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

Robert Heeter, Arielle Sanford
COMP/ELEC 576 *Introduction to Deep Learning*: Final Project Proposal
15 November 2023

Accelerating Drug Discovery Using Convolutional Graph Neural Networks

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. Furthermore, the high price tag of de novo drug design gives little incentive to treat rare diseases and cancers. In many cases, negative research outcomes go unpublished, resulting in a bias towards 'positive' data, which has widely been recognized as a critical inefficiency within the field [1]. Recent progress in machine learning has the potential to accelerate every step of the drug development pipeline, such as the computational assessment of protein 'druggability' (i.e., the ability of a protein target to bind a drug compound) without using biochemical assays or screening chemical compounds as drug candidates. 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 and reliability [2]. This project seeks to (1) explore the theory behind convolutional graph neural network algorithms and their recent application to biochemical data, and to (2) train a convolutional graph neural network on proteomic data to identify druggable cavities on novel proteins, potentially streamlining drug development.

Background & Motivation

Graph convolutional neural networks (CNNs) are particularly compelling because they extend the power of traditional CNNs beyond grid-like data (i.e., as in the case of image data) to more complex and irregular structures, like graphs and networks. Graphs are ubiquitous in the real world, representing social networks, molecular structures, or communication networks, where entities and their relationships form non-Euclidean data. Graph neural networks, through their ability to capture the dependency between nodes (entities) through edges (relationships), can harness this intricate structure, allowing for the analysis and learning of more abstract and high-dimensional relationships. Thus, learning about graph CNNs opens up a new approach to data analysis and machine learning, offering enhanced tools for tackling problems in fields like bioinformatics, cheminformatics, and social network analysis, where data is inherently graph-structured and not grid-like. In the context of computational biology specifically, graph CNNs have been proven to be a promising machine learning architecture that can identify deep

underlying trends in protein and chemical data, such as with Google DeepMind's highly accurate AlphaFold protein folding prediction model.

We want to study applications of graph CNNs to the field of drug discovery and development because it is notoriously expensive and time-consuming, with a single FDA-approved drug costing $1-2 billion and over a decade to bring to market. High attrition rates, primarily due to poor efficacy or toxicity, significantly contribute to this inefficiency. Traditional methods of drug target discovery are often like finding a needle in a haystack due to the vast number of proteins that can be screened. The recent advent of graph-based machine learning methods has introduced a novel approach to tackling this challenge.

Previous work in protein-ligand binding prediction, such as in the canSAR knowledge-base using a simpler decision tree classifier by Halling-Brown, M. D. *et al*. and Patel, M. N. *et al*, has laid the groundwork for this type of classification task but lacks the predictive accuracy and efficiency that graph-based methods hope to offer [3][4]. The limited performance of these older models can likely be attributed to their use of abstract computed physiochemical descriptors to approximately represent 3-D protein training data. In contrast, graph neural networks rely on graphical data for training; thus, protein connectivity can be encoded more naturally in graph form, where atoms are nodes, bonds or interactions are edges, and more 'raw' chemical features (i.e., element type, charge, etc.) can be assigned to each node or edge. The Geometric Interaction Graph Neural Network for binding affinity predictions developed by Yang, Z. *et al*. offers a good example of this concept with state-of-the-art performance [5]. In theory, redesigning a protein target druggability classifier using graph CNNs should yield more accurate predictions and better support protein target selection.

Proposed Experimentation & Implementation

First, we will design a graph CNN using the PyTorch Geometric library (see https://pytorch-geometric.readthedocs.io), which is built upon the PyTorch framework that we have been using in this course thus far. We will start by exploring the use of graph CNNs for node and graph classification tasks involving biochemical data using the numerous toy graph datasets available. There are many examples in the PyTorch Geometric documentation to guide this exploration. *Arielle will lead this aspect of the project*.

Next, we aim to develop a 3-layer convolutional graph attention neural network to classify the druggability of protein cavities. The network will be designed to handle a class-imbalanced dataset and learn features indicative of binding kinetics. The graph dataset that we are planning to use for this was constructed at the University of Texas M.D. Anderson Cancer Center in a prior project by the authors; as stated at the beginning of this proposal, *we request that this project in its entirety—including this proposal—never be made public until publication.* This unpublished 5,000-protein dataset consists of Protein Data Bank (PDB) structures that have been cleaned; filtered by resolution, completeness, and sequence uniqueness; and converted to PyTorch Geometric graph objects with assigned node and edge features, such as element type, charge, valency, bond type, etc. Altogether, this dataset is approximately 10 million nodes and

over 1 billion edges large. High-performance computing resources available to us at the University of Texas M.D. Anderson Cancer Center and Rice University will be employed to perform the training process with this enormous dataset. We are curious to determine what modifications will need to be made in our GNN to make it better suited to classify real-world data as opposed to toy data. *Robert will lead this deep learning model component of our project.*

Lastly, we aim to produce simple visualizations of the node classifications on the test protein structures with the PyMOL software. It will be interesting to compare the training and test accuracies of both small toy datasets and the large protein dataset over many epochs.

Feasibility & Limitations

The end goal is to develop a graph CNN model capable of predicting druggable protein cavities. Time constraints and computational resources limit the scope and potential accuracy of our project; while we do not expect to achieve state-of-the-art performance within the timeframe of this project, we hope to successfully train our model using graph proteomic data and provide insight into applying deep learning to this type of data. Challenges include handling the large dataset and overcoming the class imbalance problem. Initial potential solutions involve leveraging advanced network architectures (i.e., using class weighting) and supercomputing resources as previously described. Nonetheless, unforeseen limitations will likely arise from the complexity of protein structures that will also be discussed in our report.

Potential Impact

If validated, predicted drug-target candidates would immediately be useful for preclinical study. Even slight improvements in protein target identification and drug screening would likely save millions of dollars across the entire pharmaceutical industry. Ultimately, if successful, this advancement in therapeutic discovery can reduce the cost of drug treatments for patients, particularly those in low-resource settings and with rare diseases, and increase the arsenal of treatment options for doctors, especially by enabling the use of combinatorial therapies and more personalized medicine.

References

[1] Yu, H. (2021). Responsible use of negative research outcomes. *J. Antibiot.*, 74(9).

[2] Amann, J. *et al*. (2020). Explainability for artificial intelligence in healthcare. *BMC Med. Inform. Decis. Mak.*, 20(1).

[3] Halling-Brown, M. D., *et al*. (2012) canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res.* 40.

[4] Patel, M. N. *et al*. (2013). Objective assessment of cancer genes for drug discovery. *Nat. Rev. Drug Discov.*, 12(1).

[5] Yang, Z. *et al*. (2023). Geometric interaction graph neural network for predicting protein-ligand binding affinities from 3D structures. *J. Phys. Chem. Lett.*, 14(8).