Adapting Large Language Models to Predict Gene Interactions

Mark Heimann
heimann2@llnl.gov
Lawrence Livermore National Laboratory
Livermore, California, USA

Christine Klymko
klymko1@llnl.gov
Lawrence Livermore National Laboratory
Livermore, California, USA

Jayaraman Thiagarajan
jayaramanthi1@llnl.gov
Lawrence Livermore National Laboratory
Livermore, California, USA

ACM Reference Format:
Mark Heimann, Christine Klymko, and Jayaraman Thiagarajan. 2024. Adapting Large Language Models to Predict Gene Interactions. In *Proceedings of 23rd International Workshop on Data Mining in Bioinformatics (BioKDD '24).* ACM, New York, NY, USA, 1 page. https://doi.org/XXXXXXX.XXXXXXX

The accurate prediction of gene interactions is an important aspect in the development of many medical treatments [2, 6]. However, gene interactions are extremely complex and the training data for any prediction problem of interest is often both sparse (too limited to train an effective problem-specific model) and includes some level of experimental noise. Embedding large-scale gene interaction knowledge has been shown to help in certain context-specific predictions [4, 5], but the sheer volume of gene information used for training means that the embeddings may encode largely irrelevant information for many problems of interest and thus generalize poorly to them. Consequently, we seek to use large language models (LLMs) to produce robust, yet high-fidelity, gene interaction predictions via two crucial design choices: (a) graph-based prompting of gene LLMs and (b) noisy edge model tuning.

Given a graph G, whose nodes i are genes, predicting gene-gene interactions can be formulated as an edge prediction task. However, in specific use cases, our focus is restricted to a small subset of genes (e.g., a few hundred), and hence the resulting graphs contain orders of magnitude less information than what existing LLMs have been trained on. Thus, we seek to leverage the power of pre-trained LLMs to perform edge prediction on scenario-specific small graphs. In this work, we consider two benchmarks, namely the publicly available synthetic lethality prediction dataset [6] and an in-house dataset consisting of measurements of HIV replication in human cells after single and double gene knockdown.

Proposed Pipeline. At its core, our pipeline (Fig. 1) formulates the task of predicting gene-gene interaction as follows: (i) Learn gene embeddings using a large language model (LLM module); (ii) Combine the embeddings of two genes to predict their interaction strength (Edge predictor module).

We learn embeddings with the recently proposed Geneformer [5] for our LLM model. We employ a graph neural network (GNN) for the edge predictor module (it is also possible to simply use an MLP or an inner product decoder).

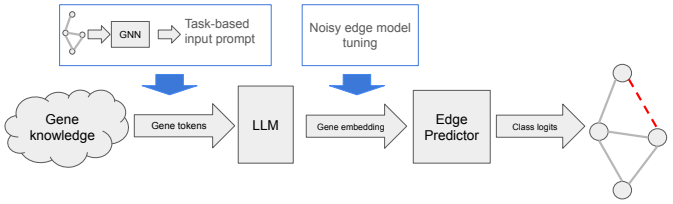

Figure 1: Proposed pipeline and our design choices.

However, we have found that naïvely such an approach works poorly, since the gene embeddings from the LLM module are not adapted for the task at hand Our ongoing work explores two different design choices: prompting the LLM with task-specific graph information, and adding noise to regularize the gene embeddings.

Graph-based prompting of gene LLM. We use a graph neural network GNN(G) to learn embeddings of each gene. Note that although the GNN is trained on much less information than the LLM as G is generally small, it is directly task-relevant as the task is defined on G. Our idea is to use these gene embeddings not as the feature vectors for the final edge prediction, but as a prompt vector that we prepend to the input embeddings of the LLM module [3]. We feed the resulting concatenated embedding through the self-attention layers of the LLM, using the output for edge prediction.

Noisy edge model tuning. We extend Noisy Model Tuning [1], a recent fine-tuning method for image classification, to pairwise prediction problems. Noisy Model Tuning introduces three forms of regularization on a predictor mapping features from a pre-trained model into a feature space for a new task. We extend these regularization methods to edge-level tasks where the features of two genes must be combined to form a prediction, and benchmark the utility of this scheme in practice across varying levels of edge noise.

REFERENCES

[1] Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. 2024. Understanding and mitigating the label noise in pre-training on downstream tasks. In *ICLR*.
[2] Pengyong Li Li, Zhengxiang Jiang, Tianxiao Liu, Xinyu Liu, Hui Qiao, and Xiaojun Yao. 2024. Improving drug response prediction via integrating gene relationships with deep learning. *Briefings in Bioinformatics* 25, 3 (2024).
[3] Zheyuan Liu, Xiaoxin He, Yijun Tian, and Nitesh V Chawla. 2024. Can we soft prompt LLMs for graph learning tasks?. In *Webconf Companion 2024*. 481–484.
[4] John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, et al. 2023. The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* 39, 2 (2023), btad080.
[5] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. 2023. Transfer learning enables predictions in network biology. *Nature* 618, 7965 (2023), 616–624.
[6] Shike Wang, Fan Xu, Yunyang Li, Jie Wang, Ke Zhang, Yong Liu, Min Wu, and Jie Zheng. 2021. KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers. *Bioinformatics* 37, Supplement_1 (2021), i418–i425.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
BioKDD '24, August 25, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX