Drug Target Interaction Prediction

ZHANDOS YESSENBAYEV, B.Sc., M.Sc., Ph.D.

Jason Andrew Hardjawidjaja
Student ID: 2702350781

Azka Dwi Putra Azhad
Student ID: 2702357926

Abstract

Predicting drug-target interactions is crucial to both drug repurposing and drug discovery. The scalability of conventional experimental methods is limited by the expense and slowness of laboratory-based interaction testing. The goal of this project is to use an open-source drug-interaction dataset to create a machine-learning-based pipeline that can predict interactions between known drugs. Aspirin is chosen as the primary reference compound, and Support Vector Machines (SVM), a Neural Network, and a simplified graph-based model are used to assess how it interacts with other medications. The findings demonstrate that computational models can greatly lessen the need for laboratory experimentation and offer insightful predictions. In order to increase accuracy, future work will expand the dataset and add molecular-level features.

Content

INTRODUCTION

Biochemical processes known as drug-target interactions take place when a drug molecule binds to a biological target, such as an enzyme or protein. These interactions determine the positive and negative effects of the medication. The development of new medications depends on identifying these interactions, but testing every possible drug-target combination is impractical because experimental screening is expensive and time-consuming.

Previous research has introduced computational techniques like machine learning, kernel methods, and graph neural networks to expedite the identification of potential drug interactions. These models reduce the cost of experiments while enabling more thorough drug-target space exploration.

This project offers a machine learning framework for predicting drug interactions using an open-access dataset. Aspirin, also known as acetylsalicylic acid, is used as a case study to look into expected drug interactions. The principal inputs are:

1. a full pipeline that includes modeling, assessment, prediction, and data processing.
2. application of three learning strategies: a simplified graph-based model, SVM, and neural networks.
3. An example of how computational techniques can improve drug interaction analysis in practice.

METHODOLOGY

A. Dataset Description

Drug names, encoded features, and interaction labels are all included in the drug-drug interactions dataset. The dataset was chosen to enable interaction prediction using aspirin as the focal drug based on the project documentation. Because the dataset is organized in CSV format, Python-based machine learning workflows can use it.

B. Data Preprocessing

The *drug-prediction.py* script performs the following tasks:

- Conversion of drug names into numerical encodings
- Cleaning of missing or duplicate entries
- Normalizing feature values
- Splitting data into training and testing subsets

This ensures that the dataset is in a suitable format for model training.

C. Machine Learning Models

1) Support Vector Machine (SVM)

The SVM model uses an RBF kernel to classify drug-interaction labels. It serves as the baseline classifier due to its robustness on moderate-sized datasets.

2) Neural Network

A fully connected neural network is implemented using TensorFlow/Keras. It captures nonlinear patterns and is trained using backpropagation. It requires more tuning but offers flexibility in learning feature representations.

3) Graph-Based Model

NetworkX is used to create a simplified graph structure in which interactions form weighted edges and drugs are nodes. Although it serves as a qualitative analysis rather than a formal classifier, this model offers relational insights.

D. Libraries and Tools

The project uses several Python libraries:

- Pandas and NumPy for preprocessing
- Scikit-learn for SVM and evaluation
- TensorFlow/Keras for neural network training
- NetworkX for graph construction
- Matplotlib for visualization

E. Evaluation Metrics

Standard classification metrics are used to evaluate model performance:

- Accuracy
- Precision
- Recall

- F1-score

  Confusion matrix

These metrics quantify how well each model predicts drug interactions.

F. Ranking and Selection of Interactions

Following model inference, the neural network's confidence scores were used to rank the anticipated aspirin-drug interactions. A mixed-selection approach was used to enhance interpretability and prevent repetitive interaction-type dominance. Results were grouped by interaction type and severity, and samples were taken proportionately from each group, rather than choosing the top-N interactions solely based on confidence.

The top 5, 10, or 15 predicted interactions could be dynamically displayed by users without restarting the pipeline thanks to the implementation of an interactive command-line interface. A more varied and clinically representative summary of anticipated interaction types, severity levels, and confidence scores is guaranteed by this method.

RESULTS

A. Dataset Statistics and Preprocessing Results

Experiments were conducted using the DDI_data.csv dataset, which initially contained 222,646 recorded drug–drug interaction entries. After preprocessing steps which include lowercase normalization, whitespace removal, and duplicate elimination, the dataset was standardized for model training.

A total of 1,868 unique drugs were identified, forming the basis for constructing a symmetric drug–drug interaction matrix. This matrix was represented as a sparse structure to reduce memory usage. Latent representations for each drug were generated using Truncated Singular Value Decomposition (SVD) with 30 latent dimensions, capturing structural interaction patterns between drugs.

To enable supervised learning, 222,271 positive interaction pairs were extracted from the dataset. An equal number of negative samples was generated through controlled random sampling of non-interacting drug pairs, resulting in a balanced dataset. The final training set was capped at 20,000 samples to ensure feasible computational time.

B. Feature Engineering Results

Each drug pair was encoded into a high-dimensional feature vector combining structural, textual, and similarity-based information. The final feature representation included:

- Latent SVD vectors for each drug
- Absolute difference and element-wise product of latent vectors
- TF–IDF representations of drug names (100 features)
- Cosine similarity between drug-name vectors
- Drug-name length difference

All features were standardized using z-score normalization. Interaction classes with fewer than two samples were removed, eliminating 12 rare samples. The final dataset contained 19,988 samples across 65 valid interaction classes, ensuring stable stratified training and evaluation.

C. Model Performance Evaluation

Two machine-learning models were trained and evaluated:

1. **Support Vector Machine (SVM)** with RBF kernel
2. **Multi-Layer Perceptron (MLP) Neural Network**

The dataset was split into 14,991 training samples and 4,997 testing samples. Due to computational constraints, SVM training was performed on a randomly sampled subset of 10,000 training instances, while the neural network utilized the full training set.

Table I

Model Performance Comparison

Model	**Precision**	**Recall**	**F1-Score**
SVM	0.3653	0.1927	0.2340
Neural Network	**0.5276**	**0.4285**	**0.4524**

The neural network significantly outperformed SVM across all evaluation metrics, indicating superior generalization capability in multi-class drug interaction prediction.

D. Confusion Matrix and Classification Accuracy

Due to the large number of interaction classes (61 classes in the test set), a full confusion matrix visualization would be impractical. Therefore, a simplified evaluation was applied.

The neural network achieved an overall classification accuracy of 83.93% on valid test samples. When feasible, the confusion matrix was visualized and saved as confusion_matrix.png, providing insight into class-wise prediction behavior while maintaining readability.

E. Aspirin Drug Interaction Prediction Results

After training and evaluation, the neural network model was applied to generate interaction predictions between **Aspirin (Acetylsalicylic Acid)** and all other drugs in the dataset. Feature vectors were constructed dynamically using the same preprocessing and feature-engineering pipeline employed during training, ensuring consistency between training and inference.

Predictions corresponding to invalid or filtered interaction classes were excluded. Model performance on valid test samples achieved an overall classification accuracy of 83.93%, indicating strong generalization across multiple interaction classes.

Rather than summarizing predictions solely by severity counts, the output focuses on confidence-ranked interaction predictions, allowing users to interactively inspect the most relevant predicted drug interactions. A simplified confusion matrix representation was used due to the large number of interaction classes (61 classes).

Table II

Aspirin Interaction Severity Summary

Metric	Value
Reference Drug	Acetylsalicylic Acid (Aspirin)
Total Interaction Classes	61
Model Used	Neural Network (MLP)
Accuracy on Valid Samples	83.93%
Prediction Filtering	Invalid / low-confidence classes removed
Output Type	Ranked interaction predictions

F. Interaction Type Distribution

The most frequently predicted interaction types for Aspirin are shown in Table III.

Table III

Top Predicted Interaction Types for Aspirin

Rank	Interaction Type	Number of Drugs
1	Serum Concentration	24
2	Risk or Severity of Adverse Effects	21
3	Absorption	1

These interaction types are associated with pharmacokinetic and pharmacodynamic effects, suggesting clinically relevant predictions.

G. Top Predicted Aspirin Interactions

A mixed-ranking method that strikes a balance between interaction-type and severity diversity and confidence scores was used to choose the top predicted aspirin interactions. An interactive menu system allows users to dynamically view the top 5, 10, or 15 anticipated interactions.

Table IV

Representative mixed interactions across top **5** types, severity, and confidence levels.

Drug	Interaction Type	Severity	Confidence
Sunitinib	Serum Concentration	Moderate	0.91
Glyburide	Serum Concentration	Moderate	0.89
Sulfamethoxazole	Risk or Severity of Adverse Effects	Moderate	0.87
Dosulepin	Serum Concentration	Moderate	0.85
Disopyramide	Absorption	Moderate	0.84

Table V

Top **10** Mixed Aspirin–Drug Interaction Predictions

Drug	Interaction Type	Severity	Confidence
Sunitinib	Serum Concentration	Moderate	0.91
Glyburide	Serum Concentration	Moderate	0.89
Sulfamethoxazole	Risk or Severity of Adverse Effects	Moderate	0.87
Dosulepin	Serum Concentration	Moderate	0.85
Disopyramide	Absorption	Moderate	0.84
Chlorpropamide	Serum Concentration	Moderate	0.83
Saxagliptin	Risk or Severity of Adverse Effects	Moderate	0.82

Sitagliptin	Serum Concentration	Moderate	0.81
Glipizide	Serum Concentration	Moderate	0.80
Alogliptin	Risk or Severity of Adverse Effects	Moderate	0.79

Table VI

Top **15** Mixed Aspirin–Drug Interaction Predictions

Drug	**Interaction Type**	**Severity**	**Confidence**
Sunitinib	Serum Concentration	Moderate	0.91
Glyburide	Serum Concentration	Moderate	0.89
Sulfamethoxazole	Risk or Severity of Adverse Effects	Moderate	0.87
Dosulepin	Serum Concentration	Moderate	0.85
Disopyramide	Absorption	Moderate	0.84
Chlorpropamide	Serum Concentration	Moderate	0.83
Saxagliptin	Risk or Severity of Adverse Effects	Moderate	0.82
Sitagliptin	Serum Concentration	Moderate	0.81
Glipizide	Serum Concentration	Moderate	0.80
Alogliptin	Risk or Severity of Adverse Effects	Moderate	0.79

Tolbutamide	Serum Concentration	Moderate	0.78
Repaglinide	Risk or Severity of Adverse Effects	Moderate	0.77
Pioglitazone	Serum Concentration	Moderate	0.76
Rosiglitazone	Risk or Severity of Adverse Effects	Moderate	0.75
Linagliptin	Serum Concentration	Moderate	0.74

By preventing a single interaction category from dominating the displayed top-N interactions, the mixed-selection strategy maintains confidence-based ranking while offering a more representative picture of anticipated interaction mechanisms.

H. Generated Outputs and Artifacts

The pipeline automatically generated the outputs:

- Aspirin_interactions.csv: complete prediction results
- aspirin_interactions_major.csv
- aspirin_interactions_moderate.csv
- aspirin_interactions_minor.csv

  Confusion_matrix.png: model performance visualization

  Serialized models: svm_model.joblib, mlp_model.joblib

  Latent feature table: drug_latent_features.csv
- Interactive command-line interface enabling dynamic top-5, top-10, and top-15 interaction exploration without re-running the pipeline

These artifacts ensure reproducibility and support future model extensions.

DISCUSSION

The findings demonstrate how well machine-learning models can recognize patterns of interactions in drug datasets. The SVM was a solid baseline because it performed well and generated reliable predictions. More hyperparameter tuning was needed for the neural network, particularly because of the dataset size and training time limitations. Although the graph-based method offered valuable structural insights, it was not as predictive as the other models.

Limitations

- Model generalization is limited by dataset size.
- There were no molecular fingerprints (like SMILES) included, training took a long time, and model tuning is still unfinished.
- Due to computational limitations, the graph model was simplified.

Despite these limitations, the study successfully demonstrates that ML-based interaction prediction is feasible and scalable. To mitigate output homogeneity in post-prediction analysis, a mixed interaction selection strategy was introduced, improving interpretability without altering model behavior.

CONCLUSION

This study presented a machine-learning-based pipeline for predicting drug–drug interactions using a large-scale interaction dataset. Using 222,646 recorded interactions from the DDI dataset, a comprehensive feature engineering approach combining latent structural representations and textual similarity features was employed. A total of 1,868 unique drugs were modeled through matrix factorization, enabling scalable interaction analysis.

Experimental results demonstrated that the Neural Network model outperformed the Support Vector Machine, achieving a precision of 0.5276, recall of 0.4285, and F1-score of 0.4524, while maintaining a high classification accuracy of 83.93% across valid interaction classes. These results indicate that neural models are better suited for capturing complex, multi-class drug interaction patterns compared to traditional kernel-based approaches.

When applied to a real-world case study, the trained model predicted 46 moderate-severity interactions involving Aspirin (Acetylsalicylic Acid). The majority of predicted interactions were related to changes in serum concentration and an increased risk of adverse effects, which aligns with known pharmacokinetic and pharmacodynamic considerations. Although no major or minor interactions were identified, the results demonstrate the model’s ability to generate clinically interpretable interaction summaries.

Overall, this work confirms that computational drug–drug interaction prediction can effectively reduce reliance on costly laboratory experiments while providing meaningful insights for drug safety and repurposing. Future work will focus on integrating interaction-type taxonomies, molecular descriptors, and advanced graph-based neural networks to improve predictive accuracy and clinical relevance further.

Reference

Yu, Hui (2020), “data of multiple-type drug-drug interactions ”, Mendeley Data, V1, doi: 10.17632/md5czfsfnd.1

Liao, Q., Zhang, Y., Chu, Y., Ding, Y., Liu, Z., Zhao, X., Wang, Y., Wan, J., Ding, Y., Tiwari, P., Zou, Q., & Han, K. (2025). Application of Artificial Intelligence in Drug-target Interactions Prediction: A review. *Npj Biomedical Innovations.*, *2*(1). https://doi.org/10.1038/s44385-024-00003-9

Peng, J., Li, J., & Shang, X. (n.d.). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-020-03677-1

Zhao, Y., Xing, Y., Zhang, Y., Wang, Y., Wan, M., Yi, D., Wu, C., Li, S., Xu, H., Zhang, H., Liu, Z., Zhou, G., Li, M., Wang, X., Chen, Z., Li, R., Wu, L., Zhao, D., Zan, P., . . . Bo, X. (2025). Evidential deep learning-based drug-target interaction prediction. *Nature Communications*, *16*(1), 6915. https://doi.org/10.1038/s41467-025-62235-6

Yuqi, C. Xiaomin, L. Wei, D. Yanchun, L. Garry, W. Liang, C. (2024). Drug–Target Interaction Prediction Based on an Interactive Inference Network. https://www.mdpi.com/1422-0067/25/14/7753

Demirsoy, I., & Karaibrahimoglu, A. (2023). Identifying drug interactions using machine learning. *Advances in Clinical and Experimental Medicine*, *32*(8), 829–838. https://doi.org/10.17219/acem/169852

Han, K., Cao, P., Wang, Y., Xie, F., Ma, J., Yu, M., Wang, J., Xu, Y., Zhang, Y., & Wan, J. (2022). A review of Approaches for Predicting Drug–Drug Interactions Based on Machine Learning. *Frontiers in Pharmacology*, *12*, 814858. https://doi.org/10.3389/fphar.2021.814858