OPENDDI: A COMPREHENSIVE BENCHMARK FOR DDI PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Drug–Drug Interactions (DDIs) significantly influence therapeutic efficacy and patient safety. As experimental discovery is resource-intensive and time-consuming, efficient computational methodologies have become essential. The predominant paradigm formulates DDI prediction as a drug graph-based link prediction task. However, further progress is hindered by two fundamental challenges: (1) lack of high-quality data: most studies rely on small-scale DDI datasets and single-modal drug representations; (2) lack of standardized evaluation: inconsistent scenarios, varied metrics, and diverse baselines. To address the above issues, we propose OpenDDI, a comprehensive benchmark for DDI prediction. Specifically, (1) from the data perspective, OpenDDI unifies 6 widely used DDI datasets and 2 existing forms of drug representation, while additionally contributing 3 new large-scale LLM-augmented datasets and a new multimodal drug representation covering 5 modalities. (2) From the evaluation perspective, OpenDDI unifies 20 SOTA model baselines across 3 downstream tasks, with standardized protocols for data quality, effectiveness, generalization, robustness, and efficiency. Based on OpenDDI, we conduct a comprehensive evaluation and derive 11 valuable insights for DDI prediction while exposing current limitations to provide critical guidance for this rapidly evolving field. Our code is available at `https://anonymous.4open.science/r/anonym1234`.

1 INTRODUCTION

With the expansion of pharmacotherapy and the prevalence of polypharmacy, alterations in drug effects due to co-administration have emerged as a common phenomenon in medication. We formally refer to such phenomena as Drug–Drug Interactions (DDIs) Zhong et al. (2024), which may produce effects like enhanced therapeutic efficacy or aggravated disease progression Takeda et al. (2017); Huang et al. (2017); Chen et al. (2016). Therefore, identifying and understanding DDIs is of paramount importance Han et al. (2017); Sun et al. (2019), which is critical to patient safety Gulikers et al. (2024), polypharmacy strategies Zitnik et al. (2018a), and drug development Qiu et al. (2021a).

However, traditional laboratory-based approaches for validating DDIs often require significant time and extensive resources Whitebread et al. (2005); Safdari et al. (2016). To address aforementioned issues, machine learning Qiu et al. (2021b), particularly graph-based approaches, offers a promising alternative by modeling drugs as nodes and interactions as edges, with their pharmacological attributes used as node features, which casts DDI prediction as a link prediction task effectively addressed by graph neural networks (GNNs) Sun et al. (2024); Defferrard et al. (2016). These methods have achieved strong performance while reducing experimental costs Li et al. (2023b); Ma & Lei (2023).

Despite these advances, studies still remain constrained by two challenges: (1) **Limitations in data quality**: Due to policy constraints and limited accessibility, DDI metadata remain difficult to collect, leading current methods to depend on small-scale datasets that contain only few drug nodes and sparse DDI edges. Besides, most methods rely on single-modal drug representations such as SMILES or path, which capture structural or relational information while overlooking rich biological and semantic information. (2) **Challenges in evaluation standardization**: Current studies face diverse real-world objectives (e.g., binary, multiclass, multilabel tasks) and research contexts. Therefore, they often adopt inconsistent experimental scenarios, varied evaluation metrics, and different baseline selections, making fair and standardized evaluation difficult. Existing benchmarks, such as DDI-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: DDI prediction benchmark comparison.

Benchmarks	Datasets	Representations	Algorithms	Tasks	Evaluations
DDI-Ben Shen et al. (2025)	2 small-scale	Single-modal	10 +	2	Generalization, Robustness
OpenDDI (Ours)	9 (3 million-level)	Multimodal	20 +	3	Quality, Effectiveness,Efficiency Generalization, Robustness

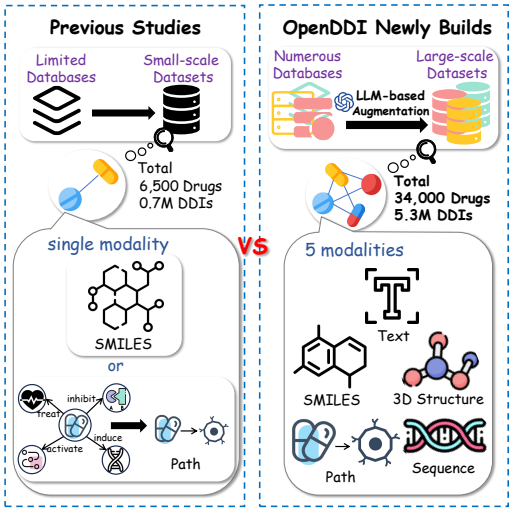

Figure 1: Comparison of previous and our newly-built datasets.

Ben Shen et al. (2025), evaluate generalization but overlook fundamental issues in data (e.g., relying on 2 small datasets). Moreover, they include few algorithms and limited evaluation scenarios.

To address these limitations, we propose OpenDDI, a comprehensive benchmark for DDI prediction. A systematic comparison with DDI-Ben is provided in Table 1, which highlights advantages of OpenDDI in data coverage, algorithm breadth, task variety, and evaluation scenarios. Following the comparison, we describe OpenDDI from two perspectives. (1) From the **data** perspective, OpenDDI integrates representative datasets and diverse drug representations. Specifically, OpenDDI unifies 6 widely used datasets and 2 existing forms of drug representation. Notably, OpenDDI additionally contributes 3 new large-scale datasets. New datasets are constructed through systematic curation and LLM-based augmentation. They contain over 5.3 million DDIs across 34000 drugs, significantly expanding the scale and diversity of available resources. In addition to new datasets, OpenDDI newly builds multimodal drug representation. The representation is constructed by integrating heterogeneous information about drugs, categorized into 5 modalities: SMILES, path, 3D conformations, amino acid sequences, and textual descriptions. Such integration enables richer understanding of drug properties and interactions. Together, these datasets and representations provide diverse data construction options for DDI prediction. They offer several advantages as demonstrated in Figure 1.

(2) From the **evaluation** perspective, OpenDDI incorporates 20 SOTA baselines and 3 representative link-level downstream tasks. It further establishes standardized evaluation protocols and metrics for fair assessment across scenarios involving predictive performance, model robustness, and computational efficiency. Based on OpenDDI, we provide a comprehensive evaluation of existing DDI prediction methods and datasets, drawing some valuable insights. For **data quality**, we evaluate existing datasets from two perspectives: the scale and coverage of datasets, as well as the performance of various models trained on them. Through comparative analyses, we demonstrate the superiority of our curated datasets and further provide an in-depth investigation of the multimodal drug representations. For **effectiveness**, we compare the performance of models across different datasets, comprehensively assessing the predictive capability of each model. For **generalization**, we simulate realistic scenarios involving unseen drugs and evaluate the performance of models on them, thereby assessing their ability to predict beyond the training distribution. For **efficiency**, we analyze model scalability with respect to both time and memory consumption. By examining empirical complexities, we highlight the computational efficiency and scalability of the models. For **robustness**, we focus on challenges arising from data noise and sparsity, assessing how well existing approaches maintain performance when exposed to different types and degrees of data perturbation.

Contributions. (1) *Comprehensive Benchmark*: OpenDDI unifies 6 existing datasets and 2 existing forms of drug representation, while additionally contributing 3 new LLM-augmented datasets and a new multimodal drug representation covering 5 biomedical modalities, broadening the current scope of DDI studies. Furthermore, OpenDDI integrates 20 SOTA algorithms and 3 downstream tasks for evaluations. (2) *Analytical Insights*: Through extensive experimentation, we identified 11 pivotal insights from five analytical perspectives, which collectively deepen our understanding of DDI prediction and inspire methodological advancements. (3) *Open-Source Library and Detailed Repository*: OpenDDI is designed as an open-access data and code platform, allowing users to easily evaluate their algorithms, while promoting innovation using the curated resources.

2 Definitions and Background

We formalize DDI prediction problem as follows.

DDI Graph Formulation. We model the DDI prediction task using an undirected attributed multi-relational graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$. The node set $\mathcal{V} = \{v_1, \ldots, v_n\}$ comprises n distinct drugs. The edge set $\mathcal{E}$ characterizes interactions, indicating that drugs v_i and v_j exhibit interaction label(s) $y_{ij} \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the label set. Furthermore, $\mathbf{X} \in \mathbb{R}^{n \times d}$ serves as the node feature matrix, with the i-th row $\mathbf{x}_i \in \mathbb{R}^d$ encoding the feature representation of drug v_i.

Drug Representation. Each drug $v \in \mathcal{V}$ is associated with a feature representation. Former studies typically rely on a single modality, such as (i) chemical structure(e.g., SMILES) or (ii) path embeddings derived from drug–entity relations. In contrast, OpenDDI introduces a unified *multimodal* representation by integrating heterogeneous evidence, including SMILES, path, 3D conformations, protein sequences, and textual descriptions. Formally, each drug is mapped to a feature tuple

$$\Phi(v) = \left(\mathbf{x}_v^{(1)}, \ldots, \mathbf{x}_v^{(M)}\right),$$

where M is the number of modalities and $\mathbf{x}_v^{(m)} \in \mathbb{R}^{d_m}$ denotes the m-th modality-specific embedding. For downstream prediction, we further construct a fused feature vector

$$\mathbf{x}_v = g\Big(\mathbf{x}_v^{(1)}, \ldots, \mathbf{x}_v^{(M)}\Big) \in \mathbb{R}^d,$$

where $g(\cdot)$ is a modality-aggregation function. The resulting fused feature vectors $\{\mathbf{x}_v\}_{v \in \mathcal{V}}$ constitute the node feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ defined in the graph formulation.

Prediction Tasks. The goal is to learn a predictor $f(v_i, v_j; \mathcal{G}) \mapsto \hat{y}_{ij} \in \mathcal{Y}$ using graph structure and node features. We consider three standard instantiations. *Binary interaction prediction*: $\mathcal{Y} = \{0, 1\}$. y_{ij} is a binary variable where $y_{ij} = 1$ denotes the presence of a DDI between drug v_i and drug v_j, and $y_{ij} = 0$ denotes its absence. *Multiclass interaction classification*: $\mathcal{Y} = \{1, \ldots, C\}$, where $y_{ij} = c$ specifies that the pair interacts through the c-th predefined interaction category. *Multilabel interaction classification*: $\mathcal{Y} = \{0, 1\}^K$, where $y_{ij} = (y_{ij}^{(1)}, \ldots, y_{ij}^{(K)})$ is a multi-hot vector. For each mechanism index k, $y_{ij}^{(k)} = 1$ indicates the presence of the k-th interaction mechanism, while $y_{ij}^{(k)} = 0$ indicates that this mechanism does not occur.

3 Benchmark Design

3.1 Algorithm Framework

DDI prediction Taxonomy. Existing DDI prediction methods can be systematically categorized into three paradigms according to their core modeling strategies: similarity-based, graph-based, and integration-based approaches. This comprehensive coverage reflects the major methodological directions in the field, as illustrated in Figure 2.

DDI prediction Algorithms. For DDI prediction algorithms, our framework incorporates 20 algorithms that are implemented according to their original repositories or interpreted from the corresponding papers. A full description of these algorithms is provided in Appendix A. In addition, we design a unified operational interface that integrates all models into a single API, enabling seamless invocation, systematic evaluation, and future extensibility. Through the efforts within OpenDDI, we offer a powerful platform that supports rigorous and trustworthy benchmarking for researchers.

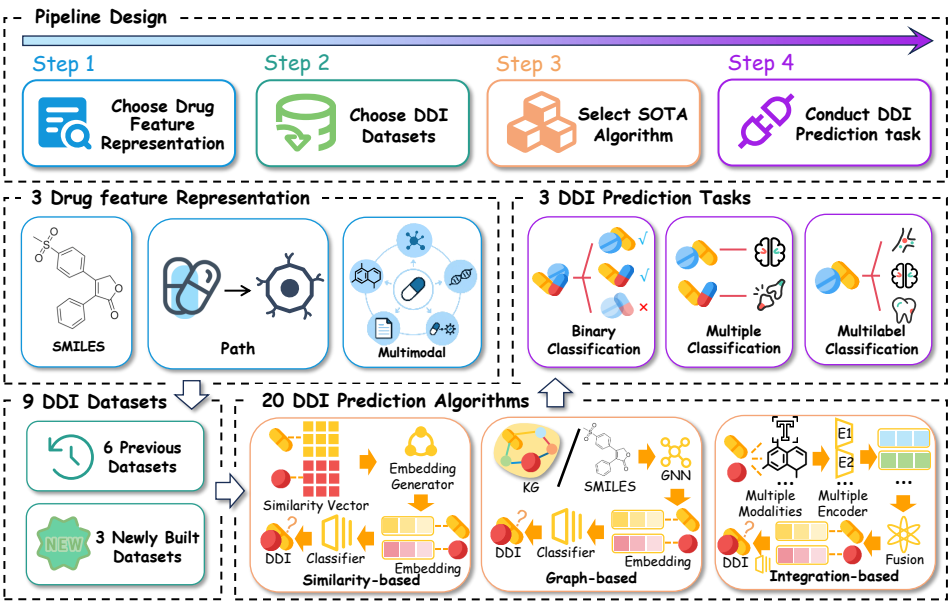

Figure 2: An overview of the OpenDDI Framework.

3.2 Dataset Curation

Data Components. To comprehensively evaluate DDI prediction methods, we compile 9 datasets along with 3 types of drug feature representations. Specifically, the dataset includes 6 previously used and 3 newly curated large-scale datasets. For drug representations, we include 3 categories, namely SMILES, path, and a newly developed multimodal representation that spans five distinct types of information. Detailed descriptions are provided in Appendix B.

Datasets newly built in OpenDDI. In this section, we introduce the newly built DDI datasets and representation. We curate drug data from 12 authoritative databases (e.g., DrugBank Wishart et al. (2017), ChEMBL Mendez et al. (2018), KEGG Kanehisa et al. (2016)), forming the large DDI datasets and the rich modality coverage for drug representation. Specifically, (1) **For DDI datasets**, OpenDDI substantially enlarges dataset scale and coverage, containing 3 task-specific datasets with more than 5.3 million interactions across 34,000 drugs. Raw interactions and metadata are systematically integrated, and then annotated into binary, multiclass, and multilabel formats to support diverse tasks. LLMs are employed to uncertain labels, ensuring their completeness. Compared with existing datasets such as ChCh-Miner Wang et al. (2024a) and Ryus Ryu et al. (2018), OpenDDI achieves an order-of-magnitude improvement, establishing a richer foundation for research. Besides, (2) **For drug feature representation**, OpenDDI achieves a comprehensive representation of drugs by integrating five modalities Gan et al. (2023); Wang et al. (2025), each providing a unique perspective on drug properties. Text contributes the semantic pharmacological knowledge, SMILES captures the structural similarity as a primary driver of DDIs, protein sequences and path encode the biochemical information, 3D molecular structures reflect the spatial conformations critical for binding affinity. Collectively, these modalities complement each other across semantic, structural, biological, and relational dimensions, thereby enabling models to capture the underlying logic of drug interactions from micro-level structures to macro-level mechanisms.

3.3 Evaluation Procedure

Pipeline Design. In previous studies, differences in datasets, drug representations, and downstream tasks posed significant challenges to achieve fair comparison. To bridge the gap, we systematically modify and extend existing methods, achieving standardized data input through unified code interfaces and ensuring comprehensive coverage of tasks. This pipeline enables thorough evaluation of methods.

Data Quality. For the datasets and drug representations included in OpenDDI, we assess their quality from two perspectives: the scale and coverage of each dataset; the model performance across different datasets and drug representations. Results highlight the advantages of our curated data. In addition, we conduct a ablation study to further analyze the contribution of multimodal drug representations.

Effectiveness. To evaluate the effectiveness of methods, we conduct multi-level comparisons across three downstream tasks—binary, multiclass and multilabel classification. We use three widely adopted evaluation metrics (ACC, F1, and AUC-ROC) to measure the model's predictive ability.

Generalization. To evaluate the generalization of methods, we consider realistic scenarios where trained models are evaluated on DDI test sets containing unseen drugs, in order to assess their ability to predict DDIs involving novel drugs.

Efficiency. In evaluating the efficiency of methods, we focus on both time and space complexity. Time complexity is assessed through empirical analyses of computational capability, while space complexity emphasizes memory efficiency. These dimensions provide a view of model scalability and enable us to determine whether a method is suitable in resource-constrained scenarios.

Robustness. We examine the robustness in two scenarios: data sparsity and noise. Noisy data are employed to evaluate the capacity of a model to remain stable under perturbed data, whereas sparse data serve to investigate how models respond to rare DDI events. Both challenges represent critical obstacles that must be addressed to ensure reliable DDI prediction in practical applications.

4 Experiments and Analyses

In this section, we present a suite of experiments that evaluate the quality of data and the effectiveness, generalization, efficiency, robustness of methods. Guided by a set of targeted questions, we aim to: (1) reveal differences in how existing datasets support model training and evaluation; (2) investigate the multimodal representation proposed in OpenDDI in depth; (3) examine methods in terms of their performance across diverse tasks and ability to cope with real-world difficulties.

For **data quality**, **Q1**: What are the scale and coverage of existing datasets? **Q2**: Can these datasets effectively support model training and evaluation? **Q3**: What insights can be further uncovered from the multimodal drug representation? For **effectiveness**, **Q4**: How do existing DDI prediction methods perform across different DDI prediction tasks? For **generalization**, **Q5**: How do existing DDI prediction methods perform when encountering unseen drugs? For **efficiency**, **Q6**: How do these methods perform regarding space and time consumption in practical scenarios? For **robustness**, **Q7**: How do DDI prediction methods handle real-world challenges such as data noise and data sparsity?

Table 2: Comparative statistics of existing DDI datasets

Dataset	Task Type	Drugs	DDIs	Label Type
OpenDDI	Binary	14,090	2,547,620	Existence of DDI
	Multiclass	14,086	2,534,350	167 DDI types
	Multilabel	6,704	211,951	200 DDI types
ZhangDDI	Binary	548	48,548	Existence of DDI
ChCh-Miner	Binary	1,514	48,514	Existence of DDI
Ryus	Multiclass	1,700	191,570	86 DDI types
Dengs	Multiclass	570	37,264	65 DDI types
ZS-DDIE	Multiclass	2,004	394,118	174 DDI types
Twosides	Multilabel	350	19,535	100 DDI types

4.1 Datasets Comparison

To address **Q1**, we conducted detailed curation and statistical analysis of existing DDI datasets.

From Table 2, we derive several key observations. (1) The DDI data compiled in previous datasets involves no more than 2,100 drugs, whereas OpenDDI encompasses DDI data covering 14,090 drugs, substantially exceeding previous datasets in terms of drug coverage. (2) Previous datasets typically contain no more than 0.4 million interactions, with most comprising around 80,000, whereas OpenDDI organizes over 2.5 million interactions, representing a substantial increase in scale. (3) Previous multiclass datasets generally include fewer than 100 labels, although ZS-DDIE includes a larger variety of DDI types, it contains a substantial number of sparse types (with frequencies

below 10). In contrast, OpenDDI provides 167 well-represented DDI labels with far fewer sparse types, ensuring that each label is supported by more than 50 records, thereby enhancing the statistical reliability and categorical balance of the dataset. (4) Previous datasets are usually designed for a single task, whereas OpenDDI includes datasets for all major downstream tasks—binary classification, multiclass classification, and multilabel classification—thus achieving the broadest task coverage.

Through the exploration of **Q1**, we draw two key conclusions. **C1**: *OpenDDI substantially surpasses previous datasets in scale, with a far greater number of drugs and a much larger volume of DDI records.* **C2**: *OpenDDI covers significantly more DDI types with reliable records, and offers broader applicability, as it provides datasets for all major DDI prediction tasks, thereby meeting a wider range of research and practical needs.* Tatonetti et al. (2012)

Table 3: Performance comparison of methods on datasets with MM (multimodal), SMILES, and Path representations. The impr. shows the average relative performance improvement of MM over SMILES and Path, with positive values (marked with ↑).

Method	ODDI-M				Ryus				ZS-DDIE			
	SMILES	Path	MM	impr. (%)	SMILES	Path	MM	impr. (%)	SMILES	Path	MM	impr. (%)
DeepDDI	35.5 ±1.9	29.0 ±1.8	53.3 ±1.6	**68.8** ↑	70.8 ±1.3	64.0 ±1.4	78.3 ±1.0	**16.9** ↑	53.2 ±1.7	47.9 ±1.8	67.5 ±1.5	**33.7** ↑
CASTER	52.7 ±1.8	52.8 ±1.7	77.2 ±1.3	**46.4** ↑	75.4 ±1.2	73.3 ±1.3	84.4 ±0.9	**13.4** ↑	65.1 ±1.6	59.4 ±1.8	66.8 ±1.5	**8.9** ↑
GoGNN	71.1 ±1.4	67.4 ±1.5	74.7 ±1.2	**8.3** ↑	84.0 ±0.9	80.6 ±1.0	88.4 ±0.7	**7.6** ↑	74.5 ±1.3	67.3 ±1.5	78.0 ±1.2	**11.0** ↑
MRCGNN	76.5 ±1.3	74.3 ±1.4	81.6 ±1.1	**8.4** ↑	85.1 ±0.9	83.3 ±1.0	92.6 ±0.6	**10.1** ↑	75.9 ±1.3	72.9 ±1.4	79.9 ±1.1	**7.8** ↑
LAGAT	71.4 ±1.5	69.2 ±1.6	73.4 ±1.4	**4.5** ↑	86.7 ±1.0	85.3 ±1.1	90.5 ±0.9	**5.3** ↑	79.2 ±1.4	75.2 ±1.6	83.5 ±1.2	**8.7** ↑
DDKG	84.9 ±1.1	83.0 ±1.3	86.1 ±1.0	**2.6** ↑	88.9 ±0.8	87.6 ±0.9	90.5 ±0.7	**2.6** ↑	80.7 ±1.3	77.3 ±1.5	81.2 ±1.2	**3.0** ↑
KGNN	67.2 ±1.7	64.7 ±1.8	71.5 ±1.5	**8.6** ↑	84.6 ±1.1	84.0 ±1.2	84.9 ±1.1	**0.7** ↑	78.0 ±1.4	75.7 ±1.6	80.1 ±1.3	**4.4** ↑
DDIMDL	44.2 ±2.0	43.0 ±1.9	56.8 ±1.8	**30.7** ↑	71.1 ±1.5	70.4 ±1.5	79.4 ±1.2	**12.1** ↑	58.1 ±1.7	55.9 ±1.8	74.0 ±1.4	**30.4** ↑
TIGER	42.1 ±2.1	37.5 ±2.0	62.8 ±1.6	**60.0** ↑	65.1 ±1.5	63.6 ±1.6	84.6 ±1.0	**31.9** ↑	51.2 ±1.9	48.8 ±1.9	56.2 ±1.8	**13.1** ↑
MVA-DDI	58.5 ±1.7	56.3 ±1.7	67.4 ±1.4	**17.5** ↑	75.5 ±1.3	74.3 ±1.4	83.6 ±1.1	**11.7** ↑	67.0 ±1.6	67.1 ±1.7	72.6 ±1.5	**8.3** ↑
MKGFENN	86.5 ±1.0	85.6 ±1.1	87.0 ±0.9	**1.1** ↑	89.7 ±0.8	87.5 ±0.9	90.3 ±0.7	**1.9** ↑	81.1 ±1.2	81.5 ±1.2	82.8 ±1.1	**1.8** ↑
MUFFIN	66.9 ±1.5	77.2 ±1.3	80.2 ±1.2	**12.6** ↑	87.5 ±0.9	87.7 ±0.8	88.4 ±0.9	**0.4** ↑	77.2 ±1.2	76.2 ±1.4	77.9 ±1.3	**0.7** ↑

4.2 Datasets and Drug Representations Evaluation

To address **Q2**, we evaluate the model performance across diverse DDI datasets and drug representations, as shown in Table 3. From the perspective of DDI datasets, when facing more challenging multiclass prediction tasks, the performance of models trained and evaluated on OpenDDI datasets drops significantly, and the differences between models become more pronounced. This indicates that OpenDDI's multiclass datasets are more challenging and better at differentiating the strengths and weaknesses of different methods. From the perspective of drug representations, methods that utilize multimodal drug features consistently outperform those based on other types of drug features, with the advantage being particularly significant in multiclass DDI datasets. This highlights the superiority of modeling drugs with multimodal features.

Through the exploration of **Q2**, we can derive several key conclusions. **C3**: *For the more challenging DDI multiclass prediction tasks, existing methods perform poorly when facing difficult datasets, highlighting the need for further advancement in DDI prediction approaches*. **C4**: *Drug multimodal representations demonstrate clear advantages in drug modeling, underscoring the necessity of constructing multimodal representations* Wu et al. (2024).

4.3 Multimodal Research

To address **Q3**, we conduct an in-depth investigation of the multimodal representations from two perspectives: (1) the necessity of each modality in drug modeling, and (2) the capability of other modalities to supplement perspectives absent in the SMILES modality.

For the necessity of each modality in drug modeling, we obtain several observations from Figure 3: (i) Compared with the complete multimodal presentations, the absence of any individual modality negatively impacts F1 scores. (ii) In particular, when SMILES, text, 3D, or path modality is removed,

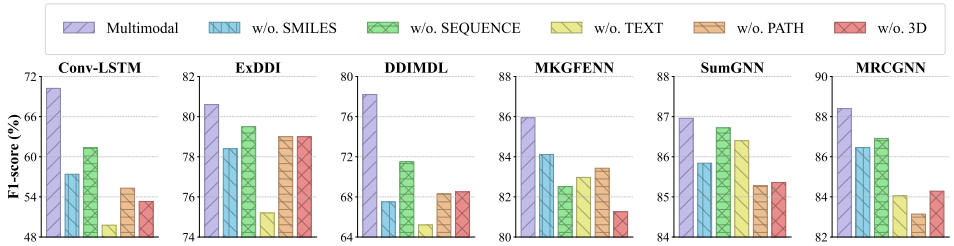

Figure 3: F1-score(%) comparison for multiclass-prediction task after ablating a modality.

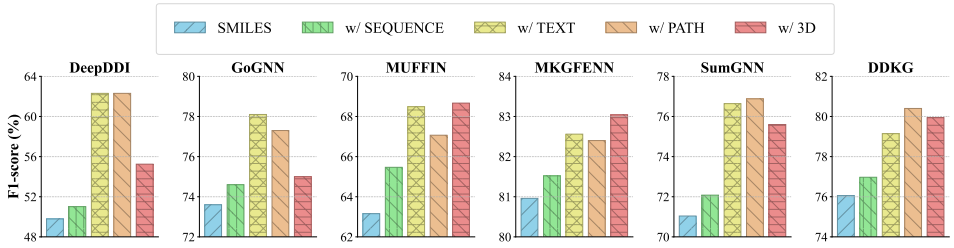

Figure 4: F1-score(%) comparison for multiclass-prediction task with SMILES and another modality.

most models exhibit a marked drop in task performance. (iii) By contrast, the absence of sequence modality causes a relatively smaller decline in performance for roughly half of the models.

For the capability of other modalities to supplement perspectives absent in the SMILES modality, we also derive several observations from Figure 4: (i) The integration of SMILES with other modalities consistently enhances F1 scores, demonstrating improved performance in the DDI prediction task. (ii) Incorporating text, 3D, or path modalities in addition to SMILES leads to substantial improvements across most models. (iii) However, compared with other modalities, the sequence modality contributes less pronounced benefits when integrated with SMILES for drug modeling learning.

Through our investigation of **Q3**, we derive three conclusions. **C5**: *All five modalities provide complementary perspectives for drug modeling, enhancing the ability of models to capture drug characteristics and improve task performance*. **C6**: *Among them, text, 3D, path, and SMILES prove most influential, with text, 3D, and path particularly reinforcing SMILES by supplying additional information and perspectives*. **C7**: *By contrast, the sequence modality contributes less, likely because drug–drug interactions are more strongly determined by structural properties and shared protein interactions, which tend to be better captured by other modalities* Su et al. (2022); Wang et al. (2025).

4.4 Model Performance Comparison

To address **Q4**, we focus on Table 3, Table 7 and Table 8. In the binary classification task, Graph-based and Integration-based methods generally perform well, with models such as DDKG and Muffin standing out. In the multiclass tasks, Graph-based methods excel by leveraging graph structures to enhance the understanding of drugs and DDIs, highlighting the importance of graph-aware modeling. Conversely, similarity-based methods exhibit limited predictive capacity. In the multilabel classification task, the differences among the three types of methods are relatively small.

Table 4: Generalization performance of methods when training set drugs are missing (w/: with training drugs; w/o.: without training drugs). The best results in each column are in **bold**.

Method	Ryus		ZS-DDIE	
	w/	w/o.	w/	w/o.
DeepDDI	77.9 ±1.2	58.5 ±2.8	72.0 ±1.8	25.9 ±3.2
CASTER	87.7 ±1.0	**68.5** ±2.4	77.8 ±1.5	**55.8** ±2.9
DSNDDI	86.4 ±1.1	56.3 ±2.7	75.1 ±1.6	40.1 ±3.1
MRCGNN	88.5 ±0.9	55.7 ±2.6	80.4 ±1.4	35.5 ±3.0
MKGFENN	**89.6** ±0.8	41.9 ±3.1	**82.1** ±1.3	37.8 ±3.3
MVA-DDI	86.6 ±1.0	61.6 ±2.5	81.9 ±1.2	47.4 ±2.8

Through the exploration of **Q4**, we draw the conclusion **C8**: *In DDI prediction tasks, Graph-based and Integration-based methods hold certain advantages. In contrast, similarity-based methods often yield sub-optimal performance, failing to capture high-order relational patterns* Lin et al. (2023).

4.5 Generalization Analyses

To address **Q5**, we withhold 30% drugs from the training set to serve as unseen drugs during testing, the detailed experimental results are summarized in Table 4. Observations indicate a universal performance drop across all methods when they encounter previously unseen drugs. Notably, DeepDDI, MRCGNN, and MKGFENN exhibit a substantial decline on the ZS-DDIE dataset, with their performance metrics dropping by more than twofold. Conversely, CASTER maintains a degree of stability across both datasets, demonstrating superior generalization capabilities.

Through the exploration of **Q5**, we draw the conclusion **C9**: *Existing methods exhibit insufficient generalization capabilities. To address the challenges of DDI prediction for unseen drugs, it is imperative for methods to bolster their generalization* Huang et al. (2020).

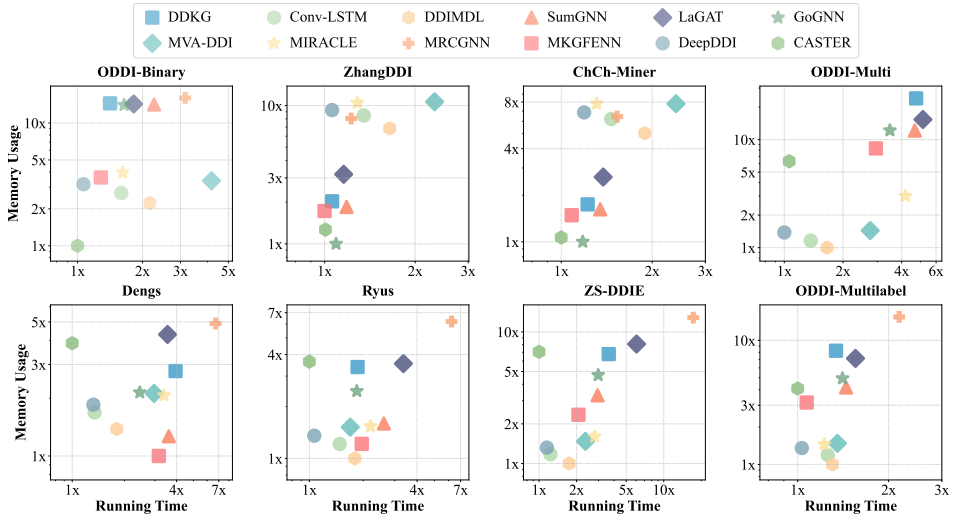

Figure 5: Efficiency Performance on Various Datasets.

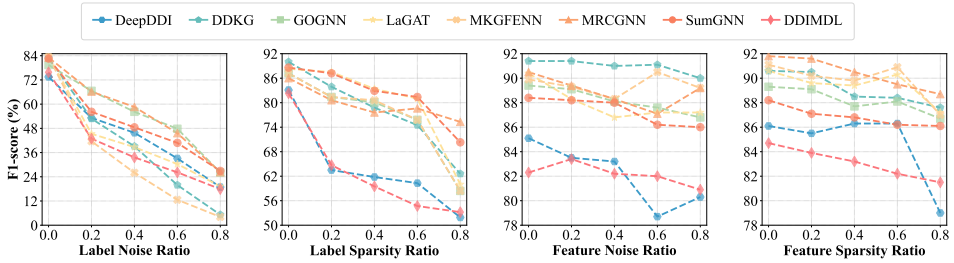

Figure 6: Performance under Different Noise and Sparsity Ratios at Label and Feature Levels.

4.6 Efficiency Analyses

To address **Q6**, we evaluated the efficiency of DDI prediction methods across a series of datasets from perspectives of memory and time, thereby analyzing their scalability.

As shown in Figure 5, we can observe several notable trends. Graph-based methods, due to their reliance on additional operations over graph structures, generally incur higher memory in most experiments. Similarity and Integration-based methods, in contrast, exhibit lower memory usage on smaller datasets such as Dengs. However, when scaled to larger datasets like OpenDDI-Multi, these methods still demonstrate instability in memory consumption. With respect to computational time, most models achieve satisfactory efficiency on smaller datasets. Yet, when confronted with large-scale datasets, certain models such as MKGFENN incur substantially higher time costs.

Based on the analysis of **Q6**, we derived conclusion **C10**: *Existing methods require further reductions in memory and time to enhance model scalability* Hong et al. (2022).

4.7 Robustness Analyses

To address **Q7**, we designed comprehensive experiments to simulate real-world challenges in DDI prediction task, including data noise and sparsity. These experiments evaluated the performance of methods under realistic conditions, ensuring a thorough assessment of their robustness.

The experimental results (Figure 6) show that noise and sparsity scenarios negatively affect the performance of all methods. In the case of label noise, increasing levels of noise consistently lead to sharp and steady declines in the performance of all algorithms. Regarding label sparsity, MRCGNN and SumGNN manage to maintain performance to some extent, whereas most other methods eventually exhibit substantial degradation; notably, DeepDDI and DDIMDL already experience significant drops under relatively low levels of label sparsity. Feature noise has a milder impact on model performance, and algorithms such as MKGFENN even show slight improvements under certain proportions of feature noise. As for feature sparsity, its effect on most methods is relatively limited, with the exception of DeepDDI, whose performance deteriorates markedly under high feature sparsity.

Based on the analysis of **Q7**, we draw the conclusion **C11**: *the robustness of existing DDI prediction methods remains insufficient, particularly in handling label-related issues. Therefore, enhancing the robustness of these methods is an urgent need that must be addressed* Wang et al. (2020).

5 Conclusion and Future Directions

In this paper, we first introduce DDI prediction, outlining its application scenarios and recent research progress. We then present OpenDDI, a comprehensive benchmark for DDI prediction research, which unifies 6 widely used DDI datasets and 2 existing forms of drug representation, while additionally contributing 3 new large-scale LLM-augmented datasets and a new multimodal drug representation covering 5 modalities. Besides, it unifies 20 baselines across 3 downstream tasks, with standardized protocols for data quality, effectiveness, generalization, robustness, and efficiency. With the standardized evaluation by OpenDDI, we conduct an in-depth investigation of DDI prediction, summarizing 11 valuable conclusions. Building upon these conclusions, we further identify promising future directions.

Advancing the data development (*C1*, *C2*, *C3*, and *C4*). Most existing DDI datasets suffer from limited scale and narrow application scenarios, while current drug representations often contain coarse or single-dimensional information. In fact, the scale and diversity of datasets determine the scope and applicability of DDI prediction tasks, whereas the quality of representations directly affects the models' understanding of DDIs. For the sustainable development of DDI prediction, it is essential to advance the data development.

Enhancing the interpretability of DDIs (*C5*, *C6*, and *C7*). Understanding how DDIs occur is of great significance both for DDI prediction and for real-world applications. The introduction of multimodal information offers additional perspectives for interpretability, enabling deeper insights into the underlying causes of DDIs. Future work on DDI prediction can further advance interpretability, providing novel perspectives on DDI mechanisms.

Improving the predictive capability of methods (*C8* and *C9*). Existing methods demonstrate promising performance on some earlier datasets, yet they often fail to sustain strong results when confronted with more challenging datasets or unseen drugs. To achieve more comprehensive and outstanding DDI prediction, and to address increasingly difficult challenges in the future, it is necessary to further strengthen both the predictive capability and adaptability of models.

Optimizing the efficiency of algorithms (*C10*). Current DDI prediction methods incur substantial time and memory costs when handling large-scale datasets. As the field continues to advance, future datasets will inevitably grow in scale, further intensifying the challenge of algorithmic efficiency. To support the continued development of DDI prediction, it is therefore imperative to optimize algorithm efficiency and enhance model scalability.

Addressing real-world DDI data challenges (*C11*). In practical DDI prediction scenarios, data noise and sparsity are unavoidable obstacles. Current DDI prediction methods remain insufficiently adapted to these real-world challenges, particularly when handling noisy and sparse label. Future methods should place greater emphasis on coping with such challenges and enhancing robustness.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript for limited purposes, including language refinement, literature retrieval and discovery, research ideation and execution support, and synthetic data generation. Specifically, AI tools assisted with improving writing clarity and organization, identifying and exploring related work and research directions, discussing methodological designs and experimental considerations, and supporting the generation, validation, and annotation of synthetic DDI data.

REFERENCES

Xing Chen, Biao Ren, Ming Chen, Quanxin Wang, Lixin Zhang, and Guiying Yan. Nllss: predicting synergistic drug combinations based on semi-supervised learning. *PLoS computational biology*, 12(7):e1004975, 2016.

Yujie Chen, Tengfei Ma, Xixi Yang, Jianmin Wang, Bosheng Song, and Xiangxiang Zeng. Muffin: multi-scale feature fusion for drug–drug interaction prediction. *Bioinformatics*, 37(17):2651–2658, 03 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab169. URL https://doi.org/10.1093/bioinformatics/btab169.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046—1051, November 2011. ISSN 1087-0156. doi: 10.1038/nbt.1990. URL https://doi.org/10.1038/nbt.1990.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36(15):4316–4322, 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa501. URL https://doi.org/10.1093/bioinformatics/btaa501.

Yanglan Gan, Wenxiao Liu, Guangwei Xu, Cairong Yan, and Guobing Zou. Dmfddi: deep multimodal fusion for drug–drug interaction prediction. *Briefings in Bioinformatics*, 24(6):bbad397, 11 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad397. URL https://doi.org/10.1093/bib/bbad397.

Judith L Gulikers, Leila-Sophie Otten, Lizza EL Hendriks, Kristien Winckers, Yvonne Henskens, Jenneke Leentjens, Michel M van den Heuvel, Rob Ter Heine, Sander Croes, Berber Piet, et al. Proactive monitoring of drug–drug interactions between direct oral anticoagulants and small-molecule inhibitors in patients with non-small cell lung cancer. *British Journal of Cancer*, 131(3): 481–490, 2024.

Kyuho Han, Edwin E Jeng, Gaelen T Hess, David W Morgens, Amy Li, and Michael C Bassik. Synergistic drug combinations for cancer identified in a crispr screen for pairwise genetic interactions. *Nature biotechnology*, 35(5):463–474, 2017.

Micheal Hewett, Diane E. Oliver, Daniel L. Rubin, Katrina L. Easton, Joshua M. Stuart, Russ B. Altman, and Teri E. Klein. Pharmgkb: the pharmacogenetics knowledge base. *Nucleic Acids Research*, 30(1):163–165, 01 2002.

Yue Hong, Pengyu Luo, Shuting Jin, and Xiangrong Liu. Lagat: link-aware graph attention network for drug–drug interaction prediction. *Bioinformatics*, 38(24):5406–5412, 10 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac682. URL https://doi.org/10.1093/bioinformatics/btac682.

Yang Hua, Zhenhua Feng, Xiaoning Song, Xiao-Jun Wu, and Josef Kittler. Mmdg-dti: Drug–target interaction prediction via multimodal feature fusion and domain generalization. *Pattern Recognition*, 157:110887, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2024.110887. URL https://www.sciencedirect.com/science/article/pii/S0031320324006381.

Degen Huang, Zhenchao Jiang, Li Zou, and Lishuang Li. Drug–drug interaction extraction from biomedical literature using support vector machine and long short term memory networks. *Information sciences*, 415:100–109, 2017.

Kexin Huang, Cao Xiao, Trong Nghia Hoang, Lucas Glass, and J. Sun. Caster: Predicting drug interactions with chemical substructure representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 702–709, 04 2020. doi: 10.1609/aaai.v34i01.5412.

Vassilis N. Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xiangxiang Zeng, and George Karypis. Drkg - drug repurposing knowledge graph for covid-19. https://github.com/gnn4dr/DRKG/, 2020.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 11 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1092. URL https://doi.org/10.1093/nar/gkw1092.

Md. Rezaul Karim, Michael Cochez, Joao Bosco Jares, Mamtaz Uddin, Oya Beyan, and Stefan Decker. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pp. 113–123, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366663. doi: 10.1145/3307339.3342161. URL https://doi.org/10.1145/3307339.3342161.

Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. Pubchem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 09 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv951. URL https://doi.org/10.1093/nar/gkv951.

Zimeng Li, Shichao Zhu, Bin Shao, Xiangxiang Zeng, Tong Wang, and Tie-Yan Liu. Dsn-ddi: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning. *Briefings in Bioinformatics*, 24(1), 2023a.

Zimeng Li, Shichao Zhu, Bin Shao, Xiangxiang Zeng, Tong Wang, and Tie-Yan Liu. Dsn-ddi: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning. *Briefings in Bioinformatics*, 24(1), 2023b.

Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. Kgnn: knowledge graph neural network for drug-drug interaction prediction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, IJCAI'20, 2020. ISBN 9780999241165.

Xuan Lin, Qi Wen, Sijie Yang, Zu-Guo Yu, Yahui Long, and Xiangxiang Zeng. Interpretable multi-view attention network for drug-drug interaction prediction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 422–427, 2023. doi: 10.1109/BIBM58861.2023.10385757.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, 35(suppl_1):D198–D201, 12 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl999. URL https://doi.org/10.1093/nar/gkl999.

Tiqing Liu, Linda Hwang, Stephen K. Burley, Carmen I. Nitsche, Christopher Southan, W. Patrick Walters, and Michael K. Gilson. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53(D1):D1633–D1644, nov 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1075. URL https://doi.org/10.1093/nar/gkae1075.

Mei Ma and Xiujuan Lei. A dual graph neural network for drug–drug interactions prediction based on molecular structure and interactions. *PLOS Computational Biology*, 19(1):e1010812, 2023.

David Mendez, Anna Gaulton, A. Patrícia Bento, Jon Chambers, Marleen de Veij, Eloy Felix, María P. Magariños, Juan Fernando Mejía Mosquera, Prudence Mutowo, Michal Nowotka, María Gordillo-Marañón, Fiona M. I. Hunter, Laura Junco, Grace Mugumbate, Milagros Rodríguez-López, Francis

Atkinson, Nicolas Bosc, Chris J. Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47:D930 – D940, 2018. URL https://api.semanticscholar.org/CorpusID:53223613.

Yang Qiu, Yang Zhang, Yifan Deng, Shichao Liu, and Wen Zhang. A comprehensive review of computational methods for drug-drug interaction detection. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(4):1968–1985, 2021a.

Yang Qiu, Yang Zhang, Yifan Deng, Shichao Liu, and Wen Zhang. A comprehensive review of computational methods for drug-drug interaction detection. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(4):1968–1985, 2021b.

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

Reza Safdari, Reza Ferdousi, Kamal Aziziheris, Sharareh R Niakan-Kalhori, and Yadollah Omidi. Computerized techniques pave the way for drug-drug interaction prediction and interpretation. *BioImpacts: BI*, 6(2):71, 2016.

Zhenqian Shen, Mingyang Zhou, Yongqi Zhang, and Quanming Yao. Benchmarking drug–drug interaction prediction methods: a perspective of distribution changes. *Bioinformatics*, 41(11): btaf569, 2025.

Xiaorui Su, Lun Hu, Zhuhong You, Pengwei Hu, and Bowei Zhao. Attention-based knowledge graph representation learning for predicting drug-drug interactions. *Briefings in Bioinformatics*, 23(3): bbac140, 04 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac140. URL https://doi.org/10.1093/bib/bbac140.

Xiaorui Su, Pengwei Hu, Zhu-Hong You, Philip S. Yu, and Lun Hu. Dual-channel learning framework for drug-drug interaction prediction via relation-aware heterogeneous graph transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):249–256, Mar. 2024. doi: 10.1609/aaai.v38i1.27777. URL https://ojs.aaai.org/index.php/AAAI/article/view/27777.

Henan Sun, Xunkai Li, Zhengyu Wu, Daohan Su, Rong-Hua Li, and Guoren Wang. Breaking the entanglement of homophily and heterophily in semi-supervised node classification. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 2379–2392. IEEE, 2024.

Xia Sun, Ke Dong, Long Ma, Richard Sutcliffe, Feijuan He, Sushing Chen, and Jun Feng. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, 21(1):37, 2019.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, and Yulan He. Exddi: explaining drug-drug interaction predictions with natural language. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i24.34709. URL https://doi.org/10.1609/aaai.v39i24.34709.

Takako Takeda, Ming Hao, Tiejun Cheng, Stephen H Bryant, and Yanli Wang. Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *Journal of cheminformatics*, 9(1):16, 2017.

Nicholas P. Tatonetti, Patrick P. Ye, Roxana Daneshjou, and Russ B. Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125):125ra31–125ra31, 2012. doi: 10.1126/scitranslmed.3003377. URL https://www.science.org/doi/abs/10.1126/scitranslmed.3003377.

Hanchen Wang, Defu Lian, Ying Zhang, Lu Qin, and Xuemin Lin. Gognn: graph of graphs neural network for predicting structured entity interactions. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2020. ISBN 9780999241165.

Jihong Wang, Xiaodan Wang, and Yuyao Pang. Structnet-ddi: Molecular structure characterization-based resnet for prediction of drug–drug interactions. *Molecules*, 29(20), 2024a. ISSN 1420-3049. doi: 10.3390/molecules29204829. URL https://www.mdpi.com/1420-3049/29/20/4829.

Lingfeng Wang, Yinghong Li, Yaozheng Zhou, Liping Guo, and Congzhou Chen. Mfe-ddi: A multi-view feature encoding framework for drug-drug interaction prediction. *Computational and Structural Biotechnology Journal*, 27:2473–2480, 2025. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2025.05.029. URL https://www.sciencedirect.com/science/article/pii/S2001037025001941.

Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 2921–2933, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449786. URL https://doi.org/10.1145/3442381.3449786.

Ziyan Wang, Zhankun Xiong, Feng Huang, Xuan Liu, and Wen Zhang. Zeroddi: a zero-shot drug-drug interaction event prediction method with semantic enhanced learning and dual-modal uniform alignment. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024b. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/671. URL https://doi.org/10.24963/ijcai.2024/671.

Steven Whitebread, Jacques Hamon, Dejan Bojanic, and Laszlo Urban. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug discovery today*, 10(21):1421–1433, 2005.

David S. Wishart, Yannick Djoumbou-Feunang, An Chi Guo, Elvis Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel L. Johnson, Carin Li, and Zinat Sayeeda. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2017. ISSN 0305-1048. doi: https://doi.org/10.1093/nar/gkx1037.

Di Wu, Wu Sun, Yi He, Zhong Chen, and Xin Luo. Mkg-fenn: A multimodal knowledge graph fused end-to-end neural network for accurate drug–drug interaction prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10216–10224, Mar. 2024. doi: 10.1609/aaai.v38i9.28887. URL https://ojs.aaai.org/index.php/AAAI/article/view/28887.

Zhankun Xiong, Shichao Liu, Feng Huang, Ziyan Wang, Xuan Liu, Zhongfei Zhang, and Wen Zhang. Multi-relational contrastive learning graph neural network for drug-drug interaction event prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5339–5347, Jun. 2023. doi: 10.1609/aaai.v37i4.25665. URL https://ojs.aaai.org/index.php/AAAI/article/view/25665.

Yue Yu, Kexin Huang, Chao Zhang, Lucas M Glass, Jimeng Sun, and Cao Xiao. Sumgnn: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics*, 37(18):2988–2995, 03 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab207. URL https://doi.org/10.1093/bioinformatics/btab207.

Yongna Yuan, Jiaqi Yue, Ruisheng Zhang, and Wei Su. Phgl-ddi: A pre-training based hierarchical graph learning framework for drug-drug interaction prediction. *Expert Systems with Applications*, 270:126408, 2025. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2025.126408. URL https://www.sciencedirect.com/science/article/pii/S0957417425000302.

Yi Zhong, Gaozheng Li, Ji Yang, Houbing Zheng, Yongqiang Yu, Jiheng Zhang, Heng Luo, Biao Wang, and Zuquan Weng. Learning motif-based graphs for drug–drug interaction prediction via local–global self-attention. *Nature Machine Intelligence*, 6(9):1094–1105, 2024.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018a.

Marinka Zitnik, Rok Sosič, Sagar Maheshwari, and Jure Leskovec. BioSNAP Datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata, August 2018b.

A BASELINE ALGORITHM DETAILS

DeepDDI Ryu et al. (2018) is a classic deep learning-based benchmark model for drug interaction prediction. It takes chemical structure information (SMILES) and names of drug pairs as input, extracts features through a structural similarity profile (SSP), and employs a deep neural network for multi-label classification, outputting human-readable sentences that describe the interaction types. The architecture of this method is relatively straightforward, providing a reliable and commonly used performance reference point for evaluating more complex drug interaction prediction models.

DDIMDL Deng et al. (2020) is a multimodal benchmark model designed to predict specific events of drug-drug interactions. It constructs individual deep neural network sub-models for different drug features (such as chemical substructures, targets, and enzymes), then fuses the outputs of these sub-models to learn cross-modal representations of drug pairs. This method demonstrates excellent performance on standard datasets. Compared to approaches that rely solely on single-source features (e.g., DeepDDI), it provides an important reference on how to effectively integrate multi-source data for fine-grained prediction.

CASTER Huang et al. (2020) is an advanced model for predicting drug interactions that emphasizes interpretability. It extracts functional substructures from drug chemical structures via sequence mining and pre-trains an autoencoder on large-scale unlabeled data to obtain generalized representations. Its core innovation lies in projecting drug pairs into a substructure space through dictionary learning, where the resulting sparse coefficients are used both for prediction and for quantifying the importance of each substructure. This achieves a unification of performance and chemical interpretability, establishing an important benchmark for such research.

DDKG Su et al. (2022) is an attention-based knowledge graph representation learning model for predicting drug interactions. It synergistically leverages the chemical structures (SMILES) of drugs and knowledge graph triples, first learning initial embeddings via an encoder-decoder and then propagating and aggregating neighbor information along high-attention paths to derive global node representations. The model serves as a strong benchmark for fusing multi-source knowledge for relational inference.

KGNN Lin et al. (2020) is a knowledge graph-based graph neural network framework for drug-drug interaction prediction. Unlike methods that rely on chemical structures or single-source features, KGNN learns high-order topological representations of drugs and their related entities (e.g., targets, pathways) from knowledge graphs through multi-hop neighborhood sampling and aggregation. The model achieves strong performance across multiple drug datasets, particularly excelling in modeling long-range semantic relationships, and serves as an important benchmark for knowledge-graph-based drug reasoning.

LaGAT Hong et al. (2022) is a graph attention network model for drug-drug interaction prediction. Its core innovation lies in a link-aware attention mechanism, which dynamically adjusts node attention pathways based on the currently predicted drug pair, thereby capturing the semantic diversity of the same drug in different interaction contexts. This method achieves excellent predictive performance and interpretability on knowledge graphs, providing an important technical reference for drug relation inference.

ExDDI Sun et al. (2025) is a model series designed to generate natural language explanations for drug-drug interaction prediction. It extends the DDI task for the first time to a joint learning task of "prediction + explanation generation." The series leverages the molecule-text pretrained model MolT5, exploring various methods such as sequence-to-sequence generation, multi-task training, retrieval, and large language model in-context prompting. Experiments demonstrate that training with more detailed explanations not only enhances interpretability but also improves prediction performance, providing an important benchmark for explainable drug safety research.

GOGNN Wang et al. (2020) is a graph-of-graphs neural network designed for drug-drug interaction (DDI) prediction. The model represents each entity as a local graph (e.g., molecular structure) and models inter-entity interactions as a global graph. It employs attention-based pooling on local graphs to extract key substructures and uses attention aggregation on the global graph to integrate interaction information. GoGNN demonstrates strong performance across multiple DDI prediction tasks, serving as a key benchmark method for graph-structured drug relationship modeling.

MIRACLE Wang et al. (2021) is a multi-view graph contrastive learning model for drug-drug interaction prediction. It encodes drug molecular structures using a bond-aware message passing network (BAMPN), integrates drug interaction network information via a graph convolutional network (GCN), and employs contrastive learning to align representations from both molecular structure and interaction relationship views. The model demonstrates robust performance across datasets of varying scales and exhibits good scalability. As a representative multi-view learning approach, MIRACLE is often adopted as a benchmark to evaluate the ability of new methods in integrating structural and relational information.

TIGER Su et al. (2024) is a dual-channel drug-drug interaction prediction model based on a relation-aware graph Transformer. It models multi-semantic relations between nodes via a relation-aware self-attention mechanism and employs subgraph sampling strategies to handle large-scale knowledge graphs. The model integrates dual-channel information from both drug molecular graphs and biomedical knowledge graphs, effectively capturing long-range dependencies and high-order structures. It demonstrates strong adaptability and predictive performance across datasets of varying scales, serving as a benchmark for evaluating Transformer-based multi-relational graph learning methods.

SumGNN Yu et al. (2021) is a graph representation learning framework based on knowledge graph summarization for multi-type drug-drug interaction prediction. It employs local subgraph extraction and layer-independent self-attention to generate sparse, interpretable reasoning pathways, and integrates multi-channel information (such as subgraph features and drug fingerprints) to enhance predictive performance—especially in low-resource scenarios. This model serves as an important benchmark for knowledge graph-enhanced DDI prediction.

PHGLDDI Yuan et al. (2025) is a hierarchical graph neural network framework designed for drug-drug interaction (DDI) prediction. The model adopts a dual-level architecture: at the molecular level, it extracts structural features of drugs using a pre-trained Graph Convolutional Network (GCN); at the network level, it learns the topological relationships between drugs in the interaction network through a Graph Isomorphism Network (GIN). By integrating molecular structural information with network structural information, PHGL-DDI demonstrates strong generalization capability in inductive learning tasks (predicting new drugs), making it a representative method in graph neural network-based DDI prediction.

MRCGNN Xiong et al. (2023) is a graph neural network model designed for predicting drug–drug interaction events. The model integrates structural information from drug molecular graphs with interaction information from DDI event graphs, employing a multi-relational contrastive learning approach and a dual-view negative augmentation strategy, which significantly enhances its ability to predict rare DDI events. MRCGNN not only demonstrates the advantages of jointly leveraging structural and interaction information, but also serves as an important example of using contrastive learning to address long-tailed distribution in graph data. It is regarded as a representative work in graph neural network-based DDI prediction.

DSNDDI Li et al. (2023a) is a state-of-the-art graph neural network-based framework for drug-drug interaction prediction that emphasizes dual-view representation learning. It models drugs as molecular graphs and iteratively learns substructures from both intra-view (within a single drug) and inter-view (between drug pairs) perspectives using a series of DSN encoders, which integrate graph attention mechanisms for local atom-level updates and self-attention pooling for global drug representations. The DSN decoder employs a co-attention scoring function to predict interactions by integrating pairwise substructure embeddings, enabling high accuracy in both transductive and inductive settings without relying on additional domain knowledge.

ConvLSTM Karim et al. (2019) is a drug-drug interaction prediction model that combines knowledge graph embeddings with a convolutional long short-term memory (Conv-LSTM) network. This method learns drug representations from multi-source knowledge graphs, extracts local features through convolutional layers, and then integrates them via LSTM layers for final prediction. Conv-LSTM excels in integrating heterogeneous drug information and handling high-dimensional sparse data, making it an early representative work in knowledge graph-based DDI prediction.

MUFFIN Chen et al. (2021) is a multi-scale feature fusion model for drug-drug interaction prediction. It integrates molecular structure features of drugs with semantic information from knowledge graphs, employing a dual-level fusion strategy (cross-level and scalar-level) to effectively combine multimodal information. MUFFIN supports binary, multi-class, and multi-label prediction tasks, demonstrating

strong performance across multiple benchmark datasets. It is regarded as an early representative approach for fusing molecular graph and knowledge graph information in DDI prediction.

MVA Lin et al. (2023) is a multi-view attention network model designed for drug-drug interaction prediction. It employs Transformer and Graph Convolutional Networks to extract sequence and structural features from drugs, respectively, and utilizes contrastive learning-based self-supervised pretraining to enhance graph representation. The model further adaptively fuses the dual-view features through an attention mechanism, generating discriminative drug representations. This approach demonstrates excellent performance across multiple benchmark tasks and serves as a representative example of integrating self-supervised pretraining with attention-based fusion in multi-view representation learning.

MKGFENN Wu et al. (2024) is an end-to-end neural network model for DDI prediction that fuses multimodal knowledge graphs. It integrates four types of knowledge graphs—drug–chemical entities, drug–substructures, drug–drug interactions, and molecular structures—and employs a four-channel graph neural network to extract features, which are then fused via a multilayer perceptron in an end-to-end manner. The model demonstrates strong performance across a variety of DDI prediction tasks, serving as a representative approach for combining multi-source knowledge through end-to-end joint learning.

MMDGDTI Hua et al. (2025) is a DTI prediction model that integrates multimodal feature fusion and domain generalization. It employs a large language model-based textual feature extractor and a hybrid graph neural network-based structural feature extractor to extract deep semantic and topological features of drugs, while enhancing generalization capability through domain adversarial training and contrastive learning. The model demonstrates strong performance in both single-domain and cross-domain tasks, serving as a multimodal prediction method with notable domain adaptability.

ZeroDDI Wang et al. (2024b) is a method designed for zero-shot drug-drug interaction event prediction. It learns discriminative features by integrating biologically enhanced semantic representations of DDI events with molecular substructure-guided attention, and introduces a dual-modal uniform alignment strategy to mitigate class imbalance. The method demonstrates strong performance in zero-shot prediction scenarios.

B Dataset details

The three core datasets in the OpenDDI benchmark were constructed by integrating and harmonizing data from 12 authoritative source databases, followed by a significant LLM-assisted augmentation process. This yielded a unified, multimodal graph encompassing over 34,000 drugs and 5.3 million interaction edges, which substantially expanded the coverage (e.g., drugs increased from 4,467 to 14,090 and DDI relations from 1.4M to 2.55M in the binary task). To elucidate the data composition, the following section details these 12 source databases, outlining their respective roles in providing multimodal node features and DDI edges, along with their inherent characteristics. Subsequently, five external DDI datasets used for comparative evaluation are introduced. A comprehensive quantitative comparison between the OpenDDI datasets and these external benchmarks is provided in Table 2.

B.1 Source Databases for Graph Construction

BindingDB Liu et al. (2006; 2024) is a public database of measured binding affinities, focusing on protein-target interactions with drug-like molecules. It provided detailed Kd, Ki, and IC50 values, which were instrumental in building the quantitative mechanism protein interaction modality, informing the model about the strength and specificity of drug-target binding.

The BIOSNAP Drug-Drug Interaction (DDI) dataset Zitnik et al. (2018b) is a network of known interactions between FDA-approved drugs, containing 1,514 drugs and 48,514 interactions extracted from drug labels and scientific publications. It provides validated DDI edges for graph construction and network statistics, serving as a valuable resource for studying drug interactions.

ChEMBL Mendez et al. (2018) is a large, manually curated bioactivity database containing over 1.8 million drug-like compounds and more than 15 million bioactivity records (as of Release 24, 2018). It serves as a primary source for molecular structures (e.g., SMILES strings) and bioactivity data, which are essential for constructing the 2D molecular graph modality.

Davis Davis et al. (2011) dataset provides quantitative binding affinity measurements (Kd values) between kinases and their inhibitors. These continuous scores were used to weight and refine the mechanism protein interaction modality, adding a layer of quantitative confidence to drug-target links and offering a granular perspective beyond binary relationships.

Drug Repurposing Knowledge Graph (DRKG) Ioannidis et al. (2020) is a large, heterogeneous biological knowledge graph integrating genes, compounds, and diseases across 107 relationship types. It was mined for relationships between drugs and biological entities, contributing to the textual description and mechanism protein interaction modalities with rich, interlinked contextual information.

DrugBank Wishart et al. (2017) provides comprehensive information on drug–drug interactions, including interaction types, affected targets, and metabolic enzymes. It served as the central pillar for data integration, providing a primary drug ID system, core DDI edges, textual descriptions, molecular structures (SMILES, SDF), and mechanism protein interactions for thousands of approved and investigational drugs.

KEGG Kanehisa et al. (2016) contains known adverse drug–drug interactions annotated as Contraindications (CI) and Precautions (P), providing systematic annotations of structural information, drug targets, and molecular mechanisms (e.g., CYP enzyme involvement) to support both 2D/3D structure-based and textual description-based modalities.

MEDLINE is a premier bibliographic database from the U.S. National Library of Medicine. Its vast corpus of life sciences literature was the foundational source for the textual description modality, used to generate structured, informative textual profiles for each drug.

OFFSIDES is a resource containing statistically associated adverse drug reactions derived from FDA Adverse Event Reporting System (FAERS) reports, after filtering out known side effects already listed on FDA-approved labels. It was used to enrich the textual description modality by adding off-label safety information to drug profiles.

Pharmacogenomics Knowledge Base (PharmGKB) Hewett et al. (2002) curates knowledge on how genetic variation impacts drug response. It provides annotated drug–gene relationships and pharmacogenomic pathway information, which was used to enrich the textual description modality and support the analysis of drug mechanism networks.

PubChem Kim et al. (2015) is a large-scale public repository of chemical substances and their biological activities. It contains over 60 million unique chemical structures (CIDs), each associated with multiple representations including SMILES, InChI, 2D/3D structures, and extensive bioactivity annotations. It is a major source for standardizing and retrieving molecular structure information, ensuring comprehensive coverage in the 2D molecular graph modality.

TWOSIDES dataset Tatonetti et al. (2012) is a resource for drug–drug–effect relationships, focusing on polypharmacy side effects. It serves as a core source for the multi-label classification dataset and providing multi-hot labels and DDI edges.

B.2 External Datasets for Benchmarking

ChCh-Miner dataset Wang et al. (2024a) is a medium-scale DDI benchmark comprising 1,514 drugs and 48,514 DDI links. It is widely adopted for evaluating the generalization ability of DDI prediction models across distinct drug networks.

ZS-DDIE dataset Wang et al. (2024b) is a manually annotated dataset for zero-shot drug–drug interaction event prediction, containing 2,004 approved drugs and 394,118 DDIs across 175 unique event types. It is used to evaluate the model's ability to generalize to unseen interaction types based on textual descriptions.

DeepDDI dataset Ryu et al. (2018) is a large-scale benchmark extracted from DrugBank, containing 192,284 DDIs across 191,878 drug pairs, annotated with 86 human-readable DDI event types. It is widely used for polypharmacy effect prediction and comparative analysis of DDI prediction models.

Dengs dataset Deng et al. (2020) is a curated benchmark for drug–drug interaction event prediction, comprising 572 drugs, 74,528 DDIs annotated with 65 fine-grained event types. It provides multi-

modal drug features (chemical substructures, targets, enzymes, pathways) and is widely used to evaluate multi-label classification models for DDI event prediction.

ZhangDDI dataset Wang et al. (2021) is a small-scale DDI dataset containing 548 drugs and 48,548 pairwise DDI links. It includes multiple types of similarity information about drug pairs, making it suitable for comparative analysis of model performance on similarity-based link prediction tasks.

C DRUG FEATURE REPRESENTATION

C.1 SMILES

SMILES (Simplified Linear Input System for Molecules) is a standard chemical notation that uses ASCII strings to linearly represent molecular structures. It precisely encodes the topological structure of molecules with a compact syntax (such as atom symbols, bond type symbols, and brackets), ensuring a one-to-one correspondence between each string and a unique molecular configuration. Due to its efficiency and explicit machine readability, SMILES has become a fundamental tool in cheminformatics for storing and exchanging molecular data and as input for computational models, and is widely used in database management, molecular visualization, and drug discovery research.

C.2 PATH

Path is a structured semantic network used to systematically organize drug-related information. Its nodes represent biomedical entities such as drugs, targets, diseases, and pathways, while edges represent the interactions between them (e.g., inhibition, activation, regulation). By integrating heterogeneous data from multiple sources (such as DrugBank and KEGG), these graphs place drugs within a rich biological context, revealing their potential mechanisms of action and relationships. In computational modeling, knowledge graphs can be transformed into low-dimensional vector representations of drug nodes using embedding techniques. These representations effectively capture the high-order semantic relationships of drugs within biological networks, providing key features for interaction prediction based on graph structure or relational reasoning.

C.3 MULTIMODAL

Multimodal drug feature representation aims to comprehensively characterize the complex properties of drugs by integrating multiple complementary biomedical evidence dimensions. This representation systematically combines chemical structure, spatial conformation, biochemical mechanisms, and semantic information to form a holistic description of drug properties and mechanisms of action. The detailed composition of each modality is as follows:

The **chemical structure modality** represents the chemical scaffold of a drug using SMILES strings and their derived two-dimensional molecular graphs. Atoms and chemical bonds constitute the nodes and edges of the graph, respectively, and are encoded via graph neural networks to capture functional group distributions, bond types, and overall topology. This enables modeling interactions directly driven by structural similarity.

The **three-dimensional conformation modality** captures the three-dimensional spatial geometry of drug molecules. Based on energetically favorable conformations generated via computational chemistry methods, three-dimensional graph neural networks extract stereochemical features such as atomic coordinates, distances, and angles. This modality supplements spatial complementarity information not reflected in 2D structures, which is crucial for understanding drug–target binding affinity.

The **target sequence modality** focuses on the primary structure of protein targets associated with drug action. Pretrained protein language models encode the amino acid sequences of these targets, generating deep semantic vectors rich in biochemical functional information. This reveals drug specificity and mechanistic relevance at the protein level.

The **textual semantic modality** integrates unstructured textual knowledge about drugs, including indications, pharmacological effects, and mechanistic descriptions. Domain-specific pretrained language models extract deep semantic representations from the text, introducing expert knowledge

and contextual information into the model. This assists in inferring interactions that go beyond structural similarity.

The **path** specifically constructs association networks between drugs and key mechanistic proteins (e.g., metabolic enzymes, transporters), forming a drug–mechanism protein bipartite graph. Encoding this network yields functional representations of drugs within specific pharmacokinetic and pharmacodynamic contexts, directly supporting the prediction of interactions mediated by specific biological mechanisms such as enzyme inhibition or metabolic competition.

D Evaluation Metrics

To ensure a rigorous and comprehensive assessment of model performance, this study employs distinct standardized evaluation metrics tailored to the specific nature of each Drug-Drug Interaction (DDI) prediction task. The definitions, calculation methodologies, and formulas for these metrics are detailed below.

D.1 Binary Classification Metrics

For the binary classification task, which determines the presence or absence of an interaction between a drug pair, **Accuracy** is utilized as the primary metric. Accuracy quantifies the overall probability of correct predictions across the entire dataset. It is defined as the ratio of correctly classified samples (both true positives and true negatives) to the total number of samples. Let TP (True Positive) denote the number of correctly predicted interacting pairs, TN (True Negative) denote the number of correctly predicted non-interacting pairs, FP (False Positive) denote the non-interacting pairs incorrectly predicted as interacting, and FN (False Negative) denote the interacting pairs incorrectly predicted as non-interacting. The accuracy is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

This metric provides an intuitive measure of the model's global performance in distinguishing between interacting and non-interacting drug pairs.

D.2 Multi-Classification Metrics

In the multi-classification scenario, where the goal is to predict specific interaction types among mutually exclusive classes, the **Macro-average F1 score (Macro-F1)** is adopted as the core metric. Given the prevalence of class imbalance in DDI datasets—where certain interaction types are far more frequent than others—simple accuracy can be misleading. The Macro-F1 score addresses this by treating all classes equally, regardless of their support size. It is calculated by first determining the Precision (P_i) and Recall (R_i) for each specific class i, and then computing the harmonic mean of these values to obtain the F1 score for that class ($F1_i$):

$$P_i = \frac{TP_i}{TP_i + FP_i}, \quad R_i = \frac{TP_i}{TP_i + FN_i}$$

$$F1_i = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i}$$

The final Macro-F1 score is the arithmetic mean of the per-class F1 scores across all C classes:

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} F1_i$$

This approach ensures that the model's performance on minority classes contributes equally to the final score, preventing majority classes from dominating the evaluation.

D.3 Multi-Label Classification Metrics

For the multi-label classification task, where a single drug pair may exhibit multiple interaction types simultaneously, the **Macro-average Area Under the Receiver Operating Characteristic Curve (Macro-AUC)** is selected as the evaluation standard. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various threshold settings. The Area Under the Curve (AUC) represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

In a multi-label context with L distinct labels, the AUC is calculated independently for each label j (treating it as a binary one-vs-rest problem) to obtain AUC_j. The Macro-AUC is then derived by averaging these values:

$$\text{Macro-AUC} = \frac{1}{L}\sum_{j=1}^{L} AUC_j$$

This metric is particularly robust for evaluating scenarios with high label sparsity, as it assesses the model's ability to correctly rank the presence of each specific interaction type independently of others.

E Experimental Setting Details

The experiments were carried out on a computational platform featuring an NVIDIA A100 80GB PCIe GPU and an Intel(R) Xeon(R) Gold 6240 CPU running at 2.60GHz, with CUDA version 12.4 activated. The programming environment consisted of Python 3.8.0 and PyTorch 2.2.0, selected for their excellent compatibility and efficiency in supporting the DDI prediction algorithms. Furthermore, the hyperparameters for each algorithm were tuned according to findings from previous studies to ensure reproducibility and dependable outcomes.

F Implementation Details of OpenDDI Framework

F.1 Framework Architecture

The framework consists of four core subsystems coordinated by a central entry point (`main.py`):
Data Abstraction Layer (`DatasetManager`): Encapsulates the heterogeneity of biomedical data sources. It dynamically instantiates specific dataset classes (e.g., for graph-based or integration-based models) based on the configuration, transparently handling feature loading (SMILES strings, Multimodal...) and preprocessing.
Model Factory (`ModelManager`): Implements a factory pattern to manage the lifecycle of deep learning models.
Training Protocol Abstraction (`Trainer`): Encapsulates the concrete optimization and evaluation logic. It provides specialized implementations (e.g., `Unified_Trainer`) to handle diverse learning paradigms, ensuring consistent metric calculation (AUC, AUPR, F1) across all comparisons.
Execution Engine (`Pipeline`): Orchestrates the experimental workflow. It integrates the dataset, model, and optimizer, serving as the central coordinator that dispatches the configured components to the appropriate training handler.

F.2 DatasetManager Flow

Representative code examples as follows:

F.3 Trainer Flow

Representative code examples as follows:

Algorithm 1 Data Abstraction Layer: Modular Dataset Management

```
class DatasetManager:
    """
    Factory class responsible for instantiating the appropriate dataset
        module
    based on the model configuration and origin flag.
    """
    def __init__(self, args):
        self.args = args
        # Mapping model names to their corresponding unified dataset
            classes
        self.dataset_mapping = {
            # ... mappings for models
        }

    def load_dataset(self):
        """
        Dynamically returns a dataset instance.
        Supports the unified OpenDDI implementation.
        """
        dataset_cls = self.dataset_mapping[self.args.model]
        return dataset_cls(self.args)


class BaseDataset:
    """
    Abstract base class encapsulating the end-to-end data processing
        pipeline.
    Composed of specialized functional modules to handle features, graphs
        , and splits.
    """
    def __init__(self, args):
        self.args = args
        # Composition of functional modules
        self.loader = DataLoadingModule()
        self.feature_processor = FeatureProcessingModule()
        self.splitter = DataSplittingModule()
        self.graph_builder = GraphConstructionModule()

    def load_data(self):
        """
        The main entry point for data preparation.
        Orchestrates the sequence: Load -> Process -> Build Graph ->
            Split.
        """
        pass
```

F.4 Pipeline Flow

G Hyperparameter Settings

General Experimental Settings. Across all DDI prediction experiments, we adopt a unified set of hyperparameters to ensure fair and consistent evaluation. The learning rate is set to 1×10^{-4} by default, and models are trained with a batch size of 512. Training is performed for 100 epochs. The hidden dimension is fixed to 512 for all models unless otherwise specified.

To stabilize optimization and mitigate overfitting, we apply a weight decay of 5×10^{-4} and a dropout rate of 0.3. All models are optimized using the Adam optimizer. These settings are used consistently across binary, multiclass, multilabel, and generalization experiments to isolate the impact of model

Algorithm 2 Training Protocol Abstraction: Unified Execution Loop

```
class BaseTrainer(ABC):
    """
    Abstract base class defining the standardized training protocol.
    Implements the Template Method pattern for the execution loop.
    """
    def __init__(self, args, logger, dataset, model, optimizer):
        self.device = self._setup_device(args.device)
        self._move_data_to_device(dataset)

    def train(self):
        """
        The Template Method defining the skeleton of the training process
            .
        """
        task_type = self._determine_task_type()
        loss_fct = self._get_loss_function(task_type)

        for epoch in range(self.args.epochs):
            # 1. Training Step
            train_metrics, train_loss = self._train_epoch(epoch, loss_fct
                )

            # 2. Validation Step
            val_metrics, val_loss = self._evaluate(self.dataset.
                val_loader)

        # 3. Final Testing
        test_metrics, _ = self._evaluate(self.dataset.test_loader)
        self._save_results(test_metrics)

    @abstractmethod
    def _train_epoch(self, epoch, loss_fct):
        """Specific implementation of the forward-backward pass."""
        pass

    @abstractmethod
    def _compute_metrics(self, y_true, y_logits):
        """Standardized metric calculation (e.g., AUC, F1, Accuracy)."""
        pass


class UnifiedTrainer(BaseTrainer):
    """
    Standard implementation for most DDI models.
    """
    def _train_epoch(self, epoch, loss_fct):
        pass

    def _get_loss_function(self, task_type):
        pass
```

design rather than hyperparameter choices. All results are reported as the mean and standard deviation over five independent runs.

Baseline Hyperparameter Tuning. For baseline methods, we perform systematic hyperparameter tuning to ensure a comprehensive and unbiased comparison. Hyperparameter optimization is conducted, with search spaces defined according to the recommended configurations in the original papers. The detailed hyperparameter ranges and tuning strategies for all baseline models are provided in our public repository. For additional implementation details and hyperparameter definitions, we refer readers to the corresponding original works.

Algorithm 3 Execution Engine: Pipeline

```
class OpenDDI_Pipeline:
    """
    Base class for implementing the unified OpenDDI benchmarking
        framework.
    """

    def __init__(self, args, logger):
        # Implementation omitted for brevity
        pass

    def run_benchmark(self):
        """Dataset processing -> Model initialization -> Pipeline
            execution."""
        pass

    def setup_data_manager(self):
        """
        Dynamically instantiates the specific dataset class (e.g., Graph/
            Sequence)
        and performs feature preprocessing (SMILES encoding, KG
            construction).
        """
        pass

    def setup_model_manager(self):
        """
        Implements the Model Factory pattern.
        """
        pass

    def dispatch_trainer(self, dataset, model):
        """
        Orchestrates the experimental workflow by selecting the
            specialized
        Trainer (e.g., MRCGNN_Trainer, ZeroDDI_Trainer) for the task.
        """
        pass

    def evaluate_performance(self):
        """Evaluates model performance using standard DDI metrics (AUC,
            AUPR, F1)."""
        pass
```

H Prompt for DDI Validation and Annotation

To validate and annotate drug–drug interactions during the LLM-based data augmentation process, we adopt a standardized prompt that guides the model to assess the existence of candidate DDIs based on known pharmacological mechanisms. The prompt is designed to elicit concise and deterministic responses, enabling scalable and consistent interaction validation.

> You are a pharmacology expert. Given the following drug pair:
> - Drug 1: {drug1}
> - Drug 2: {drug2}
>
> We want to determine if the following drug–drug interaction (DDI) exists based on known pharmacological mechanisms. {ddi_desc}
> Please confirm if this DDI exists by simply returning **yes** or **no**. Optionally, briefly explain why.

I LLM-BASED DATA AUGMENTATION

I.1 OUR THREE-STAGE DATA AUGMENTATION PIPELINE

Our augmentation process follows a rigorous three-stage pipeline that ensures data reliability through multiple safeguards:

Data Collection (Stage 1): All original data originates from authoritative pharmacological databases including DrugBank and KEGG. Curated labels from these sources are **preserved and never overwritten** by LLM outputs. This ensures that the foundation of our dataset remains reliable.

ML-Based Candidate Screening (Stage 2): Pre-trained DDI prediction models (specifically, MR-CGNN Xiong et al. (2023)) generate candidate DDIs to reduce the search space and improve the efficiency of the validation pipeline. This step leverages the predictive power of domain models for initial filtering.

Domain-Specific LLM Validation (Stage 3): Qwen3_Medical_GRPO, a medical domain-specialized large language model, validates the pharmacological validity of candidate DDIs. The model provides pharmacological reasoning with a **confidence threshold $\geq$ 0.7** to ensure high-quality outputs.

This dual-layer approach (ML screening + LLM validation) leverages both the **predictive power of domain models** for initial screening and the **rich pharmacological knowledge of LLMs** for rigorous validation.

I.2 QUALITY CONTROL MEASURES

To ensure the reliability of augmented data, we apply the following quality control measures:

Confidence threshold: Only LLM outputs with confidence ≥ 0.7 are retained. **Minimum record requirement:** Each DDI type must have at least 50 records to be included in the dataset. **Curated label preservation:** All curated labels from authoritative sources (e.g., DrugBank) are preserved without modification. **Augmentation scope:** Augmentation is applied only to expand coverage of drug and interaction pairs, not to replace or overwrite existing curated entries.

I.3 QUANTIFIABLE QUALITY METRICS

To provide transparent and objective quality assessment, we computed validation metrics by comparing LLM validation outputs against the original curated DDI labels from authoritative sources:

Table 5: LLM Validation Quality Metrics

Metric	Value
False Positive Rate (FPR)	7.6%
False Negative Rate (FNR)	11.7%

These error rates (FPR = 7.6%, FNR = 11.7%) are comparable to the inter-annotator disagreement rates commonly observed in manually curated biomedical datasets, providing empirical evidence that the quality of LLM-augmented labels is within an acceptable range for benchmark construction.

J MULTIMODAL ANALYSES DETAILS

We conduct additional multimodal analyses to examine the role of different drug modalities in DDI prediction under a unified experimental protocol. These analyses are designed to evaluate both the necessity and complementarity of multimodal representations, while remaining consistent with the evaluation settings used in the main experiments. These multimodal analyses are conducted across multiple DDI datasets with diverse data scales and modality coverage, providing a systematic evaluation of multimodal modeling strategies in DDI prediction.

J.1 Modality Ablation Setting

To assess the necessity of individual modalities in multimodal drug modeling, we perform controlled modality ablation experiments. Starting from the complete multimodal representation, we remove one modality at a time while keeping all other components unchanged. Models are trained and evaluated under the same data splits and optimization settings as in the main experiments to ensure fair comparison. Performance is reported using the same evaluation metrics as the corresponding prediction tasks.

J.2 Multimodal Combination Setting with SMILES

To evaluate the ability of non-SMILES modalities to complement structural information, we further conduct multimodal combination experiments by integrating SMILES with one additional modality at a time. In this setting, SMILES is treated as the core modality, and other modalities are incrementally incorporated to examine their supplementary effects. All models are trained and evaluated under identical experimental protocols, allowing for consistent comparison across different modality combinations.

K Model Performance Details

K.1 Overall Performance Comparison

The following table presents a comprehensive performance comparison of various DDI prediction methods across multiple benchmark datasets.

Table 6: Overall performance comparison of DDI prediction methods across multiple datasets

Method	ODDI-B	ChCh-Miner	zhangddi	ODDI-M	Dengs	Ryus	ZS-DDIE	ODDI-ML	Twosides
MVA-DDI	97.1	91.3	74.0	67.4	56.6	83.6	72.6	97.8	77.5
MMDGDTI	97.0	92.1	71.3	68.9	66.2	74.6	58.2	97.6	78.2
GoGNN	**97.5**	92.8	75.2	74.7	74.0	88.4	78.0	98.0	78.9
ExDDI	97.4	**93.2**	75.4	62.3	62.0	80.3	65.3	97.9	78.8
ConvLSTM	97.0	85.6	74.7	56.9	58.4	71.9	58.7	97.7	75.3
MUFFIN	**97.7**	92.6	70.2	80.2	52.3	88.4	77.9	**98.0**	73.1
MKGFENN	**97.7**	93.1	73.1	87.0	**72.5**	**90.3**	**82.8**	97.8	78.4
DDIMDL	96.8	91.2	74.3	56.8	57.7	79.4	74.0	97.2	79.1
TIGER	96.4	87.6	72.5	62.8	63.2	84.6	56.2	97.7	77.9
KGNN	94.2	87.0	74.7	71.5	56.2	84.9	80.1	97.8	**80.2**
MRCGNN	**97.5**	92.6	**76.0**	81.6	71.5	**92.6**	79.9	**98.0**	79.9
SumGNN	94.8	85.3	75.1	70.2	72.6	89.7	83.0	97.8	80.8
LAGAT	94.8	85.2	**75.3**	73.4	64.3	90.5	**83.5**	**97.9**	**81.9**
CASTER	96.5	93.6	75.2	77.2	67.7	84.4	66.8	97.8	71.8
DDKG	**98.0**	**93.7**	69.2	86.1	72.2	90.5	81.2	97.7	79.0
DeepDDI	94.6	90.9	74.6	53.3	68.9	78.3	67.5	97.7	79.4
MIRACLE	95.7	91.5	68.9	**85.3**	69.1	87.1	79.6	97.0	78.4
PHGLDDI	95.7	**93.4**	70.8	66.1	58.8	74.6	60.9	96.1	70.0
DSNDDI	97.4	92.0	73.7	71.2	52.6	86.5	74.9	98.1	79.7
ZeroDDI	96.3	87.3	69.4	61.6	56.8	79.8	72.3	96.3	74.5

K.2 Binary classification Experiment

We first evaluate model performance on the binary DDI prediction task, where the objective is to determine whether a given pair of drugs exhibits any interaction. This setting reflects a fundamental and widely studied formulation of DDI prediction, and serves as a basis for assessing the overall discriminative capability of different methods.

All models are trained to predict the presence or absence of interactions using the same training–validation–test splits to ensure fair comparison. Performance is assessed using standard binary classification metrics.

Experiments are conducted across multiple DDI datasets with diverse scales and interaction densities. Results demonstrate consistent trends across datasets, enabling a comprehensive comparison of model effectiveness in binary DDI prediction.

Table 7: Performance comparison of methods on ODDI-B and ChCh-Miner datasets with SMILES and MM (multimodal) features. The "impr." column shows the relative performance improvement of MM over SMILES.

Method	ODDI-B			ChCh-Miner		
	SMILES	MM	impr. (%)	SMILES	MM	impr. (%)
DeepDDI	90.7 ±0.9	94.6 ±0.7	**4.4** ↑	87.0 ±1.1	90.9 ±0.8	**4.5** ↑
CASTER	92.5 ±0.7	96.5 ±0.5	**4.3** ↑	90.0 ±0.8	93.6 ±0.6	**4.0** ↑
GoGNN	97.1 ±0.5	97.5 ±0.4	**0.4** ↑	89.9 ±0.8	92.8 ±0.7	**3.2** ↑
MRCGNN	97.2 ±0.5	97.5 ±0.5	**0.3** ↑	92.0 ±0.7	92.6 ±0.7	**0.7** ↑
LAGAT	94.4 ±0.8	94.8 ±0.7	**0.4** ↑	85.0 ±1.1	85.2 ±1.0	**0.2** ↑
DDKG	97.9 ±0.5	98.0 ±0.4	**0.1** ↑	92.4 ±0.7	93.7 ±0.6	**1.4** ↑
KGNN	93.9 ±0.8	94.2 ±0.8	**0.3** ↑	85.0 ±1.0	87.0 ±1.0	**2.3** ↑
DDIMDL	93.3 ±0.8	96.8 ±0.6	**3.8** ↑	88.4 ±1.0	91.2 ±0.9	**3.2** ↑
TIGER	94.3 ±0.7	96.4 ±0.6	**2.2** ↑	86.2 ±1.1	87.6 ±1.0	**1.6** ↑
MVA-DDI	92.4 ±0.8	97.1 ±0.5	**5.1** ↑	90.1 ±0.9	91.3 ±0.9	**1.3** ↑
MKGFENN	97.6 ±0.5	97.7 ±0.4	**0.1** ↑	92.8 ±0.7	93.1 ±0.6	**0.3** ↑
MUFFIN	92.6 ±0.8	97.7 ±0.5	**5.5** ↑	90.9 ±0.9	92.6 ±0.8	**1.9** ↑

K.3 Multiclass classification Experiment

We next evaluate model performance on the multiclass DDI prediction task, where the objective is to identify the specific interaction type associated with a given interacting drug pair. Compared with binary prediction, this setting requires models to capture finer-grained semantic distinctions among multiple interaction categories.

All models are trained to predict interaction types using the same training–validation–test splits to ensure fair comparison. Performance is assessed using standard multiclass classification metrics.

Experiments are conducted across multiple DDI datasets with diverse interaction type distributions and data scales. Results exhibit consistent patterns across datasets, enabling a comprehensive comparison of model effectiveness in multiclass DDI prediction.

K.4 Multilabel classification Experiment

We further evaluate model performance on the multilabel DDI prediction task, where each drug pair may be associated with multiple interaction types simultaneously. This setting reflects real-world clinical scenarios in which co-administered drugs can induce multiple pharmacological effects.

All models are trained to predict a set of interaction labels using the same training–validation–test splits to ensure fair comparison. Performance is assessed using standard multilabel classification metrics.

Experiments are conducted across multiple DDI datasets with diverse label co-occurrence patterns and data scales. Results demonstrate consistent trends across datasets, enabling a comprehensive comparison of model effectiveness in multilabel DDI prediction.

L Practical Generalization Analyses

We conduct practical generalization analyses to evaluate model performance in realistic scenarios where test drug pairs involve previously unseen drugs. This setting reflects real-world DDI prediction challenges, as newly approved or rarely used drugs are often absent from historical interaction data.

All models are trained using the same training–validation splits, while the test set is constructed to include drug pairs containing unseen drugs. Performance is assessed using the same evaluation metrics as in the corresponding prediction tasks to ensure consistency and fair comparison.

Table 8: Performance comparison of methods on ODDI-ML and Twosides datasets with SMILES and MM (multimodal) features. The "impr." column shows the relative performance improvement of MM over SMILES.

Method	ODDI-ML			Twosides		
	SMILES	MM	impr. (%)	SMILES	MM	impr. (%)
DeepDDI	95.9 ±0.8	97.7 ±0.6	**1.9** ↑	75.1 ±1.4	79.4 ±1.2	**5.7** ↑
CASTER	96.9 ±0.7	97.8 ±0.5	**0.9** ↑	70.6 ±1.4	71.8 ±1.3	**1.7** ↑
GoGNN	97.3 ±0.6	98.0 ±0.4	**0.7** ↑	77.2 ±1.3	78.9 ±1.1	**2.2** ↑
MRCGNN	97.3 ±0.6	98.0 ±0.5	**0.7** ↑	78.2 ±1.2	79.9 ±1.0	**2.2** ↑
LAGAT	96.9 ±0.7	97.9 ±0.6	**1.0** ↑	79.7 ±1.3	81.9 ±1.1	**2.8** ↑
DDKG	97.6 ±0.6	97.7 ±0.5	**0.1** ↑	77.9 ±1.2	79.0 ±1.1	**1.4** ↑
KGNN	96.8 ±0.7	97.8 ±0.6	**1.0** ↑	77.9 ±1.4	80.2 ±1.2	**3.0** ↑
DDIMDL	97.0 ±0.8	97.2 ±0.7	**0.2** ↑	78.1 ±1.3	79.1 ±1.2	**1.3** ↑
TIGER	96.9 ±0.7	97.7 ±0.6	**0.8** ↑	76.6 ±1.4	77.9 ±1.3	**1.7** ↑
MVA-DDI	97.5 ±0.6	97.8 ±0.5	**0.3** ↑	77.3 ±1.4	77.5 ±1.3	**0.3** ↑
MKGFENN	97.3 ±0.6	97.8 ±0.5	**0.5** ↑	76.4 ±1.3	78.4 ±1.1	**2.6** ↑
MUFFIN	95.0 ±0.8	98.0 ±0.4	**3.2** ↑	62.2 ±1.7	73.1 ±1.4	**17.5** ↑

Experiments are conducted across multiple DDI datasets with diverse data scales and drug coverage. The results provide a comprehensive assessment of the generalization ability of different methods under unseen-drug settings, highlighting their robustness and practical applicability in real-world DDI prediction scenarios.

M Practical Efficiency Analyses

We conduct practical efficiency analyses to evaluate the computational cost of different DDI prediction methods in terms of both time and memory consumption. These analyses aim to assess the feasibility of deploying the models in large-scale and resource-constrained settings.

All methods are evaluated under the same experimental environment to ensure fair comparison. We report training time and peak memory usage as primary efficiency metrics, reflecting computational overhead and resource demand during model training.

Experiments are performed on multiple DDI datasets with varying scales and graph densities. The results provide a comprehensive comparison of the practical efficiency of different methods, highlighting their scalability and suitability for real-world DDI prediction applications.

N Robustness Analyses Details

To evaluate the robustness and stability of DDI prediction models under varying interaction sparsity, we conduct a series of controlled experiments by progressively adjusting the proportion of observed drug–drug interactions. Specifically, we vary the interaction removal ratio from 0 to 0.8, where a ratio of 0 corresponds to the original, fully observed DDI graph, and larger ratios simulate increasingly incomplete or perturbed interaction scenarios. All models are trained and evaluated under identical experimental protocols to ensure fair comparison. We report results on Ryus and Dengs, providing a systematic evaluation of model performance across diverse data scales and graph characteristics. This experimental design enables a thorough analysis of the robustness and generalization ability of DDI prediction methods under varying levels of interaction incompleteness.