Benchmarking Protein Language Model Embeddings for Kinase-Substrate Link Prediction

Neel Deshmukh[1],
Hui Lin, Ph.D.[2], Jean-Philippe Coppé, Ph.D.[2], Cody Carroll[1], Ph.D.

[1]Data Institute, University of San Francisco, [2]UCSF-Radiation Oncology

Introduction

- **Background:** Kinases are specialized proteins called enzymes that perform a specific catalytic function: phosphorylation. Kinases add a phosphate group to a substrate (a macromolecule, commonly another protein) at a specific location referred to as a phosphosite. This reaction changes the shape and activity of the target, acting as an 'on/off' switch crucial in regulating many processes such as cell division and signaling as well as gene expression.
- **Problem:** Abnormal phosphorylation-related events via kinase dysfunction are associated with several diseases and are essential to understand for drug discovery efforts.
- **Current Research:** A prominent area of research at the intersection of biology and deep learning is link prediction: why, how, and if certain kinases interact with certain phosphorylation sites on substrate proteins. If a kinase develops a mutation, how does that change these protein-protein interactions?

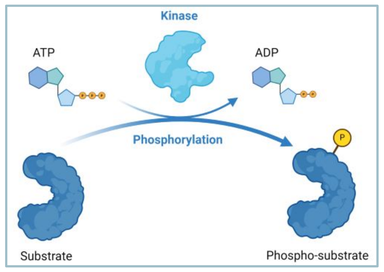

Figure 1: Depiction of phosphorylation [3]

The Project:

- With the popularity of large language models (LLMs) and the scale and complexity of biological data growing, many researchers have made advancements in natural language processing (NLP), leading to the rapid development of protein language models.[3] These protein language models (PLMs) take advantage of transformer-based architecture and represent protein sequences as strings, positing that all information about a particular protein can be gained from its sequence of amino acids.[3]
- In this project, the goal is to extract meaningful features from kinase sequences using PLMs to train a model to predict if a kinase can phosphorylate a particular phosphosite on a substrate.

Data Description

Dataset

The primary dataset for this project is PhosphoAtlas[1], an extensively curated database containing over 14,000 verified protein-phosphosite interactions. A phosphosite is a short sequence of typically 11-15 amino acids and contains a central anchor residue that is tyrosine (Y), serine (S), or threonine (T). This database is used as the training, validation, and test set for the downstream modeling.

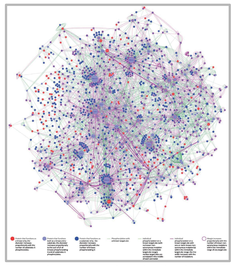

Figure 2: Map of tumor kinome from PhosphoAtlas[1]

Embedding Model Architecture

- Below is a diagram of the architecture of one of the large protein language models covered in this study, the protT5 model published by the Rost lab[4]. The main downstream use of these models is the embeddings they generate from protein sequence inputs.
- Here, embeddings are simply vectors, a numerical representation of a protein sequence that a computer model can learn from.
- For each protein, residue-level representations were extracted from the final hidden layer of a pLM and mean pooled across all positions to yield a single 1024D embedding vector.
  - This per-protein representation encodes learned patterns of residue co-evolution, sequence context, and even structural/folding propensity.

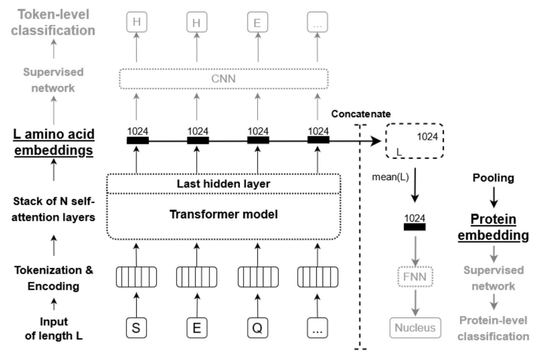

Figure 3: ProtT5 Architecture[4]

Ranking Task

- To understand how feature embedding vectors could impact existing model performance, each kinase vector was added to the feature matrix of a graph neural network (GNN) autoencoder model.
- The goal of this model was to use features such as betweenness, centrality, and contrast difference experimentally calculated through the HTKAM assay[6] to rank druggable hubs of kinase activity.
- The GNN scores each kinase-phosphosite edge based on these features and activity flow through the network (visualized in Figure 2) to determine influential hubs of activity and the kinases involved.
- Below is a graph of all the kinases scored by both methods and colored by whether they gained/lost influence with the addition of feature vectors.

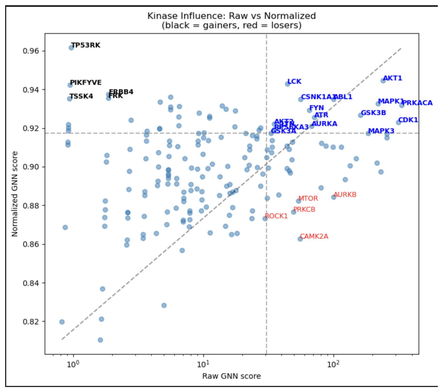

Figure 4: Raw GNN scores tracked over two different feature matrices: with and without embeddings

MIL Modeling for Link Prediction

A New Task to leverage the PhosphoAtlas database

Objective: To use a ground-truth database to create a model that can identify kinase-phosphosite interactions.

- **Data Structure:**
  a. Each kinase can bind to
     i. A single phosphosite on a single substrate
     ii. Multiple phosphosites on a single substrate
     iii. A single phosphosite on multiple substrates
     iv. Multiple phosphosites on multiple substrates
  b. In order to account for this, a multiple instance learning (MIL) model was employed
  c. A substrate protein typically contains multiple phosphorylatable residues — serine, threonine, or tyrosine sites — each of which may or may not be targeted by a given kinase. Rather than treating each site as an independent prediction target, we group all experimentally observed phosphosites belonging to a given substrate into a bag of instances.
  d. Data partition at the substrate level: all edges are assigned to train, validation, or test. The set of sites available for a substrate is fixed within each split. This is done to ensure there is no leakage between training, validation, and test sets.

Negative Sample Generation

Objective: Our database contains ground truth interactions. To train a model properly, we need to introduce negative samples into our training data. Adapted from the Phosphormer paper[7], a two-tiered negative sampling strategy was employed. One set of negatives was deemed "hard" for the model to discern: a positive kinase-phosphosite edge would have the kinase intentionally swapped out. The other set of negatives was "easy": random permutations to the kinase or the substrate and then a quick validation check to make sure a positive pair wasn't accidentally created.

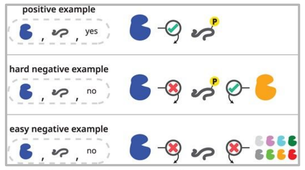

Figure 5: Phosphformer's negative sampling strategy

MIL Model Training

Overview: This is a high level overview of the MIL model training loop and how it incorporates kinase embeddings as features in its training loop.

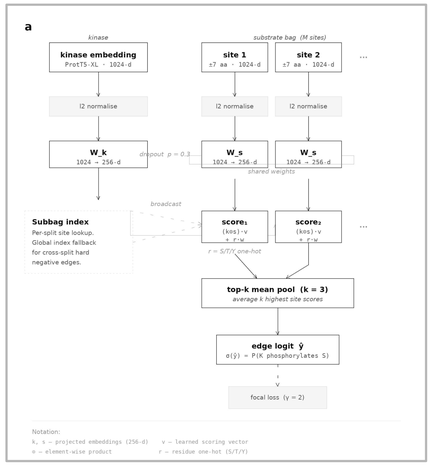

Figure 6: MIL Training Diagram

Results

ROC-AUC for MIL results

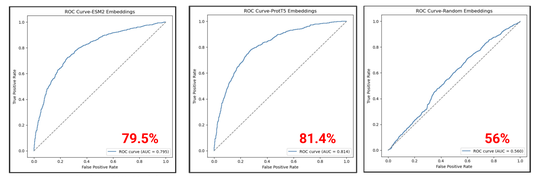

Figure 7: Results from MIL Model

Results: The MIL model can identify correct % kinase-phosphosite pairs:
- → ESM-2: **79.5%**
- → ProtT5: **81.4%**
- → Random: **56%**

- We expect the randomized 1024D embeddings to perform near 50% (random) as a baseline for model performance. Both ESM-2 and ProtT5 perform at least 23% better at the classification task. Keep in mind that no other features were fed into the model, it only used the information extracted from the embedding vectors.

Discussion

Evaluating the potential of protein embeddings in drug discovery research

- **Feature Extraction:** Based on these results, we have confidence in saying the features extracted from the embedding vectors obtained from the PLMs provide models significant information during training and can be tuned for different tasks
- **Potential for Improvement:** We could add additional features obtained from lab experiments or other resources to make these models perform even better
- **Expansion in Scope:** These PLMs were trained on hundreds of millions of protein sequences and can provide information on protein function. Some posit that we can go even further and obtain structural information on each protein as well.
- **Therapeutics:** Using wild-type protein embeddings to train models to identify correct kinase-substrate interactions can be a boon in drug discovery efforts. By introducing mutations into protein sequences/embeddings, training models to discern how function/structure changes and how signaling pathways are affected is very valuable.

Next Steps for Improvement

- **Negative Sampling Refinement**: Our negative sample generation could be more granular. Using external databases containing similar information and using correlations of kinases/substrate could provide models with better information.
- **Test More Models**: Test other PLMs at different sizes to gain a stronger idea of how well PLMs can perform on these tasks.

Acknowledgements

This work was pitched and supported at UCSF, Department of Radiation Oncology.

References

1. Olow, A., Chen, Z., Niedner, R. H., Wolf, D. M., Yau, C., Pankov, A., Lee, E. P., Brown-Swigart, L., van 't Veer, L. J., & Coppé, J. P. (2016). An Atlas of the Human Kinome Reveals the Mutational Landscape Underlying Dysregulated Phosphorylation Cascades in Cancer. Cancer research, 76(7), 1733–1745. https://doi.org/10.1158/0008-5472.CAN-15-2325-T
2. https://bpsbioscience.com/luminescence-kinase-assays-illuminate-the-path-to-inhibitor-discovery
3. https://arxiv.org/html/2502.06881v1
4. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, & Burkhard Rost. (2020). ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing
5. https://arxiv.org/html/2502.06881v1
6. Coppé, JP., Mori, M., Pan, B. et al. Mapping phospho-catalytic dependencies of therapy-resistant tumours reveals actionable vulnerabilities. Nat Cell Biol 21, 778–790 (2019). https://doi.org/10.1038/s41556-019-0328-z
7. Zhongliang Zhou, Wayland Yeung, Nathan Gravel, Mariah Salcedo, Saber Soleymani, Sheng Li, Natarajan Kannan, Phosformer: an explainable transformer model for protein kinase-specific phosphorylation predictions, Bioinformatics, Volume 39, Issue 2, February 2023, btad046, https://doi.org/10.1093/bioinformatics/btad046