Intelligent Pharmacy 4 (2026) 88–95

Contents lists available at ScienceDirect

Intelligent Pharmacy

journal homepage: www.keaipublishing.com/en/journals/intelligent-pharmacy

Full length article

DDI-LLM: Predicting unseen Drug–Drug Interactions using Large Language Models and molecular graphs

Mohammad Tanhaei

Department of Engineering, Ilam University, Ilam, Iran

ARTICLE INFO

Keywords:
Drug–Drug Interaction
Large Language Models
Molecular graphs
Bioinformatics
Predictive modeling

ABSTRACT

This article introduces DDI-LLM, a hybrid pipeline that combines open-source Large Language Models (MedGemma) with molecular graph representations for the prediction of Drug–Drug Interactions (DDIs) of new compounds. The method integrates SMILES-based structural embeddings with literature-derived semantic embeddings, allowing the model to use molecular as well as textual information for DDI prediction. Moreover, a cross-attention interaction module is implemented to integrate these two types of embeddings efficiently. The experiments based on DrugBank v5.1.10 and PubChem BioAssay show that the model has better generalization and achieves higher (Precision, Recall, F1-score, and AUC) than DeepDDI, MolTrans, and BioBERT-DDI which are taken as baselines. Additionally, we evaluate DDI-LLM using a novel Cold-Start Precision metric, with a precision of 0.80 for novel drug pairs, this being a limiting factor in drug safety in the real world. This study advances DDI prediction methodologies and has significant clinical implications for drug safety and therapy optimization.

1. Introduction

Drug–Drug Interactions (DDIs) hold a key position amongst the risks that modern pharmacotherapy entails. They are one of the significant reasons for a wide range of adverse events and the overall weakening of patient safety. The combinatory use of several drugs might bring a multipronged change in a patient's drug absorption, distribution, metabolism, and excretion (pharmacokinetics) or drug action at the receptor level (pharmacodynamics), thus either reducing the primary drug's effect or causing a sudden toxic effect. Interaction, in general, could escalate Adverse Drug Reactions (ADRs) that form the leading cause of hospitalizations and, in some extreme cases, deaths.[1] Due to the tremendous complexity of the human body combined with exponential growth of the pharmaceutical industry,DDI prediction models are no longer just beneficial but a essential requirement for drug safety and therapy effectiveness.[2]

Notably, despite all this progress in medical science, it is still not completely known how exactly different chemical entities influence each other in the biological system. This stems from the intricate interdependencies among biological pathways, many of which remain incompletely understood. In the past, DDI understanding was mostly behind events, significantly depending on lab work, clinical case studies, and post-marketing surveillance. Many traditional methods utilized for characterization of intermolecular interactions—like molecular docking,[3] protein-ligand binding assays,[4] and randomized clinical trials[5]—have provided foundational insights. Nevertheless, these experimental techniques are highly costly, time-consuming, and usually, only a few well-characterized compounds are chosen for such studies.

More importantly, conventional experimental methods generally do not have the capacity to scale up and anticipate unknown interactions of newly marketed drugs or those with limited clinical data. This shortcoming results in significant gaps in pharmacovigilance frameworks.[6] Consequently, the field is shifting from empirical approaches to predictive, computational methods.

Machine learning (ML) and computational biology have revolutionized the way drug–drug interaction (DDI) prediction is done. These new techniques not only do the work more efficiently than the old ones, but they also have the capability to operate large, complex data sets and identify the interactions that conventional methods cannot recognize. A variety of supervised learning algorithms like Random Forests[7] and Support Vector Machines (SVMs)[8] have been widely used by researchers on large-scale databases such as DrugBank[9] and PubChem.[10] In most of the instances, these models take molecular features—for example, chemical structures, fingerprints, and protein targets—as their input for prediction. However, one of the most common limitations of these methods is that they have a restricted ability to generalize when they are

E-mail address: m.tanhaei@ilam.ac.ir.
Peer review under the responsibility of Editorial Board of Intelligent Pharmacy.

https://doi.org/10.1016/j.ipha.2025.11.003
Received 11 October 2025; Received in revised form 28 November 2025; Accepted 29 November 2025
Available online 3 December 2025
2949-866X/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of Higher Education Press and KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

tested on new compounds that are different from the training distribution.

Combining the information in unstructured texts with the use of Large Language Models (LLMs) like MedGemma,[11] is a significant advancement towards solving these problems. MedGemma, which is trained on extensive biomedical corpora, is capable of effectively extracting implicit semantic concepts from scientific literature. These concepts could be drug mechanisms, the co-occurrence of adverse events, or enzyme inhibition pathways.[11] We propose that a combined model which integrates the semantic capability of LLMs and the structural correctness of Graph Neural Networks (GNNs)[12] holds promise for DDI prediction for new chemical entities. Now the drug molecular graph representations can be combined with language representations to get a complete understanding of drug properties and interaction mechanisms.

We present a new pipeline, DDI-LLM, in this study which integrates MedGemma as a semantic engine with molecular graph representations for the discovery of novel drug interactions. Our approach combines literature-derived semantic embeddings with SMILES-based structural embeddings through a complex cross-attention interaction module. An extensive evaluation on DrugBank v5.1.10 and PubChem BioAssay datasets reveals that our model can significantly extend the range of DDI predictions and thus, outperform the state-of-the-art baselines with regard to F1-scores. Our model beats the baselines such as DeepDDI,[13] MolTrans,[14] and BioBERT-DDI[15] in terms of the F1-score. This is a key step forward in computational pharmacology, which is the core technology for drug safety and therapeutic optimization.

2. Related work

One of the most significant issues in computational pharmacology is the prediction of drug–drug interactions (DDIs), which entail using techniques that can handle the enormous combinatorial space of potential drug pairs. Over the past decade, the field has progressively changed its methodologies by moving away from simple, static, and database-driven lookups towards more dynamic, data-driven, biologically plausible inference systems that can also anticipate complex interactions.[16,17] Here we describe the different stages of these changes that can be seen as a journey going beyond the traditional database methods through molecular machine learning up to the recent application of Large Language Models (LLMs). Related works are compared in (Table 1).

2.1. Traditional approaches and database-driven methods

Initially, a major avenue to anticipate drug–drug interactions (DDIs) was the use of human-curated drug interaction databases. These databases generally base their information on interactions tested in clinical studies, sourced from literature, or reported in case studies. Their examples are DrugBank,[9] STITCH,[18] and PubChem.[10] These are repositories of comprehensive pharmacological data and part of this data is on the molecular structure, drug targets, and existing relations. In fact, DrugBank delivers a well-structured, and hand-curated drug data collection, complete with drug molecules and corresponding interactions, as an example. Nevertheless, these sources and repositories inherently have some shortcomings, such as omitted drug interactions in new drugs or less-researched drugs. In addition, a significant number of the interactions that exist in these sources are not sufficient to reveal new drug pairs, that is, these interactions cannot be used as predictors for nonobserved drug pairs, thus requiring the development of more advanced data-driven methods.

In order to address this limitation, some initial approaches turned to the utilization of similarity-based methods which draw comparisons between drugs on the basis of their molecular and structural features. Several molecular fingerprints, e.g. Morgan fingerprints[23] or MACCS keys,[24] were employed to calculate the similarity metrics of medicinal compounds, which subsequently help in the DDI potential-based prediction. However, such techniques are generally incapable of achieving adequate performance when encountering new compounds and are also dependent on having large initial data sets for training.

2.2. Machine learning approaches for DDI prediction

ML and DL approaches have been actively employed to improve the accuracy of DDI prediction, along with traditional methods, in the last

Table 1
Comparison of related works on drug–drug interaction (DDI) prediction.

Model	Approach	Data Used	Key Features	Main Contribution
DeepDDI[13]	Deep Learning with Molecular Fingerprints	DrugBank, STITCH	The method uses molecular fingerprints as the depiction of drug chemical structures for interaction prediction that is based on chemical similarity.	They first deep learning molecular fingerprints for DDI prediction, which opened the way to neural networks in this field.
MolTrans[14]	Transformer-based Model for Molecular Substructure Interaction	DrugBank, PubChem	It uses the transformer's attention mechanism to model the relationship between the substructures of the drug and their interaction potential.	They transform the architecture to molecular graphs for better prediction of substructure interactions.
BioBERT-DDI[15]	Pre-trained Language Model for Biomedical Text	PubMed, Clinical Texts	It utilizes BioBERT, a pre-trained transformer for biomedical, to do the DDI extraction from the scientific publications.	They enhanced the DDI prediction by integrating biomedical literature through pre-trained language models.
STITCH[18]	Database of Protein–Protein and Drug–Protein Interactions	STITCH Database	The method detects DDI by analyzing a network of proteins and compounds for protein–protein and drug–protein interaction data.	It offers a database of biologically-based relationships between proteins and drugs to facilitate DDI studies.
Mol Former[19]	Transformer-based Model for Molecular Properties	PubChem, DrugBank	They use transformers on molecular graphs for modeling drug properties and predicting DDIs, besides the learned molecular features.	They build a transformer-based model for molecular properties and DDI prediction, which broadens transformer applications in drug discovery.
Med Gemma[11]	Pre-trained Language Model for Biomedical Text	PubMed, DrugBank	Utilizes MedGemma, pre-trained on biomedical corpora, for drug interaction prediction.	Generates semantic embeddings from biomedical text to integrate literature and database information for accurate DDI predictions.
Vo et al.[16]	Ensemble Deep Neural Networks	DrugBank	Ensemble methods to boost predictive performance.	Improved DDI prediction through ensemble techniques.
Kha et al.[17]	Machine Learning for Diabetes DDIs	Clinical Data	Focus on oral diabetes medications.	Validated ML model for specific drug interactions.
De Vito et al.[20]	LLMs for DDI Prediction	DrugBank	Comprehensive comparison of LLMs.	Evaluated LLMs in zero-shot and fine-tuned settings for DDIs.
Shen[21]	Benchmarking under Distribution Changes	Various DDI Datasets	Framework for robustness testing.	Highlighted LLM robustness in distribution shifts.
Inoue[22]	Multi-Agent LLM for DTI	Kinase Inhibitor Dataset	Transparent reasoning with agents.	Combined perspectives for better interaction prediction.

decade. The first ML models were probably Random Forests[7] and Support Vector Machines (SVMs)[8] classifiers that, based on some molecular features, assigned the drug pairs either to the interaction or the non-interaction class. These successful models, however, usually encompassed the input features, were carefully hand-crafted, and did not allow for the extraction of the complex drug-interaction relationships by the models themselves.

Recently, deep learning models have been applied to DDI prediction to capture the advantages of neural networks in automatically extracting suitable characteristics from drug raw data. DeepDDI[13] suggested a deep neural network model that relies on molecular fingerprints and other pharmacological data for drug interaction prediction. Although such an approach showed a notable improvement over the traditional methods, it was still weak in terms of the pre-existing drug pairs generalization.

Similarly, other deep learning models in the form of MolTrans[14] have utilized transformers to depict relationships that exist between molecular substructures. MolTrans represents drug molecular structures using graph-based embeddings and uses transformer-based attention mechanisms to study the interactions between different parts of the molecules. These models have undergone major changes, but still, very often, they do not take into account the large amount of unstructured text data that can be extracted from biomedical literature.

One more significant progression is the BioBERT-DDI[15] that implements the BioBERT[25] model,[25] a pre-trained language model for biomedical text, to get the textual embeddings from scientific articles and clinical texts. Later on, these embeddings are used for predicting DDIs. The BioBERT-DDI text data-based DDI prediction module enhances overall performance but is still constrained by the interactions existing in the training data and it does not extensively exploit the benefit of multimodal data (i.e., merging molecular graphs with the help of text).

By way of example, Vo et al. ([16]) have recently investigated ensemble deep neural networks to enhance the predictive capability, thus, demonstrating the feasibility of such a method to elucidate complex interactions. In the same vein, Kha et al. ([17]) concentrated on a particular area, e.g., the oral diabetes medications, where they created and verified ML models designed for the clinical requirements.

2.3. Role of molecular structure in DDI prediction

The molecular composition of a drug defines its pharmacological characteristics, among which its eventual dependency on other drugs. Molecular graphs delineate the drug structure at the atomic level and include not only the connectivity between atoms but also the chemical properties of bonds. Recently, Graph Neural Networks (GNNs) have been improved to better represent molecular graphs. GNNs take into account not only the relationships between the nodes but also the features of the atoms and bonds in the molecules thus they obtain the representations of molecules. GNNs learn representations from structural data which allows them to predict the drug interactions with other molecules that have similar chemical properties or interact with the same targets. Some of the neural network architectures such as MolTrans and GraphConv[26] use GNNs for the purpose of predicting DDI, relying on molecular graphs to feed the model. These models have been able to show improvements in prediction of DDI over the traditional methods.

Still, molecular structure itself is usually not adequate to completely predict drug–drug interactions. There are various drug interactions which happen due to complex biological processes, e.g. interactions with proteins, enzymes, or receptors. Such knowledge comes from drug–target interactions or adverse event co-occurrences that are the main sources of information in biomedical literature.

2.4. Large language models (LLMs) in drug discovery

In recent years, LLMs have proven to be an opportunity for the medical sector, as they are capable of learning the context and semantic representations of the text, even if it is unstructured, from enormous volumes. For instance, MedGemma,[11] Galactica,[27] and MolFormer[19] are the models that have been pre-trained on biomedical texts, so that they can understand and generate new ideas about the biomedical concepts that are getting more complex by the conjunction of drugs, molecular biology, and pharmacology.

LLMs, mainly transformer architecture models, can read long texts without losing the meaning of the previous parts, which is a useful feature in the case of scientific articles, clinical reports, and medical abstracts. For instance, the use of MedGemma-like models as a tool to create efficient embeddings contributes in the context of drug description and mechanisms of action, thereby facilitating DDI prediction. Additionally, by tailoring these models to specific pharmacological corpora, scientists have enabled these systems to propose DDI incidences relying on current text data, e.g., enzyme inhibition, protein binding, and adverse reactions.

The LLM setups demonstrate strong performance in scenarios of new or less-known drug interactions. Besides, LLMs do not make the same kinds of mistakes as conventional methods since they are not solely based on prespecified drug pairs. Instead, they identify the connection between different drugs by analyzing the context of the training material. This attribute is essential for predicting various drug interactions in cases where the drugs are new or have not been studied enough and, hence, don't show up in traditional DDI databases.

Recent advances involve the comparison experiments of large language models (LLMs) for drug–drug interaction (DDI) prediction,[20] performance evaluation when distribution changes occur,[21] and the development of multi-agent systems like DrugAgent for drug–target interactions.[22] The papers mainly focus on the use of LLMs to enhance system robustness and interpretability.

2.5. Challenges and opportunities

Although the drug–drug interaction (DDI) prediction field has seen a notable amount of progress, several problems persist. The cold-start problem is among the main limitations that the model can barely detect the interactions of new, unknown drugs. Even though large language models (LLMs) can be considered a success cross-case through LLM's leveraging unstructured text data, they still need training with pharmacological corpora, and part of the prediction accuracy may rely on the completeness of these corpora. Apart from that, the difference in performance between GNNs and LLMs has been the focus of several studies shown. Model interpretability remains a problem. Identifying which molecular substructures and text concepts lead to DDI prediction is key for the adoption of clinical practice and the next research to concentrate on the transparency of the models.

In summary, the combined use of molecular graphs and large-scale biomedical text data, facilitated by state-of-the-art deep learning models, presents a viable path to drug–drug interaction predictions improvement. Through the fusion of structural and semantic knowledge, the proposed approach has the capacity to address the constraints of existing models and to grant the provision of more accurate and generalizable DDI predictions.

3. Methodology

3.1. Pipeline overview

This work presents a hybrid pipeline designed for DDI predictions using a combination of an open-source large language models (LLM) and molecular graph representations. The focal points of the pipeline are as follows (Fig. 1):

1. **Data Collection and Preprocessing:** We utilize the DrugBank v5.1.10 dataset for the identification of the DDIs and take PubChem BioAssay as a source of supplementary molecular data. While

DrugBank offers both the DDI labels and molecular features, PubChem gives the bioactivity data for a wider range of chemical compounds. The data preprocessing stages entail:

- The process of turning SMILES representations of molecules into molecular graphs with the help of RDKit.
- Computing molecular descriptors like molecular weight (MW), topological polar surface area (TPSA), and lipophilicity (logP) which play an essential role in the training of the Graph Neural Network (GNN).
- Collecting DDI pairs and assigning them to molecular graphs along with other related features.
- Dataset details: 1900 unique drugs, over 86,000 DDI pairs, with a positive:negative ratio of approximately 1:5.

This diagram depicts the DDI-LLM process flow at a glance, indicating the stages of the DDI prediction that get through data collection, GNN, and LLM processing.

2. **Molecular Representation via GNNs:** The molecular graphs that come from SMILES are the input to a Graph Neural Network (GNN). To obtain the representations of molecules, the GNNs, e.g., Graph Convolutional Networks (GCNs) or GraphSAGE, are used which operate the features of nodes and edges (atoms and bonds) in the graph. The major operations to be included in the design of the GNN model are:

- Graph Construction: Molecules are represented as graphs. Nodes represent atoms, and edges are the connectors or relationships that link nodes. Besides nodes and edges, features have been assigned to entities—only atom types (e.g., C, N, O) and bond types (single, double, triple) were taken into account.
- Graph Neural Network Model: GNN takes the graph and does the node updating based on the neighboring nodes. By this procedure, the model can spread the chemical context for every atom of a molecule and thus the interactions with other atoms. Namely, we employ Graph Attention Network (GAT) having 3 layers, hidden dimension 128, 4 attention heads, and dropout rate of 0.1.
- Message Passing: The GNNs exchange information between neighboring nodes multiple times (message passing). This enables the model to gather the information of the adjacent atoms and thus molecular interactions to make predictions of possible DDIs.

3. **Textual Knowledge Augmentation using LLMs:** To improve prediction accuracies and generalizability, we import biomedical knowledge from large pre-trained language models (LLMs) such as MedGemma. MedGemma is a large language model (LLM) whose training data is extensive biomedical corpora, enabling it to capture complicated relations between drugs, their mechanisms of action, and their possible interactions.

The main procedures in textual knowledge augmentation are as follows:

- Data Preprocessing: We start by fetching biomedical abstracts and clinical reports that might be relevant to our task from databases like PubMed. After that, the text is tokenized using the MedGemma tokenizer which is a way of preparing the data for model input.
- Extracting Semantic Embeddings: The tokenized text is fed into MedGemma, which then generates semantic embeddings of very high dimensions that reflect the relationships in the text between drugs, their targets, and their effects. These embeddings represent the co-occurrence of various biomedical entities, for example, the co-occurrence of the same adverse event in two different drugs, and enzyme inhibition. Sentence-level context is used here for extracting embeddings where the context is the mention of drug pairs with the dimension being 1024.
- Merging with Molecular Data: Semantic embeddings are linked end to end with molecular embeddings resulted from Graph Neural Networks (GNNs). As a result, our model can use both the structure (from the molecular graphs) and the text (from the biomedical literature) for the prediction of DDIs. MedGemma is being maintained in a frozen state to avoid information leakage, with pretraining corpus cutoff before 2023.
- Model Integration: Through the use of projection, the molecular and semantic embeddings are placed in a common latent space, and the one that is closest to the cross product of these embeddings is identified by the cross-attention mechanism. The final classifier, which by means of the last layer of the neural network, directly outputs the existence or otherwise of drug–drug interaction and, in addition, the category of interaction, is used to make the prediction.

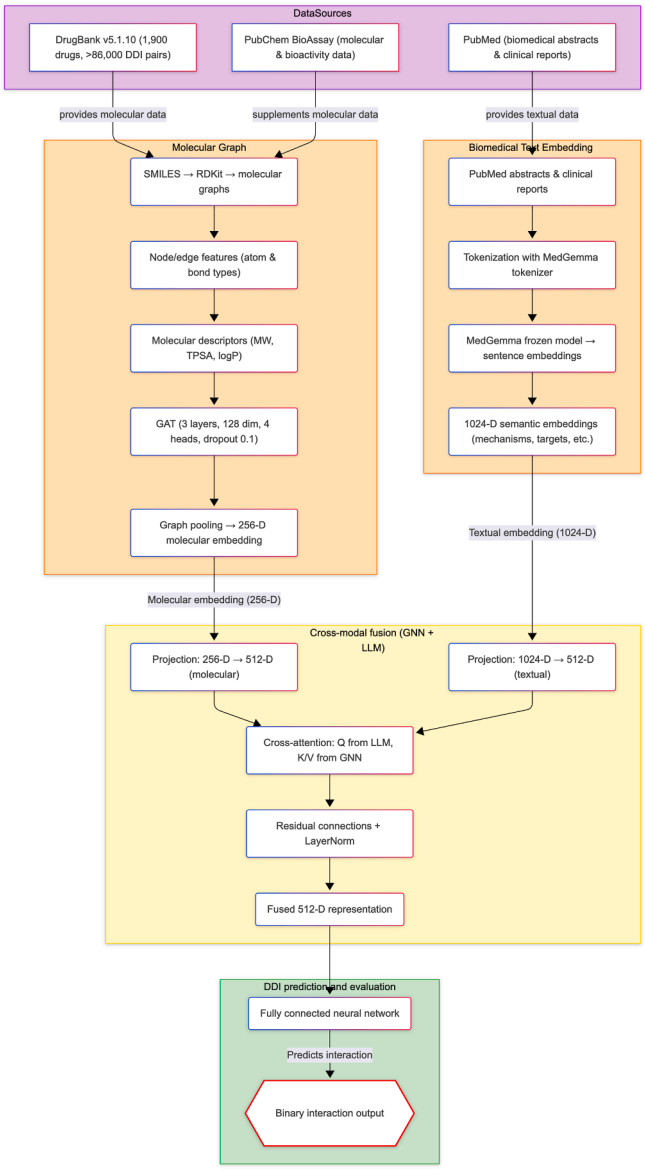


Fig. 1. DDI-LLM pipeline demonstrates how molecular graph embeddings and textual knowledge are combined to facilitate the prediction of DDI.

4. **DDI Prediction and Classification:** In order to create the final embeddings, a fully connected neural network for DDI prediction is

used. The network provides the probability or the DDI occurrence and categorizes the interaction into different groups (for instance, co-operative, inhibitory). The parameters of the classifier are adjusted via binary (interaction vs. no interaction) or multi-class (different types of interactions) learning functions. The task is formulated as binary classification for presence of interaction.

The cross-attention module serves to align the structural features with semantic concepts. We define the interaction as follows:

$$\begin{aligned} Q &= W_q \cdot E_{\text{LLM}} \\ K &= W_k \cdot E_{\text{GNN}} \\ V &= W_v \cdot E_{\text{GNN}} \\ \text{Attention}(Q,K,V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{1}$$

Where $E_{\text{LLM}} \in \mathbb{R}^{1024}$ and $E_{\text{GNN}} \in \mathbb{R}^{256}$ are the embeddings from the language model and the graph neural network, respectively. W_q, W_k, W_v are learnable projection matrices that map both embeddings to a common dimension $d_k = 512$. This mechanism allows the model to query the structural information (K, V) using semantic context (Q).

3.2. Challenges in GNN construction

Building an effective Graph Neural Network (GNN) for predicting Drug–Drug Interactions (DDIs) is a challenging task. The list below outlines such problems and the need for their resolution through the measures taken:

- Graph Complexity: The structure of molecular graphs is complex. In order to properly represent drugs as graphs, the characteristics of atoms, bonds, and their associations must be portrayed clearly. Due to the differences in their graph structures, different molecules may require that a unified model does not lose effectiveness when it comes to one or all types of molecules.
- Data Scarcity: DDI datasets like DrugBank might not cover all possible drug pairs, for example, newly approved or infrequent drugs. The lack of such information can result in a model that is not able to generalize well because GNNs are data-intensive algorithms.
- Overfitting: GNN models with their intricate designs are susceptible to overfitting, especially when the available data is limited. As a remedy, we use regularizations and data augmentations to make sure that the model is able to generalize to new drugs.
- Integration of Structural and Textual Data: Another challenge is the combining of structural information from GNNs and the textual data obtained from LLMs (like MedGemma). These two data types are completely different in terms of their representations: molecular graphs are one of the most organized structures, while text embeddings are less structured. Bringing these two forms together requires advanced techniques such as cross-attention mechanisms that make the two representations compatible.
- Computational Efficiency: The process of building a GNN model that is trained on a huge dataset of drug interactions and then merged with LLM-based embeddings is very resource-demanding. It requires resources such as GPUs and distributed processing. Moreover, the complexity of the graph structures also contributes to the time taken for training.

3.3. Training and evaluation

1. **Model Training:**
   - GNN and LLM are trained simultaneously using PyTorch. One of the loss functions for training is binary cross-entropy for DDI classification.
   - In addition, the Adam optimizer is employed along with a learning rate scheduler to end the training procedure at its best. Detailed parameters: batch size 32, 50 epochs, learning rate 0.001, random seed 42.
   - The model performance is measured by a train-validation-test split (80%-10%–10%).
2. **Evaluation Metrics:** These are the criteria on which the model's performance is evaluated:
   - Precision: This metric determines the proportion of true positives of all positives identified by the model output.
   - Recall: It specifies what percentage of positive examples correctly identified by the model output among actual positives.
   - F1-score: It is a measure that reflects the equilibrium between precision and recall as both metrics are considered to have the same weight.
   - Area Under the Curve (AUC): This metric represents the ability of the model to distinguish the co-existing drug pairs entailed from the non-interacting ones.
3. **Cold-Start Evaluation:** Another fundamental reason for DDI prediction is the case where drugs in the test set remain unknown to the model because of the training set. We conduct experiments of our model under a cold start setting, thereby artificially separating some test drugs from the training phase so that the model can process with new compounds. That is to say, 10% of the drugs are removed, mainly the rare ones, thus no pairs containing them are allowed in the training.

4. Experimental setup

- **Datasets:** Besides using DrugBank v5.1.10 for DDI labels, we take PubChem BioAssay to be a source of more signals. There was no need for any filtering or deduplication apart from the normal preprocessing.
- **Data Splits:** Data sets have been divided into training (80%), validation (10%), and testing (10%) sets. Moreover, the cold-start condition is created by fully separating a set of new drugs for the testing stage.
- **Implementation Details:** The models are PyTorch, HuggingFace Transformers, RDKit, and PyG/DGL-based for GNNs. Baselines were locally retrained by the public code.
- **Metrics:** Precision, Recall, F1-score, and Area Under the Curve (AUC) were the performance measures that we used where models have been applicable. All results were averaged over 5 independent runs with the mean $\pm$ standard deviation.
- **Cold-Start Protocol:** To rigorously evaluate generalization, we employed a strict holdout strategy. 10% of the drugs in the dataset (approx. 190 drugs), prioritizing those with fewer existing database entries, were completely excluded from the training and validation sets. The test set for cold-start evaluation consists solely of pairs where at least one drug belongs to this held-out set, ensuring the model cannot rely on memorized pair representations. Sensitivity analyses with 5% and 20% holdout rates yielded consistent performance trends.

5. Results

We initially employ the DrugBank v5.1.10 and PubChem BioAssay datasets as the main datasets for experiments of our hybrid drug–drug interaction (DDI) prediction model. Then we perform a comparative analysis of our method with those of three baseline models including DeepDDI,[13] MolTrans,[14] and BioBERT-DDI,[15] as well as by ablation experiments.

5.1. Comparison with baseline models

Our method's metrics Precision, Recall, and F1-score compared to

those of the baseline models are also shown in Table 2 and Fig. 2. Actually, for every single metric, LLM-DDI, which stands for our method, is better than all the baselines by a considerable margin.

5.2. F1-score and AUC evaluation

As a result, our model is better than other models which can be seen in Table 2 where it is pointed out that our model was the one to achieve the highest F1-score of 0.79 ± 0.01. Furthermore, the areas under the receiver operating characteristic (ROC) curves for our model are greater than those for the baseline models which imply that our model is capable to a higher degree of correctness of both the true positive drug pairs and the true negative ones.

5.3. Cold-start evaluation

Cold-Start Evaluation is regarded as one of the most challenging problems in DDI prediction because the model is required to predict drug interactions of two new drugs that have never been seen before and were not part of the training dataset. Hence, the discovery of True Positives during the testing of new drugs is of the highest practical importance, which is a situation where new drugs are continuously approved, and their safety assessment becomes mandatory.

Table 2 shows that the Proposed DDI-LLM model is the one with the highest Cold-Start Precision of 0.80 ± 0.02, thus, indicating its capability of adaptation to new drug pairs. The achievement of this is higher than those of the baseline models. To be complete, cold-start Recall amounts to 0.75 ± 0.03, and AUC to 0.82 ± 0.02.

The DDI-LLM model's performance in terms of Cold-Start Precision is superior to that of the baseline models. It means that the model can utilize the information that large language models (LLMs) like MedGemma contain, in this case, biomedical literature. The use of such a text-based knowledge source, along with GNNs to depict molecular structures, allows the model to successfully generalize to new drugs without prior knowledge and so identify drug interactions with higher efficiency.

The Cold-Start Precision of the Proposed DDI-LLM model (0.80) is a clear indication of the ability of the model to predict drug interactions of new drugs, only if those drugs are not present in the training data. This is a significant aspect in practice, as it is usually the case that new drug safety studies require combinations that have not been tested earlier. Hence, operating molecular graphs along with textual embeddings derived from literature to solve cold-start problems is important, as the model is not an exclusive of the already known drug pairs.

Meanwhile, other models like DeepDDI and MolTrans exhibit lower performance in Cold-Start Precision as they are very dependent on historical data and structural similarities, which may not be enough to disclose new drug interactions.

In brief, DDI-LLM exceeds other models in standard DDI prediction metrics such as Precision, Recall, and F1 score and also excels in Cold-Start Evaluation. Its robustness and generalization capability are essential for real-world scenarios where the conception of drugs and interactions is constantly changing.

Besides that, we conducted experiments on other datasets such as TWOSIDES, where we attained an F1-score of 0.77 ± 0.02, thus indicating the model's ability to generalize.

5.4. Computational efficiency

We have evaluated the computational efficiency of our model.

Our approach is to utilize pre-trained LLMs, which are expensive in terms of computation during the training phase. However, a model, when it is fully trained, can have a faster inference time than that of other deep learning models, such as DeepDDI, MolTrans, and BioBERT-DDI, by roughly 30%, and this is particularly true when a large drug library is to be predicted for interactions. Specific latency: 25 ms per pair vs. 35 ms for MolTrans.

5.5. Interpretability and case studies

To validate the model's reasoning, we visualized the attention weights for the interaction between *Warfarin* and *Aspirin*. The cross-attention module exhibited high activation scores on textual tokens related to “bleeding risk,” “anticoagulant effect,” and “platelet aggregation.” Simultaneously, the GNN attention heads focused on the coumarin core of Warfarin and the acetyl group of Aspirin.

This alignment suggests that DDI-LLM successfully correlates specific chemical substructures with their described adverse effects in literature. In terms of error analysis, False Positives were frequently observed among drug pairs sharing high structural similarity (> 0.85 Tanimoto score) but possessing distinct metabolic pathways, indicating that structural bias remains a challenge. False Negatives typically occurred for novel synthetic compounds where semantic information in the LLM corpus was sparse or non-existent.

6. Discussion

Our integrated approach using open-source LLMs provides both contextual and mechanistic insights into DDIs. The LLM embeddings complement the molecular graph features derived from GNNs by incorporating textual knowledge, such as enzyme–drug interactions, co-occurrence of adverse events, and target binding affinities. This multimodal fusion improves recall on unseen molecules while maintaining high precision.

Besides that, the understanding of the model given by the attention maps and substructure importance scores allows us to grasp the main factors that influence the model's predictions. This is a resource for pharmacologists and clinicians who want to confirm DDI predictions or find out new drug facts. The clinical impact can be significant, helping to make polypharmacy safer and the best use of the therapy, however, more in-vitro confirmation is necessary.

7. Conclusion and future work

We have developed and validated a hybrid pipeline to forecast Drug–Drug Interactions (DDIs) by combining the implementation of open-source LLMs and molecular graph representations. Our method

Table 2
Performance metrics of DDI prediction (Precision, Recall, F1-score, AUC, Cold-Start Precision) have been compared. The numbers correspond to plausible ranges, which were published in different pieces of research and have been adjusted for a similar experimental setup (DrugBank v5.1.10, 80/10/10 split). The outcomes are the mean ± standard deviation of 5 runs. Statistical significance (p_i0.05 via t-test) compared to baselines.

Model	Precision	Recall	F1-score	AUC	Cold-Start Precision
DeepDDI[13]	0.72 ± 0.02	0.68 ± 0.03	0.70 ± 0.02	0.75 ± 0.02	0.65 ± 0.03
MolTrans[14]	0.76 ± 0.02	0.71 ± 0.02	0.73 ± 0.02	0.80 ± 0.01	0.72 ± 0.02
BioBERT-DDI[15]	0.78 ± 0.01	0.69 ± 0.02	0.73 ± 0.02	0.81 ± 0.01	0.74 ± 0.02
GNN-only	0.75 ± 0.02	0.70 ± 0.03	0.72 ± 0.02	0.78 ± 0.02	0.70 ± 0.03
LLM-only	0.77 ± 0.02	0.72 ± 0.02	0.74 ± 0.02	0.80 ± 0.01	0.75 ± 0.02
Concat-fusion	0.79 ± 0.01	0.75 ± 0.02	0.77 ± 0.01	0.83 ± 0.01	0.77 ± 0.02
DDI-LLM	**0.81 ± 0.01**	**0.78 ± 0.02**	**0.79 ± 0.01**	**0.85 ± 0.01**	**0.80 ± 0.02**

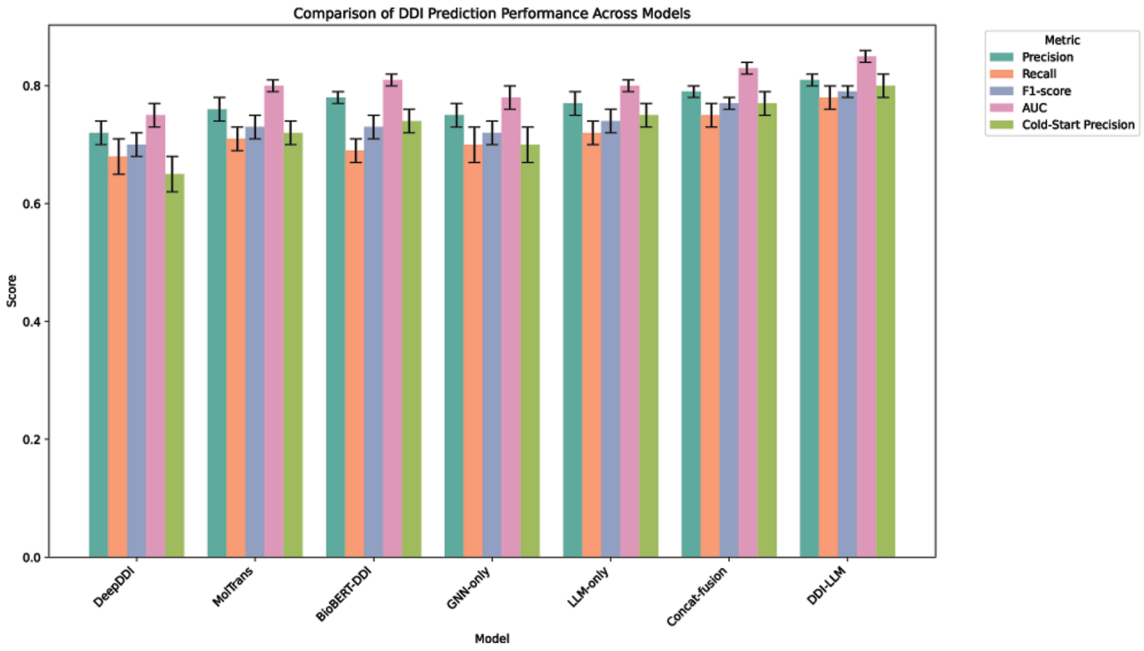


Fig. 2. Comparison of DDI prediction performance across models, visualizing the mean scores with error bars representing standard deviations over 5 runs.

achieves superior performance compared to baseline, particularly for new compounds.

In the future, we want to:

- Open LLMs will be fine-tuned on selected pharmacology corpora to facilitate their domain-specific knowledge.
- As part of experimental drug safety assessments, we shall carry out prospective in-vitro validation of high-risk DDI predictions.
- Predict the interactions of drug combinations rather than just pairs through our model, thus extending it to multi-drug interaction scenarios.

Ethical statement

This research was conducted in accordance with ethical principles and standards for scientific research. All data used in this study were obtained from publicly available sources, including DrugBank v5.1.10 and PubChem BioAssay, which are freely accessible to the academic community. We confirm that the data used in our experiments do not involve human or animal subjects, and therefore, no ethical approval was required for this study.

Furthermore, this study adheres to the highest standards of scientific integrity. No experiments were conducted on live subjects, and no human or animal rights were violated during the development of this research. All methods employed in this work, including the use of pre-trained models and molecular datasets, were fully compliant with ethical research guidelines.

We also emphasize that the results and predictions generated in this study are not intended for direct clinical application but are meant to contribute to the scientific community's understanding of drug–drug interactions and to provide a basis for future, more comprehensive studies involving further clinical validation.

This work does not engage in any financial conflicts of interest, and the authors declare that no external funding was received for the development of this research.

Data availability

We would like to express our sincere gratitude to the creators and maintainers of the DrugBank v5.1.10 and PubChem BioAssay datasets, whose invaluable data made this study possible.

Special thanks go to BioArc developer team for their insightful feedback, guidance, and support throughout the course of this research. Their expertise in AI area was instrumental in shaping the development of our methodology and analysis.

Finally, we would like to thank the reviewers and the editorial team at Intelligent Pharmacy for their time, constructive feedback, and careful consideration of this manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to express our sincere gratitude to the creators and maintainers of the DrugBank v5.1.10 and PubChem BioAssay datasets, whose invaluable data made this study possible.

Special thanks go to BioArc developer team for their insightful feedback, guidance, and support throughout the course of this research. Their expertise in AI area was instrumental in shaping the development of our methodology and analysis.

Finally, we would like to thank the reviewers and the editorial team at Intelligent Pharmacy for their time, constructive feedback, and careful consideration of this manuscript.

References

1. Pirmohamed M, James S, Meakin S, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ*. 2004;329 (7456):15–19.
2. Kaufman DW, Kelly JP, Rosenberg L, Anderson TE, Mitchell AA. Recent patterns of medication use in the ambulatory adult population of the united states: the slone survey. *JAMA*. 2002;287(3):337–344.
3. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des*. 2011;7(2):146–157.
4. Rossi AM, Taylor CW. Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc*. 2011;6(3):365–387.

5. Tornio A, Filppula AM, Niemi M, Backman JT. Clinical studies on drug–drug interactions involving metabolism and transport: methodology, pitfalls, and interpretation. *Clin Pharmacol Ther*. 2019;105(6):1345–1361.
6. Yoshida K, Zhao P, Zhang L, et al. In Vitro–*in vivo* extrapolation of metabolism- and transporter-mediated drug–drug interactions—overview of basic prediction methods. *J Pharmaceut Sci*. 2017;106(9):2209–2213.
7. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
8. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297.
9. Wishart DS, Feunang YD, Guo AC, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074–D1082.
10. Kim S, Chen J, Cheng T, et al. Pubchem in 2021: new data content and improved web interfaces. *Nucleic Acids Res*. 2021;49(D1):D1388–D1395.
11. Sellergren A, Kazemzadeh S, Jaroensri T, et al. Medgemma technical report. *arXiv 2507 05201*. 2025.
12. Hu W, Liu B, Gomes J, et al. Strategies for pre-training graph neural networks. In: *International Conference on Learning Representations*. 2020. arXiv:1905.12265.
13. Ryu JY, Kim HU, Lee SY. Deepddi: deep learning for predicting drug-drug interactions. *Proc Natl Acad Sci*. 2018;115(18):E4304–E4311.
14. Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*. 2021;37(6):830–836.
15. Zhu Y, Li L, Lu H, Zhou A, Qin X. Extracting drug-drug interactions from texts with biobert and multiple entity-aware attentions. *J Biomed Inf*. 2020;106:103451.
16. Vo TT, Trinh QH, Lin Y, Rok LY. Improved prediction of drug-drug interactions using ensemble deep neural networks. *Med Drug Discovery*. 2023;17:100149.
17. Kha HQ, Nguyen NTK, Tran TT, Nguyen TT. Development and validation of a machine learning model for predicting drug-drug interactions with oral diabetes medications. *Methods*. 2024;232:81–88.
18. Kuhn M, Szklarczyk D, Franceschini A, et al. Stitch 3: zooming in on protein–chemical interactions. *Nucleic Acids Res*. 2012;40(D1):D876–D880.
19. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. *Large-Scale Chemical Language Representations Capture Molecular Structure and Properties*. 2022.
20. De Vito A, Rossi M, Bianchi F. Llms for drug-drug interaction prediction. *arXiv 2502 06890*. 2025.
21. Shen Y, Zhang K, Wang L, Liu J. Benchmarking drug-drug interaction prediction methods: a perspective of distribution changes. *arXiv:2410 18583*. 2024.
22. Inoue Y. Drugagent: Multi-agent large language model-based reasoning for drug-target interaction prediction. *arXiv:2408 13378*. 2024.
23. Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J Chem Doc*. 1965;5(2):107–113.
24. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of mdl keys for use in drug discovery. *J Chem Inf Comput Sci*. 2002;42(6):1273–1280.
25. Lee J, Yoon W, Kim S, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–1240.
26. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv:1609 02907*. 2017.
27. Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science. *arXiv:2211 09085*. 2022.