HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

DrugSimNet: Enhancing Drug Representation Learning Through Drug Similarity for Accurate Drug-Drug Interaction Prediction

Jisu Park[1], Ryan Oh[2], Taehee Kim[3] and Devon Marks[#]

[1]North London Collegiate School Jeju, Republic of Korea
[2]Chadwick International School, Republic of Korea
[3]St. Johnsbury Academy Jeju, Republic of Korea
[#]Advisor

ABSTRACT

The field of drug discovery has garnered significant attention, particularly in light of advancements brought about by the "Homo Hundred" generation. Among the critical processes in drug discovery is drug screening, which is an important process in identifying and eliminating candidates that may pose potential side effects in the human body before in vivo experiments. The drug screening is often conducted via prediction of drug-drug Interaction, which utilizes algorithms to assess the likelihood and potential consequences of interactions between different drugs. This approach is a multifaceted process that involves the identification of potential new therapeutic entities by employing a combination of computational, experimental, translational, and clinical models. Traditional approaches to predicting drug-drug interactions in the context of drug discovery and polypharmacy heavily rely on empirical knowledge, in vitro assays, and animal experiments. However, these methods suffer from drawbacks such as being time-consuming, resource-intensive, and limited in their ability to capture the complete complexity of drug interactions. Therefore, there is a pressing need to develop automated and efficient methods that can accurately predict drug-drug interactions. To address the aforementioned problem, we proposed a novel representation learning based framework for prediction of drug-drug interaction. The proposed framework consists of two stages: representation learning, which focuses on extracting meaningful features from drugs, and transfer learning, utilized to train the drug-drug interaction prediction network. Through extensive experimentation, we have shown that the proposed drug-drug interaction prediction framework surpasses existing methods in terms of performance.

Introduction

Drug-drug interactions are a significant concern in modern healthcare, as the simultaneous administration of multiple medications can lead to unexpected adverse effects or diminish therapeutic efficacy. Accurate prediction of potential drug-drug interactions is of paramount importance to ensure patient safety and optimize treatment outcomes. Additionally, drug-drug interactions play a crucial role in the drug discovery pipeline, a complex and multifaceted process aimed at identifying and developing safe and effective therapeutic compounds. During drug discovery, numerous chemical compounds are screened and tested to identify potential drug candidates with desired pharmacological properties. Due to this reason, assessing and predicting potential drug-drug interactions early in the drug discovery process is very important. There have been numerous research efforts focused on drug-drug interactions to solve this problem.

Early methods for drug-drug interaction prediction primarily relied on knowledge-based approaches, which utilized existing knowledge about drug properties, such as chemical structures, pharmacological profiles,

and known interactions. These approaches often utilized expert-curated drug interaction databases, such as DrugBack (Wishart et al. 2008) or Medline (Greenhalgh 1997), to extract relevant information. While knowledge-based methods provided valuable insights, they suffered from limited coverage, as they were heavily based on pre-existing knowledge and were not able to capture novel drug interactions.

To address this problem, numerous machine learning-based drug-drug interaction methods have been proposed in recent years. To begin with, Feng et al. proposed a Deep Predictor for Drug-Drug Interactions (DPDDI) system which predicts drug-drug interaction using graph convolution network (Feng et al. 2020). This method demonstrates the feasibility of utilizing convolutional neural networks to solve drug-drug interaction prediction. Their method achieved an accuracy of 94.0% on the DrugBank dataset. Rozemberczki et al. proposed an unified drug pair scoring framework called ChemicalX (Rozemberczki et al. 2022). Their work provides a unification of drug-drug interaction, polypharmacy side effects and synergistic drug combination prediction tasks. Al-Rabeah et al. exploits the classic graph similarity measurement approach to find better drug representation in order to increase accuracy of the drug-drug interaction prediction system (Al-Rabeah et al. 2022). They achieved an Area Under Receiver Operating Characteristic (AUROC) of 0.990 on the DrugBank dataset. However, a major limitation of these supervised learning-based methods is their reliance on small-scale datasets. Due to the challenges associated with collecting labeled drug-drug interaction samples, researchers often have access to limited datasets, resulting in a narrow representation of drugs. These underdeveloped models are expected to exhibit poor performance in real-world applications, including the evaluation of new drugs within the drug discovery pipeline.

To solve this problem, we propose a novel approach for predicting drug-drug interactions based on drug representation learning. The proposed method is composed of two steps: drug representation learning and transfer learning. In the drug representation learning step, we employ a convolutional neural network as a drug feature extractor to capture valuable features from the input drugs. This trained drug feature extractor extracts meaningful drug characteristics by leveraging pre-measured drug similarity scores. During this stage, we utilize three different widely-recognized drug similarity measurement metrics. In the transfer learning phase, we leverage the pre-trained drug feature extractor as a starting point for training our drug-drug interaction prediction network. This approach makes the drug-drug interaction prediction network more accurate.

Related Work

Drug-Drug Interaction and Drug Discovery Pipeline

Drug-drug interaction occurs when two (or more) medications interact, or when a drug interacts. Drug interaction can alter the way a medication functions or induce undesirable side effects. Drug-drug interactions are important and at the same time fatal because they can affect treatment effectiveness, lead to adverse effects, and increase risk of toxicity or overdose.

Drug-drug interaction prediction is often employed during the initial stages of drug discovery when researchers are screening potential drug candidates. By identifying possible interactions between the candidate drug and commonly prescribed medications, researchers can prioritize drug candidates with lower drug-drug interaction risks, reducing the chances of adverse effects or contraindications.

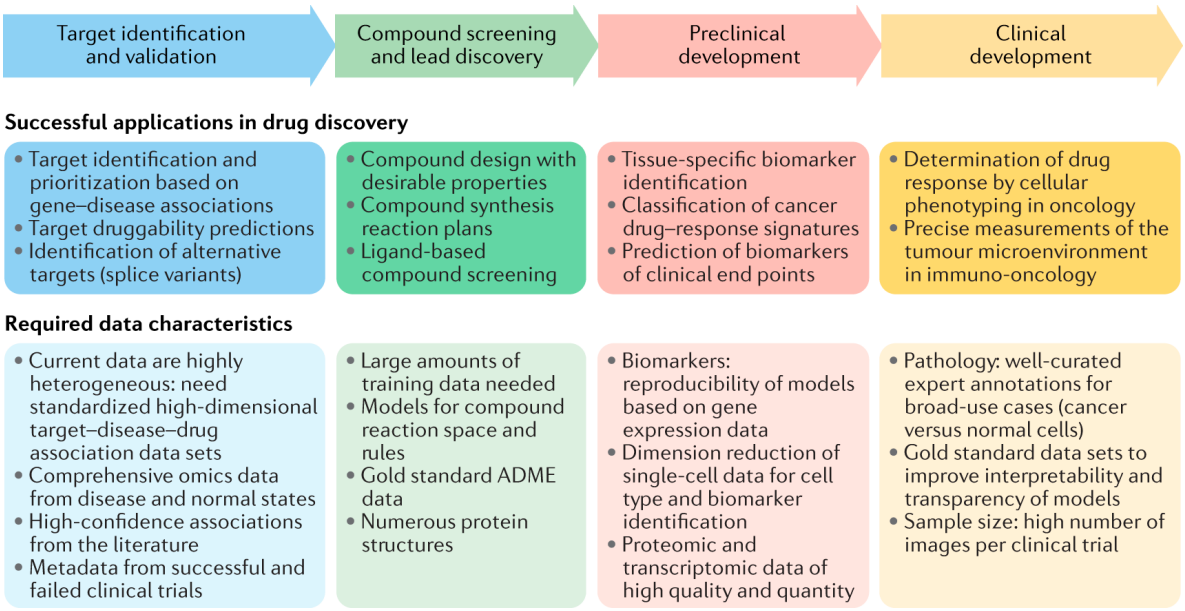

Figure 1. Flow chart of the drug discovery pipeline (Vamathevan et al. 2019)

As shown in Figure 1, the drug discovery pipeline is a complex and multifaceted process that involves the identification, development, and testing of potential new drugs before they can be approved and brought to market. This pipeline typically consists of several stages, each with its own challenges and requirements. The entire process is both time-consuming and labor-intensive, often taking many years and involving significant resources.

In this research paper, we introduce a machine learning-driven system for predicting drug-drug interactions. The proposed system takes a pair of drugs as input and predicts potential drug interactions. We conceptualize this process as an object classification system, as we categorize drug interactions into four distinct categories: mechanism, advice, effect, and interaction. The comprehensive details, including the operational framework, mathematical approaches, and experimental outcomes are explained in Chapters 3 and 4.

Machine Learning Based Object Classification

Image classification is one of the core computer vision tasks, where the computer receives an input image to classify and assigns it to one of a fixed set of categories. In order to successfully implement the image classification system, machine learning systems such as convolutional neural networks can be used. Convolutional neural networks use convolutional layers to capture spatial hierarchies in the training dataset, enabling them to recognize complex patterns. Because of these advantages, this technique finds numerous practical applications in the field of research and industry.

For instance, one promising application is face identification, where the system receives the face image as an input and the confidence score indicating the level of correctness of the identification. Another prominent example of convolutional neural networks being used in medical imaging is covid-19 detection, where the system takes the x-ray image as an input and determines whether the patient is infected to covid-19 or not. In this research, we consider prediction of drug-drug interaction as an object classification since it involves categorizing potential interactions between drugs into distinct classes (or categories) about their effects. The detailed explanation will be provided further in Chapter 3.

Proposed Method

In this chapter, we explain the detailed process of the proposed method including a thorough understanding of its implementation and effectiveness. The proposed drug-drug interaction prediction method consists of two learning stages. The first learning stage involves representation learning, which aims to train the network to learn to extract meaningful features, and transfer learning, where it utilizes pretrained weight to train the drug-drug interaction prediction network.

Drug Representation Learning

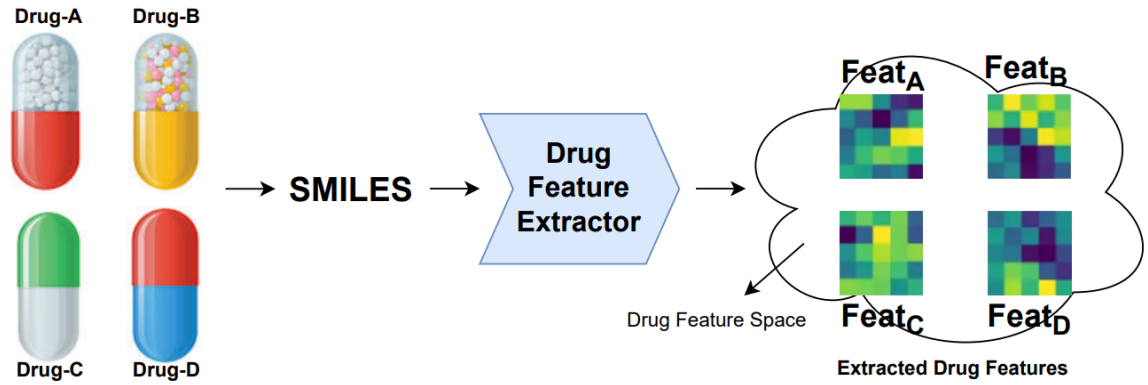

Figure 2. Architecture of the proposed drug representation learning

The ultimate goal of the proposed drug representation learning is to enhance efficiency of drug feature extractor, extracting meaningful drug features that comprehensively capture the molecular characteristics and interactions of drugs. Following procedure starts with preprocessing of chemical formulas of drug samples. All the chemical formulas go through a conversion process with the SIMILES (Toropov et al. 2005) algorithm that transforms the chemical formulas into a simplified and canonical representation of molecular structures. The proposed drug feature extractor takes these simplified formulas as input and produces feature maps that represent the chemical characteristic for each input drug, as output. Here we define this process as follows: $DFE : Drug_{onehot} \rightarrow feat$. Where, *DFE* denotes drug feature extractor, $Drug_{onehot}$ and *feat* represent the simplified formula and extracted feature maps, respectively. To quantify the similarity between two drug features, we employ the commonly utilized cosine similarity function, as expressed in Equation 1.

Equation 1: Similarity function

$$Sim_{A,B} = \frac{feat_A \circ feat_B}{|feat_A| \times |feat_B|}$$

In the equation 1, $\boldsymbol{feat_k}$ denotes the extracted drug features from input *drug-k*. $Sim_{A,B}$ represents the similarity index between two drug features, with values ranging from -1 to 1. To enhance the similarity between drug features extracted from the most similar pairs within the input set of drug pairs, we begin by transforming these similarity indices into probabilities, as illustrated in Equation 2.

Equation 2: Softmax function

$$Prob_{A,B} = \frac{e^{Sim_{A,B}}}{e^{Sim_{A,B}} + e^{Sim_{A,C}} + e^{Sim_{A,D}}}$$

Here, $Prob_{i,j}$ represents the probability of similarity index between drug features $\boldsymbol{feat_i}$ and $feat_j$. Ultimately, loss value is calculated using the cross-entropy loss function as outlined below.

Equation 3: Cross-entropy loss function

$$L_{ce} = -log_e Prob$$

The cross-entropy loss function measures the error between the target probability, computed based on the most closely related drug pairs selected from the input drug set. By employing this loss function, the drug feature extractor learns the ability to consistently extract similar features for chemically similar drugs.

Furthermore, we introduce supplementary neural networks for the direct estimation of drug similarity. While the aforementioned cross-entropy loss function relies on indirect relations between drugs, the direct drug similarity estimation approach provides the network with more robust and explicit information. To quantify the accuracy of this drug similarity estimation, we employ the mean square error function, as explained in Equation 4.

Equation 4: Mean square error function

$$L_{mse} = |y - \hat{y}|^2$$

In Equation 4, y represents the estimated drug similarity, while $\hat{y}$ denotes its actual value, which is pre-computed using conventional drug similarity measurement algorithms. A comprehensive description of these traditional drug similarity algorithms is explained in Chapter 4. The final form of loss function utilized in the drug representation learning is explained in Equation 5.

Equation 5: Final loss function

$$L_{final} = L_{ce} + L_{mse}$$

The final loss function is a linear combination of the cross-entropy loss function and mean square error function.

Drug-Drug Interaction Prediction Network

In the proposed transfer learning, the pre-trained drug feature extractor is utilized as a starting point to train the drug-drug interaction prediction network. Figure 3 illustrates the overall architecture of the transfer learning.

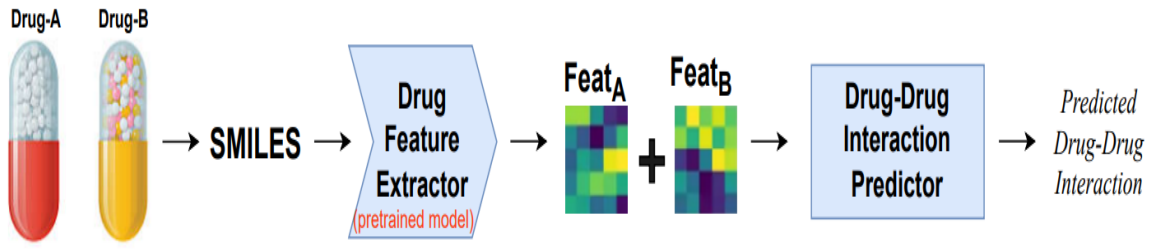

Figure 3. Architecture of the proposed transfer learning (drug-drug interaction prediction)

Similarly, when provided with two input drug pairs, the drug feature extractor generates corresponding drug features. These extracted features are then merged and input into the drug-drug interaction prediction network, which predicts potential drug interactions. This training process employs a straightforward supervised approach, relying on ground truth labels. Consequently, we apply the cross-entropy loss function, which shares the same mathematical form as Equation 3, for training.

The architecture for the drug-drug interaction prediction network is a two-layered neural network. We train the network for 80 epochs with a learning rate of 0.0001. The proposed transfer learning approach outperforms networks trained through conventional supervised methods. Detailed experimental results and comparisons are explained in Chapter 4.

Experimental Results

Dataset

In this chapter, we introduce the dataset used to train and evaluate the proposed drug-drug interaction prediction method. We utilize the DS1 (Zhang et al. 2017), DS2 (Wan et al. 2019), and DS3 (Gottlieb et al. 2012) dataset.

DS1 dataset comprises a collection of 548 unique drugs, with a total of 300,304 drug pairs. Among these pairs, a substantial number, precisely 97,168 interactions, have been identified. DS2 is a larger dataset, consisting of 707 distinct drugs and a more extensive set of drug pairs, totaling 499,849. Interestingly, this dataset contains a smaller number of interactions, specifically 34,412 interactions. DS3 encompasses 807 individual drugs and a vast collection of 651,249 drug pairs. However, this dataset contains a relatively low number of interactions, specifically 10,078 interactions.

Protocol

To assess the performance of the proposed method, we evaluate it using three metrics: precision, recall, and f-score. The calculation of each metric is detailed in Equation 2-4.

Equation 2: Precision

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

Equation 3: Recall

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

Equation 4: F-score

$$F\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

In the equation 2-4, true positive occurs when the model or test correctly identifies a positive instance as positive. In other words, it correctly classifies an example as belonging to the positive class when it does indeed belong to that class. A false positive occurs when the model or test incorrectly identifies a negative

instance as positive. A false negative occurs when the model or test incorrectly identifies a positive instance as negative.

Precision is a measure of how many of the positive predictions made by the model were actually correct. High precision indicates that the model makes positive predictions sparingly and tends to be accurate when it does so. Recall measures the ability of the model to correctly identify all positive instances in the dataset. High recall indicates that the model is effective at capturing all positive instances. Finally, the F-score is the harmonic mean of precision and recall. It provides a balance between these two metrics, helping you assess a model's overall performance.

For comparison, we choose 6 different previous drug-drug interaction prediction methods. The detailed analysis and comparison is explained in chapter 4.3.

Comparison with State-Of-The-Art Methods

Table 1. Performance comparison with previous drug-drug interaction prediction methods

	F-score on each DDI type				Overall Performance		
Method	**Advice**	**Effect**	**Mecha-nism**	**Int**	**Precision**	**Recall**	**F-score**
(Liu et al. 2016)	0.777	0.693	0.702	0.464	0.757	0.647	0.698
(Yi et al. 2017)	-	-	-	-	0.737	0.708	0.722
(Zhang et al. 2018)	0.803	0.718	0.740	0.543	0.741	0.718	0.729
(Xiong et al. 2019)	0.835	0.758	0.794	0.514	0.773	0.737	0.754
(Fatehifar and Karshenas 2021)	0.829	0.759	0.845	0.501	0.785	0.751	0.769
(Molina et al. 2023)	0.845	0.862	0.884	0.784	0.837	0.850	0.843
DrugSimNet (ours)	**0.906**	**0.895**	**0.928**	**0.798**	**0.889**	**0.893**	**0.895**

Table 1 summarizes the performance comparison of the proposed method with previous drug-drug interaction prediction approaches. The first method by Liu et al. (Liu et al. 2016) and the second by Yi et al. (Yi et al. 2017) achieved relatively low accuracy due to their shallow network depth. In contrast, the third method presented by Xiong et al. (Xiong et al. 2019) and the fourth by Fatehifar and Karshenas (Fatehifar and Karshenas 2021) exhibited improved accuracy thanks to their deeper network architectures. However, these supervised approaches did not yield comparable performance. The last method in the comparison, as introduced by Molina et al. (Molina et al. 2023), achieved state-of-the-art performance by incorporating drug similarity during the training process. Ultimately, the proposed method surpasses all of these previous approaches, demonstrating a significant performance advantage.

We attribute this superiority to the proposed approach of using drug similarity for drug representation learning. Throughout the training process, the network learns to extract consistent and meaningful drug features by leveraging drug similarity. This unique training strategy significantly enhances the accuracy of the trained network.

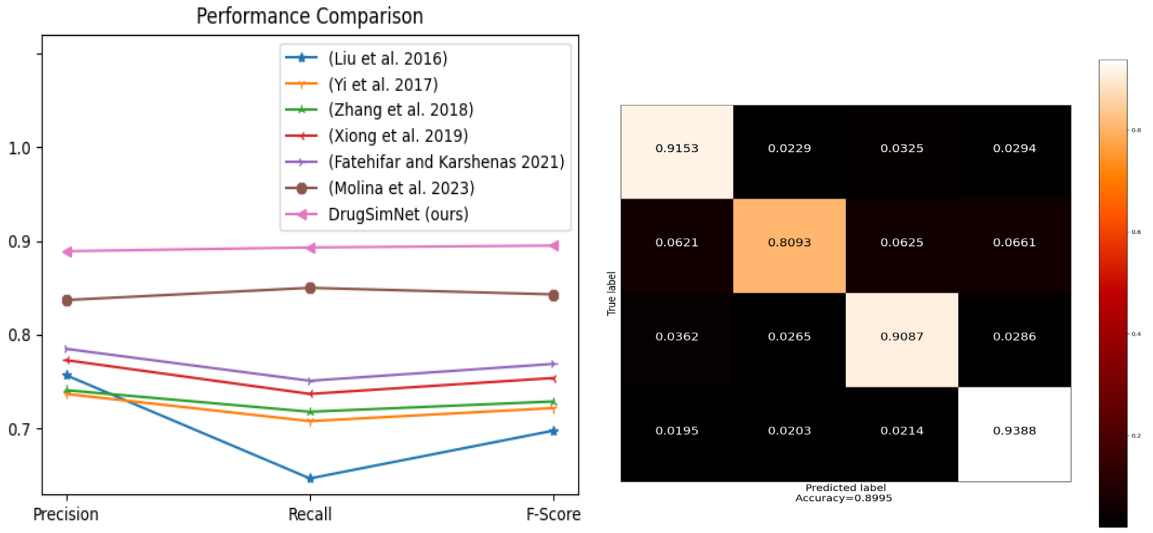

Figure 4. Performance comparison graph (left) and confusion matrix (right)

Figure 4 illustrates the performance evaluation graph and the confusion matrix for the proposed method. As detailed in Table 2, the proposed approach consistently outperforms all previously employed drug-drug interaction prediction methods by a significant margin. The substantial diagonal component values in the confusion matrix highlight the robustness and consistency of our method across all four drug-drug interaction categories.

Ablation Study

In this chapter, we perform an ablation study to assess the influence of the drug similarity prior on the final performance. Initially, we train the drug-drug interaction prediction network without incorporating drug representation learning, which we refer to as the baseline, representing a purely supervised approach. To explore the effectiveness of the drug similarity prior, we introduce three distinct measures of drug similarity: ATC code, Molecular Structure, and Gene Ontology.

For a fair comparison, we keep all training hyperparameters constant, except for the ground truth drug similarity, which is derived from the three aforementioned measurements. We start by training three models using different ground truth drug similarity values calculated from ATC codes, Molecular Structure, and Gene Ontology. These models are referred to as DrugSimNet ATC, DrugSimNet MS, and DrugSimNet GO, as indicated in Table 2.

Table 2. Ablation study results

	Overall Performance		
Method	**Precision**	**Recall**	**F-score**
Baseline	0.772	0.736	0.739
DrugSimNet ATC	0.845	0.859	0.857
DrugSimNet MS	0.852	0.858	0.860
DrugSimNet GO	0.876	0.879	0.880
DrugSimNet (GO+ATC)	0.880	0.882	0.884

DrugSimNet (GO+MS)	0.882	0.885	0.883
DrugSimNet (GO+MS+ATC)	**0.889**	**0.893**	**0.895**

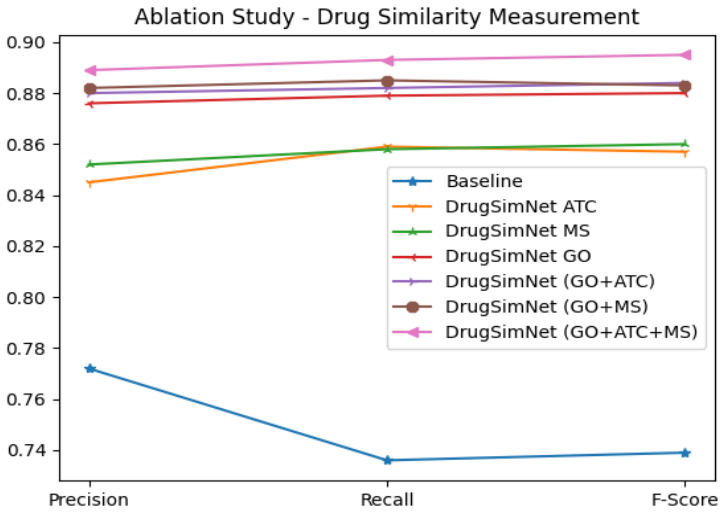

Figure 5. Ablation study performance graph

Table 2 and Figure 5 provide a summary of the results from our ablation study. When compared to the baseline, the three models that incorporate drug similarity-based drug representation learning achieved superior performance, conclusively demonstrating the benefits of utilizing drug similarity priors in classifying drug-drug interactions. In particular, the incorporation of Gene Ontology led to the most significant performance improvement. For a more comprehensive examination, we explored the combined use of multiple drug similarity measures, including Gene Ontology along with ATC code, Molecular Structure, and a combination of Molecular Structure and ATC code.

As displayed in Table 2, the incorporation of multiple combined drug similarity measures has significantly improved the overall performance. Notably, the utilization of all three drug similarity measures has yielded the highest performance. These experimental results clearly demonstrate that integrating drug similarity during the representation learning process substantially enhances the accuracy of the trained model.

Conclusion

In this research, we proposed a novel framework that incorporates representation learning and transfer learning to predict drug-drug interactions. The proposed approach allowed us to extract meaningful features from drugs and train a prediction network, ultimately leading to a significant boost in predictive accuracy. This study highlighted the use of drug similarity priors in drug representation learning. By incorporating drug similarity measures derived from ATC codes, Molecular Structures, and Gene Ontology, we achieved remarkable improvements in performance. Through experimentation, we compared the proposed method against existing approaches. The results clearly showed that the proposed framework surpasses previous methods by a significant margin. We also conducted an ablation study to examine the influence of different drug similarity priors, shedding light on the nuances of their impact. Future research will focus on expanding the scope of drug similarity

priors, incorporating more diverse data sources, and validating our findings through clinical trials and real-world applications.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Al-Rabeah, M. H., & Lakizadeh, A. (2022). Prediction of drug-drug interaction events using graph neural networks based feature extraction. Scientific Reports, 12(1), 15590.

Feng, Y. H., Zhang, S. W., & Shi, J. Y. (2020). DPDDI: a deep predictor for drug-drug interactions. BMC bioinformatics, 21(1), 1-15. https://doi.org/10.1186/s12859-020-03724-x

Gottlieb, A., Stein, G. Y., Oron, Y., Ruppin, E., & Sharan, R. (2012). INDI: a computational framework for inferring drug interactions and their associated recommendations. Molecular systems biology, 8(1), 592.

Greenhalgh, T. (1997). How to read a paper: the Medline database. Bmj, 315(7101), 180-183.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). https://doi.org/10.48550/arXiv.1512.03385

Rozemberczki, B., Hoyt, C. T., Gogleva, A., Grabowski, P., Karis, K., Lamov, A., ... & Gyori, B. M. (2022, August). Chemicalx: A deep learning library for drug pair scoring. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 3819-3828). https://doi.org/10.48550/arXiv.2202.05240

Toropov, A. A., Toropova, A. P., Mukhamedzhanoval, D. V., & Gutman, I. (2005). Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR).

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 18(6), 463-477.

Wan, F., Hong, L., Xiao, A., Jiang, T., & Zeng, J. (2019). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. Bioinformatics, 35(1), 104-111.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., ... & Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 36(suppl_1), D901-D906. https://doi.org/10.1093/nar/gkm958

Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., & Li, X. (2017). Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC bioinformatics, 18, 1-12.