino acids in the original protein can be considered as a reference state, and the probability assigned to the mutated amino acid is compared to the probability assigned to the original amino acid thereby deriving the mutation score.

Specifically, the mutation score is evaluated using a log odds ratio at the mutation site, and when multiple mutation sites are present in a protein sequence, the scores at each position are summed using an addition method:

$$\sum_{t \in T} \log p(x_t = x_t^{mt} | x_{/T}) - \log p(x_t = x_t^{wt} | x_{/T}), \quad (14)$$

Here the sum is summed over all mutation positions, and the sequence input to the model is masked at each mutation position.

Riesselman et al. (Riesselman et al., 2018) collected a set of 41 deep mutation scans and thus evaluated the model, these scans consisted of assessing the activity of a set of different proteins on various tasks. The experiments corresponded to different measurements in different tasks with different functions being tested. Each deep mutation scan dataset was treated as a separate prediction task and the model was used to score each variant in the dataset. These tasks were divided into a validation set consisting of ten mutation scan datasets and a test set consisting of the remaining datasets. Protein representation quality was assessed by comparing scores to experimental measurements using Spearman rank correlation.

Rank correlation, also known as rank correlation, is a non-parametric statistical method, but does not require a distribution of the original variable. It is suitable for information that does not follow a normal distribution, as well as information where the overall distribution is unknown and the original data is expressed in terms of rank. The Spearman rank correlation coefficient is used to calculate the correlation between two ordered sets of data. The formula for calculation is:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (15)$$

where  $d_i = (x_i - y_i)$ ,  $x_i$  and  $y_i$  denote the position of the two variables after sorting, and  $n$  is the total number of samples.

The absolute value of Spearman’s rank correlation coefficient is less than 1. The magnitude of the absolute value represents the strength of the correlation, and an absolute value of 1 indicates that the two series are perfectly correlated. When  $r_s$  is positive, then the two sequences are positively correlated; when  $r_s$  is negative, then the two sequences are negatively correlated.

The correlation between the mutation scores predicted by the model and the mutation scores generated by the real experiment is calculated, and the Pearson rank correlation is used here to return the correction. Correction is closer to 1, which means that the correlation between the two sequences is greater, and in this paper, we use the model to calculate the confusion degree of the amino acid sequences, so as to realise the zero-shot prediction of the activity of the protein. In this paper, we use the model to calculate the perplexity of amino acid sequence to achieve zero-shot prediction of protein activity, and the results are shown in Table 1.

## F Activity Prediction

Protein mutations affect the activity of protein function, and protein activity was predicted using linear regression modelling of coded proteins regression, with the mean square error used to assess the prediction. The biological activity of protein mutations is predicted by training a simple mutation prediction network, i.e., using an embedded representation of the protein coding model to the amino acid sequence. In this paper, a simple multilayer perceptron is used to predict the mutation effect of protein sequences. The embedding representations of the dataset are saved by pre-computation. Afterwards a test of the downstream task is completed using these embedding representations and the activity corresponding to each representation. In this test, the activity of  $\beta$ -lactamase variants is predicted by training a model. The data was initially obtained from deep mutation scans and published by the Envision paper (Gray et al., 2018). The dataset was taken from a fasta, and each entry in the file contains: the mutated  $\beta$ -lactamase sequence, a residue in which the mutation occurs (exchanged for another amino acid) and a target value corresponding to the activity of the mutated protein, which describes the magnitude of the strength of the protein

to act on a function and is a real-valued value. The quality of the protein representation was assessed by comparing the mean square error between the predicted activity values of the model and the experimentally measured activity values, and the results are shown in Table 1.

## G Protein-protein interactions

PPIs have so far been studied from different perspectives (e.g. signalling, biochemistry, etc.) using a variety of approaches. Protein-protein interaction networks can then be constructed using this information. The use of machine learning methods to complete the complementation of large PPI networks and to make inference predictions on PPI networks is a major research topic in computational biology. In order to explore whether the protein representation model trained in this paper can embed protein interaction relationships into protein representation, the human protein-protein interaction network (Agrawal et al., 2018) provided by Stanford University is selected in this paper. This is a protein-protein interaction network that contains physical interactions between proteins in the experimentally recorded human body, such as metabolic enzyme coupling interactions and signalling interactions. The nodes on the graph represent human proteins and the edges represent physical interactions between proteins in human cells (see Table 6).

Consider each edge in the table 6 as a positive example and use the randomly generated edges as negative examples. The final positive and negative examples are divided in the training and test sets in the ratio of 9 : 1, respectively. The two proteins were encoded into vectors with the model separately, spliced and then directly predicted by a binary neural network, with prediction accuracy as the evaluation metric, and the results are shown in Table 1.

Table 6: Statistics of human protein-protein interaction

Attribute	Quantity
Nodes	21,577
Edges	342,353
SCC Node Count	21,521
SCC Edge Count	342,316
Triangle Count	55,614,585
Closed Triangle Ratio	0.045652
Diameter	8

## H Function Prediction

The Gene Ontology resource (GO) is the most comprehensive and widely used knowledge base on gene function. Therefore, in this paper, we test whether the protein embeddings generated by the protein representation model proposed in this paper can accurately predict the functions of protein sequences by collecting the functional descriptions of proteins from the Gene Ontology database. Functional descriptions corresponding to each class of proteins can be extracted from the protein database, and a total of 39,168 human proteins with 23,666 protein functions contained in all annotations were extracted from the UniprotKB database, and the model is required to accurately predict the correct functional option for a defined protein, given its functional description and three randomly generated functional descriptions. Detailed statistical information about the dataset is given in Table 7.

In this paper, we use PubMedBERT (Gu et al., 2021) to obtain a representation of the functional description. PubMedBERT is a pre-trained model trained from scratch from a large mixed dataset containing knowledge of a range of biological literature on diseases, drugs, genes, organs, cells, etc. PubMedBERT uses masks on all of this data to predict the task. These data were pre-trained using the mask prediction task and fine-tuned and achieved state-of-the-art performance on a variety of tasks such as named entity recognition, PICO, relation extraction, semantic similarity, document classification, knowledge quizzing, and more. Each character-level encoding is obtained by feeding the functional description of the protein into PubMedBERT, and all these encodings are averaged to obtain the encoding of the sentence. The sentence representation should contain rich semantic information, and a linear layer is used to project the protein representation into the representation space of the functional description. Afterwards, the similarity between the functional description and the protein representation is computed using vector dot product and the one with the highest similarity is selected as the correct option. The accuracy of selecting the correct option using the model is used as an evaluation metric of how good the protein representation is, and the results are shown in Table 1.

Table 7: Protein function prediction dataset statistics

Attribute	Quantity
Protein Quantity	39,168
Protein function quantity	23,666
Number of Annotation Functions	45,877
Average Length of Functional Description	69.327
Functional Description Vocabulary	30,522

## I Homology Detection

We use SCOPe (Fox et al., 2014) to assess remote homology detection. According to standard practice (Söding and Remmert, 2011), Rossmann-type folds (c.2-2.5, c.27 and 28, c.30 and 31) and four- to eight-bladed propellers (b.66-b.70) were excluded from this paper. Proteins from the same superfamily belong to different families. Vector inner products were used directly to measure the structural and functional similarity of the protein representations, taking the highest 10 scored proteins to consider them as remotely homologous to the target proteins, using the Hit-10 metric, i.e., the proportion of proteins corresponding to the top 10 scores and the target proteins that do belong to the same protein superfamily. The experiments are performed by measuring the density of homologous proteins near the query sequence based on an unsupervised classifier that measures the similarity between protein representations based on the inner product of vectors. For each domain, a vector similarity query will be performed on all other domains and they will be sorted by distance to the query domain. For superfamily level assessment, any domains with the same superfamily and not belonging to the same family are considered as positive examples; any domains with different superfamilies are considered as negative samples. The metrics were measured using Hit-10, which gives the proportion of correctly predicted remote homologues among the ten highest ranked results. The experimental data is consistent with the ESM2 paper, and the final results are shown in Table 1. In order to distinguish the effect of the model more clearly, this paper also increased the difficulty of the experiments by using only protein sequences with sequence similarity below 40% to participate in the experiments, and at the same time, due to the limitation of computational resources, only proteins of the  $\alpha$  folding family were used. The experiment verifies that the protein representation can detect remote homology between proteins belonging to the same superfamily (but belonging to different



Table 8: Remote homology detection

Model	SF-Hit@10	SF-Hit@5	SF-Hit@1	F-Hit@10	F-Hit@5	F-Hit@1
ESM2-12layer	47.04%	44.33%	28.52%	14.67%	9.53%	2.27%
OurModel	50.00%	47.43%	31.85%	17.47%	10.14%	2.72%

families) as well as between proteins belonging to the same fold (but belonging to different superfamilies). The final sample size of the experiment was 2644, and the correlation between two two proteins was calculated using cosine similarity, and the top 10 proteins were taken as the homologous proteins detected by the model. The final calculations were done separately (Hit@10: the proportion of proteins ranked in the top 10 of similarity that are homologous proteins, Hit@5 and Hit@1), and the final results are shown in Table 8, where SF stands for the results on superfamily homology detection, and F stands for the results on folding homology detection.

From Table 8, it can be seen that when the sequence similarity of the proteins in the dataset used for testing is reduced to below 40%, the accuracy of the model on the remote homology detection task will be greatly reduced, and the model proposed in this paper performs significantly better than the ESM2 on this task as well. It is argued that this result arises due to the fact that the model proposed in this paper is introduced into the training of the protein family data, the protein representation itself spatially conforms to the reticulation distribution of protein families, in other words, the protein representation contains the functional information of proteins, so that functionally similar proteins can still be recognised better when sequence similarity decreases. Both models perform poorly in protein folding homology detection when sequence similarity decreases, probably because the folding structure of proteins itself is highly correlated with sequence similarity.