Exploring Large Language Models for Protein Interaction Prediction in the Presence of Cancer

Ryan Engel and Gilchan Park

Computational Science Initiative
Brookhaven National Laboratory

Abstract

Protein interactions are central to the complex machinery of cellular processes and hold significant promise in unravelling the mysteries associated with various diseases. A detailed understanding of these interactions can provide insights into molecular biology and facilitate breakthroughs in research, further advancing therapeutic discovery for many diseases. Current literature predominantly concentrates on predicting general Protein-Protein Interactions (PPIs), with limited research dedicated to disease-specific PPI predictions. Additionally, emerging LLMs have shown significant promise in various biological applications. This study aims to combine these existing research ideas by leveraging LLMs specifically for predicting protein interactions in the context of certain diseases. Our approach involves fine-tuning various LLMs such as Llama-2, Galactica, Falcon, MPT, BioMedLM, and BioGPT for predicting whether two proteins interact in the presence of cancer. Additionally, we evaluate the performance of Low Rank Adaptation (LoRA), a Parameter Efficient Fine-Tuning (PEFT) algorithm, and analyze its ability to enhance the computational efficiency of LLMs for this task. The results show that between each of the models studied, the Falcon model has the best performance, predicting protein interactions in the presence of cancer to approximately 84% accuracy. This marks a significant enhancement when fine-tuning with LoRA, compared to the 53% accuracy without any fine-tuning.

1 Introduction

Recent years have witnessed transformative advancements in Natural Language Processing (NLP), marked by the advent of Large Language Models (LLMs). Equipped with remarkable capabilities for processing and understanding text data, these models have begun to redefine the boundaries of various domains, including molecular biology.

A main goal of this study is to see how well LLMs work in understanding

complex relationships between molecules, like those present in cancer diagnoses. The hypothesis is that the LLMs will learn distinctions between different protein names through the thousands of data samples we provide during training. Given that the naming conventions in molecular biology and chemistry inherently encode relationships and characteristics within the syntax of protein names, it is anticipated that these trained language models will demonstrate proficiency in learning and predicting protein interactions.

By combining these advanced language models with knowledge of molecular biology, we hope to find new ways to better understand and predict interactions between proteins linked to diseases. This could help improve cancer research and our overall knowledge in molecular biology.

In this paper, our contribution is threefold:

1. Fine-tuning various LLMs for prediction of protein interactions when prompted with two target proteins.
2. Determining which LLMs are most suitable for this task.
3. Comparing the original pre-trained models to the models trained with LoRA, to determine which strategy is more effective and computationally efficient.

2 Related Works

With recent advancements in computational biology, many have begun to harness the power of deep learning for biological tasks. AlphaFold [5], for example, represents a significant breakthrough by employing a novel machine learning approach that accurately predicts protein structures with atomic precision.

Additionally, attention-based language models like ProGen2 have been introduced [6]. ProGen2 demonstrates state-of-the-art performance in tasks such as capturing evolutionary sequences, and predicting protein fitness without additional fine-tuning, underlining the increasing importance of large-scale language models in computational biology.

In a related study [7], a new method for identifying protein-protein interactions (PPIs) has been introduced. This method uses the Graph-BERT, ProtBERT, and SeqVec language models to better understand and represent the details within a PPI network graph. By applying these models, it manages to effectively use protein sequence information to improve the classification of PPIs, performing better than previous methods. This approach yields the highest results in general protein interaction prediction, demonstrating the promising potential of using language models for this task.

A different approach that has been gaining popularity is the use of convolutional neural networks (CNNs) for PPI prediction. Wang et al. [2] investigated this by combining a feature representation method, Bio2Vec, with a CNN to develop a language processing model aimed at predicting PPIs from protein sequences. Another study [3] introduced a deep learning framework called DPPI, which also employs a CNN to predict PPIs using sequence information.

Moreover, [4] present a new method called PIPR for predicting PPIs using only protein sequences. Unlike traditional models that depend on complicated and hard-to-obtain features, PIPR makes the process simpler by directly using the information from protein sequences. This new approach does a better job than many existing systems, not only in basic PPI predictions but also in more complex tasks like figuring out the type of interaction and estimating binding strength between proteins.

Many existing methods primarily concentrate on predicting and discovering general protein interactions, with limited exploration of deep learning for PPI prediction in the context of specific diseases. The NECARE model [1] stands out in this regard, exhibiting state-of-the-art performance in PPI prediction specifically related to cancer. It employs a deep learning framework, utilizing a Relational Graph Convolutional Network (R-GCN), enabling it to surpass all of the top models in predicting cancer-associated protein interactions.

In this paper, we leverage the carefully curated dataset presented in the NECARE paper, combined with the advanced capabilities of modern LLMs and fine-tuning strategies, to create a protein interaction prediction model that excels in the domain of cancer biology. This approach aims to improve the precision of cancer-related PPI predictions, setting the stage for the development of more refined prediction models tailored to specific diseases.

3 Dataset

The dataset used for the fine-tuning process of the LLMs in this study is the training dataset used by the NECARE model [1], made up of protein-protein interactions in the presence of cancer. For the positive training set, data were collected from three sources: the KEGG, Reactome, and OncoPPI databases. These databases provide information on known cancer PPIs. For the negative training set, the dataset incorporated pairs from the KEGG database that were marked with "disassociation/missing interaction" or other indicators signifying a lack of interaction in cancer-related pathways.

In total, there are 933 positive interactions and 1308 negative interactions, adding up to a total of 2241 data points. We used this set to train the LLMs, and formulated the problem as a binary classification task, teaching the models to predict either positive or negative interaction between two target proteins.

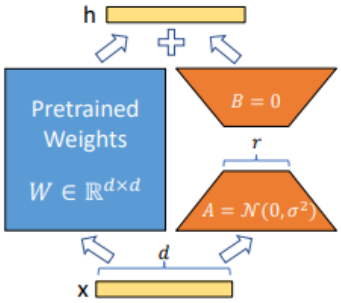

Figure 1: LoRA Reparameterization. Source (9)

4 PEFT & LoRA

There has been extensive research in the domain of LLMs, specifically in developing parameter efficient algorithms for fine-tuning them. Reference [8] goes over 30 PEFT strategies, and their tradeoffs. They conclude that LoRA [9], IA3 [10], and Prompt Tuning [11] are among the most scalable techniques. In this paper, we will make use of the LoRA strategy to minimize the computational load of the training process.

Parameter Efficient Fine-Tuning (PEFT) is an approach that focuses on fine-tuning large pre-trained models without adjusting all the parameters. The idea is to only modify a small subset of the model's parameters, making the process more efficient, especially for very large models. PEFT allows for customizing the model to specific tasks or datasets while requiring significantly less computational resources compared to traditional fine-tuning methods that update all parameters.

Low-Rank Adaptation (LoRA) is a specific type of PEFT algorithm where the adaptation is achieved by introducing low-rank matrices into the pre-trained model. In this method, the original weight matrices of the model are not directly modified. Instead, small rank matrices are introduced that interact with the original weights. This allows for efficient and effective fine-tuning, as only these low-rank matrices are updated during training. LoRA is particularly useful for adapting LLMs, as it offers a balance between adaptability and computational efficiency.

In conventional deep learning architectures, dense layers typically perform matrix multiplications using fully-ranked weight matrices. However, during adaptation, these matrices often exhibit a naturally low rank. The modification of a pre-existing weight matrix, $W_0 \in \mathbb{R}^{d \times k}$, is encapsulated via a low-rank decomposition formula:

$$W_0 + \Delta W = W_0 + BA$$

Here, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ represent the newly introduced low-rank matrices, where r—significantly smaller than the dimensions of W_0—denotes the rank. In this setup, W_0 remains static during the training phase, while A and B are the dynamic elements subject to training, thus allowing the model to adapt efficiently with minimal parameter adjustments.

5 Training

5.1 Procedure

Six different LLMs were trained, including Llama-2-7b, Galactica-7b, Falcon-7b, MPT-7b, BioMedLM-2.6b, and BioGPT-1.6b. The data was formatted into a prompt and a label, where the prompt asks "Is there a protein interaction between ProteinA and ProteinB?" and the label is either "yes" or "no". The goal was to train the language models to accurately predict whether there is an interaction of the two proteins.

To give a more robust generalization of each model's performance on the dataset, we employed 5-fold cross validation to help understand how well the model was performing on unseen data. To elaborate, we split the dataset into 5 subsets or folds, then for each fold, 60% of the data is used for training, 20% is the validation set, and the last 20% is the test set. This process is repeated for each fold, where we rotate the data split such that no two folds have the same test set. Furthermore, the testing set was never used during the training or validation procedures. Using each of the 5 folds, we train a separate model, and evaluate that model's performance on the validation set of its respective fold. At the end, we load the model that achieved the lowest validation loss and evaluate this model on the test set from its respective fold.

5.2 Hyper-Parameters

The hyperparameters used for all models during training are as follows: learning rate = 1e-5, number of epochs = 20, batch size = 16, weight decay = .05, prompt length = 18, number of new tokens generated = 2.

Furthermore, the LoRA configuration parameters remained constant through these experiments as well and these consist of r = 16, lora alpha = 8, lora dropout = .1. The LoRA target modules for Galactica, Llama-2, and BioGPT used ['q_proj', 'v_proj']. For the Falcon model, I used ['query_key_value'], for MPT I set this to ['Wqkv'], and for BioMedLM this was set to ['c_attn', 'c_proj'].

5.3 Hardware

All experiments were conducted using the Nvidia A100 GPUs at the Computational Science Initiative's High Performance Computing (HPC) cluster. This

state-of-the-art HPC environment is engineered to handle extensive computational loads and complex data analyses, making it an ideal setting for computationally intensive tasks.

6 Results

Model	Acc.	MCC	AUC	Spec.	Prec.	F1
Falcon-7b	**0.5268**	-0.1912	0.4591	**0.9183**	0.2763	0.3450
Galactica-7b	0.4397	**0.0592**	**0.5097**	0.0350	**0.5906**	0.3333
MPT-7b	0.2344	-0.5544	0.2211	0.3113	0.2245	0.2227
Llama-2-7b	0.4330	-0.1112	0.4447	0.3658	0.4442	**0.4329**
BioMedLM-2.7b	0.0067	-0.9864	0.0058	0.0117	0.0077	0.0067
BioGPT-1.6b	0.0647	-0.8776	0.0564	0.1128	0.0659	0.0608

Table 1: Comparison of original pre-trained LLMs

Model	Acc.	MCC	AUC	Spec.	Prec.	F1
Falcon-7b	**0.8438**	**0.6795**	**0.8383**	0.8755	**0.8413**	**0.8396**
Galactica-7b	0.8326	0.6562	0.8259	0.8716	0.8304	0.8278
MPT-7b	0.8103	0.6094	0.7990	0.8755	0.8106	0.8028
Llama-2-7b	0.7924	0.5743	0.7740	0.8988	0.8009	0.7799
BioMedLM-2.7b	0.6629	0.3211	0.6121	**0.9572**	0.7299	0.5842
BioGPT-1.6b	0.5871	0.1721	0.5870	0.5875	0.5851	0.5839

Table 2: Comparison of LLMs trained with LoRA

The results of these experiments give us a comprehensive analysis of the performance of each model under a variety of different conditions. To get a quantitative understanding of each model's performance, we gathered metrics of accuracy, matthews correlation coefficient (MCC), area under the roc curve (AUC), specificity, macro precision, and macro F1 score for each task.

First, I analyzed each of the original pre-trained model on the task. It is clear that the Falcon LLM performed best without any fine-tuning, predicting protein interactions with about 53% accuracy. Similarly, the Galactica and Llama-2 models also performed well, achieving close to 43% accuracy. The other LLMs, like MPT, BioMedLM, and BioGPT did not perform well without any fine-tuning.

After conducting experiments by fine-tuning these models using LoRA, it is evident that the Falcon model is superior, achieving an impressive 84% accuracy. Galactica, MPT, and Llama all performed similarly, achieving close to 80% accuracy. It is clear that as we fine-tune smaller models, there is a significant performance decrease, for example the 2.7b parameter BioMedLM model and the 1.6b parameter BioGPT model are significantly outperformed by each

of the 7b parameter models. This observation indicates that even larger models will yield better results.

The loss graphs below show that each model learns the task effectively, converging quickly and then becoming more specialized with more training epochs. The comparison of each LoRA trained model, along with the original pre-trained models, is presented in the tables above. Overall, it appears that optimizing the Falcon-7b model through fine-tuning emerges as the most effective approach for this particular task. Furthermore, it is noteworthy that even in the absence of fine-tuning, the Falcon model demonstrates superior performance. However, it is evident that fine-tuning the LLMs with LoRA shows significant improvements in all metrics when compared to the original pre-trained models, improving accuracy by more than 30%.

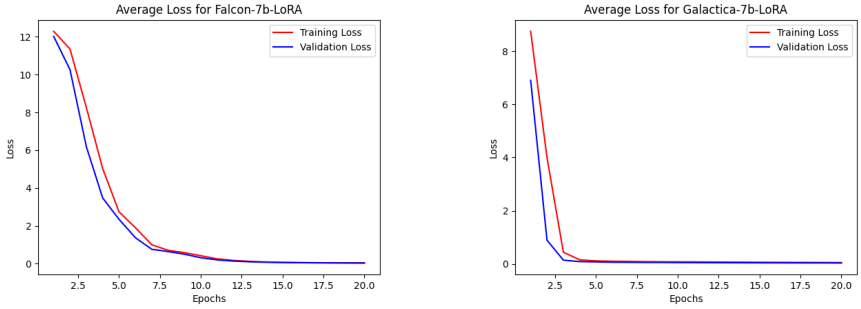

Figure 2: Loss for Falcon-7b and Galactica-7b models.

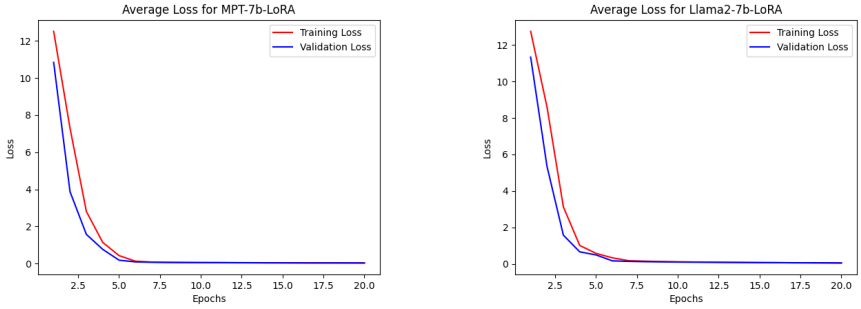

Figure 3: Loss for MPT-7b and Llama2-7b models.

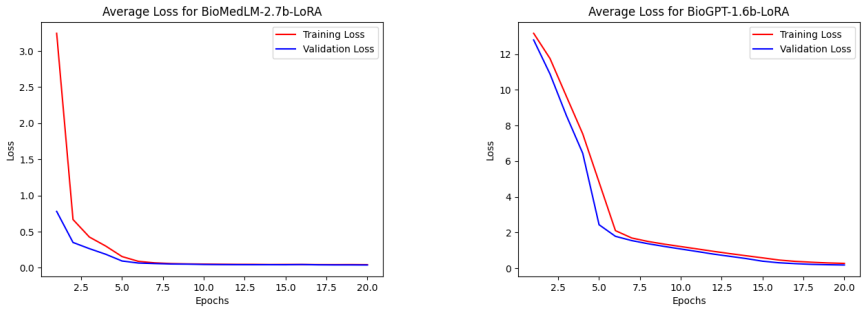

Figure 4: Loss for BioMedLM-2.7b and BioGPT-1.6b models.

7 Conclusion

Predicting protein interactions within the presence of cancer represents a relatively unexplored research frontier in the realm of natural language processing. This study demonstrates the advantages of employing LLMs for this task, showing promising results. Our findings indicate that fine-tuning LLMs can be an effective strategy for protein interaction prediction, highlighting both their performance, and their potential for future research. In addition, we investigate the ability of LLMs to perform well during fine-tuning in terms of enhanced computational efficiency, through the use of LoRA. Utilizing this technology is a big step in helping us understand the syntactic relationships between proteins, and their interactions in the context of other molecular processes.

8 Future Work

In future research endeavors, we aim to extend this study by evaluating additional datasets, for example the S. cerevisiae-benchmark dataset [14], and the H. sapiens-benchmark dataset [14]. The incorporation of benchmark datasets to our experiments would show a more robust demonstration of these models on this task. Furthermore, there are other datasets that include more specific disease-focused protein interactions, which can show how these techniques can be used for PPI prediction in other contexts, like neurodegenerative disorders [14] and metabolic disorders [14].

Another path we might take in future works would be to evaluate more PEFT algorithms, like IA3 [10], Prompt-Tuning [11], Prefix Tuning [12], and P-Tuning [13]. Evaluating the performance and efficiency of other algorithms would allow us to explore better training procedures for LLMs, demonstrating their computational efficiency in low-resource environments. Additionally, these strategies would become more valuable as we scale up the size of the LLMs, which could significantly enhance performance.

9 Acknowledgements

I wish to express gratitude to my mentor, Dr. Gilchan Park, whose extensive expertise, understanding, and patience have added immeasurably to my experience throughout the SURP program. His guidance has been crucial to the success of this project.

I would also like to extend my thanks to Brookhaven National Laboratory and its Office of Educational Programs for offering an effective research environment. Additionally, I am grateful for the opportunity to work in the Computational Science Initiative at Brookhaven National Laboratory. Their vast computational resources have provided me with the means to take on this endeavor.

References

[1] J. Qiu, K. Chen, C. Zhong, S. Zhu, X. Ma. Network-based protein-protein interaction prediction method maps perturbations of cancer interactome. 2021. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009869

[2] Y. Wang, Z.-H. You, S. Yang, X. Li, T.-H. Jiang, and X. Zhou. A High Efficient Biological Language Model for Predicting Protein–Protein Interactions. Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China; University of Chinese Academy of Sciences, Beijing 100049, China. 3 February 2019. https://www.mdpi.com/2073-4409/8/2/122

[3] S. Hashemifar, B. Neyshabur, A. A. Khan, J. Xu. Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics, Volume 34, Issue 17, September 2018. https://doi.org/10.1093/bioinformatics/bty573

[4] Muhao Chen, Chelsea J -T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, Wei Wang, Multifaceted protein–protein interaction prediction based on Siamese residual RCNN, Bioinformatics, Volume 35, Issue 14, July 2019, Pages i305–i314, https://doi.org/10.1093/bioinformatics/btz328

[5] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2 https://www.nature.com/articles/s41586-021-03819-2)

[6] E. Nijkamp, J. Ruffolo, E. N. Weinstein, N. Naik, A. Madani. "ProGen2: Exploring the Boundaries of Protein Language Models." arXiv:2206.13517 [cs.LG], Submitted on 27 Jun 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.13517

[7] Jha, K., Karmakar, S. & Saha, S. Graph-BERT and language model-based framework for protein–protein interaction identification. Sci Rep 13, 5663 (2023). https://doi.org/10.1038/s41598-023-31612-w

[8] Lialin, V., Deshpande, V., & Rumshisky, A. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. arXiv:2303.15647 [cs.CL], 2023. [Online]. Available: https://arxiv.org/abs/2303.15647

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, & Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv:2106.09685 [cs.CL], Submitted in 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[10] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. Raffel. "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning." arXiv:2205.05638 [cs.LG], 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.05638

[11] Lester, B., Al-Rfou, R., & Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv:2104.08691 [cs.CL], 2021. https://doi.org/10.48550/arXiv.2104.08691

[12] Li, X. L., & Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv:2101.00190 [cs.CL], Submitted on 01 Jan 2021. https://arxiv.org/abs/2101.00190

[13] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. GPT Understands, Too. arXiv:2103.10385 [cs.CL], Submitted on 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2103.10385

[14] Pei, F., Shi, Q., Zhang, H., & Bahar, I. Predicting Protein–Protein Interactions Using Symmetric Logistic Matrix Factorization. Journal of Chemical Information and Modeling, 61(4), 1670-1682 (2021). https://doi.org/10.1021/acs.jcim.1c00173