Predicting Drug-Drug Interactions Using Meta-path Based Similarities

Farhan Tanvir
Department of Computer Science
Oklahoma State University
Stillwater, OK 74078, USA
farhan.tanvir@okstate.edu

Muhammad Ifte Khairul Islam
Department of Computer Science
Oklahoma State University
Stillwater, OK 74078, USA
ifte.islam@okstate.edu

Esra Akbas
Department of Computer Science
Oklahoma State University
Stillwater, OK 74078, USA
eakbas@okstate.edu

***Abstract*—Drug-drug interaction (DDI) indicates the event where a particular drug's desired course of action is modified when taken together with other drugs (s). DDIs may hamper, enhance, or reduce the expected effect of either drug or, at the worst possible scenario, cause an adverse side effect. While it is crucial to identify drug-drug interactions, it is quite impossible to detect all possible DDIs for a new drug during the clinical trial. Therefore, many computational methods are proposed for this task. In this paper, we propose a novel method, `HIN-DDI` for discovering DDIs. This method considers drugs and other biomedical entities like proteins, pathways, and side effects, for DDI prediction. We design a heterogeneous information network (HIN) to model relations between these entities. Afterward, we extract the rich semantic relationships among these entities using different meta-path-based topological features. An extensive set of features are fed to different classifiers for DDI prediction. Moreover, we run extensive experiments to compare and evaluate the effectiveness of `HIN-DDI` with other methods. Results exhibit that `HIN-DDI` is quite effective in predicting new drugs as well as existing drugs. Unlike existing works, `HIN-DDI` can predict new drugs, and more importantly, it can impressively outmatch baseline methods by up to 63%.**

***Index Terms*—Drug-drug interaction, Link prediction, Meta-path, Similarity-based, Topological features**

I. Introduction

Adverse drug reactions are becoming a significant health concern in the USA. One-third of adverse drug reactions interactions occur due to drug-drug interactions (DDI), which refers to an event when the desired effect of one drug is altered when taken together with multiple drugs.

Before entering the market, each drug has to undergo tests during drug development and clinical trials. However, performing experiments for many drugs is impractical due to the massive amount of possible drug combinations and various comorbidities. Nowadays, different computational methods have been developed to predict DDIs. Similarity-based approaches presume that drugs that possess similar characteristics interact with the same drug. Moreover, biomedical literature, FAERS, and medical records are used for DDI prediction. However, these prior works have some limitations as follows:

- **Utilizing fewer data sources:** Most previous studies have focused on fewer or single data sources for predicting DDIs. However, one data source may not contain complete information for all drugs. To accurately predict possible DDIs, we need to incorporate multiple data sources [1], [2].
- **Imbalanced dataset:** Drug-drug interaction data is imbalanced and skewed, which we need to consider in the experiments [1], [3].
- **Lack of predicting for new drugs:** Most previous works do not assess their model's capability to predict new drugs. Model tested on current drugs may not give the same effectiveness for new drugs [1], [3].

In this work, we propose a novel DDI prediction model, `HIN-DDI`, to overcome these limitations. First, in `HIN-DDI`, we integrate rich drug-centric interactions from various data sources thanks to the network structure. Networks, often representing real-life systems, are graphs that capture the complex structure of interactions between related objects. In a network, vertices represent the objects, and edges represent the relations between objects. Networks exist in multiple disciplines such as social networks [4], [5], citation networks [6], [7], and biological networks [8].

Next, to model drugs and their interaction with other biomedical entities, we create a heterogeneous information network (HIN), which comprises different types of entities and relations. Lastly, meta-paths on the HIN, which are used to capture the meta-structure of HIN, are used to measure higher-level similarity and relation among two drugs. After creating different meta-paths, topological features are extracted using meta-paths. Finally, few classifiers are trained utilizing these topological features for finding out the accurate model for our features. The main contributions of `HIN-DDI` are as follows:

- **Combining multiple data sources:** Various data sources are incorporated to generate complete, enriched drug-based interaction data. While focusing on limited sources will not provide an accurate representation of drug interactions, blending more features will prevent this issue.
- **Incorporating drug-based interactions on HIN:** From multiple data sources, drugs and their interactions with other entities are represented by a heterogeneous information network. In doing so, we consider drugs' interaction with other biomedical entities like proteins, diseases, and pathways. In addition to that, inhabitant species of proteins and pathway subjects are also considered

978-1-6654-0112-8/21/$31.00

in the drugs' interaction with proteins and pathways, respectively.

- **Meta-path topological features:** Meta-path is used in HIN for measuring connection among entities. We create several meta-paths that indicate different relations between drug pairs. We measure topological features based on constructed meta-paths and utilize these features for the prediction task.
- **Addressing imbalanced and skewed data distribution:** This model addresses imbalanced data distribution through a set of controlled experiments. We determine the actual percentage of interacting drug pairs among all drug pairs through experiments between 30% and 50%.
- **Prediction of new drugs:** Experimental results demonstrate that our model can predict drugs with no known interactions.
- **Evaluation with different accuracy measures:** We perform extensive experiments acknowledging imbalanced data distribution and utilizing appropriate F_1-score, Recall, Precision, AUROC, and AUPR.

The structure of this paper is outlined as follows. Furthermore, we analyze related works in Section II. Section III describes how data from various sources are integrated and explains our methodology. Moreover, we describe our experiments in Section IV. Finally, we conclude in Section V.

II. Related Works

Different computational methods have been applied to predicting DDIs. We will discuss similarity-based and neural network-based approaches.

A. *Similarity-based Approaches*

As per prior studies, similarity-based approaches have proven effective in predicting drug-drug interactions (DDIs). These methods operate under the assumption that similar drugs will interact with the same drugs. Different research works [9], [10] have employed different numbers and types of similarity measures for predicting DDIs. Another research work worthy of mention is [1], which incorporated various data sources to compute numerous local and global similarity measures among drugs. Furthermore, they acknowledged that their dataset is imbalanced and skewed and devised several experiments to address these issues. However, most of these approaches consider limited datasets and fewer drug-centric interactions.

B. *Neural Network-based Approaches*

Recently, a growing number of research based on neural networks, especially graph neural networks, predict drug-related interactions. Graph neural network-based approaches construct knowledge graphs based on drug-centric interactions. Afterward, they employ a neural network to extract relations among drugs. [2] constructed knowledge graph based on protein-protein interaction, drug-drug interaction, and drug-protein interaction. Afterward, they developed a graph convolutional network consisting of encoding, decoding, and model training phases for DDI prediction. [11] utilized GNN to learn drugs and their neighborhood embedding from the knowledge graph and DDI. Finally, they predict potential DDIs using binary classification. [12] predicts DDI leveraging the molecular structure of drugs and type of side effects. Drugs are represented as nodes comprising of atoms, with bonds among them represented as edges. Internal messages are sent to each other by nodes/atoms. Furthermore, atoms of different drugs can communicate with one another via outer messages. They calculated an attentional co-efficient for each atom pair, where each atom belongs to a different drug.

Meanwhile, [13] addressed a similar problem, drug-target prediction, using heterogeneous information network and meta-path topological features. Meta-path has been widely researched and is used to demonstrate relation among entities. It is used in a wide range of applications including the study of research publication networks [14], recommendation [15], multi-network link prediction [16]–[18], co-authorship prediction [19], and information graph-based document analysis [20], [21]. Additionally, meta-paths and similar structures, namely meta-graphs and meta-structures, have been used to detect opioid addicts from social media [22]–[24]. Regarding the prediction of DTIs, [13] constructed a network comprising of drugs, different types of biomedical entities, and interaction among them. Then, they generated meta-paths for computing similarity among drugs. In contrast to existing works, experiments were performed to address the issue of the dataset being imbalanced and predict new DTIs. In addition to that, they used a feature ranking algorithm to select important features.

However, meta-path-based similarity has not been applied to predict DDIs. Though, HIN has been utilized to predict DDIs based on similarity among drugs [25]. Nevertheless, our proposed method, `HIN-DDI`, integrates various datasets comprising drug-based interactions. As a result, our dataset consists of numerous drugs and drug-centric interactions, which addresses prior studies' limitations. Furthermore, we have utilized meta-path to measure relations among drugs.

III. Method

In this section, we present our model. The system architecture of `HIN-DDI` is outlined in Figure 1. Our proposed model `HIN-DDI` for drug-drug interaction prediction, consists of four steps. We explain them in the following sections.

A. *Data Integration*

In this section, we will discuss how we have integrated data into our model. We extract interaction data among different node types from many publicly available datasets. For integrating all these datasets, first, we should understand the complex interactions of bio-molecules and events to detect DDI successfully. Since there are different types of drugs, they can interact with different entities that result in different events. Types of entities in our datasets are drugs, proteins, pathways, chemical substructures, ATC code, and diseases. We also have different interactions between these entities, such as having the same anatomical group in ATC code,

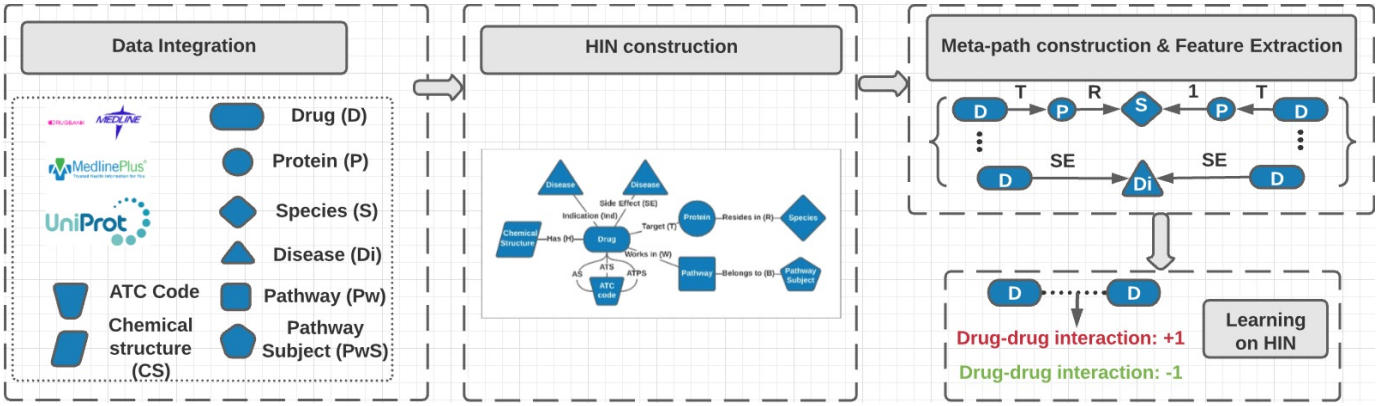

Fig. 1: System Architecture of HIN-DDI

having the same anatomical and therapeutic subgroups in ATC code, and indications/side effects. However, we need to meet specific requirements for a complete, accurate, and rich network. Firstly, appropriate columns or attributes in a dataset for a particular entity must be identified. This task will help in constructing relations among different entities. Secondly, data for a distinct entity may be inconsistent in different datasets. For complete and accurate network construction, we must use one single identifier for an entity. Thirdly, one entity instance can be connected to multiple instances of another entity. We can pre-process data to accommodate one entity instance and concatenate multiple entity instances in a single row for efficiency.

Dataset: We construct our network by integrating a wide range of data sources, including DrugBank [1], KEGG [2], and DEB2 [3] (it combines DrugBank, MedLine, MedLinePlus, Sider2, and NDRFT) and TWOSIDES. We have uploaded our pre-processed datasets in the following GitHub repository: https://github.com/farhantanvir1/HIN-DDI.

In our combined dataset, we have the following interactions.

- *Drug-Protein Interactions:* Different drugs target different proteins in the human body, known as the target protein, and make positive, therapeutic changes. Interaction data among target proteins, species that carry the protein, and drugs are obtained from DrugBank. It contains relational data among 1266 target proteins, 89 species, and 481 drugs.
- *Drug-Pathway Interactions:* The pathway of drugs conveys valuable information about the mechanism of action and metabolism of drugs. Additionally, it bears information on pathway subjects, i.e., disease, protein, physiological. Interaction data among drug pathways, subjects, and drugs are obtained from DrugBank. It consists of relational data among 481 drugs, 48703 pathways, and seven pathway subjects.
- *Drug-Indication Interactions and Side effect data:* Relational data among drugs, indications, and side effects are obtained from DEB2, a publicly available dataset containing an association among 481 drugs and 1602 indication/side effect instances. DEB2 integrates five datasets-DrugBank, MedLine, Med-LinePlus, Sider2, and NDRFT.
- *The chemical substructure of Drugs:* A drug's physiochemical characteristics are determined by its chemical substructure. Each drug can be encoded by the simplified molecular-input line-entry system (SMILES). We extract SMILES string of 481 drugs from DrugBank and KEGG. Then, we convert the SMILES string into MACCS keys, a binary fingerprint consisting of 167 keys. Every bit position corresponds to a chemical substructure, indicating its presence or not
- *ATC code of drugs:* The Anatomical Therapeutic Chemical (ATC) classification system classifies drugs into three categories: operating organs and chemical, therapeutic, and pharmacological characteristics. ATC codes of drugs are obtained from DrugBank and KEGG.
- *Drug-drug interactions:* TWOSIDES database contains information on drug-drug interactions among 481 drugs. Usually, this database is developed by collecting adverse drug effect reports from doctors, patients, and healthcare professionals.

[1]https://go.drugbank.com
[2]https://www.kegg.jp
[3]https://www.vumc.org/cpm/deb2

B. HIN Construction

This section explains how we create a heterogeneous information network to integrate multiple, distinctive entities and their relations.

Definition 1 (Heterogeneous information network): A heterogeneous information network (HIN) is defined as a graph G = (V,E) with an entity type mapping ϕ: V $\rightarrow$ A and a relation type mapping $\psi : \epsilon \rightarrow$ R, where V denotes the entity set and E is the relation set, A denotes the entity type set and R is the relation type set and the number of entity types $|A| > 1$ or the number of relation type $|R| > 1$.

The network schema of HIN-DDI is denoted in the Network Construction part of Figure 1. We utilize datasets described in subsection III-A to construct relations among these entities, which are are explained elaborately below.

- **I1:** T matrix represents the drug-target protein interaction where each element $t_{i,j}$ states whether drug i targets protein j.
- **I2:** R matrix represents which species possess which protein where each element $r_{i,j}$ states whether target protein i can be found in species j.
- **I3:** W matrix represents the relationship among drugs and pathways where each element $w_{i,j}$ describes whether pathway j is responsible for drug i.
- **I4:** The type of activities of the drug pathway may vary, such as metabolic, protein, and drug action. B matrix describes the association of pathway subjects with pathways, where each element $b_{i,j}$ shows whether pathway subject j is related to pathway i.
- **I5:** Ind matrix depicts drug-indication relation where each element $ind_{i,j}$ shows whether drug i cures indication j.
- **I6:** SE matrix represents drug-side effect relation where each element $se_{i,j}$ describes whether drug i causes side effect j.
- **I7:** H matrix outlines drug-chemical substructure relation where each element $h_{i,j}$ refers to whether drug i have chemical substructure j.
- **I8:** AS matrix demonstrates drug-anatomical subgroup of ATC code relation where each element $as_{i,j}$ refers to whether drug i affects organ or system j.
- **I9:** ATS matrix shows the interaction between drugs and the anatomical and therapeutic subgroup of ATC codes. In the matrix, each element $ats_{i,j}$ refers to whether a specific organ and its corresponding therapeutic subgroup j is impacted by drug i.
- **I10:** $ATPS$ matrix illustrates the relation among drugs and anatomical, therapeutic, and pharmacological subgroup of ATC code.Each element $atps_{i,j}$ refers to whether drug i acts on a particular organ and possesses its corresponding therapeutic and pharmacological subgroup j.

C. Meta-path Based Topological Features

After we construct the HIN, we utilize *meta-paths* to extract features for depicting diverse entities and relations. Meta-paths for HIN-DDI are depicted in Figure 2. Meta-paths are used in HINs for measuring relations and similarities between entities. Moreover, meta-paths are represented in the form of a commuting matrix.

Definition 2 (Meta-path): A meta-path P is a path on the network schema diagram $T_G = (A, R)$, and is represented in the shape of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{L+1}$, describing a composite relationship $R = R1 \circ R2 \circ \cdots \circ R$ between entities A_1 and A_{L+1}, where $\circ$ denotes composition operator association, and length of P is L.

Definition 3 (Commuting matrix): Given a network G, a commuting matrix M_P for a meta-path $P = (A_1 A_2 \cdots A_{L+1})$ is defined as $M_P = (G_{A_1 A_2} G_{A_2 A_3} \cdots G_{A_l A_{L+1}})$, where $G_{A_i A_j}$ is the adjacency matrix between types A_i and A_j. $M_P(i, j)$ represents the number of path instances between entity $x_i \in A_1$ and entity $y_i \in A_{L+1}$ under meta-path P.

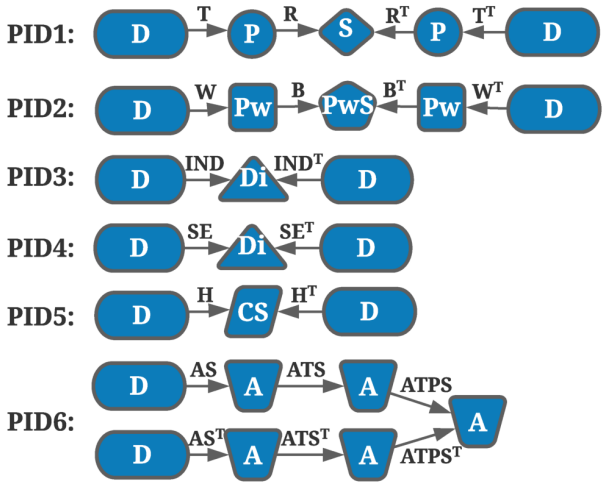

Fig. 2: Meta-paths used in HIN-DDI. Symbols used in this figure are abbreviated forms defined in Figure 1

For example, a meta-path between two drugs can be created as *drug* $\xrightarrow{target}$ *protein* $\xrightarrow{resides}$ *species* $\xrightarrow{resides^T}$ *protein* $\xrightarrow{target^T}$ *drug*. Based on the example mentioned above of meta-path, the commuting matrix for this meta-path is computed by $T * PS * PS^T * T^T$, where T is the adjacency matrix between drugs and target protein; PS is the adjacency matrix between protein and species. Based on ten interactions described in the previous subsection, we construct seven meta-paths and their commuting matrices for measuring similarity among drugs. For example, meta-path PID 1 measures the relation between two drugs based on drugs targeting the same protein in the same species.

After constructing meta-paths, we extract topological features of drug pairs using these meta-paths. Extracted features are later used for predicting interaction between drug pairs. With meta-paths, while predicting interaction, we take the structure and connectivity of the network into account. We utilize four topological features of meta paths on heterogeneous networks. The features are stated below.

- **Path count:** The number of path counts calculates the number of path instances between two entities for a meta-path R referred to as $\mathrm{PC_R}$. The path count can be determined by the commuting matrix that is connected with each meta-path relationship.
- **Normalized path count**: The normalized path count discounts the number of paths between two network entities through their total communication and determines the paths between two network entities. It is defined as

$$NPC_R(a_i, a_j) = \frac{PC_R(a_i, a_j) + PC_{R^{-1}}(a_j, a_i)}{PC_R(a_i, \circ) + PC_R(\circ, a_j)}. \quad (1)$$

In (1), R^{-1} denotes the inverse relationship of R, $PC_R(a_i, \circ)$ denotes the total number of paths beginning with a_i after R, and $PC_R(\circ, a_j)$ denotes the total number of paths ending with a_j after R. $PC_R(a_i, \circ)$ and $PC_R(\circ, a_j)$ can be interpreted as degrees of a_i and a_j in the network relative to R and R^{-1}.

- **Random walk based normalized path count**: The random measure of the walk along a meta-path, which is a generalized version of PropFlow, is specified as

$$RW_R(ai, aj) = \frac{PC_R(a_i, a_j)}{PC_R(a_i, \circ)}. \quad (2)$$

- **Symmetric random walk based normalized path count**: The symmetric random walk takes the random two-way walk and describes it as

$$SRW_R(ai, aj) = RW_R(a_i, a_j) + RW_{R^{-1}}(a_j, a_i). \quad (3)$$

D. Learning on HIN

After extracting topological features with each meta paths for drug pairs' relations, our objective is to predict whether two drugs interact or not. To accomplish this, we can use any machine learning (ML) algorithm to learn the DDI interaction between drug pairs. We experiment with various ML models, including Support Vector Machine (SVM), Logistic regression, Random Forest, and Neural Network. Our purpose is to find out the appropriate ML model capable of predicting DDIs based on our extensive meta-path topological features. Moreover, to find out the actual percentage of DDIs in our imbalanced data, we modify our training and testing data. We perform experiments with various percentages of DDI prevalence in both training and testing data.

IV. Experiment

Performance of HIN-DDI and baseline methods are assessed using different accuracy measures, which are F1-score, Recall, Precision, AUROC, and AUPR.

In addition to the testing on existing drugs, we do experiment for new drugs. We split our dataset to predict new drugs so that 20% of drugs do not appear in the training set and only appear in the testing set. In this case, our testing set consists of 20% of drugs not featuring in the training set. Instead of hiding 20% of the drug-drug interactions [9], [10], 20% of the drugs, which occur as the first component of a pair of drugs in the set $DrugPairsHIN-DDI$ of all recognized drug pairs are concealed. Therefore, these hidden drugs can be considered as newly developed drugs. $Drug_{test}$ comprises hidden drugs for which no DDIs are identified during training. Information for these drugs will only be revealed during testing. $DrugPairsHIN-DDI$ can be categorized into two sets, $DrugPairsHIN-DDI_{train}$ and $DrugPairsHIN-DDI_{test}$. $DrugPairsHIN-DDI_{train}$ comprises drug pairs intended to be used in training. As a result, each pair of $DrugPairsHIN-DDI_{train}$ belongs to $DrugPairsHIN-DDI$, but neither of them belongs to $Drug_test$. On the other hand, $DrugPairsHIN-DDI_{test}$ is constructed for testing, and so, one drug in each pair belongs to $Drug_{test}$. Drug pairs of $DrugPairsHIN-DDI_{test}$ must belong to $DrugPairsHIN-DDI$.

We perform experiments with different machine learning algorithms to see their effect on the results and select the best learning algorithm for our model. We present the detailed results for four different machine learning algorithms: SVM, Logistic regression, Random Forest, and Neural Network in Table I. According to the results, while the neural network can generate extraordinary accuracy measures compared to machine learning methods, random forest produces almost similar results to the neural network. Regarding neural networks, Non-linear and sophisticated relationships can be learned and modelled by these. Additionally, these can infer unseen relationships on unseen data after learning from the input data and their relationships. Moreover, neural networks have the potential to learn hidden relationships in the data without enforcing any fixed associations in the data. Since the neural network achieves more nuanced results on our dataset, we perform further experiments with neural network.

Also, we acquire results of our baseline methods for both existing and new drugs and then evaluate these results against that of HIN-DDI. In addition, we calculate the predictive capability of various features and demonstrate how they influence system performance. Details of experiments and results are discussed in the following sections.

A. Comparison with Baseline Methods

We perceive the issue of predicting drug-drug interactions as similar to a link prediction task. We use 80% of drug data for model training, and 20% of drug data is preserved for testing. The performance of our model is then compared with the following approaches:

- **Concatenated drug features (ConDF):** This approach generates a feature vector for each drug, using Principal component analysis (PCA) representations of the drug-target protein interaction matrix, drug-chemical substructure matrix, drug-anatomical, the therapeutic and pharmacological subgroup of ATC code matrix, drug-pathway matrix, drug-indication matrix, and drug-side effect matrix. Drug pairs are represented by concatenating the corresponding vectors of the drug, and these are used as an input to neural network models, which then predict whether drug pairs interact.
- **Embedding-based Method:** We use two different graph embedding methods: DeepWalk [26] and Node2Vec [27]. DeepWalk is a 2-phase graph embedding technique. It learns d-dimensional node embedding by generating random walks of fixed length from all vertices of a graph. Node2Vec provides an upgrade to DeepWalk [26]. Node2Vec incorporates DFS-like and BFS-like neighborhood discovery with return parameter p and in-out parameter q. First, we have learned the embeddings of drugs using DeepWalk. After that, generated embeddings of drugs are concatenated to represent pairs of the drug. We

TABLE I: Performance comparisons of different machine learning methods and neural network (at 50% DDI prevalence)

	Existing Drugs					New Drugs				
	F_1 score	**Recall**	**Precision**	**AUROC**	**AUPR**	**F_1 score**	**Recall**	**Precision**	**AUROC**	**AUPR**
SVM	53.08	41.89	72.46	62.98	59.41	33.26	21.11	78.38	57.64	55.99
Logistic Regression	58.53	50.55	69.51	64.19	59.86	43.47	30.76	74.09	60	57.41
Random Forest	73.72	**74.36**	73.09	73.49	**67.17**	63.6	58.2	**70.1**	**66.69**	**61.7**
Neural network	**74.91**	74.02	**74.15**	**74.05**	66.98	**63.61**	**64.54**	**66.21**	64.54	59.63

TABLE II: Comparison with baseline methods (with 50% DDI prevalence at training and testing data)

	Existing Drugs					New Drugs				
	F_1 score	**Recall**	**Precision**	**AUROC**	**AUPR**	**F_1 score**	**Recall**	**Precision**	**AUROC**	**AUPR**
HIN-DDI	**74.91**	**74.02**	**74.15**	**74.05**	**66.98**	**63.61**	**64.54**	**66.21**	**64.54**	**59.63**
ConDF	70.45	69.47	70.56	70.71	64.45	57	58.26	59.35	58.26	54.9
Node2vec	45.92	51.48	50.26	50.27	50.68	32.92	31.52	47.58	32.11	42.68
DeepWalk	46.54	44.92	50.35	45.63	50.47	34.20	39.57	40.82	38.19	41.94

have set the dimension of drug embedding to d= 64. After concatenating features of two drugs, the representation size of drug pairs is 128. Parameters of the methods are given as the number of random walks γ = 40, walk length t =10, and window size w=10. p and q in Node2Vec both are set to 1.

In Table II, results for each baseline method and HIN-DDI at 50% DDI prevalence are illustrated. From the results in the table, we observe that HIN-DDI outperforms all baseline methods by a large margin. Meta-paths are utilized in determining meaningful interactions among entities. So, representations obtained from meta-paths on the constructed HIN allow the model to extract semantic relations among drugs accurately. Since other methods do not adopt meta-paths, they are unable to retrieve proper interactions among drugs.

B. Imbalanced Data Analysis

Prior works assumed that drug-drug interaction data is balanced, and the ratio of positive to negative/unlabelled examples is 1:1. To contradict that, we consider DDI prevalence in our data ranging from 30% to 50% [30, 40, 50]. DDI prevalence refers to the percentage of DDI in a data set. With a given DDI prevalence in training data, we generate accuracy measures for testing data with a given DDI prevalence. In addition to that, we consider varying DDI prevalence in training and testing data for baseline methods to compare results.

Our experimental results show that HIN-DDI outperforms other methods in all cases of varying DDI prevalence in the training and testing set. Detailed comparisons of HIN-DDI and baseline methods are shown in Figure 3. For the training set with DDI prevalence ranging from 30 to 50%, our average F_1-score is 72.09% for the testing set with DDI prevalence ranging from 30 to 50%. The average Recall, Precision, AUROC, and AUPR score of HIN-DDI are 70.33, 73.07, 70.22%, and 65.35%, respectively, for varying DDI prevalence in training and testing data.

Similar to our experiment for existing drugs, we consider varying DDI prevalence in training and testing set to predict new drugs. Detailed results are outlined in Figure 4. Experimental results outline that HIN-DDI can achieve effective accuracy measures when it is tasked to predict for newly developed drugs. For DDI prevalence ranging from 30% to 50% in training and testing data, our average F_1-score, Recall, Precision, AUROC, and AUPR score are 60.28%, 61.13%, 67.96, 61.72, and 57.54%.

For predicting both existing and new drugs, HIN-DDI exceeds baselines in all accuracy measures regardless of the DDI prevalence in training and testing data. Moreover, it is worthy of mention that HIN-DDI achieves better accuracy results than the best baseline method (Concatenated drug features) by almost 11%. So, the mixture of meta-path topological features is superior to baseline approaches, even when predictions are made on drugs lacking known interactions.

C. Detailed Analysis on Meta-paths

We compute four kinds of topological features-path count, normalized path count, random walk, and symmetric random walk. Furthermore, we assess the impact of each type of topological feature on HIN-DDI performance. According to our experiments, all types of topological features produce similar accuracy measures. Combining all types of topological features generate far superior evaluation metrics. However, excluding any kind of topological feature except path count has a limited effect on HIN-DDI's predictive efficiency.

We attempt to analyze the impact of each feature on HIN-DDI by performing extensive experiments. After analyzing the results, we consider drug targets, pathways, chemical sub-structures, and drug side effects as core drug features (CDF). Then, we assess the results of adding multiple features to CDF. Table III outlines the result of these combinations. It can be deduced that these combinations differ slightly in generating accuracy measures. The use of more drug features does not enhance the efficiency of predictions. The results indicate that the drug properties, including ATC code and indication, play a minor role in DDI predictions, while the core drug properties determine the prediction efficiency.

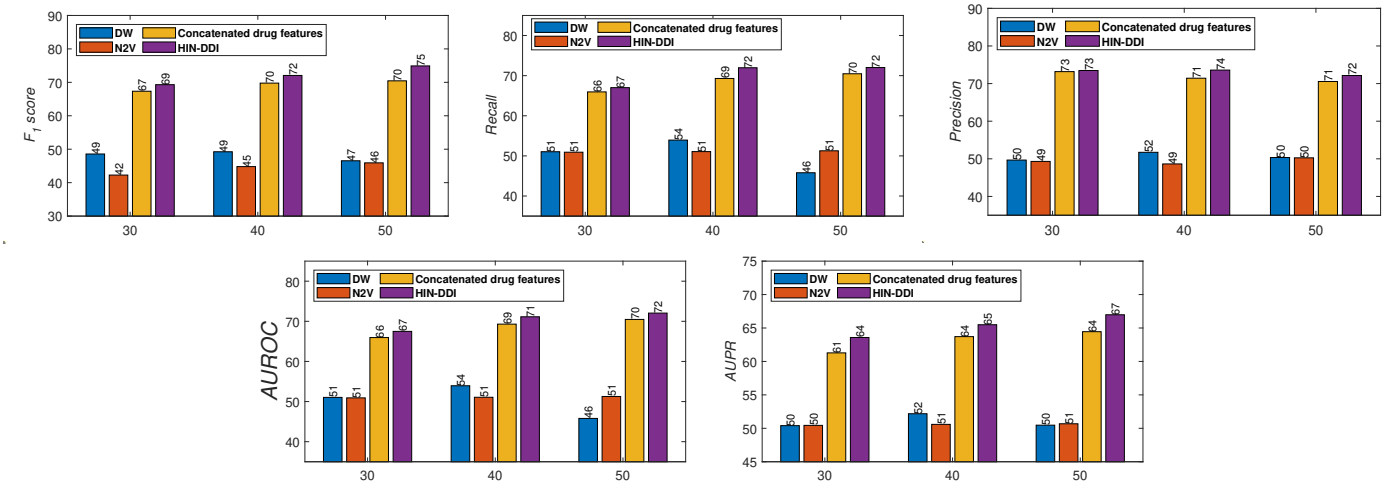

Fig. 3: Evaluating HIN-DDI performance by comparing with baseline methods for existing drugs scenario. For each figure, the x-axis refers to the percentage of DDI prevalence in training and testing data

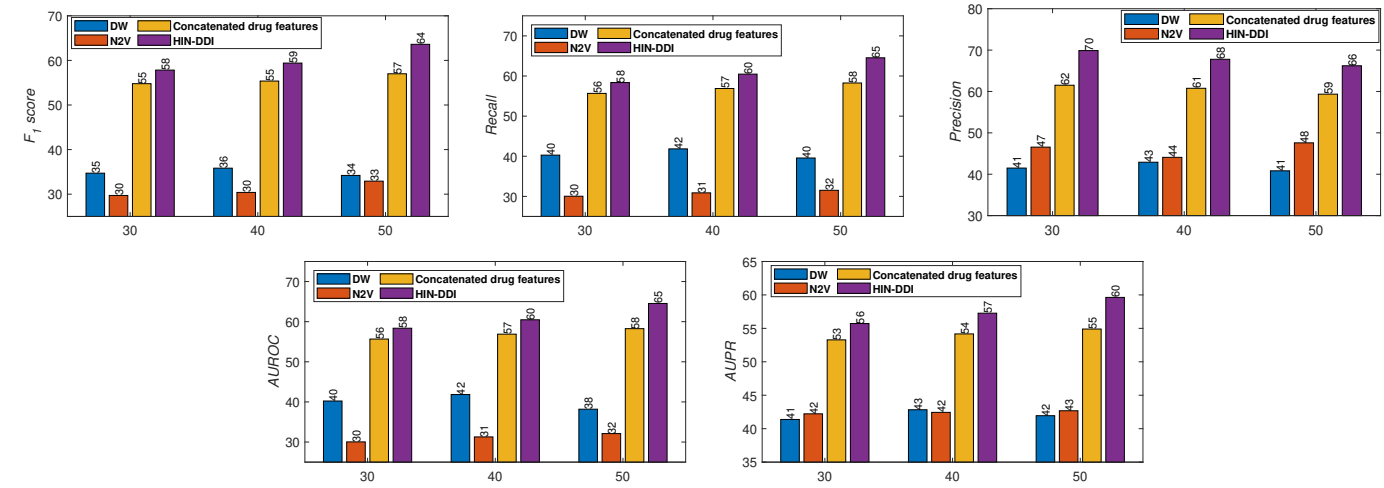

Fig. 4: Evaluating HIN-DDI performance by comparing with baseline methods for new drugs scenario. For each figure, the x-axis refers to the percentage of DDI prevalence in training and testing data

V. Conclusion

In this paper, we propose HIN-DDI, a novel method to predict drug-drug interactions. Firstly, we construct a heterogeneous information network to leverage different entities and their diverse relations. After that, we employ meta-path topological features to denote interaction and relation among drugs. We apply a neural network in HIN to facilitate the prediction of DDIs. Finally, we perform extensive experiments to demonstrate that our method exceeds other baseline methods and addresses existing shortcomings in other works. Our main results are summarized below:

- Prior works performed extensive experiments considering this balanced data distribution. Our experimental results exhibit that our method can predict superbly in the case of imbalanced datasets. The results illustrate the significance and usefulness of the methodology we suggest for treating skewed datasets.
- We outline that the superior prediction efficiency of HIN-DDI stems from a few notable meta-path topological features. Meta-path-based topological features can predict DDIs effectively.
- HIN-DDI performs superior results to best performing baseline methods by a margin of 10% for current drugs and 11% for new drugs. For existing drugs, it surpasses baseline methods by a remarkable 63%.
- The predictive ability of HIN-DDI for new drugs is close to its result for existing drugs' scenarios. It overcomes existing methods' inability to predict in the case of new drugs.

This study can be extended in various directions. Representation learning is known to be very effective in detecting

TABLE III: Combining features to HIN-DDI at 50 percent DDI prevalence at training and testing

		F_1-score	Recall	Precision	AUROC	AUPR
1 feature	Target protein	62.15	63.08	62.91	62.47	58.11
	Side effect	66.16	66.98	66.01	66.47	61.13
	Indication	46.08	55.37	67.27	55.37	55.61
	Pathway	61.78	62.95	63.97	62.49	58.68
	ATC Code	48.19	56.24	66.5	56.24	56.15
	Chemical substructure	60.61	61.17	60.77	60.51	57.11
Combining 4 features	CDF	71.42	71.66	70.34	71.33	65.4
Combining 5 features	CDF + Indication	72.2	71.98	73.01	72.59	66.21
	CDF + ATC Code	72.31	72.22	73.36	72.92	66.59

relationships among biomedical entities [28]. So, future work might accommodate representation learning using a heterogeneous graph neural network. Although meta-path is quite efficient in computing similarity among entities, it can fail to portray complicated relations. Meta-graph is known to be capable of depicting these relations [23]. We can utilize meta-graphs to represent these relationships and evaluate the method.

REFERENCES

[1] I. Abdelaziz, A. Fokoue, O. Hassanzadeh, P. Zhang, and M. Sadoghi, "Large-scale structural and textual similarity-based mining of knowledge graph to predict drug-drug interactions," *J. Web Semant.*, vol. 44, pp. 104–117, 2017.

[2] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinform.*, vol. 34, no. 13, pp. i457–i466, 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty294

[3] Y. Zheng, H. Peng, X. Zhang, Z. Zhao, X. Gao, and J. Li, "Ddi-pulearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions," *BMC Bioinformatics*, vol. 20, 2019.

[4] E. Akbas and P. Zhao, "Truss-based community search: a truss-equivalence based indexing approach," *Proceedings of the VLDB Endowment*, vol. 10, no. 11, pp. 1298–1309, 2017, vLDB Endowment.

[5] M. I. Islam, F. Tanvir, G. Johnson, E. Akbas, and M. E. Aktas, "Proximity-based compression for network embedding," *Frontiers in Big Data*, vol. 3, p. 54, 2020.

[6] W. Tanner, E. Akbas, and M. Hasan, "Paper recommendation based on citation relation," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3053–3059.

[7] G. R. Lopes, M. M. Moro, L. K. Wives, and J. P. M. De Oliveira, "Collaboration recommendation on academic social networks," in *International conference on conceptual modeling*. Springer, 2010, pp. 190–199.

[8] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, no. 1, p. 88, 2007.

[9] A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin, and R. Sharan, "Indi: a computational framework for inferring drug interactions and their associated recommendations," *Molecular Systems Biology*, vol. 8, pp. 592 – 592, 2012.

[10] S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman, and N. Tatonetti, "Similarity-based modeling in large-scale prediction of drug-drug interactions," *Nature Protocols*, vol. 9, pp. 2147–2163, 2014.

[11] X. Lin, Z. Quan, Z. Wang, T. Ma, and X. Zeng, "Kgnn: Knowledge graph neural network for drug-drug interaction prediction," in *IJCAI*, 2020.

[12] A. Deac, Y.-H. Huang, P. Velickovic, P. Lio', and J. Tang, "Drug-drug adverse effect prediction with graph co-attention," *ArXiv*, vol. abs/1905.00534, 2019.

[13] G. Fu, Y. Ding, A. Seal, B. Chen, Y. Sun, and E. Bolton, "Predicting drug target interactions using meta-path-based semantic network analysis," *BMC Bioinformatics*, vol. 17, 2016.

[14] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proc. VLDB Endow.*, vol. 4, pp. 992–1003, 2011.

[15] H. Zhao, Q. Yao, J. Li, Y. Song, and D. L. Lee, "Meta-graph based recommendation fusion over heterogeneous information networks," 2017.

[16] J. Zhang, P. S. Yu, and Z.-H. Zhou, "Meta-path based multi-network collective link prediction." New York, NY, USA: Association for Computing Machinery, 2014.

[17] J. Zhang and P. S. Yu, "Integrated anchor and social link predictions across social networks," in *IJCAI*, 2015.

[18] S. Sajadmanesh, H. R. Rabiee, and A. Khodadadi, "Predicting anchor links between heterogeneous social networks," *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 158–163, 2016.

[19] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," *2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 121–128, 2011.

[20] C. Wang, Y. Song, H. Li, M. Zhang, and J. Han, "Knowsim: A document similarity measure on structured heterogeneous information networks," in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 1015–1020.

[21] C. Wang, Y. Song, H. Li, Z. Zhang, and J. Han, "Text classification with heterogeneous information network kernels," in *AAAI*, 2016.

[22] Y. Fan, Y. Zhang, Y. Ye, X. li, and W. Zheng, "Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies," 2017.

[23] Y. Fan, Y. Zhang, Y. Ye, and X. Li, "Automatic opioid user detection from twitter: Transductive ensemble built on different meta-graph based similarities over heterogeneous information network," in *IJCAI*, 2018.

[24] Y. Fan, Y. Zhang, Y. Ye, X. Li, and W. Zheng, "Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies," 11 2017, pp. 1259–1267.

[25] K. Lee, S. Lee, M. Jeon, J. Choi, and J. Kang, "Drug-drug interaction analysis using heterogeneous biological information network," *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1–5, 2012.

[26] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations." New York, NY, USA: Association for Computing Machinery, 2014.

[27] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[28] M. M. Li, K. Huang, and M. Zitnik, "Representation learning for networks in biology and medicine: Advancements, challenges, and opportunities," *ArXiv*, vol. abs/2104.04883, 2021.