A Multi-Modality Framework for Drug-Drug Interaction Prediction by Harnessing Multi-source Data

Qianlong Wen
qwen@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Jiazheng Li
jiazhengli@brandeis.edu
Brandeis University
Waltham, Massachusetts, USA

Chuxu Zhang
chuxuzhang@brandeis.edu
Brandeis University
Waltham, Massachusetts, USA

Yanfang Ye*
yye7@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

ABSTRACT

Drug-drug interaction (DDI), as a possible result of drug combination treatment, could lead to adverse physiological reactions and increasing mortality rates of patients. Therefore, predicting potential DDI has always been an important and challenging problem. Owing to the extensive pharmacological research, we can access various drug-related features for DDI predictions; however, most of the existing works on DDI prediction do not incorporate comprehensive features to analyze the DDI patterns. Despite the high performance that the existing works have achieved, the incomplete and noisy information generated from limited sources usually leads to sub-optimal performance and poor generalization ability on the unknown DDI pairs. In this work, we propose a holistic framework, namely **M**ulti-modality Feature **O**ptimal **F**usion for **D**rug-**D**rug **I**nteraction Prediction (MOF-DDI), that incorporates features from multiple data sources to resolve the DDI predictions. Specifically, the proposed model jointly considers DDIs literature descriptions, biomedical knowledge graphs, and drug molecular structures to make the prediction. To overcome the issue induced by directly aggregating features in different modalities, we bring a new insight by mapping the representations learned from different sources to a unified hidden space before the combination. The empirical results show that MOF-DDI achieves a large performance gain on different DDI datasets compared with multiple state-of-the-art baselines, especially under the inductive setting.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; *Molecular structural biology*; • **Computing methodologies** → **Information extraction**; **Neural networks**.

*Corresponding author.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614765

KEYWORDS

Drug-drug interactions, Multi-source data, Knowledge graph

ACM Reference Format:
Qianlong Wen, Jiazheng Li, Chuxu Zhang, and Yanfang Ye. 2023. A Multi-Modality Framework for Drug-Drug Interaction Prediction by Harnessing Multi-source Data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3614765

1 INTRODUCTION

The combined medications are commonly used to treat patients with complicated diseases. However, there is always a possibility for the occurrence of adverse drug-drug interactions (DDIs). To be more specific, the metabolisms of different drugs taken by patients at the same time may interfere with each other and thereby could increase the mortality rate. For instance, the combined use of Alprazolam and opioids can significantly increase the risk of opioid-related death [28]. According to the estimation [31], more than 30% of drug side effects are associated with adverse DDIs and it has become one of the leading serious health threats. It has also been reported that there are nearly 74,000 emergency room visits and 195,000 hospitalizations are DDI-induced events in the United States each year, not to mention the statistics of the whole world [26, 29]. Therefore, predicting potential drug-drug interactions has always been an important application in pharmacology and medical health. Owing to the comprehensive research and clinical experiments on drugs, various drug-related features can be leveraged in the DDI research, which enables the application of computational models on DDI predictions. Specifically, drug features like side effects [9], pathway similarity [7], and target gene [17] have been proven to be effective in finding potential DDI pairs. However, those features are mainly annotated by domain experts after extensive clinical experiments, thereby limiting their application to drugs with incomplete features.

To alleviate this issue, previous works propose to employ deep learning models to automatically resolve DDI-related tasks (e.g., DDI prediction task[21, 50] and DDI extraction task [2, 15]) with the easy-accessible raw features. For the DDI extraction task, the medical literature text corpus are usually utilized as the input. In this work, we focus on the DDI prediction task, in which the drug molecular structures converted from SMILES strings [1] or Morgan

fingerprints [30] are commonly employed [25, 41]. Particularly, the sub-structures (functional groups) of a drug molecule usually determine certain properties of the drug, therefore, two drugs with similar molecular structures could have the same impact on humans. In this case, the co-administration of the two drugs may generate overly strengthened effects, which is also counted as a kind of DDI. Meanwhile, the knowledge graphs (KGs) mined from the biomedical literature corpus have been also proven to be a powerful predictor [16, 21]. Besides DDI pairs, the large biomedical graph also includes many other triples which reveal the metabolism activities of the drugs (e.g., drug-gene and drug-disease), thereby benefiting the prediction of potential DDI pairs from the pathway perspective. For example, drugs can only come into effect with the help of some proteins in human bodies (e.g., enzymes), therefore two drugs with sharing the same kind of enzyme will compete for it and interfere the metabolism activities of each other. In spite of the input and label differences, the objectives of the DDI prediction task and DDI extraction task are similar. Given a text description of about co-administration of pair of drugs, DDI extraction task requires training a model to accurately classify the interaction types described in the text.

Most of the existing works only rely on a single kind of feature source to predict the potential DDIs, but it is usually the case that the feature in one source is incomplete and noisy. Inspired by the remarkable performance of multi-modality models, some recent works incorporate biomedical KGs and molecular structures for the DDI prediction task [5, 50] and achieve obvious performance improvement on corresponding tasks compared with those single-source baselines. However, the limitations of those methods still exist, which leaves space for improvements. On one hand, the inherent information missing issue in KGs is still not solved, hurting the model generalization ability. Here, we provide a concrete example in Figure 1 to illustrate the advantage of incorporating contextualized information. During the construction of the biomedical KG and DDI graph, all the drug entities with DrugBank ID will be retrieved from the literature text. Therefore, the DDI pair in the first description (iopidine, opiates) will not appear in the constructed graph because opiates is a general name that can refer to multiple drugs, including morphine, codeine, naltrexone, etc. Meanwhile, the edges that are supposed to exist between iodine and other opiate-related drugs (e.g., morphine) are missed, causing incomplete graph structure for DDI predictions. Fortunately, the contextualized information in the second description can somehow align the representations of "opiates" and "morphine", thereby passing the relation between iopidine and opiates to the representation of "morphine" so that the possible interaction between iopidine and morphine can be detected. Hence, the trained KG embedding is expected to own higher expressive power and reasoning ability if we can inject the rich context information into it. To do so, we need to propose a framework that can jointly optimize the representations of text and KG entity nodes. On the other hand, existing methods usually aggregate the features learned from multiple sources via simple concatenation or weighted aggregation. Though it is computation-efficient, they ignore the fact that these features could be in different embedding spaces. Consequently, such naive fuse operations will lead to meaningless aggregated results and fail to guarantee semantic consistency. Thus,

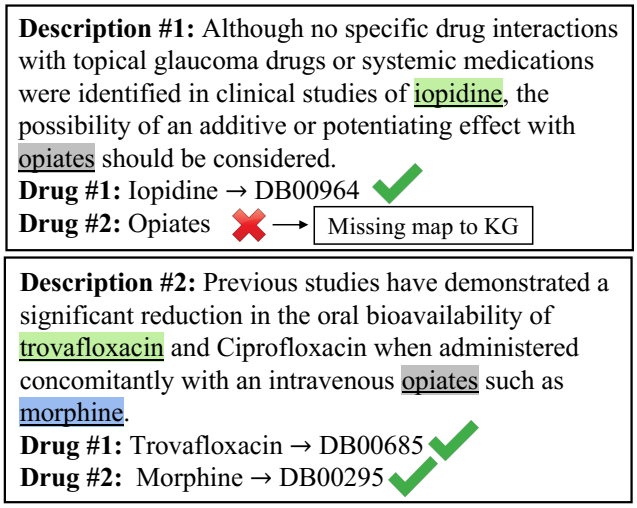


Figure 1: Two DDI description samples from the Pubmed. The DDI pair in the first description is missed in the biomedical KG since opiates can not be mapped to DrugBank ID.

a more reasonable fusion method is promising to further improve the performance.

To fill the gap, we propose a novel method, named **M**ulti-modality **F**eature **O**ptimal **F**usion for **D**rug-**D**rug **I**nteraction (MOF-DDI), to resolve the DDI prediction task. In particular, we incorporate the large biomedical knowledge graph, biomedical literature text corpus and drug molecule structures converted from SMILES string together in our framework. Due to the reason that the biomedical literature text corpus are originally utilized for the DDI extraction task, the features learned from raw texts will not be directly aligned with the features learned from other sources. Instead, we are inspired by the recent work [53] to utilize a bidirectional cross-modality text-graph encoder to pass the feature in text and KG to each other, thereby improving the reasoning ability of learned entity structural embeddings with the contextualized information in texts. Then, the contextualized entity embeddings will be saved and aligned with the embeddings learned from the drug molecule structure. To overcome the spatial heterogeneity between their feature space and maintain semantic consistency, an optimal transport map will be computed by minimizing the p-th Wasserstein distance between the distributions of the two feature spaces. Finally, a unified drug representation will be obtained for the final prediction task. To summarize, our contributions to this work are:

- We propose a novel model, namely MOF-DDI, which enables the integration of features from multiple sources, including literature text from Pubmed, biomedical KG from Hetionet and drug molecular features from Drugbank. To the best of our knowledge, our model is the first work to incorporate text features in DDI prediction task.
- We introduce a new feature fusion method to maintain the semantics consistency: (i) The contextualized features are firstly injected into structural KG features via a cross-modality text-graph encoder for the text-augmented KG representation; (ii) The contextualized KG entity embeddings are fused with the

drug molecule graph embeddings after they are projected into the same hidden space with an optimal transport map.
- Comprehensive experiments are conducted on two DDI datasets. The proposed MOF-DDI achieves state-of-the-art performance by comparison with many baseline methods on both the transductive settings and inductive settings, demonstrating its effectiveness and promising future in predicting potential DDIs.

2 RELATED WORK

2.1 Graph Neural Networks

As a commonly-used and powerful graph learning paradigm, graph neural networks (GNNs) map the non-Euclidean graph-structured data into lower-dimensional hidden spaces for further utilization of the node-level [12, 18, 37] and graph-level tasks [42, 46, 48]. Although different taxonomies can apply existing GNN methods, the key mechanism behind GNNs is message passing, where GNNs learn node representations by transforming and aggregating the information along the edges in graphs [18, 33], and information from multi-hop neighbors can be captured by stacked layers. Therefore, the learned node representations are generally optimized to preserve the proximity features, i.e., the node representations can reflect their neighborhood distributions [23]. GNN variants mainly focus on improving the transformation or aggregation functions [12, 37, 46, 51], achieving better effectiveness or scalability. Due to the remarkable performance of GNN methods on graph data, they have been broadly used in many applications, such as molecule analysis [14, 45], social network analysis [8] and medical health applications [43, 52].

2.2 Text-KG Augmentation

In the training of language models (LMs), external knowledge is commonly employed to augment the input data in many NLP tasks [11, 19, 54]. Specifically, the structural information in the knowledge graph has been proven to be effective in grounding the sequential token inputs and further improving the model reasoning ability. Though most of them aim to enrich context features with the information in KGs, those methods can be divided into two lines based on the optimization strategy. One widely adopted way is to directly add the pre-trained KG embeddings to the LM input and keep them frozen afterward. While the other line proposes to use the cross-modality encoder to jointly optimize the features in two modalities to learn better embeddings for both of them [47, 53]. Despite these works are initially proposed to improve the reasoning ability of LMs, like question answering [20, 39] and text generation [49], the second line opens a door to augment the KG embeddings and alleviate information incomplete issue in KGs with the contextualized feature in raw text. Inspired by this, we aim to study the effect of incorporating text features to enhance the entity embeddings learned from the incomplete KGs.

2.3 Drug-Drug Interaction Prediction

The applications of machine learning models in predicting Drug-Drug Interactions are widely studied. We generally assort the previous methods into three categories according to the features they use: (1) Expert features that require professional verification, like side-effect, molecular structure similarity [30], genomics similarity [17] and pathway similarity [7]. Those features reveal the working mechanisms of DDIs from different perspectives and thereby provide valuable information to analyze the probability of potential DDIs. However, an important drawback of using these features for predictions lies in their accessibility. Such professional knowledge usually needs the huge human cost of domain experts, which limits the construction of large-scale datasets used for model training; (2) Handcrafted features from existing observation, like biomedical knowledge graph [16, 21], drug molecular structures [32, 41], etc. Though human efforts are still necessary to construct those datasets, the expertise levels that they require are relatively low because we only need to set up a specific rule to convert this information to structural datasets from existing experimental observations. Previous works usually employ state-of-the-art deep model architectures in different fields to learn drug representations from the raw features and combine them in pairs for DDI predictions. (3): Fused handcrafted features. It is usually the case that there is incomplete and noisy information in a handcrafted feature, e.g., the missing links in KGs due to the mapping mistake during KG construction. Consequently, the trained deep learning models can not achieve satisfying performance equally on all the drugs. Therefore, some previous works [2, 5, 15, 50] try to incorporate two kinds of features mentioned above for better generalization ability. In spite of their success, existing DDI predictions still can not propose a framework to overcome the incomplete issue in KG and find an optimal way to fuse the features in different modalities. In light of this, we study the problem of using contextualized information in the raw biomedical literature corpus to supplement the incomplete knowledge in KG and introduce a new fusion strategy to maintain semantic consistency.

3 PRELIMINARY

For the convenience of understanding our paper, we first introduce some basic concepts used throughout this paper, then formally define the problem.

Definition 1. **Biomedical Knowledge Graph** Given the biomedical entity set $\mathcal{V}$ and the relation type set $\mathcal{R}_{KG}$, the corresponding biomedical knowledge graph (KG) can be defined as a set of triplets, i.e., $\mathcal{G}_{KG} = \{(v_h, v_t, r) | \, v_h, v_t \in \mathcal{V}, r \in \mathcal{R}_{KG}\}$. Similarly, the drug-drug interaction pairs can be expressed in the same form, i.e., $\mathcal{G}_{DDI} = \left\{(D_i, D_j, r) \middle| \, D_i, D_j \in \mathcal{D}, r \in \mathcal{R}_{DDI}\right\}$, where each triplet (D_i, D_j, r) represents the co-administration of drug D_i and drug D_j could lead to pharmacological effect r. Specifically, there are also drug-drug triplets in $\mathcal{G}_{KG}$ representing the drug-drug interaction pairs. However, unlike the triplets in $\mathcal{G}_{DDI}$ with diversified interaction types, all of the drug-drug triplets in $\mathcal{G}_{KG}$ are categorized into the same interaction type, i.e. the two drugs will interact with each other when taken at the same time. We align the drugs and DDI pairs within the two graphs, i.e., $\mathcal{D} \in \mathcal{V}$ and $\mathcal{G}_{DDI} \in \mathcal{G}_{KG}$, and remove the overlapping drug-drug interaction pairs in them to avoid information leakage.

Definition 2. **Local KG construction and Text Retrieval.** In this work, the fusion of text features and KG features is operated on the local KG of a pair of drugs in $\mathcal{G}_{KG}$ and the text description. Given a pair of drugs, we first construct the local KG of them by returning their 2-hop subgraph $\mathcal{G}_S \in \mathcal{G}_{KG}$. We represent the

node set of $\mathcal{G}_S$ as $\mathcal{V}_S = \{v_1, v_2, \ldots, v_N\}$. Then the corresponding literature reaction description in which the name of the two drugs are included will be retrieved from the PubMed corpus, denoted as $S = \{w_1, w_2, \ldots, w_M\} \in \mathcal{S}$. Then, the combination of the local knowledge graph G_S and the sentence S will be taken as an input instance, i.e., $X = (\mathcal{G}_S, S)$.

Definition 3. **Multi-relational DDI Prediction** The DDI prediction problem is to learn a machine learning model which can be further utilized to predict the pharmacological effect of any co-administrated drug pair (D_i, D_j). It can be formulated as:

$$\mathcal{F} : \mathcal{D} \times \mathcal{D} \rightarrow \mathcal{R}_{DDI}. \tag{1}$$

In this work, we evaluate the trained model under both the transductive setting and inductive setting, where any DDI pair in the testing set includes at least one drug not existed in the training set under the inductive setting.

4 METHODOLOGY

In this section, we present details of the proposed method MOF-DDI, which is shown in Figure 2. The pipeline of MOF-DDI can be divided into two phases and each of them is designed to incorporate different kinds of features. In the first phase, we aim to augment the KG embedding with the contextualized information in the DDI description text. The contextualized KG node embeddings of the drug entities will be saved to combine with the drug molecular embeddings learned from their chemical structure and the DDI graph $\mathcal{G}_{DDI}$ during the second phase. Instead of directly minimizing the embedding distance between them, an optimal transport is iteratively computed to map the two kinds of drug features into the same spaces for semantic consistency. Finally, the aligned embeddings are aggregated and utilized for the DDI prediction. More details about our proposed MOF-DDI are illustrated next.

4.1 Contextualized KG Representation

As we mentioned in the introduction, the rich context information in the literature description about DDI pairs can contextualize the structured information within the biomedical graph and further enhance the reasoning ability. Since the raw literature text and knowledge graph are in different modalities, we are inspired by recent work [53] to employ a cross-modality text-graph encoder to model the interaction between text and KG. Given an input instance $X = (\mathcal{G}_S, S)$ defined in Definition 2 in Section 3, we first map the raw input into initialized representations with a pre-trained BERT model [10] and KG entity embeddings (e.g., TransE [3]). This initial mapping procedure is formulated as:

$$\begin{aligned} \left(\mathbf{v}_{\text{int}}^{(0)}, \mathbf{v}_1^{(0)}, \ldots, \mathbf{v}_N^{(0)}\right) &= \text{KG-Embedding } (v_{\text{int}}, v_1, \ldots, v_N), \\ \left(\mathbf{w}_{\text{int}}^{(0)}, \mathbf{w}_1^{(0)}, \ldots, \mathbf{w}_M^{(0)}\right) &= \text{BERT } (w_{\text{int}}, w_1, \ldots, w_M), \end{aligned} \tag{2}$$

where we follow [53] to add a specific node v_{int} to G_S that is connected with all the entity nodes in the original G_S and insert a special token w_{int} at the beginning of S. Their representations will serve as the bridge of the feature fusion between the two independent modalities in our design.

The cross-modality text-graph model is stacked by L text-graph fusion layers and each of them is equipped with a Transformer encoder $f_{\text{LM}}(\cdot)$ and a GNN encoder $f_{\text{GNN}}(\cdot)$ to process the input text representation and KG representation, respectively. Then, the representation of w_{int} and v_{int} will be concatenated and go through a modality interaction module InfoEx$(\cdot)$ to exchange the information with each other,

$$\begin{aligned} \left(\widetilde{\mathbf{v}}_{\text{int}}^{(l)}, \mathbf{v}_1^{(l)}, \ldots, \mathbf{v}_J^{(l)}\right) &= f_{\text{GNN}}\left(\mathbf{v}_{\text{int}}^{(l-1)}, \mathbf{v}_1^{(l-1)}, \ldots, \mathbf{v}_N^{(l-1)}\right), \\ \left(\widetilde{\mathbf{w}}_{\text{int}}^{(l)}, \mathbf{w}_1^{(l)}, \ldots, \mathbf{w}_I^{(l)}\right) &= f_{\text{LM}}\left(\mathbf{w}_{\text{int}}^{(l-1)}, \mathbf{w}_1^{(l-1)}, \ldots, \mathbf{w}_M^{(l-1)}\right), \\ \left[\mathbf{w}_{\text{int}}^{(l)}; \mathbf{v}_{\text{int}}^{(l)}\right] &= \text{InfoEx}\left(\widetilde{\mathbf{w}}_{\text{int}}^{(l)} \oplus \widetilde{\mathbf{v}}_{\text{int}}^{(l)}\right), \end{aligned} \tag{3}$$

where $\oplus$ is the concatenation operation and the input of the first fusion layers is the initialized node/token representations in Equation 2. Particularly, we adopt a 2-layer MLP network as the InfoEx$(\cdot)$ module and its output is equally split into $\mathbf{v}_{\text{int}}^{(l)}$ and $\mathbf{w}_{\text{int}}^{(l)}$ as the final representation of two special node/token produced by the l-th layer. In this case, the information from one modality is enabled to propagate to another one through the interaction node/token even though other tokens/nodes are not directly involved in the modality interaction process. After L-layer iteration, the contextualized KG representations of the drug entities in the biomedical KG will be selected and saved as $\mathbf{Z}_{\text{KG}} \in \mathbb{R}^{|\mathcal{D}| \times d_k}$ for further use.

4.2 Drug Molecular Structure Representation

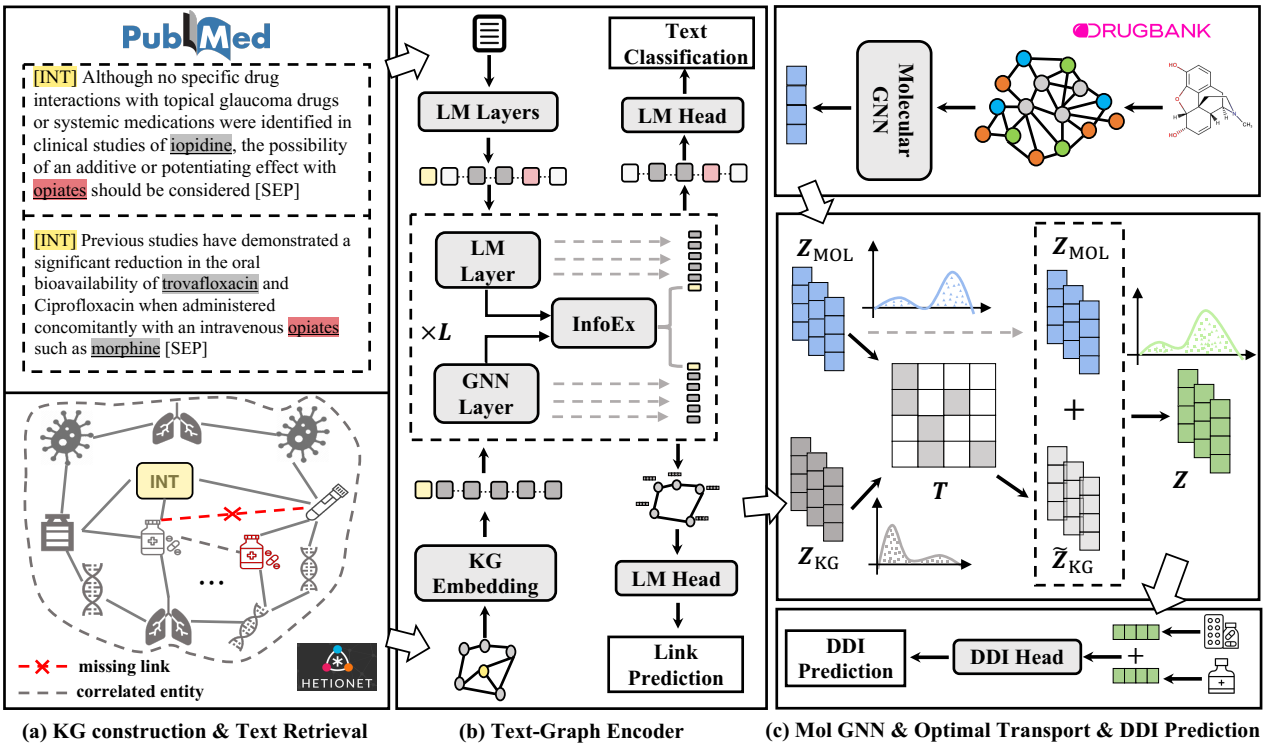
In this section, we present the process of learning drug molecular representations. For each drug $D \in \mathcal{D}$, its molecular structure can be converted to a graph when we represent each atom and chemical bond with a node and edge, respectively. The constructed molecular graph is denoted as $g = \{V, E\}$, where V and E are the atom set and chemical bond set within D. To map each molecular graph with various atoms and chemical bonds into low-dimensional embedding space, we employ a K-layer molecular GNN to process the raw molecular graph. Due to the reason that the initial features of atoms and chemical bonds are discrete, we first construct the input node feature matrix of the molecular GNN by mapping the discrete atom feature vectors $\mathbf{a}_i$ of node i into continuous space with trainable embedding matrices $\mathbf{W}_{atom} \in \mathbb{R}^{d \times d_a}$, where d_a is the dimension of $\mathbf{a}_i$. The initialization of chemical bonds is similar to the atoms except each bond $\mathbf{b}_{ij}$ that connects atom i and atom j will be initialized at the beginning of every GNN layer. The initialization matrix $\mathbf{W}_{bond}^{(l)} \in \mathbb{R}^{d \times d_b}$ can be either independent or shared across different GNN layers. So, the initialization and message-passing procedures are,

$$\begin{aligned} \mathbf{h}_i^{(0)} = \text{ReLU}\left(\mathbf{W}_{atom} \cdot \mathbf{a}_i\right) \; ; \; \mathbf{h}_{ij}^{(l)} &= \text{ReLU}\left(\mathbf{W}_{bond}^{(l)} \cdot \mathbf{b}_{ij}\right), \\ \mathbf{h}_i^{(l)} &= \sum_{\text{j} \in \mathcal{N}(\text{i})} \mathbf{W}_{\text{M}}^{(l)} \cdot \left(\mathbf{h}_{\text{j}}^{(l-1)} \oplus \mathbf{h}_{\text{ij}}^{(l)}\right), \end{aligned} \tag{4}$$

where $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$, $\mathbf{W}_{\text{M}}^{(l)} \in \mathbb{R}^{d \times 2d}$, and $\mathcal{N}(\text{i})$ represents the neighbor set of atom i. After K-layer iterations, the drug molecule representation can be obtained by a readout function,

$$\mathbf{h}_g = \frac{1}{|V|} \sum_{i \in V} \left(\mathbf{h}_i^{(I)}\right). \tag{5}$$

The learned drug molecular representations above are saved as $\mathbf{Z}_{\text{MOL}} \in \mathbb{R}^{|\mathcal{D}| \times d}$ and benefit the DDI prediction from the view of fundamental chemistry mechanisms.



(a) KG construction & Text Retrieval (b) Text-Graph Encoder (c) Mol GNN & Optimal Transport & DDI Prediction

Figure 2: The overall Framework of MOF-DDI. (a) The construction of local KG and the retrieval of text description given a DDI pair. An interaction node/token will be added to the text/KG for the information exchange. (b) The procedure to fuse the information in the biomedical knowledge graph and raw literature text. (c) The hidden space alignment between the text-KG embeddings and the molecular structure embeddings (learned from drug molecular graph) via the optimal transport map. Finally, the combination of mapped text-KG embeddings and molecular structure embeddings is used for the DDI prediction.

4.3 Multi-modality Representation Alignment

Though it has been proved that the combination of the representations learned from multiple data sources can significantly boost the performance of DDI prediction [5, 50], it might be sub-optimal to directly combine them for the final prediction. To be more specific, $\mathbf{Z}_{\text{KG}}$ is learned from the structured knowledge graph which reflects the drug metabolism pathway information, while $\mathbf{Z}_{\text{MOL}}$ is based on the drug molecular structure. Therefore, it is safe to say $\mathbf{Z}_{\text{KG}}$ and $\mathbf{Z}_{\text{MOL}}$ are learned from the features in different modalities. As stated in [4], the direct fusion of representations in different modalities could lead to meaningless intrinsic distribution and thus fail to maintain the semantic consistency in the separated representations. For example, there will be certain dimensions of the $\mathbf{Z}_{\text{KG}}$ indicate what kinds of transporter proteins are involved in the metabolism activity of the drugs, meanwhile, the corresponding dimensions of the $\mathbf{Z}_{\text{MOL}}$ may represent the whether the drugs molecule contain an aromatic ring or not. In this case, the aggregation of those dimensions will be meaningless and destroy the semantic consistency, decreasing the expressive power of learned representations.

To alleviate this issue, we would like to map drug KG embeddings and drug molecular embeddings into the unified embedding space before combining them for the DDI prediction task. This problem can be formulated into the optimal transport problem [38] in which we need to find a transport map $\mathbf{T}$ to project one probability distribution to another with minimized cost evaluated by p-th Wasserstein Distance. Specifically, given the learned $\mathbf{Z}_{\text{KG}}$ and $\mathbf{Z}_{\text{MOL}}$, we first need to compute the ground distance between each dimension of them, where the ground distance between the i-th embedding dimension of $\mathbf{Z}_{\text{KG}}$ to the j-th embedding dimension of $\mathbf{Z}_{\text{MOL}}$ can be formulated as:

$$\mathbf{C}_{ij} = \left\| \mathbf{Z}^i_{\text{KG}} - \mathbf{Z}^j_{\text{MOL}} \right\|^p_p, \forall i \in [d_k], j \in [d], \tag{6}$$

where we set $p = 2$ in this work. Assuming the empirical probability distribution of $\mathbf{Z}_{\text{KG}}$ and $\mathbf{Z}_{\text{MOL}}$ being μ and ν, given the computed ground distance $\mathbf{C}$ above, the optimal transport map is defined as:

$$OT(\mu, \nu, \mathbf{C}) = \min_{\mathbf{T} \in \Pi(\mu,\nu)} \sum_{i=1}^{d_k} \sum_{j=1}^{d} \mathbf{T}_{ij} \mathbf{C}_{ij}, \tag{7}$$

where $\mathbf{T} \in \mathbb{R}^{d_k \times d}$, T_{ij} is the transported probability from $\mathbf{Z}^i_{\text{KG}}$ to $\mathbf{Z}^j_{\text{MOL}}$ and $\Pi(\mu, \nu)$ denotes the joint probability of μ and ν. The optimal transport map $\mathbf{T}$ is iteratively computed by the Sinkhorn algorithm [6]. Then, we can use the optimal transport map $\mathbf{T}$ to project $\mathbf{Z}_{\text{KG}}$ into the unified space with $\mathbf{Z}_{\text{MOL}}$, the aligned drug KG representations are computed as,

$$\widetilde{\mathbf{Z}}_{\text{KG}} = \text{diag}\left(\frac{\mathbf{1}_d}{d}\right)\left(\mathbf{T}^\top + \Delta_{\mathbf{T}}\right)\mathbf{Z}_{\text{KG}}, \tag{8}$$

where $\mathbf{1}_d \in \mathbb{R}^d$ is all one vector, $\Delta_{\mathbf{T}}$ is the adjustable parameter with same size as $\mathbf{T}$. After obtaining the aligned drug KG embeddings,

we then carry out the fusion procedure by:

$$\mathbf{Z} = \lambda \cdot \widetilde{\mathbf{Z}}_{\text{KG}} + (1 - \lambda) \cdot \mathbf{Z}_{\text{MOL}}, \tag{9}$$

where λ is the balanced hyper-parameters, the fused drug representation $\mathbf{Z}$ will be utilized for the final DDI predictions.

4.4 Objective Function

In this section, we introduce the optimization objectives of our method. Concretely, there are two phases in our pipeline, and they own different optimization objectives.

4.4.1 Pre-training Objective. During the optimization of the text-augmented KG representation procedure, we jointly optimize the representations of text and KG with two tasks, i.e., the DDI extraction task and the link prediction task. In particular, the DDI Extraction task is to predict the DDI interaction type based on the given text description, therefore it can be also treated as a multi-class text classification task. However, unlike the DDI prediction task, there are only five kinds of interaction types in the label set, i.e., $C = \{\text{Mechanism}, \text{Effect}, \text{Advice}, \text{Int}, \text{Negative}\}$. In particular, we will append a text classification head $f_{\text{TC}}(\cdot)$ and a link prediction head $f_{\text{LP}}(\cdot)$ on top of the cross-modality text-graph encoder introduced in Section 4.1 to perform the two tasks, where both of $f_{\text{TC}}(\cdot)$ and $f_{\text{LP}}(\cdot)$ are MLP-based networks. Given the text embedding $\{\mathbf{w}_{\text{int}}, \mathbf{w}_i\}_{i=1}^{M}$ of S, it will be fed into $f_{\text{TC}}(\cdot)$ to predict the DDI interaction probability score of text S,

$$\hat{\mathbf{y}}_S = \text{SoftMax}\left(f_{\text{TC}}\left([\mathbf{w}_{\text{int}} \oplus \mathbf{w}_1 \oplus \cdots \mathbf{w}_M]\right)\right), \tag{10}$$

where the $\hat{\mathbf{y}}_S$ is the predicted probability score of DDI extraction task. Thus, the objective of the DDI extraction task is:

$$\mathcal{L}_{\text{TextCls}} = -\sum_{S \in \mathcal{S}} \sum_{k=1}^{|C|} [\mathbf{y}_S]_k \log [\hat{\mathbf{y}}_S]_k, \tag{11}$$

where $\mathbf{y}_S$ is a one-hot vector that indicates the DDI relation types described in text S.

On the other hand, the link prediction task is still similar to that in the knowledge graph models. For any triplet $(v_h, v_t, r) \in \mathcal{G}_S$, we retrieve the embedding of entity from $\{\mathbf{v}_i\}_{i=1}^{N}$ as $\mathbf{v}_h$ and $\mathbf{v}_t$. Meanwhile, we retrieve the embedding of relation r with trainable relation embeddings $\{\mathbf{r}_i\}_{i=1}^{|\mathcal{R}_{KG}|}$. Then a free-to-choose scoring function $\phi_r(\cdot, \cdot)$ is leveraged to discriminate the positive triplets from negative ones. In this work, we employ the scoring function in TransE for general consideration. So, the link prediction objective is formulated as:

$$\mathcal{L}_{\text{LinkPred}} = \sum_{(v_h, v_t, r) \in \Omega \cup \Omega^-} -y \log \sigma\left(\phi_r(\mathbf{v}_h, \mathbf{v}_t) + \gamma\right), \tag{12}$$

where σ is the sigmoid function, Ω is the positive triplets set observed in $\mathcal{G}_S$, Ω^- represents the sampled negative triplets set and $y \in \{1, -1\}$ indicates the label of the triplet (v_h, v_t, r). Thus, the model is optimized by the joint of the two objectives in Equation 11 and 12, i.e., $\mathcal{L} = \mathcal{L}_{\text{TextCls}} + \mathcal{L}_{\text{LinkPred}}$ in the first phase. By doing so, the text embeddings are enabled to make predictions with the extra structural information in the biomedical KG and the rich contextualized information can be leveraged to enhance the pathway reasoning ability of the learned KG representations, especially when there are missing links in the original KG.

4.4.2 Fine-tuning Objective. During the second phase, the model is optimized by the DDI prediction task. Depending on the dataset, the task can be either multi-class predictions or multi-label predictions. Given a DDI pair $(D_i, D_j, r) \in \mathcal{G}_{DDI}$, the corresponding drug representation $\mathbf{z}_i$ and $\mathbf{z}_j$ will be retrieved from fused drug embedding $\mathbf{Z}$ in Equation 9. Then, their concatenation is fed into an MLP network $f_{\text{DDI}}(\cdot)$ to predict the DDI probability score $\hat{\mathbf{y}}_{ij}$ of the two drugs. Finally, the cross-entropy loss between $\hat{\mathbf{y}}_{ij}$ and the ground truth $\mathbf{y}_{ij}$ will be computed as the training objective:

$$\begin{aligned} \hat{\mathbf{y}}_{ij} &= \text{SoftMax}\left(f_{\text{DDI}}\left(\mathbf{z}_i \oplus \mathbf{z}_j\right)\right), \\ \mathcal{L}_{\text{DDI}} &= -\sum_{(D_i, D_j, r) \in \mathcal{G}_{DDI}} \sum_{k=1}^{|\mathcal{R}_{DDI}|} \left[\mathbf{y}_{ij}\right]_k \log \left[\hat{\mathbf{y}}_{ij}\right]_k. \end{aligned} \tag{13}$$

5 EXPERIMENT

In this section, we first introduce the experimental settings of this work. Then, comprehensive experimental results on different DDI datasets and settings are provided to demonstrate the advantages of MOF-DDI over other baselines. Ablation studies and other analysis experiments are also included to justify the designs of MOF-DDI.

5.1 Experimental Settings

5.1.1 Datasets. (1) **Drugbank** [44]: Drugbank dataset is a huge online dataset widely used in industry and academia. It contains detailed drug information (e.g., generic name, chemical formula) and corresponding drug targets. **Hetionet** [13]: Hetionet dataset is a heterogeneous network of biomedical knowledge assembled from 29 different databases covering genes, drugs, diseases, and more. Hetionet contains 47,031 nodes of 11 types (e.g., gene, symptom) and 2,250,197 edges of 24 types (e.g., treat, cause). According to Drugbank Accession Number, we map drugs in Drugbank to compounds in Hetionet which outputs 1,414 approved drugs and 315,684 interaction pairs among these drugs. Next, we filter these interactions by removing the interaction types which occur less than 100 times and we filter these 1,414 drugs by removing the drugs without SMILES sequence. In the end, we construct 310,049 drug-drug interaction pairs consisting of 1,307 drugs and 89 types of pharmacological relations. As for the knowledge graph, we follow the same filtering strategy and construct a heterogeneous network consisting of 44,770 nodes out of 11 types with 2,248,814 edges from 24 relation types. (2) **TWOSIDES** [36]: TWOSIDES is a database describing drug-drug side effect information, which includes 1,318 side effects types (e.g., hypotension and nausea) across 63,473 drug combinations. In this work, we select 645 drugs and 46,221 drug–drug pairs following [50]. We also sample 200 medium-frequency edge types ranging from Top-600 to Top-800, ensuring every DDI type occurs at least 1,000 times in the TWOSIDES dataset. Please note that edges may have more than one DDI type. In our experiments, we follow the previous work [24, 35] to omit the inductive setting experiment on TWOSIDES since it is a smaller dataset with less rich drugs compared with DrugBank.

5.1.2 Baselines. To thoroughly evaluate the effectiveness of our propose MOF-DDI, we compare it with 11 baselines, including MLP, two network embedding models (DeepWalk [27] and LINE [34]), three graph neural network models (GraphSage [12] and GIN[46])

Table 1: Overall performance comparison between MOF-DDI and baselines for both transductive and inductive settings on the DrugBank dataset. Results are reported as mean±std%, the best performance is bolded and runner-ups are underlined. "-" means the model is not applicable to the setting.

Dataset	Drugbank (transductive)				Drugbank (inductive)			
Metric	Accuracy	Macro F1	Macro Presicion	Macro Recall	Accuracy	Macro F1	Macro Presicion	Macro Recall
MLP	70.14±0.37	57.32±0.58	65.15±0.30	53.81±0.42	58.92±0.60	40.06±0.70	54.74±0.81	35.62±0.55
DeepWalk	58.62±0.18	49.24±0.45	52.51±0.44	50.74±0.50	-	-	-	-
LINE	62.57±0.49	33.27±0.35	44.34±0.25	28.17±0.32	-	-	-	-
DeepDDI	82.29±0.18	76.58±0.18	79.75±0.20	76.12±0.18	68.43±0.48	60.87±0.70	63.62±0.58	58.90±0.65
GraphSage	81.21±0.35	76.25±0.34	81.83±0.28	72.90±0.29	71.39±0.41	65.62±0.37	68.24±0.48	65.16±0.52
GIN	83.62±0.40	80.13±0.32	83.28±0.30	78.15±0.40	73.41±0.53	68.20±0.62	70.85±0.39	67.42±0.68
KG-DDI	85.14±0.32	78.14±0.22	83.47±0.51	76.75±0.27	74.85±0.62	64.82±0.65	67.36±0.70	64.43±0.45
MIRACLE	86.54±0.51	79.90±0.36	84.22±0.20	77.87±0.42	78.12±0.37	70.43±0.26	74.40±0.16	68.36±0.20
KGNN	86.39±0.20	81.37±0.21	85.28±0.28	80.15±0.48	77.30±0.33	71.25±0.45	75.51±0.32	67.35±0.52
SSI-DDI	87.10±0.45	84.95±0.26	83.76±0.49	82.86±0.28	79.60±0.67	74.52±0.55	78.81±0.62	71.65±0.60
MUFFIN	88.31±0.30	86.69±0.27	85.83±0.18	85.31±0.40	80.68±0.45	75.57±0.52	79.90±0.38	73.88±0.49
SumGNN	88.85±0.21	85.20±0.44	86.62±0.31	84.32±0.16	81.54±0.44	76.21±0.29	79.13±0.40	74.54±0.34
MOF-DDI	**91.15±0.26**	**89.04±0.32**	**90.38±0.42**	**88.42±0.14**	**84.69±0.51**	**80.86±0.46**	**84.45±0.48**	**78.28±0.67**

and six DDI-specified models (DeepDDI [32], KG-DDI [16], MIRACLE [41], SSI-DDI [25], MUFFIN [5] and SumGNN [50]). For the general baselines, we adopt the implementations from DGL [40] and modify them to our datasets. For all the other DDI-specified baselines, we simply adopt the code released by original authors.

5.1.3 Evaluation Strategy. We spilt the DDI dataset into 7:1:2 as training, validation and test set. Because of the high class-imbalance property of the dataset, we use the stratified split method to ensure samples from every class would be equally separated in training, validation, and test set. For every experiment, we run 3 times on 3 differently split datasets and report the average and standard deviation. For the DrugBank dataset which is a multi-class classification problem, we employ Accuracy, Macro-Precision, Macro-Recall and Macro-F1 scores as the metrics for evaluation. For the TWOSIDES dataset which is a multi-label prediction problem, we consider ROC-AUC, PR-AUC and Average Precision as the metrics.

5.1.4 Implementation Details. During the optimization of text-augmented KG embeddings, we initialized the KG embeddings with TransE. Meanwhile, we follow the implementation of GreaseLM [53] to employ the PubmedBERT [10] to initialize the text token embeddings. To prevent label leakage during this phase, we only retrieve documents related to DDI pairs that appear in the training set (including positive and negative DDI pairs) and exclude any other DDI pairs from the retrieval document corpus. When a text summary contains multiple DDI pairs, we mask drug name tokens that may cause information leakage. Additionally, it is worth noting that many drug entities are not referred to by their generic name in the text corpus. To address this issue, we leveraged the drug brand names and synonyms provided in DrugBank to facilitate the mapping process and increase the mapping ratio. For the implementation of the molecular GNN, we adopt the GIN [46] encoder with mean readout function to learn the molecular representations. We adopt the Adam optimizer for the optimization. Besides, we find the best value of some hyper-parameters through grid search. Specifically, the search spaces of learning rate l, embedding dimension d, batch size B the weighted aggregation parameter λ are $\{0.005, 0.001, 0.0005\}$, $\{128, 256, 512, 1024\}$, $\{256, 512, 1024, 2048\}$ and $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, respectively.

5.2 Overall Performance Comparison

In Table 1, we report the performance of our proposed MOF-DDI and the eleven baselines on the DrugBank dataset, including the transductive setting and inductive setting. From the table, we can reach three observations: (1) The models leveraging external knowledge (i.e., biomedical KGs and drug molecular structures) in the prediction task can significantly outperform the baselines that solely rely on the DDI graph (DeepWalk and LINE). Without external knowledge, the trained models can not explain the mechanism behind any DDI pair and only make predictions based on the proximity of two drugs in the DDI graph. Besides, these models can only make predictions for those drug pairs that both of the two drugs are existing in the input DDI graph, which limits their application to the drug pairs including new drugs (i.e., inductive setting); (2) The large biomedical KG can boost the DDI prediction performance by a large margin. Apart from MOF-DDI, there are also three other baselines (i.e., KG-DDI, KGNN and SumGNN) incorporating the biomedical KGs and all of them achieve relatively high performance than other baselines. Though the drug molecular structures reflect the mechanisms from the view of fundamental chemistry, it is more challenging to train a model to capture all the intrinsic features of all the drugs because the drug molecules are in the 3-D dimension and represent them with 2-D graphs will unavoidably lose some information [22]. However, the advantages of methods without considering drug molecular structures are less obvious under the inductive setting due to the incomplete edges for the under-explored drugs in KGs. (3) Our proposed MOF-DDI achieves state-of-the-art performances in different settings compared with other baselines. Particularly, the performance gain of MOF-DDI over the state-of-the-art baseline is 2.30%, 2.25%, 3.76% and 3.13% under the transductive setting. The improvement is much more obvious under the inductive setting, where our method outperforms

Table 2: Overall performance comparison between MOF-DDI and baselines on the TWOSIDES dataset. Results are reported as mean±std%, the best performance is bolded and runner-ups are underlined.

Dataset	TWOSIDES		
Metric	Accuracy	ROC-AUC	Average Precision
MLP	70.28±0.43	81.44±0.86	80.65±0.82
DeepWalk	80.08±0.56	87.67±0.49	84.09±0.36
LINE	80.50±0.79	88.20±0.54	87.25±0.48
DeepDDI	77.45±0.31	86.12±0.70	84.96±0.62
GraphSage	77.89±0.50	87.12±0.48	85.65±0.50
GIN	78.94±0.90	88.14±0.25	87.15±0.66
KG-DDI	80.47±0.39	89.75±0.97	87.28±1.05
MIRACLE	82.45±0.70	90.86±0.63	88.35±0.86
KGNN	80.54±0.61	90.26±0.81	87.74±0.79
SSI-DDI	82.12±0.54	90.80±0.55	88.20±0.79
MUFFIN	82.87±0.69	90.65±0.48	88.26±0.60
SumGNN	83.98±0.82	91.42±0.58	90.16±0.65
MOF-DDI	**85.28±0.66**	**92.50±0.72**	**91.59±0.68**

the state-of-the-art baseline by 3.15%, 4.65%, 5.32% and 3.74% on the four metrics, respectively. These results not only demonstrate the effectiveness of MOF-DDI but also support our claim mentioned in Section 1 that the rich contextualized information in biomedical literature context is a helpful supplementary to the incomplete and noisy information in KGs and will enhance the model generalization ability on the unobserved drugs. Meanwhile, we demonstrate the experimental results on the multi-label dataset TWOSIDES in Table 2, from which we can find very simialr observations as we analyzed above. Our proposed MOF-DDI achieves 1.30%, 1.08% and 1.43% improvements over the runner-up methods in the three metrics. We noticed that the performance gain on the TWOSIDES dataset is less obvious than that on the DrugBank dataset. One possible explanation is that the TWOSIDES is less imbalanced than the DrugBank dataset and our method can better handle the imbalanced scenario.

5.3 Ablation Study

In our proposed method, we incorporate various kinds of features and add the optimal transport map to align them into the same embedding spaces. To verify the effectiveness of each component in MOF-DDI, we design three model variants and compare their performance with MOF-DDI. The details of these model variants are illustrated below and the comparison results on DrugBank and TWOSIDES are shown in Table 3 and Table 4.

- **w/o Text.** The raw text features are discarded and the output of the vanilla KG model is taken as $\mathbf{Z}_{KG}$ for the next step.
- **w/o InfoEx.** The InfoEx component is discarded. KG entity embeddings are concatenated with the corresponding pre-trained LM token embedding for prediction.
- **w/o OT.** The optimal transport map is skipped. $\mathbf{Z}_{KG}$ and $\mathbf{Z}_{MOL}$ are combined with weighted aggregation.

From the performance of the model variant w/o Text, one can clearly see that the deficiency of text features will significantly decrease overall performance and generalization ability. The phenomenon indicates the contextualized information in texts can serve as a supplement to the incomplete information in KGs. Meanwhile, the model variant w/o InfoEx can only achieve comparable or even worse performance than w/o Text, the experimental results can also serve as evidence that the deficiency of a proper information fusion module may fail fully taking advantage of the information in extra features source and may deteriorate the learned embeddings. Last but not least, the model variant w/o OT is outperformed by w/o Text under the transductive setting but achieves the best performance in the inductive setting among the three variants. On one hand, the result further proves the importance of multi-source data in enhancing the model generalization ability. On the other hand, the improvement of overall performances induced by the optimal transport reveals space heterogeneity existing among the features from different sources.

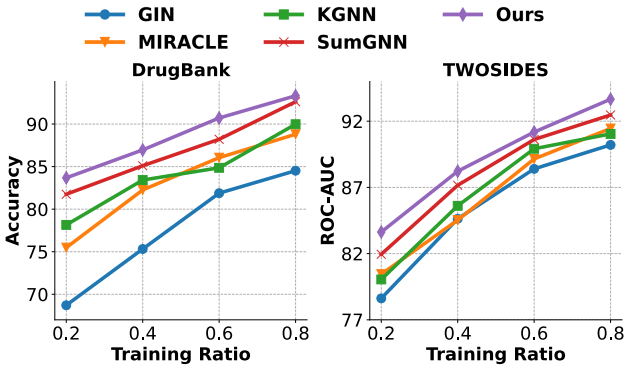


Figure 3: Performance with different training ratios.

5.4 Effects with Different Training Ratios

To further demonstrate the advantage of our method over other baselines in handling incomplete knowledge. We conduct experiments to test the effectiveness of our method and the other four strong baselines on the two datasets with different training ratios (i.e., 20%, 40%, 60% and 80%). From the experimental results in Figure 3, one can clearly find that the performances of all five methods decrease monotonously with lower training ratios, however, their performance deterioration amounts vary from each other. Taking the experimental results on DrugBank as an example, the prediction accuracy descends from 93.32% to 83.68, causing 9.64% reduction. On the contrary, the performance decreases of the four baselines from 80% training ratio to 20% training ratio are 10.87%, 11.83%, 13.32% and 15.91%, respectively. Also, we can find similar observations from the experimental results on the TWOSIDES dataset. Hence, it is safe to claim our method achieves the best performance and gains the highest robustness compared with other baselines.

5.5 Analysis on Feature Source

To thoroughly analyze the advantage of multi-source data on the DDI prediction performance, we design three more model variants (i.e., **w/o Mol**, **w/o KG**, and **w/o Text+KG**) along with **w/o Text** above to demonstrate the effectiveness of different features sources. The experimental results of the four variants are shown in Figure 4 to compare with our proposed MOF-DDI. It is easy to find that MOF-DDI can consistently beat the four variants on the two datasets, which indicates the DDI prediction performance

Table 3: Performance comparison between MOF-DDI and its variants for both transductive and inductive settings on the DrugBank dataset. Results are reported as mean±std%, the best performance is bolded.

Dataset	Drugbank (transductive)				Drugbank (inductive)			
Metric	Accuracy	Macro F1	Macro Presicion	Macro Recall	Accuracy	Macro F1	Macro Presicion	Macro Recall
w/o Text	90.06±0.30	86.31±0.35	88.75±0.30	85.20±0.34	81.76±0.49	76.35±0.60	79.20±0.72	75.32±0.50
w/o InfoEx	89.85±0.30	86.83±0.52	87.06±0.47	85.26±0.50	81.90±0.63	76.70±0.39	79.51±0.42	74.81±0.73
w/o OT	89.46±0.21	84.94±0.40	86.33±0.40	83.72±0.16	82.54±0.44	78.52±0.48	80.31±0.30	77.54±0.34
MOF-DDI	**91.15±0.26**	**89.04±0.32**	**90.38±0.42**	**88.42±0.14**	**84.69±0.51**	**80.86±0.46**	**84.45±0.48**	**78.28±0.67**

Table 4: Performance comparison between MOF-DDI and its variants on the TWOSIDES dataset. Results are reported as mean±std%, the best performance is bolded.

Dataset	TWOSIDES		
Metric	ACC	ROC-AUC	Average Precision
w/o Text	85.06±0.79	91.92±0.66	90.77±0.80
w/o InfoEx	84.16±0.72	91.47±0.65	89.88±0.81
w/o OT	84.08±0.82	91.20±0.75	89.27±0.59
MOF-DDI	**85.28±0.66**	**92.50±0.72**	**91.59±0.68**

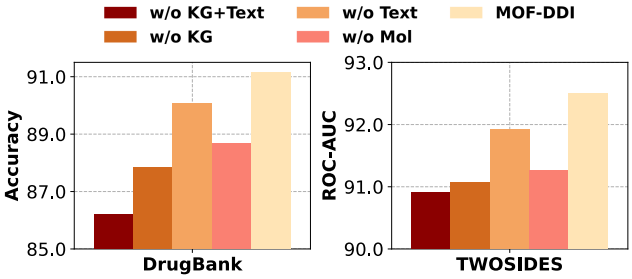


Figure 4: The impacts of different feature sources.

will benefit from every feature source involved in our framework. Among the four model variants, the deficiency of both the biomedical KG and the text feature (w/o KG+Text) leads to the largest performance deterioration (4.94% and 1.58%) compared with MOF-DDI. When considering the impact of any single feature source, one can see that biomedical KG tends to play the most important role in DDI prediction. Though less performance deterioration was gained without Text features, it still boosts the overall performance by an obvious margin. Meanwhile, the input of w/o Text is the same as some baselines, (i.e., MUFFIN and sumGNN), however, the model variant can achieve even better performance than MUFFIN and sumGNN, which further prove the effectiveness of our optimal transport component.

5.6 Analysis on Optimal Transport

To demonstrate the impacts of the optimal transport map on the prediction performance, we conduct experiments to evaluate the effectiveness of six fusion strategies. The model variants **OT-Mean**, OT-Concat and **OT-Weighted** fuse $\widetilde{\mathbf{Z}}_{\mathrm{KG}}$ and $\mathbf{Z}_{\mathrm{MOL}}$ with mean aggregation, concatenation and weighted aggregation, where the weighted aggregation is the fusion operation employed in our methods. The other three variants, **Mean**, **Concat** and **Weigted**, are their counterparts without the optimal transport map that transform $\mathbf{Z}_{\mathrm{KG}}$ to $\widetilde{\mathbf{Z}}_{\mathrm{KG}}$. We show the experimental results of the six variants on

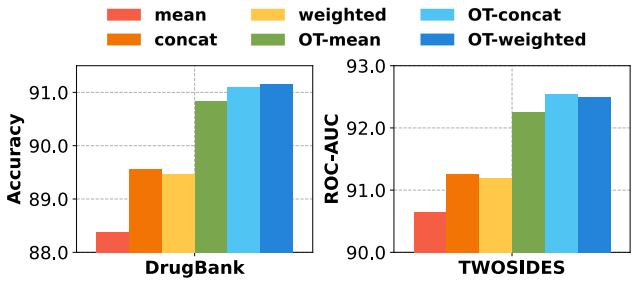


Figure 5: The impacts of different fusion strategies.

the two datasets in Figure 5. From the result, we can observe that all of the variants with optimal transport map consistently outperform their counterparts. Though the concatenation and weighted aggregation can somewhat alleviate the issue induced by the heterogeneity of feature spaces, they still suffer from relatively large performance deterioration. Meanwhile, the performance variance brought by different aggregation operations is less obvious than their counterparts. One possible explanation behind this phenomenon is that the features with similar intrinsic distribution are easy to combine without extra effort, it can also be taken as another advantage of the incorporating optimal transport map.

6 CONCLUSION

In this paper, we study the problem of predicting potential drug-drug interactions, which is of great importance in the research of pharmacology and medical health. Inspired by the promising capabilities of multi-source data on enhancing model generalization ability and further improving the overall performances, we propose **M**ulti-modality Features **O**ptimal **F**usion for **D**rug-**D**rug **P**rediction, namely **MOF-DDI**, a DDI prediction framework that can integrate the information from the raw biomedical literature corpus, large biomedical KG and drug molecular structure at the same time. We conduct extensive experiments on well-known DDI datasets from both transductive setting and inductive setting to compare our method with multiple state-of-the-art baselines. Experimental results show that MOF-DDI can consistently improve the DDI prediction performances on all the datasets and settings.

ACKNOWLEDGMENTS

This work is partially supported by the NSF under grants IIS-2334193, IIS-2321504, IIS-2209814, IIS-2203262, IIS-2214376, IIS-2217239, OAC-2218762, CNS-2203261, and CMMI-2146076. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

REFERENCES

[1] Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. 2019. Randomized SMILES strings improve the quality of molecular generative models. *Journal of cheminformatics* 11, 1 (2019), 1–13.

[2] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2021. Using drug descriptions and molecular structures for drug–drug interaction extraction from literature. *Bioinformatics* 37, 12 (2021), 1739–1746.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*.

[4] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport. In *NeurIPS*.

[5] Yujie Chen, Tengfei Ma, Xixi Yang, Jianmin Wang, Bosheng Song, and Xiangxiang Zeng. 2021. MUFFIN: multi-scale feature fusion for drug–drug interaction prediction. *Bioinformatics* 37, 17 (2021), 2651–2658.

[6] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*.

[7] Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. 2020. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* 36, 15 (2020), 4316–4322.

[8] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *WWW*.

[9] Assaf Gottlieb, Gideon Y Stein, Yoram Oron, Eytan Ruppin, and Roded Sharan. 2012. INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular systems biology* 8, 1 (2012), 592.

[10] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[11] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *ICML*.

[12] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*.

[13] Daniel S Himmelstein and Sergio E Baranzini. 2015. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS computational biology* 11, 7 (2015), e1004259.

[14] Biaobin Jiang, Kyle Kloster, David F Gleich, and Michael Gribskov. 2017. AptRank: an adaptive PageRank model for protein function prediction on bi-relational graphs. *Bioinformatics* 33, 12 (2017), 1829–1836.

[15] Xin Jin, Xia Sun, Jiacheng Chen, and Richard Sutcliffe. 2022. Extracting Drug-drug Interactions from Biomedical Texts using Knowledge Graph Embeddings and Multi-focal Loss. In *CIKM*.

[16] Md Rezaul Karim, Michael Cochez, Joao Bosco Jares, Mamtaz Uddin, Oya Beyan, and Stefan Decker. 2019. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. In *BCB*.

[17] Eunyoung Kim and Hojung Nam. 2022. DeSIDE-DDI: interpretable prediction of drug-drug interactions using drug-induced gene expressions. *Journal of cheminformatics* 14, 1 (2022), 1–12.

[18] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

[20] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *EMNLP*.

[21] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. 2020. KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction.. In *IJCAI*.

[22] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training Molecular Graph Representation with 3D Geometry. In *ICLR*.

[23] Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. 2022. Is Homophily a Necessity for Graph Neural Networks?. In *ICLR*.

[24] Arnold K Nyamabo, Hui Yu, Zun Liu, and Jian-Yu Shi. 2022. Drug–drug interaction prediction with learnable size-adaptive molecular substructures. *Briefings in Bioinformatics* (2022).

[25] Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. 2021. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics* (2021).

[26] Bethany Percha and Russ B Altman. 2013. Informatics confronts drug–drug interactions. *Trends in pharmacological sciences* 34, 3 (2013), 178–184.

[27] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*.

[28] Chaim G Pick. 1997. Antinociceptive interaction between alprazolam and opioids. *Brain research bulletin* 42, 3 (1997), 239–243.

[29] Catrin O Plumpton, Daniel Roberts, Munir Pirmohamed, and Dyfrig A Hughes. 2016. A systematic review of economic evaluations of pharmacogenetic testing for prevention of adverse drug reactions. *Pharmacoeconomics* 34 (2016), 771–793.

[30] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 5 (2010), 742–754.

[31] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences* 115, 18 (2018), E4304–E4311.

[32] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences* 115, 18 (2018), E4304–E4311.

[33] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* (2008).

[34] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *WWW*.

[35] Zhenchao Tang, Guanxing Chen, Hualin Yang, Weihe Zhong, and Calvin Yu-Chian Chen. 2023. DSIL-DDI: A Domain-Invariant Substructure Interaction Learning for Generalizable Drug–Drug Interaction Prediction. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[36] Nicholas P. Tatonetti, Patrick P. Ye, Roxana Daneshjou, and Russ B. Altman. 2012. Data-Driven Prediction of Drug Effects and Interactions. *Science Translational Medicine* 4, 125 (2012), 125ra31–125ra31.

[37] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.

[38] Cédric Villani et al. 2009. *Optimal transport: old and new*. Vol. 338. Springer.

[39] Kuan Wang, Yuyu Zhang, Diyi Yang, Le Song, and Tao Qin. 2020. GNN is a Counter? Revisiting GNN for Question Answering. In *ICLR*.

[40] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* (2019).

[41] Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. 2021. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *WWW*.

[42] Qianlong Wen, Zhongyu Ouyang, Chunhui Zhang, Yiyue Qian, Yanfang Ye, and Chuxu Zhang. 2022. Graph contrastive learning with cross-view reconstruction. In *NeurIPS 2022 Workshop: New Frontiers in Graph Learning*.

[43] Qianlong Wen, Zhongyu Ouyang, Jianfei Zhang, Yiyue Qian, Yanfang Ye, and Chuxu Zhang. 2022. Disentangled dynamic heterogeneous graph learning for opioid overdose prediction. In *SIGKDD*.

[44] David Scott Wishart, Yannick Djoumbou Feunang, Anchi Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig K. Knox, and Michael Wilson. 2017. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* 46 (2017), D1074 – D1082.

[45] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* (2018).

[46] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

[47] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022. Deep Bidirectional Language-Knowledge Graph Pretraining. In *NeurIPS*.

[48] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph Contrastive Learning with Augmentations. In *NeurIPS*.

[49] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *Comput. Surveys* 54, 11s (2022), 1–38.

[50] Yue Yu, Kexin Huang, Chao Zhang, Lucas M Glass, Jimeng Sun, and Cao Xiao. 2021. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics* 37, 18 (2021), 2988–2995.

[51] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *SIGKDD*.

[52] Jianfei Zhang, Ai-Te Kuo, Jianan Zhao, Qianlong Wen, Erin Winstanley, Chuxu Zhang, and Yanfang Ye. 2021. Rxnet: Rx-refill graph neural network for overprescribing detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2537–2546.

[53] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2021. GreaseLM: Graph REASoning Enhanced Language Models. In *ICLR*.

[54] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL*.