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

Beth Farr, Sreeja Vepa, Wren Warren

Northeastern University

DS 5500 Capstone: Application in Data Science

Spring 2025

4/26/2025

Chapter 1

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 aim is to predict drug-drug interactions (DDIs) using the SMILES codes.

1.2.1 What is the issue or challenge you aim to solve?

In this project, we aim to create a model that shows an efficient and accurate way of predicting DDIs using SMILES codes.

1.2.2 Why is it important to address this problem?

The classic method of finding DDIs using clinical experimental validation can be time-intensive and expensive. By using machine learning techniques in combination with research efforts, a more time and cost-effective solution can be created. This would allow for earlier identification of harmful DDIs and thus improve patient safety.

1.2.3 Provide a real-world context to explain its significance.

Accurate DDI predictions are critical as they can help healthcare providers ensure patient safety by making affirmed prescribing decisions. Existing methods can be time-intensive and thus may lead to a retroactive identification of harmful interactions, leading to higher rates of unwanted adverse DDIs. This project aims to find a proactive way of determining DDIs.

Our team chose this project as we all found the intersection between data science and medical science to be intriguing. Additionally, we all saw a plethora of methods to apply the findings.

1.3 Team and Student Objectives

1.3.1 Team Objectives

- Identify the best Machine Learning/Artificial Intelligence algorithms to predict drug-drug interactions, given two drugs' SMILES codes.
- Successfully extrapolate a smaller-scale dataset of drug-drug interactions to a larger dataset.
- Create a Graphical User Interface that provides a user with an easy-to-use experience.
- Identify shared goals that the team aims to accomplish as part of this project.
- Collaborate effectively, ensuring all members contribute equally to the project's success.

1.3.2 Individual Objectives

- Beth Farr
  - Develop my skills with big data and improve on my previous skills with interface design and visualizations.
  - Learn how to compile a scientific report which I believe will help me further with my studies.
- Sreeja Vepa
  - Grow my knowledge in PyTorch and TensorFlow.
  - Learn the basics of the R language.
  - Gain a deeper understanding of supervised machine learning models.
  - Learn more generally about large language models and consider implementing a task suited to this project if time allows.
- Wren Warren

- Gain hands-on experience in building machine learning predictive algorithms for drug discovery in a python environment.
- Learn to integrate a SQL database with python for convenient data access and increased functionality.
- Increase my understanding of how to operate random forest and deep learning models on less theoretical data.

Chapter 2

Dataset Description

2.1 Input 1: Therapeutic Data Commons

This source 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.

- **Link:** `https://tdcommons.ai/multi_pred_tasks/ddi`
- **Format:** tab file
- **Size:** 191,808 DDI pairs with 1,706 drugs, 6 features
- **Features:** ID of drug 1, ID of drug 2, Type of interaction, Meaning of interaction, Compound structure of drug 1, compound structure of drug 2.
- **Cleaning:** The data may need decoding to separate chemical compound structures into its SMILES counterparts.

2.2 Input 2: Kaggle SMILES dataset

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

- **Link:** `https://www.kaggle.com/datasets/art3mis/chembl22?resource=download`
- **Format:** text file
- **Size:** 1,048,576 rows of data, 2 features/columns
- **Features:** Column 1 relays the SMILES String. The second column is a reference to the full ChEMBL entry for that particular molecule.

- **Cleaning:** None

2.3 Input 3: CureFFI.org

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

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

Chapter 3

Methodology

3.1 Data Collection and Preprocessing

The tools and libraries to be used are Pandas, NumPy for pre-processing, SMILES string preprocessing using RDKit in order to handle chemical structure representations. Using DrugBank, Kaggle and CureFFI datasets.

3.2 Exploratory Data Analysis (EDA):

- Matplotlib and Seaborn to create visualizations of the drug interaction trends and distributions. Ggplot, dplyr and tidyverse to wrangle data and create statistical visualizations in R. Kernlab and Caret for machine learning. RandomForest for showing feature importance. Using deepnet and neuralnet for natural language processing.
- Using Random Splitting to split the dataset into training and test sets. Using k-fold cross validation for performance assessment.

3.3 Modeling:

Graph Neural Network (GNN) and Deep Neural Network (DNN) using TensorFlow & PyTorch

3.4 Evaluation:

Our models will be evaluated by accuracy, specificity, precision, sensitivity and recall scores, as well as F-1 score, Area under the receiver operating characteristic curve (AUC-ROC), and Precision-Recall Curve.

3.5 Deployment:

deployment will be done in Flask/Streamlit to create user interface and MySQL to store results of the predictions.

Chapter 4

Expected Output

The deliverables that we intend to create will help make drug to drug interaction prediction easier. The TDC dataset categorizes each mentioned drug pairing into one of 86 types of interactions. First, by creating a predictive machine learning model from these pairings, we aim to identify if any two drugs within the kaggle dataset can fit into at least one of these 86 interaction types. We hope to store all the drug-drug interactions listed in the TDC dataset in a created SQL database. This database will also contain which interaction type the drug pairing showed. Along with the SQL database, a GUI will be created. The GUI will allow a user to input two drugs and interact with the database to see if an interaction was recorded. If not, the newly produced interaction finding will be added to the SQL database. If time permits, we also hope to explore large-language models in this project. One consideration is to use LLMs to generate human-readable explanations for the interactions by utilizing BioGPT. For example by utilizing the interface while discovering new drugs this prediction will make it easier to see the relationship between two drugs. Furthermore, by creating the LLM, the aim is to make it easier to see commonalities between the interaction between two drugs and the make up of their SMILES codes.

4.1 Summary

- **Predictive Model:** Development of a predictive machine learning model using the TDC dataset.
- **SQL Database:** Creation of a comprehensive SQL database to store drug pairs and their corresponding interaction types.
- **User Interface:** A GUI enabling users to input two drugs and query the SQL database for known interactions; if no known interactions are recorded, the predictive model's findings will be added to the database.
- **LLM:** Utilizing a LLM to create explanations for drug-drug interactions.

Chapter 5

Tools and Programming Languages

We plan on both improving existing skills along with the tools that we will be using with learning new ones, in regards to the languages and libraries that we use.

5.1 Languages:

Python, R, and SQL

5.2 Libraries:

Pandas, NumPy, Scikit-learn, TensorFlow, Keras, PyTorch, Matplotlib, Seaborn, RDKit, DeepChem, DGL, Flask/HTML

5.3 Tools:

GitHub, Google Docs, Jupyter Notebook.

Chapter 6

Related Work

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. One study utilized Random Forests and XGBoost to achieve a 74% accuracy in predicting interactions specific to osteoporosis and Paget's disease. In addition to the type of model, various datasets such as KEGG, SIDER, HIPPIE, and PharmGKB were explored along with the DrugBank dataset. In each study, these helped improve the predictive accuracy of their respective models.

Deep learning models such as graph-based neural networks were also widely used as their focus on graph structures better captured the complex relationships between drugs. One network, SmileGNN, was used for binary classification of difficult drug relationships. This network outperformed K-Nearest Neighbors (KNN). Additionally, Chemprop, a message-passing neural network, was used to distinguish between known and unknown drug interactions through adjacency matrix factorization with propagation (AMFP) and lookup adjacency matrix factorization with propagation (LAMFP). In this particular study, it was found that the combination of the unique models helped improve the overall accuracy. Another type of neural network model considered were deep neural networks (DNNs) as these focus on feature-based representations of drugs rather than the relationships between them.

Our approach is to utilize the DrugBank dataset as a training set and apply the created model to a kaggle dataset named Drug Designs with Small Molecule SMILES. The model we aim to create will be a combination of a deep neural network and graph neural network. The GNN is expected to capture the topological structure of the molecules while the DNN is expected to capture feature-based representations of the drugs.

Related work clearly outlining the combination of these neural network models was not found in our topic at the moment. Additionally, while most sources utilized the DrugBank dataset, none combined the dataset with the Kaggle dataset this project lists.

6.1 Sources

Han, Xueting, et al. “SmileGNN: Drug–Drug Interaction Prediction Based on the SMILES and Graph Neural Network.” *Life*, vol. 12, no. 2, 21 Feb. 2022, p. 319, https://doi.org/10.3390/life12020319.

Hou, Xinyu, et al. “Predicting Drug-Drug Interactions Using Deep Neural Network.” *Association for Computing Machinery*, Proceedings of the 2019 11th International Conference on Machine Learning and Computing, 1 Jan. 2019, https://doi.org/10.1145/3318299.3318323.

Hung et al., “An AI-based Prediction Model for Drug-drug Interactions in Osteoporosis and Paget’s Diseases from SMILES,” *Molecular Informatics*, vol. 41, no. 6, pp. 2100264–2100264, Jan. 2022, doi: https://doi.org/10.1002/minf.202100264.

Mei, Suyu, and Kun Zhang. “A Machine Learning Framework for Predicting Drug–Drug Interactions.” *Scientific Reports*, vol. 11, no. 1, 2 Sept. 2021, https://doi.org/10.1038/s41598-021-97193-8.

Shtar, Guy, et al. “A Simplified Similarity-Based Approach for Drug-Drug Interaction Prediction.” *PLoS ONE*, vol. 18, no. 11, 9 Nov. 2023, pp. e0293629–e0293629, https://doi.org/10.1371/journal.pone.0293629.