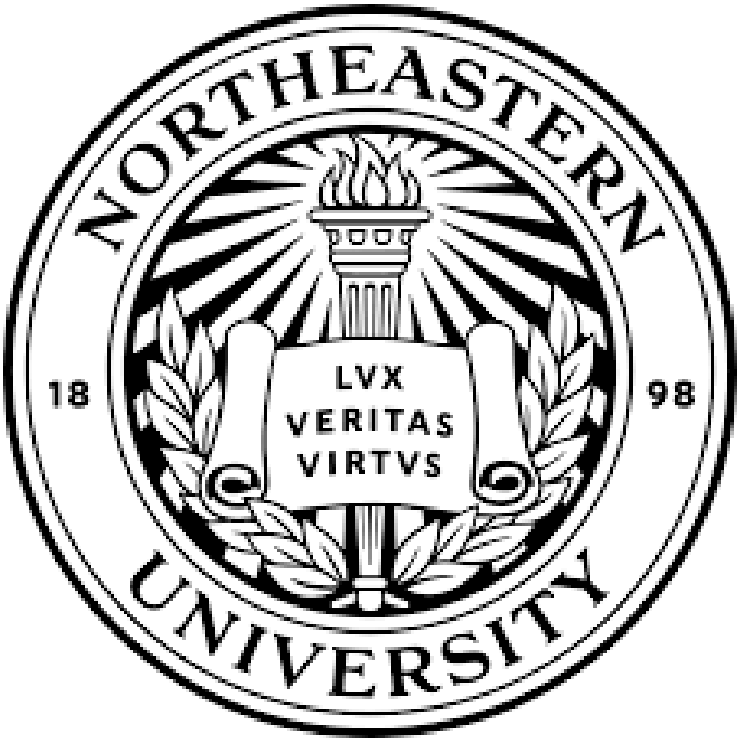


AI in Drug Discovery: Predicting Drug-Drug Interactions using SMILES codes for Safer Prescriptions

Iteration 03

Beth Farr, Sreeja Vepa, Wren Warren

Northeastern University

DS 5500 Capstone: Application in Data Science

Spring 2025

3/10/2025

Github repository link: `https://github.com/drewcark/DS5500-drug-discovery`

I. Introduction

1.1 Background

Drug interactions can be hard to fully understand, and due to the number and complexity of drugs in current medical use, there are many drugs that may have some interactions with other drugs that are not currently known to the medical community. Because these drug interactions can have many effects on patient health, both beneficial and detrimental, it is important to predict drug interactions when prescribing medications.
SMILES (Simplified Molecular Input Line Entry System) codes attempt to standardize a drug's chemical formula into a string to represent a drug's composition and thus its properties.

1.2 Problem Statement

The objective of this project is to predict the type of interaction between two drugs based on SMILES codes. Working with a labeled dataset, we use supervised machine learning to address a multi-class classification problem in the domain of machine learning. The input includes pairs of drug molecules, each represented by SMILES strings. The strings are transformed into various feature representations for the different models including as molecular descriptors, Morgan Fingerprints, and Molecular Graphs. The model will produce a classification label that indicates the type of interaction between the two drugs.

The goal of the classification model is to accurately classify how two drugs interact which is a critical concern in effective patient care and drug development. Without proper consideration of drug interactions, patients are at a higher risk of experiencing adverse drug reactions (ADRs) that can lead to serious health complications, prolonged hospital stays, or even mortality. Having tools to quickly understand implications of drug combinations can allow medical professionals to make fast and accurate patient-care oriented decisions. In addition to healthcare settings, understanding drug interactions is key in drug development. The inability to predict potential interactions during development can result in delayed timelines and higher cost-expenditures.

Several studies have utilized the DrugBank dataset and machine learning to predict drug-drug interactions (DDIs). Developed models include traditional

ML models and neural networks and many of these models relied on the DrugBank dataset. We focused on using supervised machine learning to work with an associated DrugBank dataset from Therapeutic Data Commons (TDC). This dataset contains drug pairs and their respective interactions. While using algorithms such as Random Forests and XGBoost achieved moderate accuracies, deep learning models such as graph-based neural networks showed to be a better fit for working with SMILES codes.

Related work clearly outlining the combination of these neural network models in relation to our dataset was not found. This project aims to explore the creation of neural network models which will best represent our dataset.

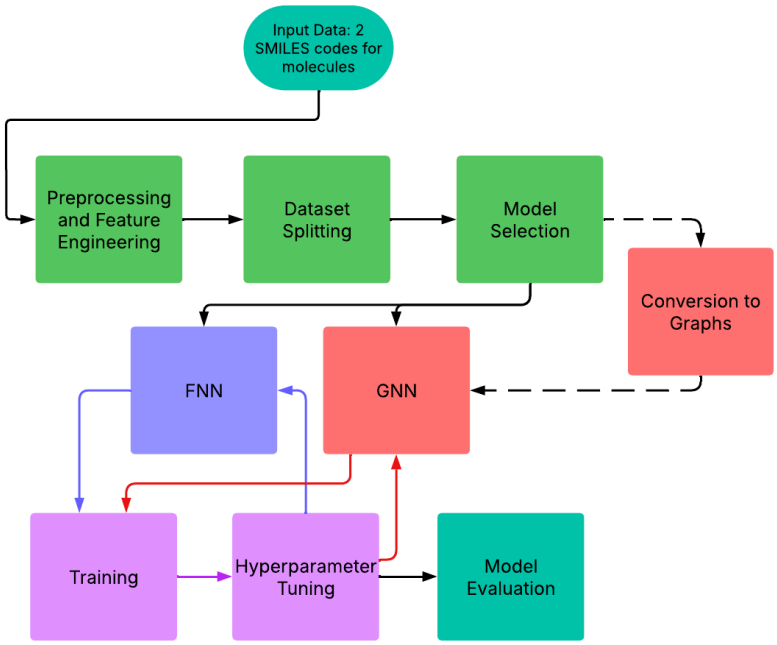
1.3 Primary Approaches and Techniques

Our primary approach to model drug-drug interactions (DDIs) was to build multiple different types of neural networks and compare their effectiveness at categorizing two drugs as having a certain interaction. We did this by processing and featurizing the data, splitting the datasets depending on the kinds of features we wanted each model to receive, and then training the selected models on the selected data, tuning their hyperparameters, and repeating until the models were optimized. After this, we evaluated the different models. This process is illustrated in the above flowchart.

The three models considered all focus on different aspects of the physicochemical and structural elements of molecules to best capture accurate interaction classification.

The models considered are:

1. Graph Neural Network utilizing molecular descriptors
2. Feed-Forward Neural Network utilizing Morgan fingerprints
3. Graph Neural Network utilizing molecular graphs



II. Methods

2.1 Purpose of the Methodology

In this project, we aim to predict DDIs by leveraging SMILES codes as representations for chemical compounds. The work in this project relates to how neural networks capture the complex information given by the SMILES codes. We aim to use a combination of Graph Neural Networks (GNNs) and Feed-Forward Neural Networks (FNNs) to find a computationally efficient and accurate measure of interaction prediction. During model development, we looked to compare models which would identify the trade-offs between interpretability, computational cost, and predictive accuracy. We explored three distinct approaches, each designed to capture different aspects of chemical structure and properties.

1. Feature-Based GNN: This model utilizes several molecular descriptors de-

rived from the SMILES codes of the drug pairs. From this we will learn how effective individual descriptors will aid in interaction prediction.

2. Morgan Fingerprint FNN: The second model utilized Morgan Fingerprints alongside molecular descriptors in a FNN, with a comparison to a GNN trained on the same data. This comparison of two different kinds of models on the same data will show us how well a FNN performs compared to a GNN on non-graph data.

3. Molecular Graph GNN: Utilizing the structural and relational information found in molecular structure, this model evaluates the compounds as molecular graphs. This model will allow us to learn if relational elements of the compound's structure are critical to capture.

Through comparison of these models we aim to uncover which strategies generalize best across DDI datasets and can scale to larger drug libraries.

2.2 Development Environment

This project uses Python 3.12 and the following python libraries: csv, math, numpy, matplotlib, pandas, scipy, scikit-learn, deepchem, RDKit, pyarrow, tensorflow, and pytorch. a full list of dependencies is available on GitHub under "requirements.txt".

2.3 Data Collection and Preparation

The datasets used for creating the models came from TDCommons, which provides information about the drug-drug interactions the size of the dataset is 191,808 rows with 6 columns. There are 191,808 drug-drug pairs with related interactions. The API is provided by DrugBank. There are 86 different interaction types, with a high imbalance in frequency of different interaction types. In order to address a portion of the imbalance, the 86 interactions were sorted and filtered to the 20 most frequently occurring interactions. This change led to an improved accuracy and class balance.

The feature engineering included using SMILES codes provided by the DrugBank dataset, and using RDKit to extract topological features about the drugs as well as one-hot encoding categorical variables. The features extracted included the Molecular Weight, LogP, Number of Hydrogen Bond Donors, Number of Hydrogen Bond Acceptors, Topological Polar Surface Area, Morgan Fingerprints.

2.3.1 Input 1: Therapeutic Data Commons

The Therepeutic Data Commons contains numerous datasets and associated machine learning and artificial intelligence tasks for single instance prediction,

multiple instance prediction, and generation within the healthcare field. The dataset we'll be using is comprised of a large list of drug-drug interactions. No permissions are necessary.

Source:

- **Link:** `https://tdcommons.ai/multi_pred_tasks/ddi`
- **Format:** tab file
- **Cleaning:** No cleaning
- **Availability:** Available to all, free
- **Purpose:** Provides a large dataset of drugs with SMILES codes and their interactions with each other.

Structure and Metadata:

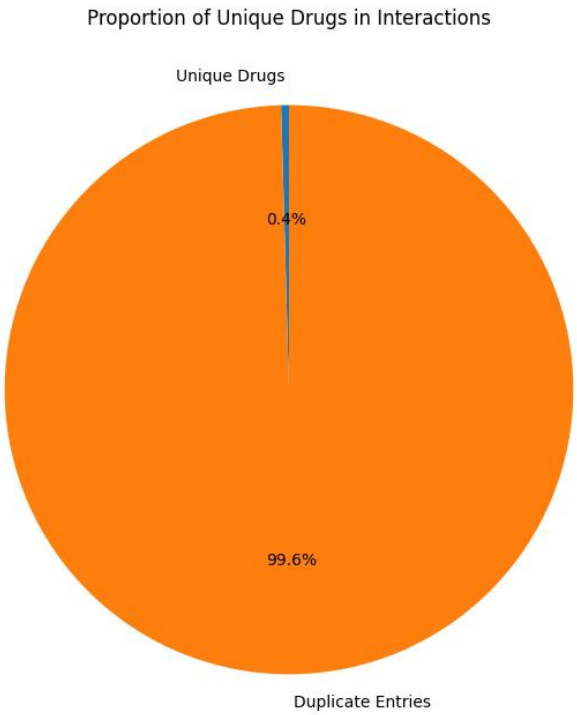
- The size of the dataset is 191,808 rows with 6 columns. There are 191,808 drug-drug pairs with related interactions.
- There are 6 features: ID1 (ID of drug 1), ID2 (ID of drug 2), Y (Type of Interaction), Map (Meaning of Drug Interaction), X1 (Compound Structure of drug 1), X2 (Compound Structure of drug 2)
- No direct API, but data originates from DrugBank API.
- There are 86 different interaction types, some are highly imbalanced.
- Some drugs only have listed interactions with one other drug.

Missing Data:

- There are no missing values in this dataset.

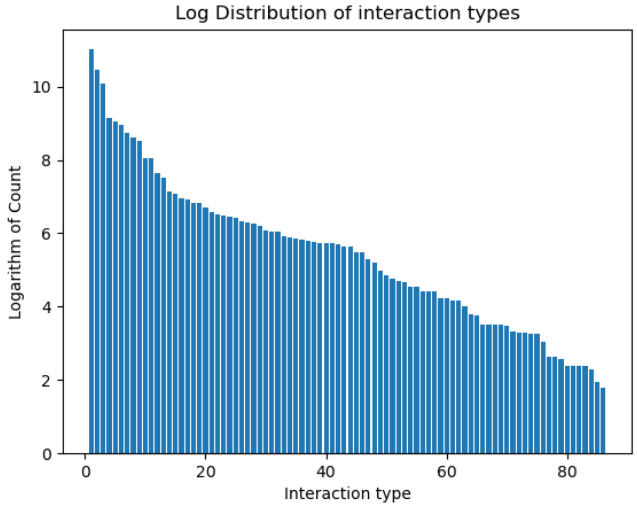
Anomalies:

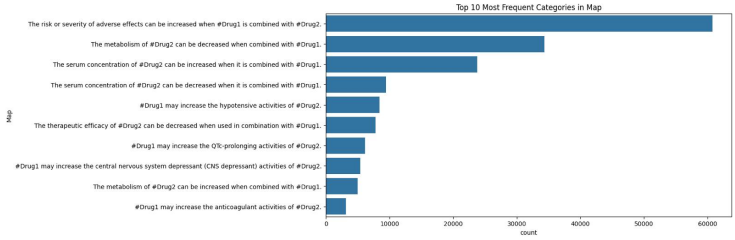
- There are some drugs that only appear interacting with one other drug, or a very small number of other drugs, these interactions will be dropped from analyses
- Data is mostly unique pairings, and thus consistency is not applicable, the 0.4% of unique drugs will be dropped from the dataset as they would not aid in training.



Bias:

- The dataset adequately represents the data we are interested in (structures of drugs).
- Since certain drugs and types of interactions are overrepresented in the dataset, we will have to be proactive to avoid representational bias in our sample.





Distributions:

- There are no numerical features, nor any features strongly correlated. Thus multicollinearity is not a concern.

Categorical Data:

- There is one categorical variable that is essential to this dataset, the type of interaction between the two drugs, of which there are 86 different categories. The mode of interaction type is type 49, which is an unspecified set of adverse side effects from the interaction of the two drugs. The categories are highly imbalanced, which will have to be countered.

Ethical Considerations:

- There is no use of Personally Identifiable Information. There are no ethical concerns.

Alignment with Goals:

- The data does align with the project's goals, because it contains a set of SMILES codes we are using to assess a target feature that it also includes (the interaction type).

Scalability:

- Manageable on GPUs.
- Could benefit from distributed processing if a GNN approach is used.

Transformations:

- The interaction types (Y column) needed encoding.
- There are opportunities to create new features by turning the chemical structure in SMILES format into a graph of interaction groups, but we are still exploring this possibility.

Data Encoding:

- Categorical variables should be given one-hot encoding.
- There are no sequential/temporal aspects necessary to be encoded.
- For the Feedforward Neural Network (DNN) we would need to encode into molecular fingerprints. For the Graph Neural Network (GNN) we would need to encode into molecular graphs.

Predictive Power:

- The features do create sufficient predictive information as we will be using the ID labels to pull information from the ChEMBL database for the molecular fingerprints and molecular graphs.
- Feature selection and dimensionality reduction are not necessary.

Target Variable:

- The target variable is type of interaction, and it is one of 86 different possibilities and they are categorized by the Y column.
- The target is not well balanced and will require resampling.

Validation Strategy:

- Data from the DrugBank DDI file will be split into training and validation sets as well as a testing set, via a random sampling, in an 80%/10%/10% distribution.
- There are no temporal or spatial dependencies that need to be preserved.

Data Leakage:

- There is no risk of data leakage if the split is performed properly.

Interpretability:

- Results will be communicated via metrics and visualizations.
- Confusion matrices from model performance.

Limitations:

- One limitation is that all training and testing is only on one dataset, and the number of drugs in DrugBank is not large. The data is also imbalanced in type of interaction and in the number of times some drugs are mentioned. The interactions are also limited in type and only drugs that interact with each other are included in the dataset.
- Additional datasets with interaction type would improve the model, as would a dataset that breaks SMILES codes out into their components.

2.3.2 Input 2: Kaggle SMILES dataset

This dataset from Kaggle contains a list of drugs and their SMILES codes. No permissions are necessary.

Source:

- **Link:** `https://www.kaggle.com/datasets/art3mis/chembl22?resource=download`
- **Format:** text file
- **Cleaning:** None
- **Availability:** Available to all, free.
- **Purpose:** To provide more SMILES codes connected to their corresponding ChEMBL name.

Structure and Metadata:

- The size of the dataset is 1048576 rows of data, 2 features/columns.
- The Features: Column 1 relays the SMILES String. The second column is a reference to the full ChEMBL entry for that particular molecule. As this dataset is not being used for analysis and more for additional information, there are not any key metrics or summaries for this dataset.
- There is a related API available to this dataset named ChEMBL API. This allows data retrieval on compounds and cheminformatics uses.

Missing Data:

- There are no missing values in this dataset.

Anomalies:

- Because of the nature of the dataset, there are no outliers or anomalies.

Bias:

- The dataset adequately represents the data we are interested in (structures of drugs). The dataset includes a broad range of molecules.

Distributions:

- There are no numerical features, nor any features strongly correlated. Thus multicollinearity is not a concern.

Categorical Data:

- The two columns in this dataset are categorical variables. However, both of them consist of all unique values. Balancing is not applicable for this database as it is a comprehensive dataset with key-value pairs.

Ethical Considerations:

- There is no sensitivity of PII data in this dataset.
- There are no ethical concerns that arise by publishing the insights from this data.

Alignment with Goals:

- The dataset is useful for generating molecular embeddings for drugs in the DDI database.
- It can provide better feature representations with the molecular fingerprints.
- May help predict interaction for unseen drugs.

Scalability:

- The dataset is very large. For now, as we do not plan on using it in its entirety it is manageable with the available resources.
- If we decide to use the dataset in its entirety, we would need dimensionality reduction if used in a DNN.

Transformations:

- SMILES codes will need to be converted into numerical fingerprints or graphs. However, this is not preprocessing and more related to model development.
- Yes, there are opportunities due to the linked ChEMBL identifiers to the ChEMBL database. However, we are not considering doing this at this moment.

Data Encoding:

- Data encoding is not required as SMILES is already string-based.
- There are not any temporal or sequential features which require specific transformations at the moment.

Predictive Power:

- Although this dataset does not directly contain predictive information. If utilizing the ChEMBL database it will have predictive power in the long-run and can improve generalisation in the DDI model.
- Feature selection or dimensionality reduction is not needed for this dataset at the moment.

Target Variable:

- There is no direct target variable as it is a feature dataset.

Validation Strategy:

- Not directly applicable to this dataset.

Data Leakage:

- No risk for this dataset.

Interpretability:

- This dataset can provide interpretable insights for stakeholders due to its connection to the ChEMBL database. Through connection to the ChEMBL database, the dataset can indirectly provide interpretable insights for stakeholders.
- The results will not need to be communicated for this particular dataset. Instead, the final analysis using the main dataset and additional information will be combined to be shown through histograms, plots, accuracy analyses, and more.

Limitations:

- This dataset does not contain interactions, only molecular structures.
- In order to have predictive power, this dataset needs to be combined with the DDI dataset.

2.3.3 Input 3: CureFFI.org

This source contains a list of FDA approved drugs by generic name, and their SMILES codes. No permissions are necessary.

Source:

- **Link:** `https://www.cureffi.org/2013/10/04/list-of-fda-approved-drugs-and-cns-drugs-with`
- **Format:** text file
- **Cleaning:** Adding a few missing SMILES codes.
- **Availability:** publicly available

Structure and Metadata:

- **Size:** 1691 rows, 3 features/columns.
- **Features:** Column 1 relays the generic drug name, column 2 relays whether it is a CNS drug, column 3 relays the preferred SMILES code.
- **Statistical Interest:** This dataset contains drugs' generic names and SMILES codes, which may prove useful in mapping common names to SMILES codes.
- There is no API for this dataset.

Missing Data:

- There are 194 missing values in SMILES codes. These will be excluded from our project.

Anomalies:

- Because of the nature of the dataset, there are no outliers or anomalies.

Bias:

- The dataset adequately represents the data we are interested in (structures of drugs).
- However there is a bias towards FDA-approved drugs.

Distributions:

- There are no numerical features, nor any features correlated. Thus multicollinearity is not a concern.

Categorical Data:

- The first two variables: generic_name and cns_drug are categorical. The generic_name is a nominal variable and has all unique names without order. The cns_drug column is binary with Yes or No labels.
- Balancing is not necessary to examine in this dataset as we are planning to use this dataset for its connection to generic_names only.

Ethical Considerations:

- This data does not contain any PHI. There are also not any ethical concerns related to this data.

Alignment with Goals:

- The dataset does align with the project's objectives.
- The dataset has the necessary features for the intended analysis.
- It will be useful for GUI and human-readable outputs.
- Provides common names for drugs in the DDI database.

Scalability:

- Yes, the dataset is manageable with the available resources and does not require advanced techniques.

Transformations:

- No, the data does not need preprocessing, normalization, standardization, or scaling.
- We will not be using the dataset features beyond the common names.

Data Encoding:

- Categorical variables will not be encoded as only one column will be used for mapping purposes.
- There are not any temporal or sequential features that require specific transformations.

Predictive Power:

- For this dataset feature selection and dimensionality reduction are not necessary.

Target Variable:

- This does not apply to this dataset as we are only using it for one column. The column's data will be added to the SQL database.

Validation Strategy:

- These do not apply to this dataset.

Data Leakage:

- This does not apply to this dataset as it is not being used for analysis purposes, is instead being used for more reference purposes instead.

Interpretability:

- The main interpretable insight from this dataset for stakeholders would be the wider audience that the GUI could reach in our project. This is because connecting the database to the generic name column in this dataset will allow users to check interactions using common names rather than SMILES codes.
- Results will not be communicated for this dataset.

Limitations:

- The limitations of this dataset is that it is smaller and does not have the generic names of all the drugs.
- A larger dataset could potentially enhance the final product and analysis.

2.4 Motivation

These datasets were chosen because they are publically available and contain comprehensive data on drug-drug interactions (DDI), as well as being used in numerous other papers on drug interactions. The datasets are relevant to our project goals because we are attempting to create and evaluate a model to accurately predict interactions between two drugs given their chemical structures, and these datasets contain chemical structure information and the presence and typology of any interactions between the two drugs. The inclusion of the CureFFI data improves the practical applicability of our results by linking the molecular interactions and their common names, meaning that the GUI would be more comprehensive. The ChEMBL dataset provides the information needed

to help train the model beyond just the known interactions thus making it more scalable.

2.5 Selection of Machine Learning Models

This project uses both FNNs and GNNs. The GNNs are used to model drug interactions using the pre established interactions within DrugBank. FNNs are used in order to create high-dimensional representation of drug feature vectors.

The GNN uses the molecules which are typically represented as graphs with nodes as atoms and edges as the chemical bonds. GNNs can capture both local chemical properties and global molecular structures through message passing. They have the ability to represent the full molecular structure and keep relational information. Therefore there are higher levels of generalisability when molecular graphs are processed using a GNN.

The FNN uses a sequential model using the Morgan Fingerprints (or Circular Fingerprints) which are fixed-length representations of structures which keep the chemical and structural features of the drug. The investigation used 2048 circular bits which represented chemical structures numerically.

Comparing computation complexity and accuracy was important to model selection due to the intention to create a SQL database and GUI. If the model was slow and computationally complex, this would not be the most efficient to connect to a GUI and therefore in order to promote more effective usage of the model.

2.6 Model Development and Training

First, a GNN model was built on a dataset of 12 features: 2 sets of 6 other molecular properties: molecular weight, the logarithm of the partition coefficient of a molecule between water and an organic solvent (logP), number of proton donors in the molecule, number of proton acceptors in the molecule, and topological polar surface area of the molecule. Each pair of molecules was represented as a graph where nodes corresponded to individual molecular representations. The edge connections represent potential interactions between the two molecules and a fully connected edge scheme is used so each molecule in a pair is linked to the other. Features were selected for their influence on combining with other molecules, no dimensionality reduction was performed. Each node had 6 input features, the molecular descriptors.

A FNN and a GNN were built to model a dataset consisting of 4108 features. The features in question were 2 molecular fingerprints consisting of 2048 features each, and the previous molecular features. The FNN consisted of an input layer of 4109 nodes, an output layer of 20 nodes representing the 20 possible interaction types, and a hidden layer consisting of the average between the two layers (2064 nodes). For the GNN, the same setup was used as before, but each

node had 2054 input features, which combined the 2048-bit Morgan fingerprint and 6 molecular descriptors.

A GNN for molecular graphs was built to model molecular graphs for predictive tasks. In this representation, each molecule is treated as a graph, where nodes correspond to atoms and edges represent chemical bonds. This mode utilized the atomic number of each atom as the node feature allowing each element with the molecule to be represented. This model has an input of one feature and an output of 20 possible interaction types. Grid search on a subset of the data indicated the optimal epochs and hidden dimensions to be 10 and 8 respectively.

The training set was 80% of the dataset and the testing set was the other 20%. The stratify method was used to ensure relatively proportional class representation within both sets. Within the training set, 20% of rows were used for validation. 20% dropout after the input and hidden layer was performed to handle overfitting. No transfer learning was used. Batch size and number of epochs to train the model were tuned via grid search to find optimal model performance via basic grid search.

III. Results

3.1 Evaluation and Comparison

The key performance metric used to evaluate the models was accuracy, providing a direct measure of the proportional of correctly classified instances. Additionally, metrics such as AUC-ROC, precision, and recall were also measured to offer a more comprehensive evaluation of model performance.

The analysis showed that feed-forward neural networks consistently outperformed graph neural networks with a 6-8% higher accuracy when evaluated on the same dataset. The highest accuracy was achieved by the FNN at 94%. There are various possibilities for this difference, but one key consideration is that the simplicity of both the GNN molecular graph model, as only atomic numbers were employed as node features, and the GNN feature model may be the reason for their lesser accuracies. While it is true that GNNs remain a promising method of learning structural and physicochemical elements of graph structure, the model with the most effective and consistent performance is the Feed-Forward Neural Network using Morgan Fingerprints as input.