Please note that this project is an extension of current research at the University of Texas M.D. Anderson Cancer Center. We request that this project never be made public until publication.

Applications of Convolutional Graph Neural Networks for Proteomic Analysis

COMP/ELEC 576 Introduction to Deep Learning: Final Project Report

Robert Heeter
Department of Bioengineering,
Rice University
rch5@rice.edu

Arielle Sanford
Department of Physics,
Rice University
ams31@rice.edu

Abstract

Drug discovery and development is an expensive process—it typically takes $1-2 billion and more than a decade to produce an FDA-approved drug—and nearly 90% of therapeutics fail during clinical trials. Recent progress in machine learning has the potential to accelerate every step of the development pipeline. However, applications of artificial intelligence in medicine often fail to appropriately address the latent imperfections in experimental data or implement convoluted and unexplainable machine learning models, leading to poor prediction interpretability. Convolutional graph neural networks (CNNs) as applied to proteomic prediction tasks offer better interpretability by more naturally representing molecular data. This project will (1) explore the use of graph CNNs by modeling the interaction between proteins using the popular Protein-Protein Interaction (PPI) dataset and (2) attempt to identify druggable cavities in proteins using graph representations of structural data in the Protein Data Bank (PDB) to train a novel graph CNN.

1. Background & Motivation

1.1. Protein Biology (Proteomics)

Looking at the drug development pipeline holistically (*Figure 1*), this project focuses on the identification of protein targets in the early drug discovery stage. Understanding protein-protein and protein-drug interactions is essential for designing small-molecule therapeutics. These interactions are influenced by both physical and chemical properties; in particular, the binding of a drug (ligand) to a protein cavity (*Figure 2*) is heavily impacted by the volume, 'enclosure', surface protrusions, and depth of the cavity, as well as adhesive forces due like hydrogen bonding, electrostatic interactions, hydrophobic interactions, and van der Waals interactions.

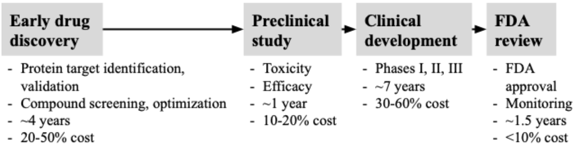


Figure 1. Typical drug development steps, timeline, and costs.

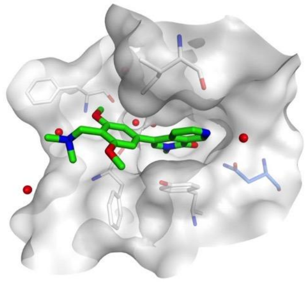


Figure 2. Drug/ligand compound (colored) binding to protein cavity (gray region).

1.2. Graph Neural Networks

Graphs are one means of storing data through the use of nodes/vertices and edges. Research over the past decade has enabled the training of graph neural networks using graphical data, where 'message passing' (i.e., feed-forward and backpropagation) occurs along edges between two connected nodes (*Figure 3*) [1]. PyTorch Geometric is one of the most popular libraries for building these graph-based neural networks.

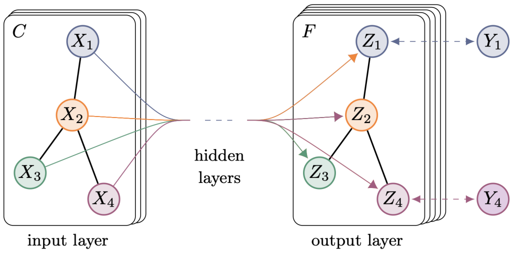


Figure 3. Graph neural network architecture [1].

Protein data (and molecular data in general) can be most 'naturally' encoded in a graphical format, where individual atoms are nodes and covalent and non-covalent interactions between atoms are edges (*Figure 4*). In prior research, many machine learning tasks involving 3-D protein data have relied on abstracted calculated or estimated structural features or properties using traditional classification techniques (i.e., logistic regression, decision trees) [2]. This approach results in poor model interpretability and an enormous number of dependencies from algorithms and software used for computing dozens of features. Instead, with graph neural networks, more 'raw' chemical features like element type, number of bonds, and charge can be assigned to each atom (node), resulting in 'purer' data and greater performance.

Before diving into applying graph neural networks to 3-D structural protein druggability classification, the Protein-Protein Interaction toy dataset is used to explore the rational design of graph neural network architectures for proteomic analysis.

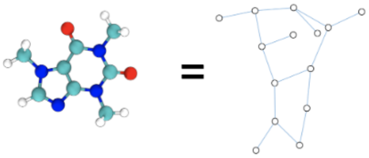


Figure 4. Atomic view of a molecule and an equivalent graph structure.

1.3. Protein-Protein Interaction Task Dataset

The Protein-Protein Interaction (PPI) dataset is a collection of data that details interactions between proteins, obtained from the "Predicting Multicellular Function through Multi-layer Tissue Networks" paper [3]. It contains 20 graphs, with 50 features and 21 labels, averaging 2,245.3 nodes and 61,318.4 edges per graph. The graphs represent various types of human tissues (*Figure 5*). The nodes of the graph represent proteins while the edges represent whether there is an interaction between the proteins. The node features are composed of positional gene sets, motif gene sets, and immunological signatures. Positional gene sets represent groups of genes based on their physical locations on chromosomes. The motif gene sets are groupings based on commonly shared patterns in the genes' DNA and RNA sequences. Immunological signatures group genes based on their relation to the immune system. The labels are based on gene ontology, which is a comprehensive set of vocabulary describing the roles of genes and proteins.

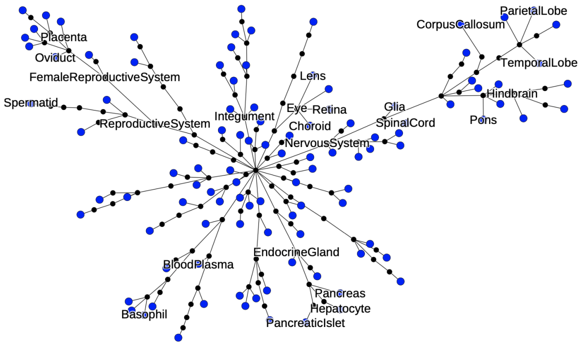


Figure 5. Network of the tissue types represented in the PPI dataset.

1.4. Protein-Ligand Binding Task Dataset

The novel protein-ligand binding dataset used for druggability classification consists of roughly 3,600 disconnected graphs, each of which represents a protein where nodes are atoms (without hydrogen atoms) and edges are covalent and non-covalent atomic interactions. Each node has 39 physiochemical features (i.e., element type, atom degree, charge, number of radical electrons, hybridization) and each edge has a single binary feature indicative of whether it represents a covalent or non-covalent interaction. Altogether, these graphs sum to over 1 million nodes and 1 billion edges.

Each node is labeled as whether or not it contributes to drug (ligand) binding (*Figure 6*) based on whether it lies within a 5 Angstrom distance of a drug-like molecule.

This dataset was generated by filtering an exhaustive list of over 40,000 proteins by structural completeness, resolution, and uniqueness (to prevent sampling multiple highly similar structures). The 3,600 filtered proteins were then cleaned and converted to PyTorch graph objects with 39 assigned node features using PyMOL, the RDKit cheminformatics Python library, and the NetworkX Python library. Only a subset of 30 proteins (graphs) was used for this project due to the enormous size of the complete dataset.

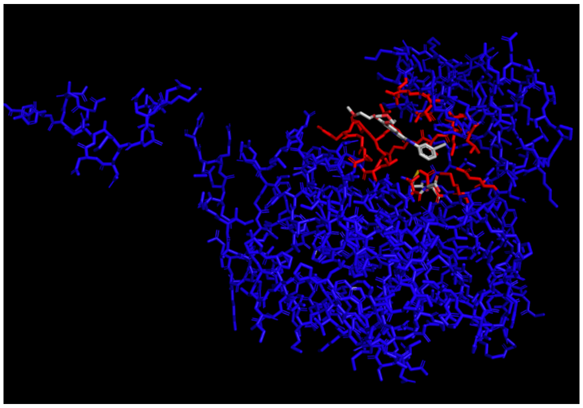


Figure 6. Node (atom) labels for an example protein; the red region is labeled positive (contributes to drug binding) and the blue region is labeled negative (does not contribute to drug binding). Created with PyMOL.

2. Experimentation & Model Design

2.1. Protein-Protein Interaction Task

In the development of a graph neural network (GNN) for the classification of the PPI dataset, we experimented with several features of the model. We found that 3 sets of Graph Attention Convolution layers (GATconv) and Linear layers, gave a high accuracy. The GATConv layers leverage the attention mechanism to allow the model to focus on the most important proteins (nodes) and edges (protein interactions) when aggregating information from neighbors [4]. Linear layers combine the node features into higher-level representations. This is important for capturing the complex relationships in the data.

Next, we compared the performance of the activation functions ELU and ReLU (*Figure 7*). Over many trials, ELU converges to ~3% higher accuracy than ReLU with equal elapsed time. In theory, ELU is able to learn more complex patterns than ReLU for a few reasons. First, ELU avoids dead neurons by allowing small values for negative inputs. Second, ELU has a smoother gradient than ReLU which aids in training stability. However, ELU is more computationally intensive to evaluate than

ReLU, but this may be negated by the fact that ELU has a zero-centered output, which speeds up learning.

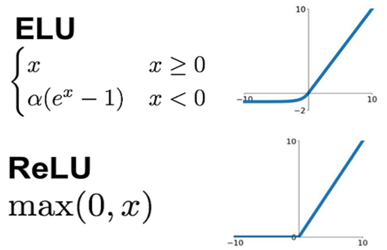


Figure 7. ELU vs. ReLU activation functions tested for the PPI task.

Next, we would have to find a Loss function that works well for multi-label classification tasks. Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss) accomplishes this, combining a logistic (sigmoid) layer with binary cross-entropy loss, to determine whether or not a protein has a label. This combination is more stable than applying a sigmoid layer and then BCELoss separately, as it handles the numerical issues arising when dealing with very small or large inputs.

2.2. Protein-Ligand Binding Task

The classification of protein druggability using the novel protein-ligand binding dataset used a model architecture inspired by the PPI task described previously and similar work in research [5]. Three GATv2Conv convolutional graph attention layers were used with ELU activations and Binary Cross Entropy with Logits Loss function following similar logic to the PPI task. Using the GATv2Conv layer operations allowed the separate or independent treatment of covalent and non-covalent interactions in the “heterogenous interaction layer” (*Figure 8*). Non-covalent interactions are less impactful than covalent interactions, so a radial basis function was applied to the non-covalent edges to reduce the weight of more long-distance interactions between atoms.

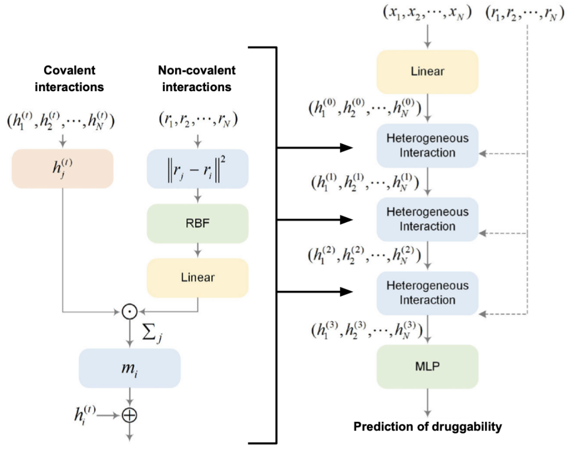

Covalent interactions
Non-covalent interactions
$(h_1^{(t)}, h_2^{(t)}, \cdots, h_N^{(t)})$
$(r_1, r_2, \cdots, r_N)$
$h_j^{(t)}$
$\|r_j - r_i\|^2$
RBF
Linear
$\odot$
Σ_j
m_i
$h_i^{(t)}$
$\oplus$
$(x_1, x_2, \cdots, x_N)$
$(r_1, r_2, \cdots, r_N)$
Linear
$(h_1^{(0)}, h_2^{(0)}, \cdots, h_N^{(0)})$
Heterogeneous Interaction
$(h_1^{(1)}, h_2^{(1)}, \cdots, h_N^{(1)})$
Heterogeneous Interaction
$(h_1^{(2)}, h_2^{(2)}, \cdots, h_N^{(2)})$
Heterogeneous Interaction
$(h_1^{(3)}, h_2^{(3)}, \cdots, h_N^{(3)})$
MLP
Prediction of druggability

Figure 8. Protein druggability classification model architecture, adapted from [5].

3. Results

3.1. Protein-Protein Interaction Task

After 150 epochs, taking 11 minutes to run on a laptop, the model achieved a 95.72% validation accuracy, a training loss of 0.0294 and a validation loss of 0.1289 (*Figure 9* and *Figure 10*). The disparity between the training and validation loss indicates some overfitting, which is to be expected. However, the validation accuracy is sufficiently high, considering the complexity of the multi-label protein classification task.

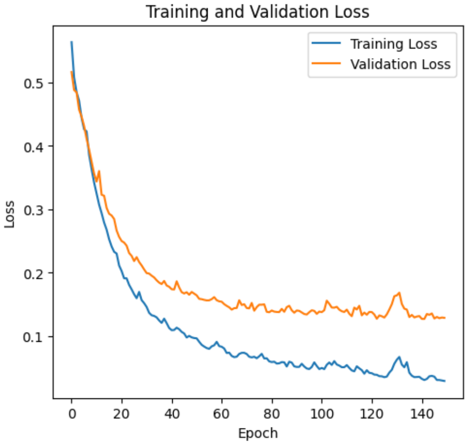


Figure 9. PPI training and validation loss.

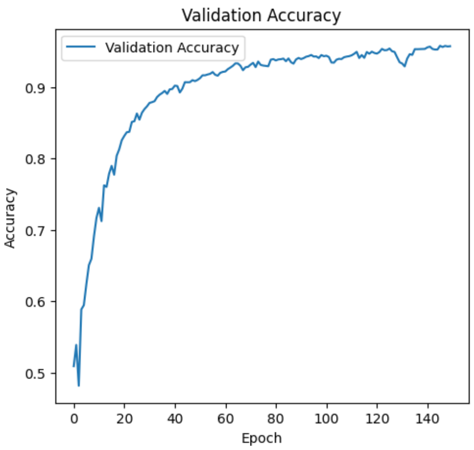


Figure 10. PPI validation accuracy.

We produced a graphic to illustrate the connections between proteins of an individual graph in the PPI dataset, and the GNNs success in correct node classification (*Figure 11*).

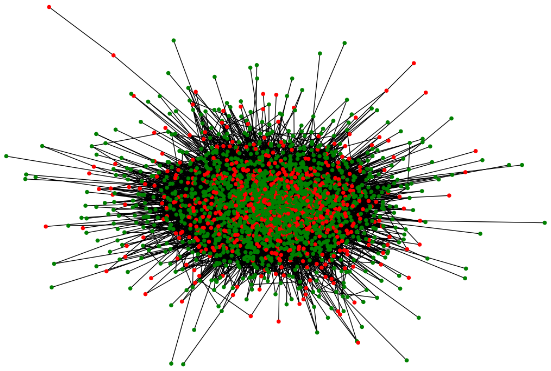
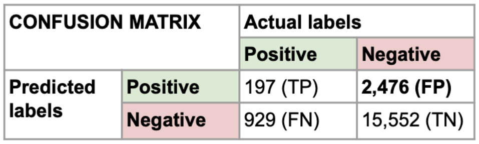


Figure 11. Visual of predicted node label accuracy. Nodes with one or more incorrectly predicted labels are marked in red. Nodes with all labels predicted correctly are marked in green.

3.2. Protein-Ligand Binding Task

Using a supercomputing cluster to train with a 30-protein subset of the protein-ligand dataset, after 800 epochs the training loss reached about 0.35 and the test accuracy reached 81.8%, with a poor recall of about 17.5%. (*Figure 12* and *Figure 13*). The training loss gradually decreased with this small dataset, though it became roughly stagnant and noisy past 50 epochs. The test accuracy is not a good performance metric for this dataset, however, since it is very class-imbalanced (<10% positively-labeled nodes). The poor recall can likely be attributed to the small size of the dataset. Current work is being done to train with a significantly larger dataset and tune the hyperparameters of this model, though the training process takes dozens of hours with the available computing resources.

Overall, from the predicted node labels, the druggability of a protein cavity can be concluded from the cavity's constituent atom labels; a cavity with mostly conducive to drug-binding' atoms is likely more suitable than one with mostly 'non-conducive to drug-binding' atoms.



CONFUSION MATRIX		**Actual labels**	
		Positive	**Negative**
Predicted labels	**Positive**	197 (TP)	**2,476 (FP)**
	Negative	929 (FN)	15,552 (TN)

Figure 12. Confusion matrix from test data.

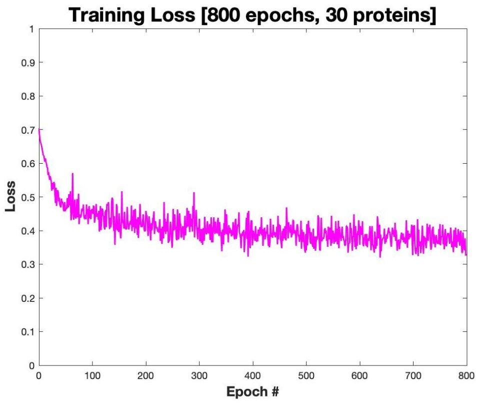


Figure 13. Training loss.

4. Discussion & Future Work

Graph Neural Networks (GNNs) are particularly well-suited for predictive tasks related to biochemical data, owing to their inherent graph-like structure. This adaptability makes GNNs ideal for handling complex relationships and interactions within biochemical networks.

We could continue our exploration of GNNs by aiming for even higher accuracy. When applied to the Protein-Protein Interaction (PPI) dataset, enhancing the GNN architecture by adding more hidden channels and layers could increase its accuracy. This improvement would stem from the ability of the network to facilitate longer-range message passing, thereby capturing more intricate patterns in the data.

Furthermore, training a druggability classifier on the complete protein-ligand dataset can lead to an increase in the recall score and overall performance. Additionally, exploring methodologies like positive unlabeled learning and self-supervised representation learning for atom embeddings could be pivotal in achieving this improvement.

The successful application of such a refined model in identifying valid drug target candidates holds considerable promise. It could significantly accelerate preclinical studies and reduce treatment costs. This advancement is particularly beneficial for rare diseases, where the need for efficient and cost-effective treatment options is paramount. By harnessing the full potential of GNNs in the proteomic context, there is a path toward more effective and accessible healthcare solutions.

5. Code Availability

The code for this project is available at github.com/roberttheeter/COMPELEC576/tree/main/project.

6. Contributions

Arielle Sanford led the experimentation of using graph CNNs for the PPI dataset. Robert Heeter focused on building a novel protein-ligand graph dataset and applying Arielle's insights to a novel graph CNN with the new dataset. Both members contributed equally to creating the poster and this report; they focused on writing sections pertaining to their respective datasets and deep learning tasks, and divided up the other general sections.

References & Acknowledgements

The authors would like to thank Dr. Phillip Gingrich and Dr. Bissan Al-Lazikani at the U.T. M.D. Anderson Cancer Center for their invaluable advice and support.

[1] B. Sanchez-Lengeling, E. Reif, A. Pearce, and A. B. Wiltschko, "A gentle introduction to graph neural networks," *Distill*, vol. 6, no. 9, p. e33, Sep. 2021, doi: 10.23915/distill.00033.

[2] M. N. Patel, M. D. Halling-Brown, J. E. Tym, P. Workman, and B. Al-Lazikani, "Objective assessment of cancer genes for drug discovery," *Nat Rev Drug Discov*, vol. 12, no. 1, pp. 35–50, Jan. 2013, doi: 10.1038/nrd3913.

[3] M. Zitnik and J. Leskovec, "Predicting multicellular function through multi-layer tissue networks," *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, Jul. 2017, doi: 10.1093/bioinformatics/btx252.

[4] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," 2017, doi: 10.48550/ARXIV.1710.10903.

[5] Z. Yang, W. Zhong, Q. Lv, T. Dong, and C. Yu-Chian Chen, "Geometric interaction graph neural network for predicting protein-ligand binding affinities from 3d structures(Gign)," *J. Phys. Chem. Lett.*, vol. 14, no. 8, pp. 2020–2033, Mar. 2023, doi: 10.1021/acs.jpclett.2c03906