Predicting Toxicity of a Molecule Using Graph Neural Networks: Project Proposal

Isaiah Gocool
York University
Toronto, Canada
218918052
goisaiah@my.yorku.ca

Vince Flores
York University
Toronto, Canada
218801779
vincegab@my.yorku.ca

ACM Reference Format:
Isaiah Gocool and Vince Flores. 2025. Predicting Toxicity of a Molecule Using Graph Neural Networks: Project Proposal. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 Motivation

Research and development in the pharmaceutical industry is oftentimes a very long and resource-intensive process, making the process of discovering new medicines difficult and costly. A survey conducted by the Congressional Budget Office in 2021 found that the expected cost to develop a new drug ranges less than $1 billion to over $2 billion, and only 12% of drugs entering clinical trials are approved by the FDA [1]. The extremely high failure rate is due to a variety of factors, including the drug not working, but most notably toxicity [2]. These challenges highlight the need for predictive models that can accurately and efficiently predict molecular properties based on their structure, helping to streamline the drug development process. We hope to leverage graph neural networks to detect one key property affecting this problem, which is toxicity.

1.2 Related Work

One of the traditional methods for predicting molecular compounds is through relying on “molecular fingerprints”, such as Extended-Connectivity Fingerprints (ECFP). Molecular fingerprints are fixed-length vectors that represent a molecule’s structure. These vectors can then be used as inputs for a machine learning algorithm like Random Forests for analysis [3]. A major downside of this method is that the conversion of a three-dimensional structure into a two-dimensional one inherently leads to a loss of information, something graphs can prevent from occurring.

Zhang et. al (2025), provides an enhancement in molecular property predictions by introducing MTSSMol, a multitask self supervised deep learning framework. The tool involves a GNN encoder which takes in molecules represented in Simplified Molecular Input Line Entry System (SMILES) notation [4] as input and outputs latent feature representations or embeddings which are then useful for performing predictive tasks. The nodes of the GNN represent atoms while the edges represent bonds. The advantage of using a GNN is that it allows aggregation and integration operations to extract features from nodes and its neighbors. The study found improvements over classical fingerprint methods like ECFP [5].

There are two types of graph neural networks that are relevant to our project. Graph Convolution Networks (GCN) as proposed by Kipf & Welling (2017) focus on aggregating information from a node’s neighbours, helping it to detect simple patterns [6]. For each node in the graph, a GCN creates a new feature vector by averaging the feature vectors of itself and its neighbours. This will not only give us information on an atom but also its immediate chemical environment in the molecule, such as functional groups. The second GNN of interest is Graph Attention Networks (GAT) by Veličković et al (2018). GATs are a more complex version of GCNs that assign a weight to each node, so information obtained from a node’s neighbours also has a level of importance assigned to it [7]. This can help us detect which atoms have a greater influence on factors like toxicity.

1.3 Problem Definition

The core problem we are trying to solve is to learn the relationship between a molecule’s structure (atoms and bonds) and its function through the use of graph neural networks. Some questions this project aims to answer include the following:

- Are GNNs better at predicting molecular properties than standard machine learning methods, like random forests?
- Which GNN is better for this task?
- What are the limitations of GNNs when predicting molecular properties?

Although the primary goal of this project is to predict the toxicity of a molecule, we hope to expand it to predicting other molecular properties as well if time permits.

2 Methodology

The following steps will be followed to build, compare, and evaluate predictive graph models.

2.1 Data Acquisition and Processing

We will be using a dataset from the website MoleculeNet, introduced by Wu, et. al (2017), which curates large scale bench marks for machine learning [8]. The Tox21 dataset will be used to train the GNNs, since it contains molecules that have been deemed toxic [9]. The ClinTox dataset can then be used to validate how accurately

Conference’17, Washington, DC, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the GNNs classify molecules, since it contains both toxic and non-toxic drugs [9].

2.2 Baseline Model Implementation

We will implement a baseline non-graph model to verify how its performance compares to the use of GNNs for this task. Molecules from the datasets will be converted into ECFPs using the RDKit library. A classifier, such as the Random Forest classifier, will be trained using the ECFPs as input to make predictions on toxicity.

2.3 GNN Model Implementation

Molecules are represented by SMILES strings, and they will be converted into graph data structures using the PyTorch Geometric library. Nodes will represent atoms, and edges will represent bonds. Node features include atom type, charge, and hybridization, while edges only have the bond type as their sole feature. The two GNNs (GCNs and GATs) will aggregate the data on the molecules to produce a feature summarizing whether a molecule is toxic.

3 Evaluation

3.1 Datasets

The Tox21 dataset from MoleculeNet will be used to train the GNNs and baseline models on what makes a molecule toxic. This dataset contains thousands of chemical compounds that have been verified to be toxic by U.S. federal agencies [9], so this will serve as the “ground truth” for our graphs. Another data set we can use is the ClinTox dataset, also on MoleculeNet, which contains both drugs approved by the FDA, and drugs that were rejected due to toxicity [9].

3.2 Experiments

There are two different approaches we can take to using the datasets. The first approach involves only using the Tox21 dataset. We can split the dataset into two halves, one for training, and the other for prediction. The second approach is to use the Tox21 dataset only for training, and then use the ClinTox dataset for prediction. The latter approach is likely more helpful, since the ClinTox dataset includes both toxic and non-toxic molecules. However, this method will also be more difficult because the data from the Tox21 and ClinTox datasets will need to be standardized, so extra data processing will need to be done to ensure consistent results.

The first experiment will compare the performance of the baseline model to the GNNs. This will be done by running each model separately on the same data. The second experiment will compare the performance of the two GNN models, the GCN and the GAT. An error analysis will be performed during all experiments, as well as standard machine learning metrics like accuracy, precision, recall, and F1-score.

References

[1] David Austin and Tamara Hayford. Research and development in the pharmaceutical industry, 04 2021.

[2] Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12, 02 2022.

[3] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50:742–754, 04 2010.

[4] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28:31–36, 02 1988.

[5] Xin Yang, Yang Wang, Ye Lin, Mingxuan Zhang, Ou Liu, Jianwei Shuai, and Qi Zhao. A multi-task self-supervised strategy for predicting molecular properties and fgfr1 inhibitors. *Advanced Science*, 12, 02 2025.

[6] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv (Cornell University)*, 01 2016.

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv (Cornell University)*, 10 2017.

[8] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning, 10 2018.

[9] Zhenqin Wu, Bharath Ramsundar, Evan Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh Pappu, and Karl Leswing. Datasets.