Predicting Protein-Protein Interaction via Protein Textual Description using Large Language Model

Stanford CS224N Custom Project

Khoa Hoang
Department of Bioengineering
Stanford University
khoang99@stanford.edu

Abstract

Protein-protein interactions (PPI) are pivotal for various biological functions. Traditionally, experimental techniques like mass spectrometry and immunoprecipitation, as well as methodologies like the yeast two-hybrid system, have been employed to determine PPI. However, computational methods have also emerged as valuable tools for PPI prediction. While many computational approaches rely solely on protein sequence and structure data, they often overlook the valuable contextual information available in textual descriptions of proteins. This paper introduces the utilization of a pre-trained large language model, PubMedBERT, to generate embeddings of protein functional descriptions sourced from UniProt. Through fine-tuning the pre-trained model under various conditions, such as fixing pretrained weights or fine-tuning them with LoRA, notable differences of 0.4-0.8 standard deviations between embeddings of interacting and non-interacting proteins are observed. In contrast, methods like TF-IDF yield insignificant disparities between the embeddings. An analysis of the results reveals that the fine-tuned model not only captures keywords but also identifies conceptually similar terms when generating embeddings, showcasing its ability to grasp the nuanced context of protein functional descriptions.

1 Key Information to include

- Mentor: Tathagat Verma
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Determining Protein-Protein Interactions (PPI) has long been a significant research domain, with numerous experimental and computational methodologies developed for this purpose (Rao et al. (2014)). Experimental techniques for determining PPI can be resource-intensive and time-consuming. Computational methods, such as Rosetta (|Lyskov and Gray (2008)), Alphafold (Jumper et al. (2021)), or ESM (Lin et al. (2023)) for docking simulation or complex prediction, often require considerable computational time. For instance, Colabfold typically takes about 30 to 45 minutes to predict a protein complex. Given the computational overhead involved in these methods, it becomes crucial to limit the number of candidate proteins for analysis. One approach to achieve this is by leveraging the functional descriptions of proteins to generate embeddings. These embeddings can then be compared

[1] Please ensure to verify the time required for Colabfold predictions, as it may vary depending on factors like hardware specifications and the size of the protein complex.

and filtered to identify likely interacting proteins. Advancements in natural language processing have made this task feasible through the use of large language models (LLMs) pretrained on extensive scientific text corpora, such as PubMedBERT. If successful, such text-based models can serve as effective filters for protein candidates before employing other sequence-based or structure-based models for PPI prediction, thereby enhancing throughput of computational methods in PPI.

3 Related Work

Generating text embeddings using Large Language Models (LLMs) to represent biological entities is not a novel concept, as evidenced by prior research such as BioTranslator (Xu et al. (2023)). BioTranslator introduces a multilingual translation framework capable of converting various biological modalities into text representations. Its applications range from identifying novel cell types to predicting protein functions and learning drug-phenotype-gene relationships in an unsupervised manner. Notably, BioTranslator surpasses traditional text embedding methods like TF-IDF, Doc2Vec, and ClusDCA in protein functional prediction and stands out as a pioneering approach capable of predicting relationships between different biological modalities without relying on paired data. However, while BioTranslator excels in several tasks, its exploration in predicting Protein-Protein Interactions (PPI) remains absence. This omission might stem from the inherent complexity of PPI prediction, where proteins with divergent textual descriptions can still interact, necessitating models to capture interaction-based ontological distances. To the best of our knowledge, this study marks the first attempt to utilize LLMs for predicting PPIs through text embeddings, presenting an avenue for further exploration in this area.

4 Approach

A Siamese network architecture is employed, leveraging the PubMedBERT model as an encoder to derive embeddings from functional description text Gu et al. (2020). Accompanying this architecture is a Head module, comprising two linear layers, tasked with embedding projections and reducing the encoder's output dimensionality from a 768-dimensional vector to 128 dimensions (Figure 1). Throughout each training iteration, a fixed number of proteins are designated as anchors. For each anchor, a positive example is selected alongside several negative examples, all of which were used as input for embeddings generation. These embeddings are then subjected to negative sampling loss within the Siamese network framework. During training, two strategies are employed: either pre-trained weights are kept fixed (head fine-tuning), or they undergo fine-tuning with LoRA (Chen (2021)). As a baseline, TF-IDF is utilized to generate embeddings for each protein text description. Ultimately, four distinct sets of embeddings for functional texts are acquired: TF-IDF, pretrained embeddings, Head fine-tuned embeddings, and LoRA fine-tuned embeddings. To assess the efficacy of each method, cosine similarity calculations are performed between normalized embeddings of a query protein (the anchor protein) description and other protein candidates descriptions. Then, the difference in mean embeddings between interacting and non-interacting proteins is computed. The trained model is anticipated to yield higher cosine similarity scores among interacting proteins and lower cosine similarity scores among non-interacting proteins compared to the TF-IDF baseline.

$$\text{Negative Sampling Loss} = -\log\left(\sigma(\mathbf{v}_{p_p} \cdot \mathbf{v}_{p_c})\right) - \sum_{p_j \in \text{Negatives}} \log\left(\sigma(-\mathbf{v}_{p_j} \cdot \mathbf{v}_{p_c})\right) \tag{1}$$

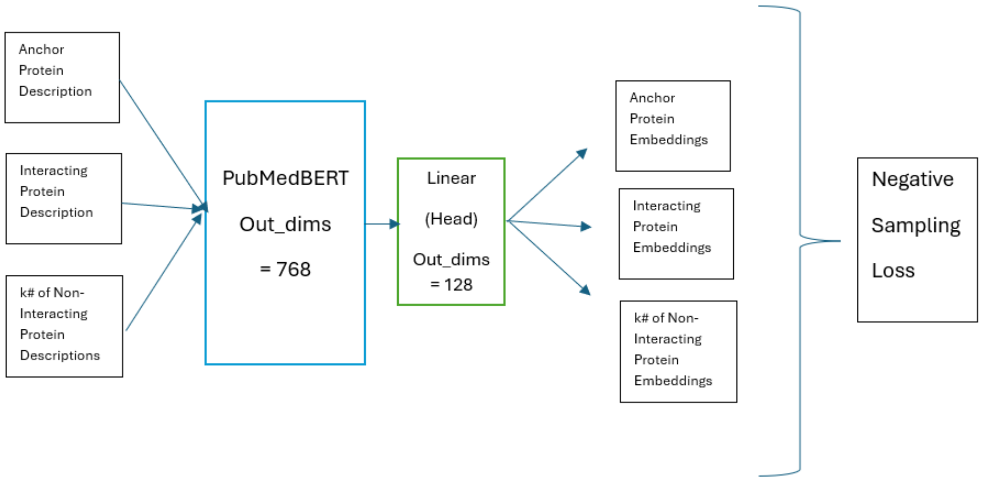

Anchor Protein Description
Interacting Protein Description
k# of Non-Interacting Protein Descriptions
PubMedBERT
Out_dims
= 768
Linear
(Head)
Out_dims
= 128
Anchor Protein Embeddings
Interacting Protein Embeddings
k# of Non-Interacting Protein Embeddings
Negative
Sampling
Loss

Figure 1: Simplified model architecture

5 Experiments

5.1 Data

The dataset was sourced from `https://www.uniprot.org/`, focusing on human proteins with accompanying textual descriptions. Following filtering, the dataset comprised approximately 17,000 proteins, each accompanied by a textual description and information regarding interacting proteins. On average, these descriptions spanned around 90 words. Notably, approximately 40% of the proteins lacked interacting partners, potentially attributed to database curation. Consequently, these proteins were omitted from the dataset, resulting in a refined set of around 10,000 proteins with descriptions and at least one interacting partner. To ensure uniformity, any PubMed reference numbers present within the functional descriptions were removed during preprocessing, and all text was converted to lowercase.

Example of protein descriptions data

Anchor protein: glycosyltransferase that generates the core 1 o-glycan gal-beta1-3galnac-alpha1-ser/thr (t antigen), which is a precursor for many extended o-glycans in glycoproteins. plays a central role in many processes, such as angiogenesis, thrombopoiesis and kidney homeostasis development

Interacting protein: probable chaperone required for the generation of 1 o-glycan gal-beta1-3galnac-alpha1-ser/thr (t antigen), which is a precursor for many extended o-glycans in glycoproteins. probably acts as a specific molecular chaperone assisting the folding/stability of core 1 beta-3-galactosyltransferase (c1galt1)

Non-interacting protein: forms a water-specific channel that participates in distinct physiological functions such as glomerular filtration, tubular endocytosis and acid-base metabolism

5.2 Evaluation method

To assess the efficacy of the four embedding methods – TF-IDF, pretrained PubMedBERT, Head fine-tuned, and LoRA fine-tuned – embeddings were generated for all protein descriptions using each respective approach. Subsequently, pairwise cosine similarity calculations were conducted for all proteins. To ensure comparability, the similarity values were normalized, ensuring that each anchor protein exhibited a mean similarity of 0 and a standard deviation of 1 when compared to all other proteins. Following normalization, the difference in mean similarity was computed for interacting proteins and non-interacting proteins associated with each anchor protein

5.3 Experimental details

- Data inputs: 50 anchor proteins were selected for each train, validation, and test set out of 10,000 proteins. The splits were ensured so that there were no overlapping proteins between sets to avoid information leakage
- Batch size: Batch size of 12 were used
- Number of negative examples: 6 non-interacting proteins were used in negative sampling.
- Number of epochs: Each model was run for 500 epochs. Models were saved every 100 epochs and models with lowest loss values were used for each training strategy
- Learning rate: 0.001
- Optimizer: Adam optimizer was used
- LoRA parameters: peft module of Hugging Face was used with settings

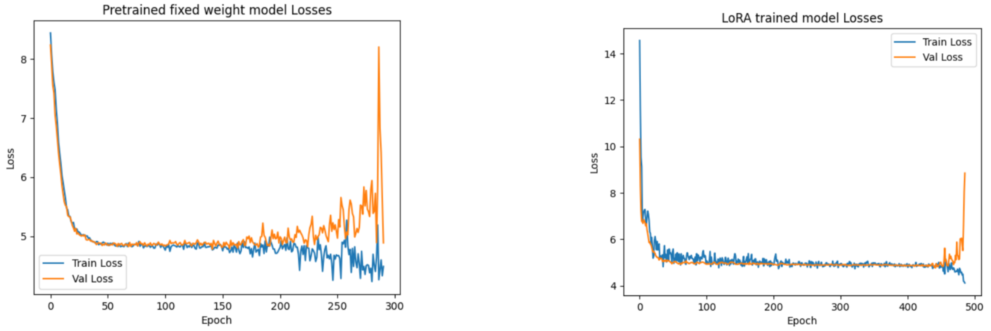
```
lora alpha=8, lora dropout=0.05, r=6, bias="none", target modules="all-linear"
```

5.4 Results

Table 1: Loss value for different embeddings

	Train Loss	**Validation Loss**	**Test Loss**
Pre-trained	64.14	63.72	63.71
Head fine-tuned	4.96	4.95	4.95
LoRA fine-tined	**4.91**	**4.91**	**4.91**

[1]



(a) Head-fine-tuning

(b) LoRA fine-tuning

Figure 2: Train and Validation Loss of fine tuning models

[1] Loss for TF-IDF not report since TF-IDF produce different embedding dimensions, not comparable to other models loss

Table 2: Similarity Difference for different embeddings

	Train Similarity Difference	**Validation Similarity Difference**	**Test Similarity Difference**
TF-IDF	-0.120	0.065	-0.013
Pre-trained	0.337	0.434	0.775
Head fine-tuned	0.410	**0.600**	0.868
LoRA fine-tined	**0.432**	0.474	**0.869**

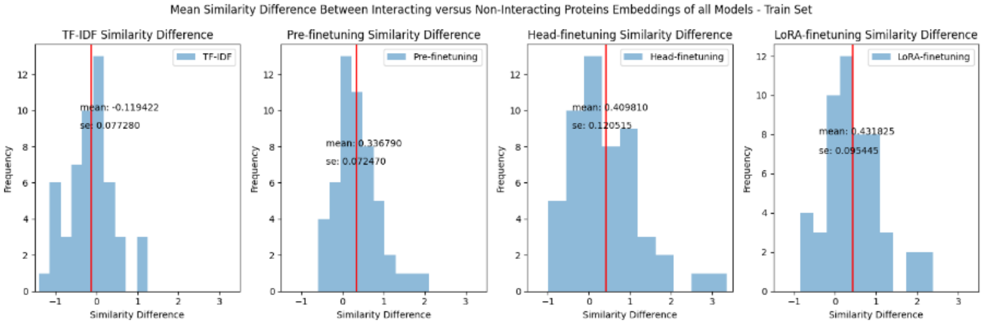


Figure 3: Train data similarity difference

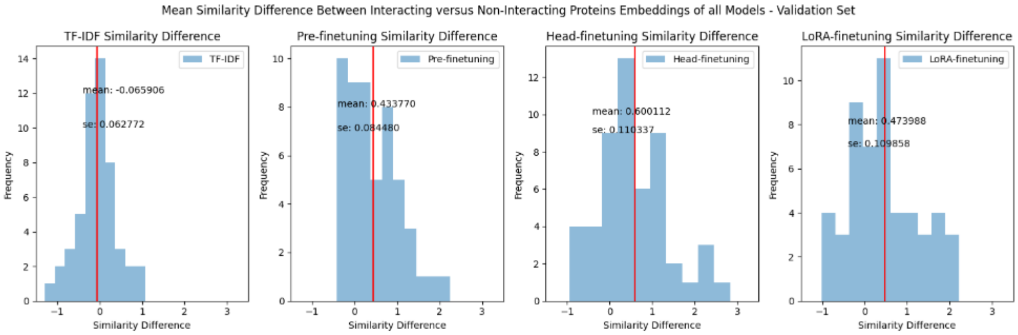


Figure 4: Train data similarity difference

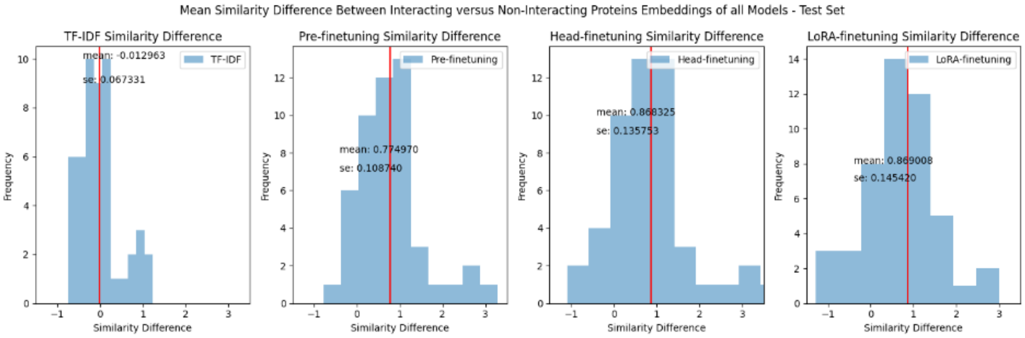


Figure 5: Train data similarity difference

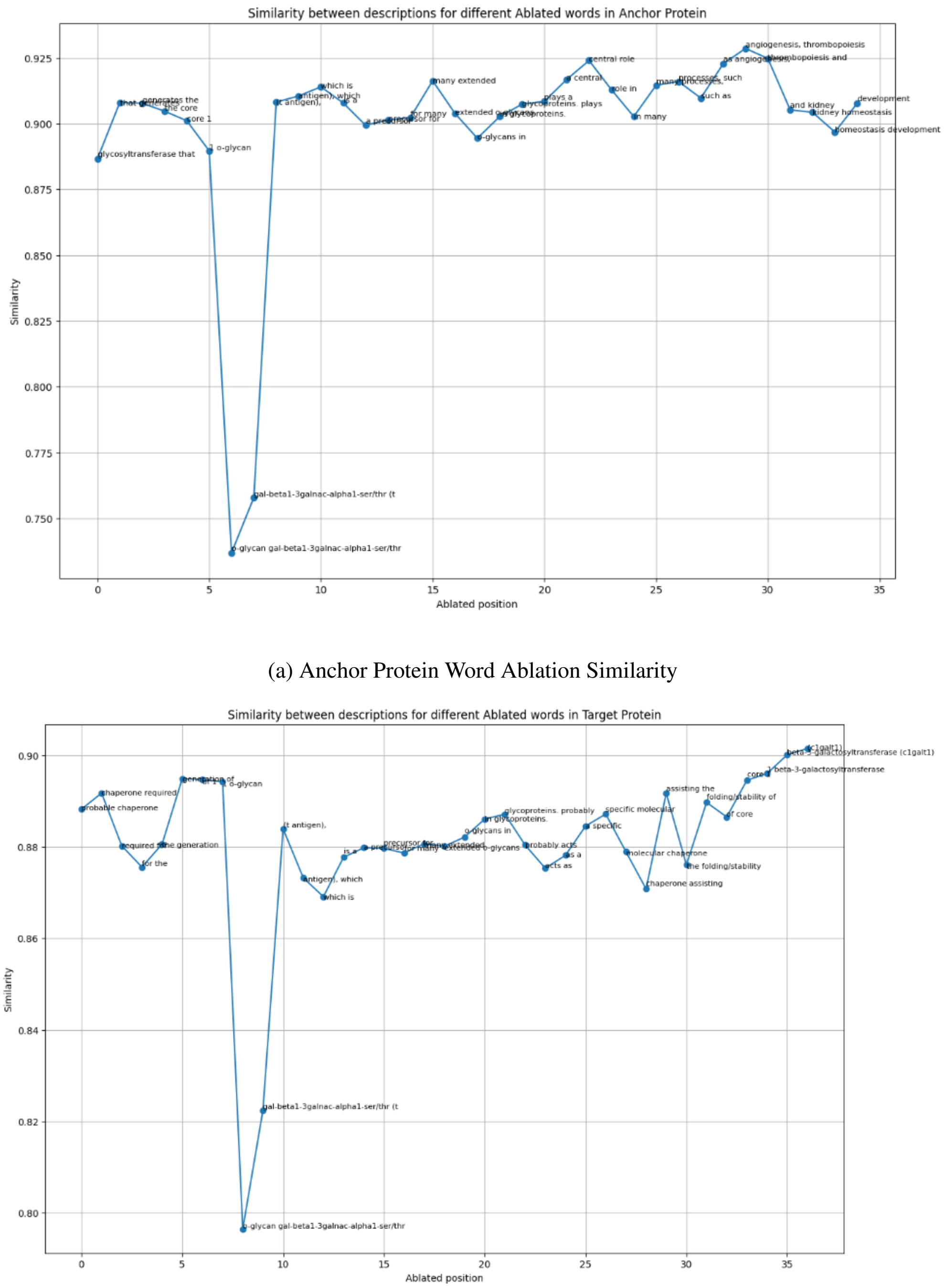
LoRA fine-tuned model performs best in loss and similarity difference. However, there are large deviations in metrics between train, validation, and test set which may due the size of data input. Although LoRA-fine-tuned model performs best, there was not much improvement between fine-tuning the entire model or just the head module, suggesting that either the pretrained embeddings already capture the notion of interatcions or the data input was not sufficiently large to see the difference.

6 Analysis

To delve into the information encapsulated within the embeddings, we closely examined two instances of interacting protein descriptions. In the first scenario, both descriptions contained the substring "1 o-glycan gal-beta1-3galnac-alpha1-ser/thr (tantigen), which is a precursor for many extended o-glycans in glycoproteins." Consequently, it wasn't surprising to observe a high degree of similarity between these descriptions. We conducted a text ablation experiment where certain words in either the anchor or interacting protein were replaced with "()" to discern the pivotal components contributing to embedding similarity. Interestingly, we found that removing the phrase "1 o-glycan gal-beta1-3galnac-alpha1-ser/thr" resulted in a substantial drop in similarity (6), indicating its significant role. Confirming this observation, removing "1 o-glycan gal-beta1-3galnac-alpha1-ser/thr" from both anchor and interacting protein led to a decrease in similarity from above 0.9 to just above 0.6. Conversely, when a longer substring "which is a precursor for many extended o-glycans in glycoprotein" was removed from both descriptions, the similarity remained in the 0.9 range. This illustrates that the Language Learning Model (LLM) pays particular attention to specific keywords rather than merely comparing similarity based on word frequency, such as the TF-IDF method.

Simple case of high similarity embeddings

Interacting protein
glycosyltransferase that generates the core 1 o-glycan gal-beta1-3galnac-alpha1-ser/thr (t antigen), which is a precursor for many extended o-glycans in glycoproteins. plays a central role in many processes, such as angiogenesis, thrombopoiesis and kidney homeostasis development
Anchor protein
probable chaperone required for the generation of 1 o-glycan gal-beta1-3galnac-alpha1-ser/thr (t antigen), which is a precursor for many extended o-glycans in glycoproteins. probably acts as a specific molecular chaperone assisting the folding/stability of core 1 beta-3-galactosyltransferase (c1galt1)



(a) Anchor Protein Word Ablation Similarity

(b) Interacting Protein Word Ablation Similarity

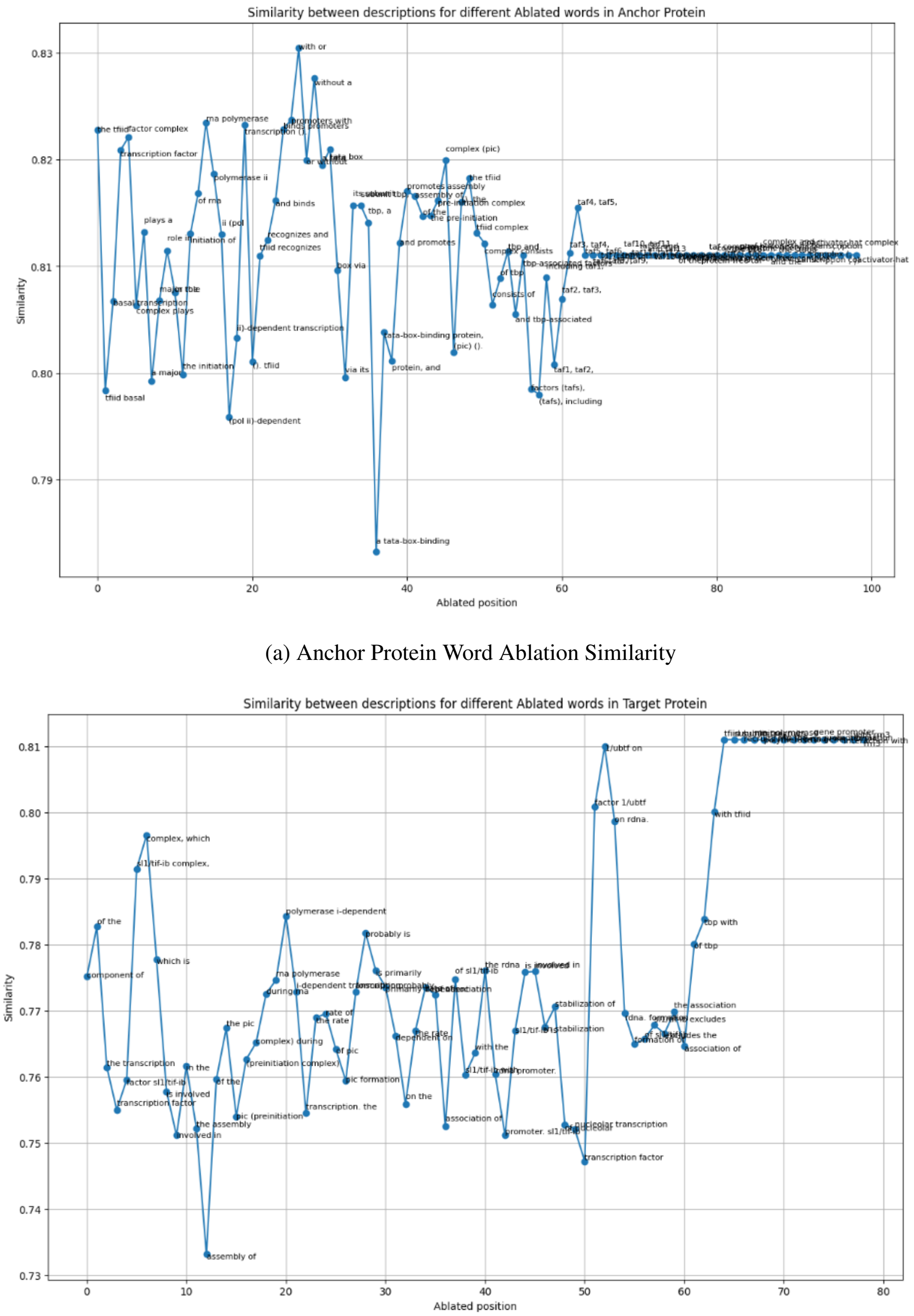
Figure 6: Word ablation analysis on high similarity proteins - simple case

In another scenario where two descriptions exhibit high similarity without apparent common substrings (7), the ablation analysis revealed intriguing insights. The model demonstrated attention towards specific words, such as "tata-box" in one protein description and "assembly of pic" in another, implying the acquisition of certain conceptual relationships between PIC and tata-box. However, further comprehensive analysis is required to validate this hypothesis.

Subtle case of high similarity embeddings

Interacting protein
component of the transcription factor sl1/tif-ib complex, which is involved in the assembly of the pic (preinitiation complex) during rna polymerase i-dependent transcription. the rate of pic formation probably is primarily dependent on the rate of association of sl1/tif-ib with the rdna promoter. sl1/tif-ib is involved in stabilization of nucleolar transcription factor 1/ubtf on rdna. formation of sl1/tif-ib excludes the association of tbp with tfiid subunits. recruits rna polymerase i to the rrna gene promoter via interaction with rrn3

Anchor protein
the tfiid basal transcription factor complex plays a major role in the initiation of rna polymerase ii (pol ii)-dependent transcription (). tfiid recognizes and binds promoters with or without a tata box via its subunit tbp, a tata-box-binding protein, and promotes assembly of the pre-initiation complex (pic) (). the tfiid complex consists of tbp and tbp-associated factors (tafs), including taf1, taf2, taf3, taf4, taf5, taf6, taf7, taf8, taf9, taf10, taf11, taf12 and taf13 (). component of the tata-binding protein-free taf complex (tftc), the pcaf histone acetylase complex and the staga transcription coactivator-hat complex (, , , , ,)



(a) Anchor Protein Word Ablation Similarity

(b) Interacting Protein Word Ablation Similarity

Figure 7: Word ablation analysis on high similarity proteins - subtle case

7 Conclusion

This paper demonstrates the utility of Language Learning Models (LLMs) in generating text embeddings for protein function, facilitating the selection of potential interacting proteins. However, a notable limitation of this study lies in the relatively small sample size of protein descriptions used for model fine-tuning. Nonetheless, there is promising potential for enhancing model performance through fine-tuning on a larger dataset.

References

Weizhu Chen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, and et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, and et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

S. Lyskov and J. J. Gray. 2008. The rosettadock server for local protein-protein docking. *Nucleic Acids Research*, 36(Web Server).

V. Srinivasa Rao, K. Srinivas, G. N. Sujini, and G. N. Kumar. 2014. Protein-protein interaction detection: Methods and analysis. *International Journal of Proteomics*, 2014:1–12.

Hanwen Xu, Addie Woicik, Hoifung Poon, Russ B. Altman, and Sheng Wang. 2023. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14(1).