Journal of Molecular Graphics and Modelling 148 (2026) 109530

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/jmgm

DSMV-DDI: A dual-level pharmacological semantic and stereochemical visual representation learning framework for drug–drug interaction prediction

Fangni Chen [a], Tingting Jiang [a,b,*], Shuai Yang [a,b], Xiaohui Yuan [a,c], Chao Wang [a], Lichuan Gu [a]

[a] School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, Anhui, China
[b] Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Hefei, Anhui, China
[c] Department of Computer Science and Engineering, University of North Texas, TX 76203, Denton, USA

ARTICLE INFO

Dataset link: https://github.com/CvChen-111/dsmv_ddi_model

Keywords:
Drug–drug interactions
Multimodal representation
Dual-level pharmacological semantics
Molecular visual representation
Knowledge graph

ABSTRACT

Drug–drug interactions (DDIs) are a major cause of adverse drug events in clinical practice, especially under polypharmacy settings where patients receive multiple medications simultaneously. Reliable computational prediction of DDIs is therefore essential for improving medication safety and supporting clinical decision-making. Despite recent advances in computational DDI prediction, existing methods often struggle to jointly model multi-granularity pharmacological semantics and stereochemical molecular characteristics, limiting their ability to generalize to previously unseen drugs under cold-start scenarios. To address these limitations, we propose DSMV-DDI, a multimodal representation learning framework for drug–drug interaction prediction that integrates biomedical knowledge graph topology, chemical substructure features, dual-level pharmacological semantic representations, and stereochemical molecular visual representations derived from three-dimensional molecular conformations. In particular, the proposed dual-level semantic strategy jointly characterizes interaction-level pharmacological associations and intrinsic single-drug functional properties, enabling complementary modeling of pharmacological information across different semantic granularities. Furthermore, molecular visual representation learning captures geometric and spatial characteristics beyond topology-based molecular representations, improving generalization to topologically unseen drugs. Extensive experiments on real-world DDI datasets demonstrate that DSMV-DDI outperforms state-of-the-art methods, achieving an accuracy of 0.967 and an AUPR of 0.992 under the conventional setting. The proposed framework also maintains strong performance under both partial and complete cold-start settings. Ablation analyses show that dual-level pharmacological semantics contribute most to overall performance, while molecular visual representations provide complementary geometric information that further improves prediction accuracy.

1. Introduction

Combination therapy is widely adopted in clinical practice to improve therapeutic efficacy and reduce the risks associated with monotherapy, including drug resistance and dose-related adverse reactions [1,2]. However, the increasing prevalence of polypharmacy substantially elevates the risk of drug–drug interactions (DDIs), which may result in altered pharmacological effects, toxicity accumulation, therapeutic failure, or severe adverse drug reactions [3–5]. Previous studies have shown that DDIs contribute to a considerable proportion of adverse drug events, representing a critical challenge for medication safety and clinical decision-making [6]. Although experimental and clinical validation strategies are generally reliable, they remain labor-intensive, time-consuming, and economically expensive, rendering comprehensive evaluation of the rapidly expanding space of potential drug combinations practically infeasible [7]. Consequently, computational approaches for DDI prediction have attracted increasing attention as scalable and efficient alternatives for identifying potential DDIs across large drug libraries [8].

Recent advances in deep learning and multimodal representation learning have significantly promoted biomedical prediction tasks, in-

* Corresponding author at: School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, Anhui, China.
E-mail addresses: chenfangni@stu.ahau.edu.cn (F. Chen), jiangtt@ahau.edu.cn (T. Jiang), yangs@ahau.edu.cn (S. Yang), xiaohui.yuan@unt.edu (X. Yuan), warton_wang@qq.com (C. Wang), glc@ahau.edu.cn (L. Gu).

https://doi.org/10.1016/j.jmgm.2026.109530
Received 9 June 2026; Received in revised form 13 July 2026; Accepted 25 July 2026
Available online 1 August 2026

cluding drug–disease association prediction and molecular generation [9,10]. Benefiting from these developments, computational DDI prediction has also witnessed rapid progress in recent years. Existing computational approaches for DDI prediction can generally be categorized into three groups according to how drug-related representations are constructed and integrated. Methods based on drug feature representation utilize intrinsic drug properties, including molecular fingerprints, physicochemical descriptors, and SMILES-derived structural profiles, to characterize drugs and their potential interactions [11–13]. These approaches provide effective molecular descriptors and have achieved competitive performance by capturing molecular composition and local chemical characteristics. However, they generally overlook higher-order biological associations among drugs, targets, enzymes, and transporters. Graph-based methods employ graph neural networks and biomedical knowledge graph representations to model relational dependencies among biological entities [14,15]. By propagating information through large-scale biomedical networks, these methods substantially enhance relational representation learning and improve the modeling of biological interaction topology. Nevertheless, most existing graph-based approaches remain predominantly topology-centric and insufficiently capture complementary structural, pharmacological, and stereochemical information beyond relational associations. More recently, hybrid representation methods [16,17] have attempted to integrate complementary molecular and pharmacological information from multiple perspectives, demonstrating improved capability in modeling complex DDIs [18,19]. Inspired by the success of multimodal representation learning in related biomedical prediction tasks, similar fusion strategies have demonstrated promising performance in toxicity prediction [20], peptide toxicity prediction [21], mutagenicity prediction [22,23], and cardiotoxicity prediction [24], highlighting the potential of multimodal representation learning for biomedical prediction tasks. These studies collectively demonstrate the effectiveness and generalizability of multimodal representation learning across diverse biomedical prediction tasks.

Motivated by these advances, recent multimodal DDI prediction methods have substantially improved predictive performance. However, achieving robust generalization under realistic clinical scenarios remains challenging. In particular, constructing drug representations that simultaneously capture complementary pharmacological knowledge and transferable molecular characteristics is still an open problem. Recent semantic-enhanced DDI prediction methods have demonstrated that pharmacological textual knowledge provides valuable complementary information beyond structural representations. Existing methods, however, typically emphasize either interaction-level pharmacological context or intrinsic single-drug functional properties, while rarely modeling both levels jointly within a unified representation framework [18,25]. This limitation restricts the ability of current approaches to simultaneously characterize intrinsic drug properties and higher-order pharmacological relationships emerging from drug co-administration.

Second, stereochemical and spatial characteristics encoded in three-dimensional molecular conformations remain insufficiently integrated with multi-granularity pharmacological semantics in current DDI prediction studies [19]. Two-dimensional topology-based molecular graph representations effectively describe atomic connectivity and relational structure, yet cannot fully characterize geometric information encoded in three-dimensional molecular conformations, including stereochemical patterns and spatial arrangements of functional groups. Such topology-independent stereochemical characteristics are particularly valuable under cold-start scenarios, where previously observed interaction topology is unavailable for novel drugs. Unlike SMILES-based sequence descriptors or molecular fingerprints, three-dimensional conformer renderings explicitly encode transferable geometric features such as ring conformations, steric bulk distributions, and pharmacophoric spatial arrangements that are shared across structurally related drug families, thereby facilitating generalization beyond the observed interaction topology. These observations motivate the development of a unified multimodal representation learning framework that jointly models multi-granularity pharmacological semantics and stereochemical molecular characteristics. Unlike existing multimodal DDI prediction methods, such a framework should explicitly integrate dual-level pharmacological semantics with stereochemical molecular visual representations, enabling complementary modeling of pharmacological knowledge and transferable geometric information within a unified representation learning framework.

To address these challenges, we propose DSMV-DDI, a multimodal representation learning framework that integrates multi-source biomedical knowledge graph topology, chemical substructure patterns, dual-level pharmacological semantic representations, and stereochemical molecular visual representations derived from force-field-optimized three-dimensional conformers. These complementary modalities are jointly optimized to improve both conventional DDI prediction and topology-independent generalization under cold-start settings. The main contributions of this work are summarized as follows:

- **Dual-level pharmacological semantic modeling.** We jointly characterize interaction-level associations and single-drug functional properties, providing complementary multi-granularity pharmacological information beyond single-perspective approaches.
- **Stereochemical visual representation.** We encode 3D molecular conformational geometry from multi-view renderings, delivering transferable spatial features that substantially improve cold-start generalization for unseen drugs.
- **Unified multimodal integration.** We propose DSMV-DDI that fuses knowledge graph topology, chemical substructures, dual-level semantics, and molecular images. Extensive experiments show consistent SOTA performance across conventional, partial, and complete cold-start settings.

2. Related work

This section reviews three major categories of DDI prediction methods: methods based on drug feature representation, methods based on graph representation, and methods based on hybrid representation, providing the background and motivation for our proposed framework.

2.1. Methods based on drug feature representation

These methods leverage intrinsic drug properties, such as molecular fingerprints, physicochemical descriptors, and SMILES-based structural similarity profiles, to learn drug representations for DDI prediction [26]. DeepDDI [11] constructs representations from SMILES-derived structural similarity. MDF-SA-DDI [12] integrates multi-source drug attributes with self-attention mechanisms, while MDDI-SCL [13] introduces structural contrastive learning to improve discrimination on rare interaction types. Despite their effectiveness, these approaches overlook topological relationships among drugs, targets, enzymes, and transporters, limiting their ability to capture higher-order biological associations. To address this, our framework incorporates multi-source knowledge graph structural encoding alongside intrinsic molecular features.

2.2. Methods based on graph representation

Methods based on graph representation model drugs and biological entities as nodes in relational networks, capturing topological dependencies through techniques such as matrix factorization [27], random walk-based embedding [28], and graph neural networks [29,30]. Knowledge graphs further enrich representations by encoding

multi-relational interactions among drugs, targets, and enzymes. Representative models include KGNN [14], which propagates information along biomedical relationships, and SumGNN [15], which employs graph summarization for large-scale biomedical KGs. However, these methods remain predominantly topology-centric and lack effective mechanisms for integrating semantic knowledge from textual descriptions and visual features from molecular structures. Our framework extends graph-based representations by incorporating semantic, chemical substructure, and molecular image modalities.

2.3. Methods based on hybrid representation

Hybrid representation methods improve DDI prediction by integrating complementary molecular and pharmacological information from multiple perspectives. Representative studies such as MKG-FENN [16] and MDG-DDI [17] enhance DDI modeling through multi-source graph representations and enriched biomedical associations. Other studies further introduce textual or visual molecular representations into DDI prediction. For example, SubGE-DDI [18] enriches biomedical graph representations using interaction-related textual information, and ImageDDI [19] improves molecular representation learning through image-enhanced molecular motif modeling. In addition, LSFC [31] demonstrated the effectiveness of local substructure feature combinations for DDI classification. Despite these advances, existing hybrid approaches still exhibit several limitations. Pharmacological semantic information is often modeled at a single granularity, making it difficult to jointly characterize intrinsic drug properties and interaction-level pharmacological relationships. In addition, stereochemical and spatial characteristics encoded in molecular conformations remain insufficiently integrated with multi-granularity pharmacological semantics, particularly under cold-start scenarios where topology-dependent information becomes limited. To address these limitations, we propose DSMV-DDI, which integrates multi-source knowledge graph topology, chemical substructure representations, pharmacological semantic representations, and stereochemical molecular visual representations to improve both conventional DDI prediction and cold-start generalization.

3. Methods

3.1. Overview of the method

We propose DSMV-DDI, a multimodal framework for drug–drug interaction prediction that integrates four complementary modalities. As illustrated in Fig. 1, the framework consists of: (i) multi-source knowledge graph structural encoding (M1), which captures topological relationships among drugs and biological entities from multiple biomedical perspectives; (ii) chemical substructure encoding (M2), which extracts local functional group patterns and atomic configurations from molecular graphs derived from SMILES sequences; (iii) dual-level pharmacological semantic encoding (M3), which hierarchically integrates drug-pair interaction semantics with intrinsic single-drug pharmacological properties; and (iv) molecular visual encoding (M4), which encodes stereochemical spatial features from multi-view three-dimensional conformer renderings. These four modalities are projected into a unified embedding space and fused to predict drug–drug interaction types.

3.2. Problem formalization

Let $D = \{d_1, d_2, \ldots, d_N\}$ denote the set of $N = 572$ drugs, and $Y = \{1, 2, \ldots, C\}$ denote the set of $C = 65$ predefined DDI event types. Given a drug pair (d_i, d_j), the task is formulated as a multi-class classification problem that predicts the interaction label $y_{ij} \in Y$.

For each drug d_i, we construct four modality-specific representations: structural $\mathbf{h}_i^{(1)}$, chemical substructure $\mathbf{h}_i^{(2)}$, semantic $\mathbf{h}_i^{(3)}$, and molecular image $\mathbf{h}_i^{(4)}$, corresponding to the four modalities M1–M4 introduced in Section 3.1. For a drug pair (d_i, d_j), the pairwise representation under modality m is derived as:

$$\mathbf{h}_{ij}^{(m)} = f^{(m)}\left(\mathbf{h}_i^{(m)}, \mathbf{h}_j^{(m)}\right), \quad m \in \{1, 2, 3, 4\} \tag{1}$$

where $f^{(m)}(\cdot)$ comprises concatenation, element-wise absolute difference, and element-wise product to capture both individual drug properties and their pairwise interaction patterns. The final fused representation is obtained by projecting each modality into a shared dimensionality d_p and concatenating:

$$\mathbf{h}_{ij}^{\text{fusion}} = \phi^{(1)}\left(\mathbf{h}_{ij}^{(1)}\right) \| \phi^{(2)}\left(\mathbf{h}_{ij}^{(2)}\right) \| \phi^{(3)}\left(\mathbf{h}_{ij}^{(3)}\right) \| \phi^{(4)}\left(\mathbf{h}_{ij}^{(4)}\right) \tag{2}$$

where $\|$ denotes vector concatenation, and $\phi^{(m)}(\cdot)$ denotes the modality-specific projection function that maps each pairwise representation into the shared embedding space. The model is trained end-to-end by minimizing the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|\mathcal{T}|} \sum_{(i,j)\in\mathcal{T}} \sum_{c=1}^{C} y_{ij}^{(c)} \log \hat{y}_{ij}^{(c)} \tag{3}$$

where $y_{ij}^{(c)}$ is the one-hot encoding of the ground-truth label and $\hat{y}_{ij}^{(c)}$ is the predicted probability for class c.

3.3. Multi-source knowledge graph structural encoding

Topological relationships among drugs and biological entities provide essential context for DDI prediction that cannot be derived from molecular attributes alone. We construct four complementary knowledge graphs to capture structural information at different levels: (1) a drug-chemical entity graph derived from DrugBank [32] interactions, encoding associations among drugs, enzymes, transporters, and protein targets; (2) a drug-substructure graph based on SMILES-derived molecular fragments; (3) a DDI graph constructed from the known DDI adjacency matrix; and (4) a molecular structure graph based on MACCS fingerprints. These four graphs are encoded through a four-channel graph neural network to obtain multi-perspective structural representations.

Each knowledge graph $G^{(k)}$ is represented as a collection of (d_i, r, t), where d_i denotes a drug entity, r a relation type, and t a tail entity. For each drug d_i, a fixed-size neighbor set $\mathcal{N}(d_i)$ of size $M = 6$ is uniformly sampled from each graph. For channel k at propagation layer l, the relational feature score between drug d_i and neighboring tail entity t_n is computed as:

$$\pi_{(d_i, r_{in})}^{(l)} = \sum \left[\left(\mathbf{e}_{d_i}^{(l-1)} \odot \mathbf{e}_{r_{in}}^{(l-1)}\right) W_1 + b_1\right] \tag{4}$$

where $\mathbf{e}_{d_i}^{(l-1)}$ and $\mathbf{e}_{r_{in}}^{(l-1)}$ are the drug and relation embeddings at layer $l-1$, $\odot$ denotes element-wise multiplication, W_1 and b_1 are learnable parameters. The neighbor representations are aggregated by these scores:

$$\mathbf{e}_{\mathcal{N}(d_i)}^{(l)} = \sum_{t_n \in \mathcal{N}(d_i)} \pi_{(d_i, r_{in})}^{(l)} \mathbf{e}_{t_n}^{(l-1)} \tag{5}$$

The drug representation is then updated by fusing its own embedding with the aggregated neighbor information:

$$\mathbf{e}_{d_i}^{(l)} = \sigma\left(\left[\mathbf{e}_{d_i}^{(l-1)} \oplus \mathbf{e}_{\mathcal{N}(d_i)}^{(l)}\right] W_2 + b_2\right) \tag{6}$$

where σ is the ReLU activation, $\oplus$ denotes concatenation, and W_2, b_2 are learnable parameters. The structural representations obtained from the four KG channels are concatenated to form the comprehensive structural feature $\mathbf{h}_i^{(1)} = \mathbf{h}_i^{(1,1)} \| \mathbf{h}_i^{(1,2)} \| \mathbf{h}_i^{(1,3)} \| \mathbf{h}_i^{(1,4)}$.

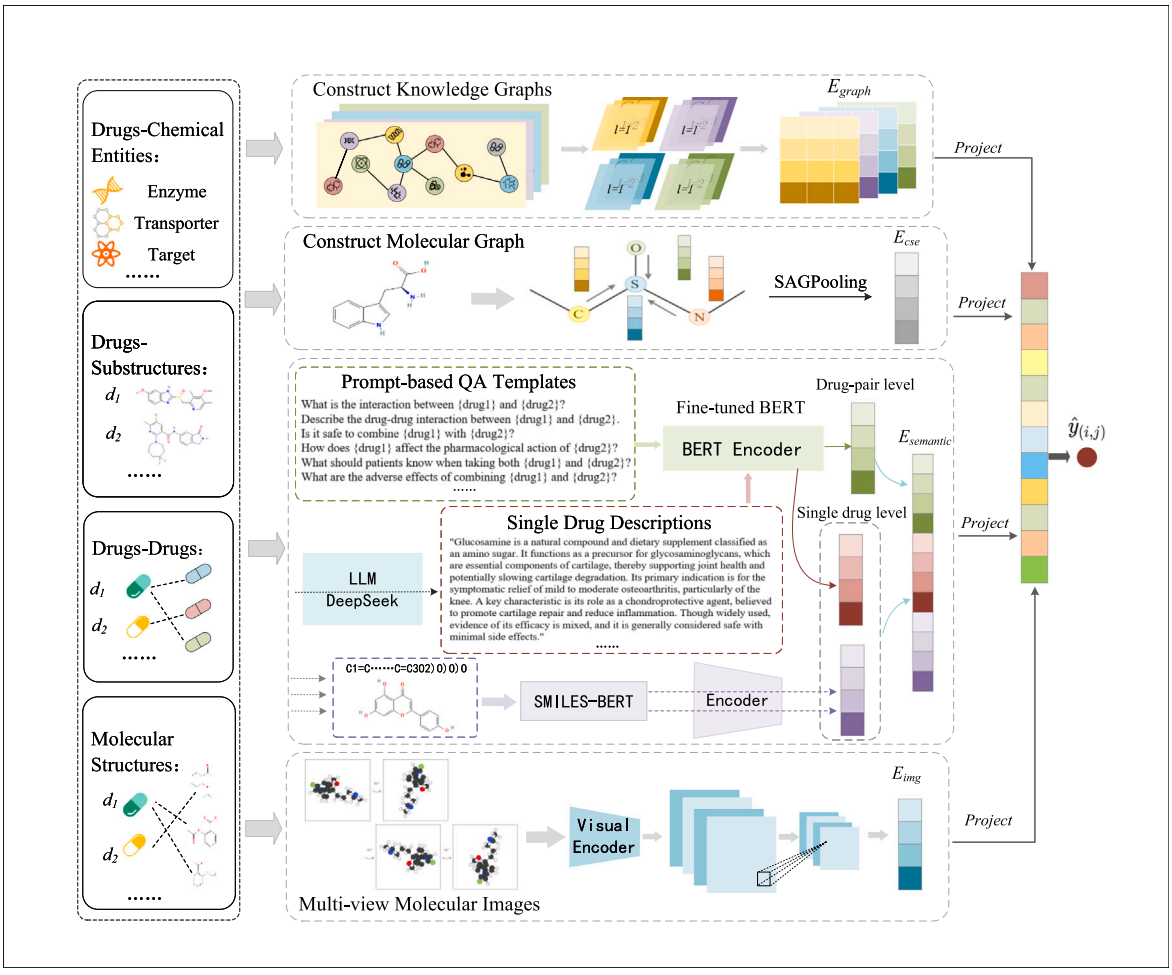

Fig. 1. Overview of the proposed DSMV-DDI framework. The model integrates four complementary modalities, including knowledge graph structure, chemical substructures, dual-level semantic representations, and molecular visual features. A unified multimodal fusion module is then employed to generate discriminative drug pair representations for DDI prediction.

3.4. Chemical substructure encoding

Stereochemical features, including functional group identity, bond hybridization, and local atomic configurations, are critical determinants of drug metabolic susceptibility and interaction specificity that are not fully captured by knowledge graph topology alone. We employ a dedicated molecular graph convolutional network to encode these intrinsic chemical patterns directly from molecular structure.

For each drug d_i, RDKit [33] is used to construct a molecular graph $\mathcal{G}(d_i) = (\mathcal{V}, \mathcal{E})$ from its SMILES sequence, where $\mathcal{V}$ denotes atoms and $\mathcal{E}$ denotes chemical bonds. Each atom node $v \in \mathcal{V}$ is characterized by a 58-dimensional feature vector encoding atomic symbol, degree, total hydrogen count, implicit valence, aromaticity, hybridization state, formal charge, and radical electron count, as implemented in the accompanying preprocessing pipeline. Each bond $e \in \mathcal{E}$ is represented by a 17-dimensional feature vector encoding bond type (single, double, triple, aromatic), conjugation, ring membership, and stereochemistry configuration.

The molecular graph is encoded through a multi-architecture graph neural network combining three convolutional structures at successive layers. For atom node v at layer l, message passing and aggregation proceed as:

$$\mathbf{h}_v^{(l)} = \phi^{(l)}\left(\mathbf{h}_v^{(l-1)}, \bigcup_{u \in \mathcal{N}(v)} \psi^{(l)}\left(\mathbf{h}_v^{(l-1)}, \mathbf{h}_u^{(l-1)}, \mathbf{e}_{(v,u)}\right)\right) \tag{7}$$

where $\mathcal{N}(v)$ is the neighborhood of node v, $\psi^{(l)}$ is the message function, $\phi^{(l)}$ is the update function, and $\bigcup$ is a commutative aggregation operator. GINConv [34] is applied at the first layer for its strong sensitivity to graph isomorphism, enabling precise discrimination of chemically distinct substructures including ring systems and branching patterns. GCNConv [35] captures global topological patterns at the second layer, providing a holistic view of atomic connectivity that reflects overall molecular shape. TransformerConv [36] at the third layer models long-range atomic dependencies through attention mechanisms, particularly relevant for pharmacophoric features, such as spatially separated hydrogen bond donors and acceptors, which jointly determine binding specificity.

After L layers of message passing, a graph-level substructure representation is obtained through attention-weighted pooling:

$$\alpha_v = \frac{\exp\left(\mathbf{w}^\top \mathbf{h}_v^{(L)}\right)}{\sum_{u \in \mathcal{V}} \exp\left(\mathbf{w}^\top \mathbf{h}_u^{(L)}\right)}, \quad \mathbf{h}_i^{\text{sub}} = \sum_{v \in \mathcal{V}} \alpha_v \cdot \mathbf{h}_v^{(L)} \tag{8}$$

where α_v represents the attention weight of atom v and $\mathbf{w}$ is a learnable parameter vector. This mechanism adaptively amplifies atoms and functional groups that most distinctively characterize the drug's chemical identity — such as reactive centers, aromatic systems, and stereocenters — while suppressing less discriminative structural regions.

For a drug pair (d_i, d_j), the pairwise chemical interaction representation $\mathbf{h}_{ij}^{(2)}$ is constructed via $f^{(2)}$ applied to their single-drug embeddings,

Table 1
Prompt-based Q&A template used to elicit clinically relevant interaction information between drug pairs for semantic representations.

Type	Prompt
Direct inquiry	What is the interaction between {drug1} and {drug2}? Describe the DDI between {drug1} and {drug2}. What happens when {drug1} is taken with {drug2}?
Safety-related	Is it safe to combine {drug1} with {drug2}? What are the adverse effects of combining {drug1} and {drug2}? What are the risks of concomitant use of {drug1} and {drug2}?
Mechanism-focused	How does {drug1} affect the pharmacological action of {drug2}? Explain the mechanism of interaction between {drug1} and {drug2}. How does {drug2} alter the metabolism of {drug1}?
Clinical advice	What should patients know when taking both {drug1} and {drug2}? What are the clinical recommendations for combining {drug1} with {drug2}? How should clinicians manage the coadministration of {drug1} and {drug2}?
Severity assessment	How severe is the interaction between {drug1} and {drug2}? What are the potential clinical consequences of {drug1}-{drug2} interaction? Evaluate the clinical significance of the interaction between {drug1} and {drug2}.

capturing structural complementarity between the two molecules at the functional group level.

3.5. Dual-level semantic encoding

3.5.1. Drug-pair level interaction semantics

To capture relational semantics between specific drug pairs, we construct prompt-based templates covering five clinical perspectives: direct inquiry, safety assessment, mechanism of action, clinical recommendations, and severity evaluation, as summarized in Table 1. These templates are used to construct interaction-oriented natural language queries for semantic representation learning. For a drug pair (d_i, d_j), a natural language query such as "What is the interaction between d_i and d_j?" is constructed, then encoded by a fine-tuned BERT [37] model adapted to interaction-oriented text:

$$\mathbf{h}_{ij}^{\text{pair}} = \text{BERT}_{\text{pair}}(\text{Prompt}(d_i, d_j)) \in \mathbb{R}^{768} \tag{9}$$

This embedding encodes higher-order pharmacological context inherent in drug co-administration, enabling the model to leverage textual interaction knowledge that transcends structural features.

3.5.2. Single-drug level attribute semantics

At the single-drug level, two complementary semantic sources are integrated. First, functional descriptive text $T(d_i)$ is generated by DeepSeek [38] to capture pharmacological properties, mechanisms of action, and clinical profiles. This text is encoded by a fine-tuned BERT model to obtain functional semantics $\mathbf{h}_i^{\text{desc}} \in \mathbb{R}^{768}$. Second, the SMILES sequence $S(d_i)$ is encoded by ChemBERTa [39], a domain-specialized model pre-trained on molecular corpora, to extract chemical language semantics $\mathbf{h}_i^{\text{chem}} \in \mathbb{R}^{768}$ capturing fine-grained bonding patterns and atomic configurations. The two sources are fused through linear projection and layer normalization:

$$\mathbf{h}_i^{\text{sem}} = f_{\text{align}}\left(\mathbf{h}_i^{\text{desc}} \parallel \mathbf{h}_i^{\text{chem}}\right) \tag{10}$$

where $f_{\text{align}}(\cdot)$ denotes the alignment function incorporating linear projection and layer normalization. The comprehensive semantic representation for a drug pair integrates both levels:

$$\mathbf{h}_{ij}^{(3)} = \mathbf{h}_{ij}^{\text{pair}} \| \mathbf{h}_i^{\text{sem}} \| \mathbf{h}_j^{\text{sem}} \tag{11}$$

This hierarchical design captures interaction semantics at the macro level while preserving intrinsic pharmacological attributes at the micro level, forming a unified semantic embedding spanning multiple granularities.

3.6. Molecular visual encoding

Stereochemical features, including chirality, conformational geometry, three-dimensional pharmacophore arrangement, and spatial distribution of hydrogen bond donors and acceptors, are critical determinants of drug-receptor complementarity and metabolic susceptibility. These features are largely inaccessible through two-dimensional graph representations, which capture atomic connectivity but cannot encode the three-dimensional geometry governing molecular behavior in biological environments. Molecular images of three-dimensional conformers provide an information-rich encoding of these stereochemical properties, complementing the topological and chemical information captured by the preceding modalities.

For each drug d_i, RDKit generates a 3D conformer from its SMILES sequence using MMFF94 [40] force field optimization to obtain a geometrically relaxed molecular structure. The conformer is rendered from four orthogonal viewpoints using a custom visualization pipeline based on RDKit and Matplotlib [41], with atoms represented as colored spheres and bonds as line segments connecting adjacent atoms, producing four RGB images per drug. Four viewpoints are selected to ensure comprehensive coverage of the molecular surface, capturing steric bulk distribution, ring conformational geometry, hydrophobic patch geometry, and the three-dimensional positioning of pharmacophoric elements governing drug-target complementarity. The resulting images are resized to the standard ImageNet input resolution and normalized with ImageNet statistics prior to feature extraction.

A ResNet50 [42] backbone with its classification head removed serves as the visual encoder. The 2048-dimensional feature from the final convolutional layer is projected to a 256-dimensional embedding through a linear layer followed by L2 normalization. The multi-view image representation for drug d_i is obtained by averaging embeddings across all valid views:

$$\mathbf{h}_i^{(4)} = \frac{1}{K}\sum_{k=1}^{K} \text{ResNet50}\left(I_i^{(k)}\right) \tag{12}$$

where $K \in \{1, 2, 3, 4\}$ is the number of valid views and $I_i^{(k)}$ denotes the kth view image. If no valid conformer can be generated, a learnable missing-image embedding is used as a fallback.

3.7. Multimodal fusion and prediction

Upon obtaining the four modality-specific pairwise representations, each is projected into a shared dimensionality $d_p = 512$ through an independent linear layer with ReLU activation and layer normalization:

$$\mathbf{z}_{ij}^{(m)} = \text{LayerNorm}\left(\text{ReLU}\left(W^{(m)}\mathbf{h}_{ij}^{(m)} + b^{(m)}\right)\right) \tag{13}$$

Table 2
Details of the dataset.

Dataset	Drug number	Entity number	Relationship number	Triple number
Part 1	572	825	235	6541
Part 2	572	583	1	70 350
Part 3	572	572	65	74 528
Part 4	572	20	13	11 440

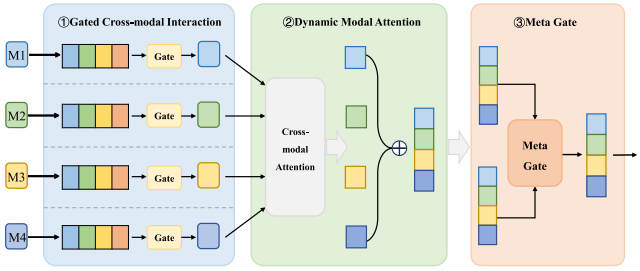

Fig. 2. Overview of the proposed multimodal fusion framework for DDI prediction. M1, M2, M3, and M4 denote the multi-source knowledge graph structural, chemical substructure, dual-level semantic, and molecular visual modalities, respectively.

The aligned modality representations are subsequently integrated through a three-stage fusion module consisting of gated cross-modal interaction, dynamic modal attention, and a meta-gated fusion strategy, as illustrated in Fig. 2. Here, $W^{(m)}$ and $b^{(m)}$ denote the learnable projection parameters for modality m. First, each aligned modality feature is refined by an independent gating unit to obtain gated modality representations. These representations are then fed into a dynamic modal attention module, which computes attention-enhanced modality features by modeling interactions across different modalities. Finally, the attention-enhanced representations are combined with the original aligned features through a meta gate to produce the final fused representation $\mathbf{h}_{ij}^{\text{fusion}}$.

The fused representation is subsequently processed by a two-layer MLP with ReLU activation, batch normalization, and dropout ($p = 0.3$), followed by a softmax classifier for DDI type prediction. The model is optimized using the cross-entropy loss with Adam [43], a learning rate of 1×10^{-2}, and a weight decay of 1×10^{-8}.

4. Results

4.1. Datasets

To evaluate the effectiveness of DSMV-DDI, we constructed a real-world drug dataset by integrating four complementary data sources. First, DDI event information was obtained from DrugBank (version 5.1.7). Second, drug molecular representations derived from SMILES were incorporated by transforming them into knowledge graph structures, following the preprocessing strategy introduced in DDIMDL [26]. Third, DDI relationship matrices were further utilized and similarly converted into graph-based representations. Finally, MACCS (Molecular ACCess System) fingerprints were employed to characterize molecular structural properties, with 13 informative MACCS keys and 7 other molecular features adopted following established studies [44].

By integrating these diverse data sources, four distinct yet complementary knowledge graphs were constructed, allowing the model to capture drug information from multiple perspectives. Detailed statistics of the dataset are summarized in Table 2.

As illustrated in Fig. 3, the 65 DDI interaction types exhibit severe class imbalance: the top two categories together account for over 50% of all samples, while the majority of rare interaction types each contribute less than 0.5%.

Table 3
Experimental task setup.

Task	Description
Task 1	Predict interactions between known drugs
Task 2	Predict interactions between known drugs and novel drugs
Task 3	Predict interactions between novel drugs

4.2. Evaluation metrics

To evaluate model performance in predicting drug–drug interactions, we employed multiple metrics, including accuracy (ACC), precision (PRE), recall (REC), F1 score, area under the ROC curve (AUC), and area under the precision–recall curve (AUPR). For precision, recall, and F1 score, macro-averaging was used to provide balanced assessment across classes under imbalanced data distributions. For AUC and AUPR, micro-averaging was applied to reflect overall predictive performance by aggregating across all classes.

4.3. Experimental task setup

To validate the model's generalization capability and robustness across different scenarios, we designed three experimental settings: (i) DDIs between known drugs (Task 1), (ii) DDIs between known and novel drugs (Task 2), and (iii) DDIs between novel drugs (Task 3). Task 1 evaluates model performance in conventional DDI prediction. Task 2 simulates a clinical scenario in which novel drugs are co-administered with known drugs, assessing generalization under partial cold-start conditions. Task 3 further examines the predictive capacity of DSMV-DDI under complete cold-start conditions. The task setup is summarized in Table 3.

In this study, novel drugs refer to drugs whose DDI interaction records are excluded from the training set. Although DDI interactions involving these drugs are not observed during training, auxiliary information such as molecular structures, semantic descriptions, molecular images, and knowledge graph associations remains available for representation learning.

4.4. Baselines

To comprehensively evaluate the effectiveness of DSMV-DDI, we compared our framework with a wide range of representative baseline methods under the Task 1 setting. These methods can be broadly categorized into four groups, including traditional machine learning methods, drug feature representation-based methods, graph representation-based methods, and hybrid representation-based methods. Detailed descriptions of the compared methods are provided below.

Traditional machine learning methods. Several conventional classification algorithms were adopted as baseline models, including DNN, RF, KNN, GNN [35], and LR [26]. These methods mainly rely on manually engineered drug features or shallow interaction representations for DDI prediction.

Methods based on drug feature representation. These methods primarily focus on extracting molecular structural descriptors, chemical properties, or multimodal drug attributes for interaction prediction.

- **DeepDDI** [11] learns latent structural patterns directly from SMILES sequences in an end-to-end manner for DDI prediction.

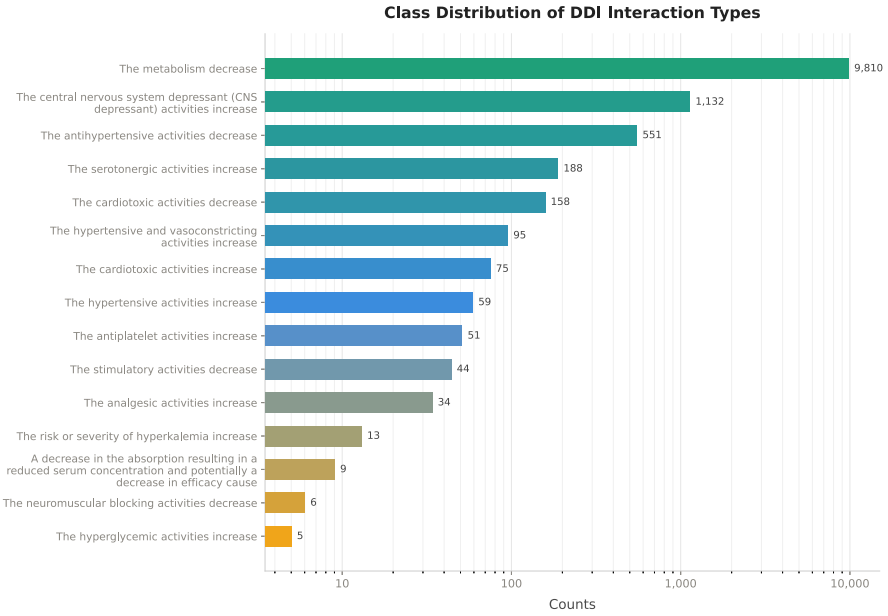

Fig. 3. Class distribution of DDI interaction types. Fifteen representative types are sampled at regular intervals from 65 total categories, illustrating the severe imbalance between high-frequency and rare interaction types.

- **MDF-SA-DDI** [12] integrates multidimensional drug features with a self-attention mechanism to emphasize informative interaction patterns.
- **MCNN-DDI** [45] employs heterogeneous convolutional kernels within a multi-channel CNN framework to capture multi-granularity drug representations.
- **DDIMDL** [26] combines multiple biomedical drug features under a multimodal deep learning framework for DDI prediction.
- **MDDI-SCL** [13] introduces structural contrastive learning to refine multimodal drug representations and improve discriminative capability for complex interaction categories.

Methods based on graph representation. These approaches mainly model molecular structures or biomedical relationships through graph neural networks and graph-based representation learning.

- **MDNN** [46] adopts a multi-branch deep neural architecture to encode heterogeneous graph-based drug features.
- **CNN-Siam** [47] utilizes a Siamese convolutional architecture with weight-sharing branches to model pairwise drug interactions.
- **CNN-DDI** [48] applies convolutional neural networks to extract local structural patterns from molecular representations.
- **DM-DDI** [49] combines deep matrix factorization and multi-task learning for multi-type DDI prediction.
- **NMDADNN** [50] incorporates neighbor multi-view data augmentation into deep neural networks to improve representation robustness.
- **MOLGAECL** [51] integrates graph autoencoders with contrastive learning to enhance molecular representation learning.

Methods based on hybrid representation. These methods integrate multiple heterogeneous modalities, such as molecular structures, biomedical knowledge graphs, semantic information, and multimodal drug descriptors.

- **MFDA** [52] employs a dual-attention fusion mechanism to adaptively integrate heterogeneous drug modalities.
- **MKG-FENN** [16] constructs a multi-perspective drug knowledge graph and adopts a graph-enhanced neural framework for interaction prediction.
- **Lee et al.** [53] combines chemical substructure similarity with protein target information to construct multi-view structural representations.
- **MSDF** [54] jointly models structural, textual, and functional drug descriptors through multimodal fusion.

4.5. Comparison on known drugs

In Task 1, where all drugs in both the training and test sets are observed during training, we evaluate the performance of DSMV-DDI under the conventional DDI prediction setting. The quantitative results are summarized in Table 4.

To further evaluate the stability of DSMV-DDI, we conducted an additional five-fold cross-validation experiment and reported the mean and standard deviation of four representative evaluation metrics, as shown in Table 5. The results show that DSMV-DDI maintains comparable performance across different data splits, with relatively small standard deviations.

Several important observations can be derived from the experimental results. First, DSMV-DDI consistently achieves the best performance across all evaluation metrics, demonstrating the effectiveness of the proposed multimodal fusion framework under fully observed conditions. In particular, the full model achieves the highest ACC (0.967) and AUPR (0.992), substantially outperforming existing baseline methods.

Second, even without molecular image features, the three-modality variant (w/o image) still surpasses all competing methods, indicating that the proposed dual-level semantic enhancement and structural encoding modules already provide strong discriminative capability for DDI prediction.

Third, traditional machine learning methods such as KNN and logistic regression exhibit relatively limited performance, while multimodal and graph-based approaches achieve better results by leveraging richer structural and semantic representations. Among these methods, hybrid frameworks such as MKG-FENN and MSDF show competitive performance, confirming the importance of heterogeneous biomedical information for DDI prediction.

Finally, incorporating molecular image features further improves the overall performance, particularly in recall, which increases from 0.916 to 0.935. This result suggests that molecular visual representations provide complementary stereochemical and spatial information

Table 4
Performance comparison between our method and baselines on Task 1.

Type	Method	ACC	AUPR	AUC	F1	PRE	REC
Traditional Methods	KNN	0.721	0.772	0.981	0.483	0.717	0.408
	DNN	0.880	0.913	0.996	0.722	0.805	0.703
	RF	0.778	0.835	0.996	0.594	0.789	0.516
	LR	0.792	0.840	0.996	0.595	0.744	0.524
	GNN	0.914	0.969	0.999	0.833	0.894	0.801
Methods based on drug feature representation	DeepDDI	0.837	0.890	0.996	0.685	0.728	0.661
	MDF-SA-DDI	0.930	0.974	0.999	0.888	0.909	0.876
	MCNN-DDI	0.900	0.948	0.998	0.829	–	–
	DDIMDL	0.885	0.921	0.998	0.759	0.847	0.718
	MDDI-SCL	0.938	0.978	0.998	0.876	0.880	0.877
Methods based on graph representation	MDNN	0.918	0.967	0.998	0.830	0.862	0.820
	CNN-Siam	0.924	0.963	0.999	0.924	0.924	0.924
	CNN-DDI	0.887	0.925	0.998	0.750	0.856	0.722
	DM-DDI	0.908	0.964	0.999	0.852	0.879	0.839
	NMDADNN	0.898	0.959	0.999	0.807	0.875	0.781
	MOLGAECL	0.930	–	–	0.852	0.871	0.850
Methods based on hybrid representation	MFDA	0.902	0.963	0.998	0.851	0.857	0.853
	MKG-FENN	0.941	0.979	0.999	0.896	0.913	0.888
	Lee et al.	0.909	0.956	0.996	0.839	0.851	0.834
	MSDF	0.943	0.981	0.999	0.863	0.889	0.854
Proposed	DSMV-DDI (w/o image)	0.961	0.989	0.999	0.929	0.945	0.916
	DSMV-DDI (w/ image)	**0.967**	**0.992**	**0.999**	**0.936**	**0.945**	**0.935**

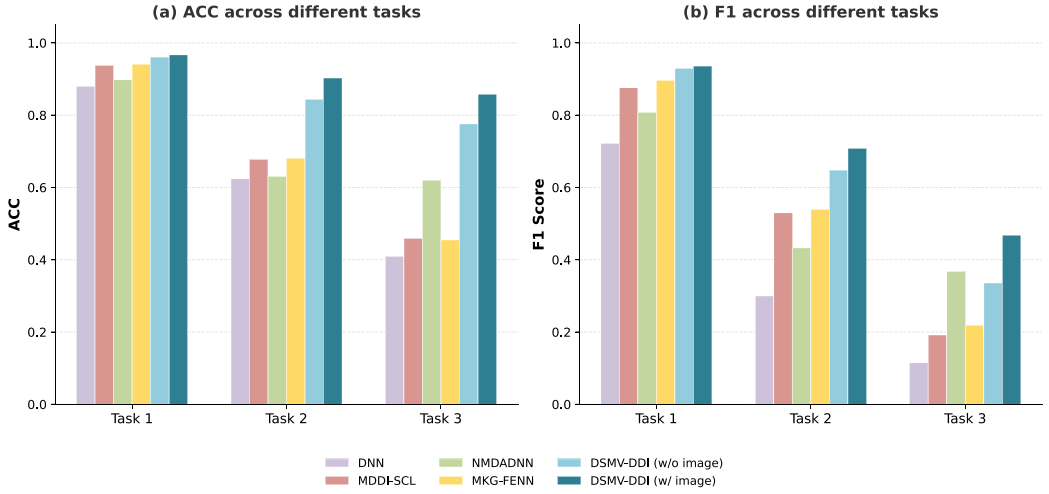

Fig. 4. Performance comparison across three tasks. The bar chart shows ACC and F1 scores of DSMV-DDI on Task 1, Task 2, and Task 3, demonstrating the model's robust performance across conventional and cold-start scenarios.

Table 5
Additional five-fold cross-validation stability analysis of DSMV-DDI on Task 1.

Metric	Mean ± Std
Accuracy	0.9647 ± 0.0019
Macro-F1	0.9318 ± 0.0143
Macro-Precision	0.9394 ± 0.0204
Macro-Recall	0.9329 ± 0.0119

that is difficult to capture using graph- or sequence-based representations alone. Overall, these results validate the effectiveness of integrating structural topology, semantic knowledge, chemical substructures, and molecular visual representations for comprehensive DDI prediction.

4.6. Comparison on novel drugs

To further evaluate the generalization capability of DSMV-DDI under realistic cold-start scenarios, experiments were conducted on Task 2 and Task 3, where one or both drugs in the test set are unseen during training. As illustrated in Fig. 4, DSMV-DDI maintains strong predictive performance across increasingly challenging settings. The quantitative results are summarized in Table 6.

DSMV-DDI consistently outperforms all competing methods under both partial and complete cold-start conditions, demonstrating superior robustness for unseen drug interaction prediction. Specifically, the full model achieves ACC scores of 0.903 and 0.858 on Task 2 and Task 3, respectively. Compared with the strongest baseline MKG-FENN, DSMV-DDI improves ACC by 22.2% on Task 2 and 40.3% on Task 3, indicating a substantially stronger generalization capability when interaction topology information is limited or unavailable.

Most baseline methods exhibit severe performance degradation as the prediction scenario shifts from Task 2 to Task 3. For example, the ACC of MKG-FENN decreases from 0.681 to 0.455, whereas DSMV-DDI only declines from 0.903 to 0.858. This observation suggests that many existing methods rely heavily on previously observed interaction patterns and biomedical associations, which become increasingly sparse under complete cold-start settings.

Table 6
Performance comparison under cold-start settings (Task 2 and Task 3).

Method	Task 2						Task 3					
	ACC	AUPR	AUC	F1	PRE	REC	ACC	AUPR	AUC	F1	PRE	REC
MDDI-SCL	0.677	0.695	0.963	0.530	0.625	0.481	0.459	0.394	0.905	0.192	0.259	0.168
MDF-SA-DDI	0.663	0.678	0.950	0.558	0.655	0.508	0.434	0.387	0.863	0.233	0.272	0.223
DDIMDL	0.642	0.656	0.980	0.446	0.561	0.432	0.408	0.364	0.951	0.159	0.241	0.145
NMDADNN	0.630	0.662	0.981	0.433	0.541	0.382	0.620	0.650	0.982	0.367	0.396	0.401
Lee et al.	0.641	0.624	0.925	0.504	0.539	0.489	0.410	0.318	0.830	0.202	0.222	0.203
DeepDDI	0.577	0.559	0.958	0.342	0.363	0.389	0.360	0.278	0.906	0.137	0.159	0.145
DNN	0.624	0.636	0.980	0.300	0.424	0.284	0.409	0.378	0.955	0.115	0.184	0.109
MKG-FENN	0.681	0.705	0.967	0.539	0.606	0.511	0.455	0.416	0.915	0.219	0.275	0.213
DSMV-DDI (w/o image)	0.844	0.898	0.993	0.648	0.735	0.604	0.776	0.829	0.986	0.336	0.403	0.320
DSMV-DDI (w/ image)	**0.903**	**0.956**	**0.997**	**0.708**	**0.765**	**0.685**	**0.858**	**0.916**	**0.994**	**0.468**	**0.513**	**0.455**

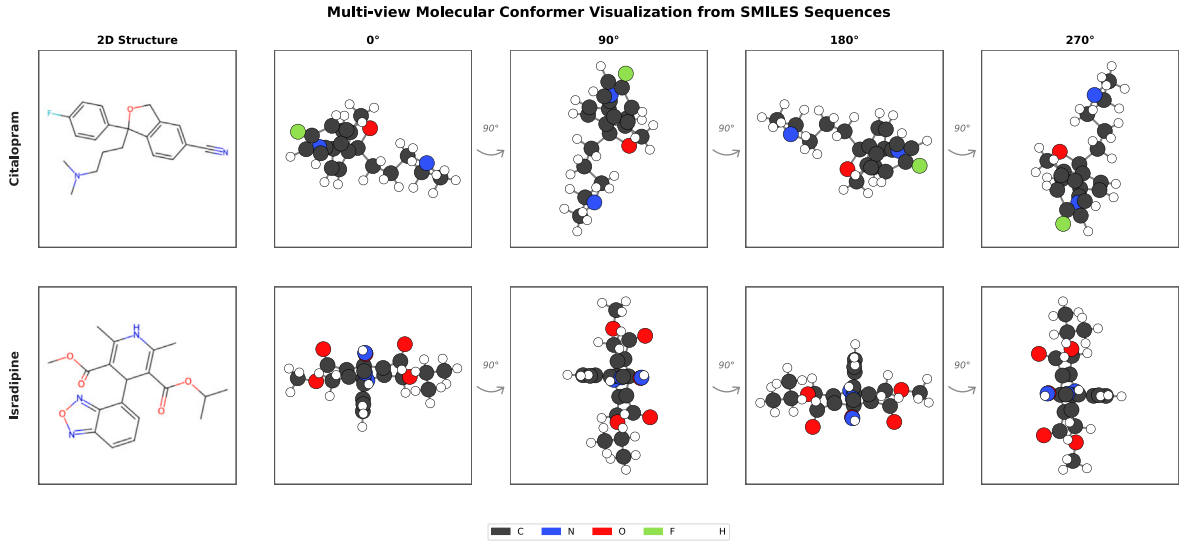

Fig. 5. Multi-view molecular conformer visualization for two representative DrugBank drugs. From left to right: 2D molecular structure, and four orthogonal views at 0°, 90°, 180°, and 270° around the Y-axis.

Molecular visual representations contribute increasingly larger performance gains as structural interaction information becomes less accessible. Compared with the w/o-image variant, the full model improves ACC by 5.9% on Task 2 and 8.2% on Task 3. Similar trends can also be observed for F1 score and AUPR, highlighting the complementary value of molecular visual information for unseen drug prediction.

4.7. Molecular conformational analysis

To better understand the stereochemical information encoded by the molecular image modality, Fig. 5 visualizes the multi-view conformer rendering pipeline for two representative DrugBank drugs, Citalopram and Isradipine. For each drug, the energy-minimized 3D conformer is rendered from four orthogonal viewpoints at 0°, 90°, 180°, and 270° around the Y-axis, corresponding to the same rendering protocol used in the image encoding module.

Several observations support the chemical validity of the image modality. First, the orthogonal views reveal complementary stereochemical characteristics that a single 2D representation cannot simultaneously capture. For Citalopram, the overall molecular shape and the spatial distribution of heteroatoms, including the nitrogen and fluorine atoms, vary substantially across the four viewpoints, revealing three-dimensional structural information that is absent from the planar 2D representation. For Isradipine, the distinct spatial arrangement of oxygen-rich functional groups relative to the carbon scaffold becomes progressively apparent as the viewing angle rotates, demonstrating that multiple orthogonal views together convey conformational information that no single projection can fully represent. Second, the ball-and-stick rendering explicitly preserves atomic positions and bond geometry in three-dimensional space, making the spatial relationships between atoms and functional groups directly visible across different viewpoints—information that is abstracted away in two-dimensional graph representations. Third, the multi-view strategy ensures that the ResNet50 encoder receives spatially comprehensive molecular information, mitigating the loss of stereochemical detail that would occur if only a single projection were used.

These visualizations qualitatively illustrate the stereochemical information captured by the molecular image modality and provide an intuitive interpretation of the proposed multi-view representation. To further evaluate the effect of incorporating multiple molecular conformers, we constructed a multi-conformer representation using an energy-based conformer selection strategy. For each drug molecule, multiple candidate conformations were generated using a standard conformer generation and energy minimization procedure. The Top-3 lowest-energy conformers were selected, and each conformer was rendered from four predefined viewpoints. We then compared the single-conformer representation with the multi-conformer representation. As shown in Table 7, the two configurations achieve comparable performance across all evaluation metrics, with minor differences observed in individual metrics.

4.8. Ablation study

To systematically evaluate the contribution of each modality, ablation experiments were conducted under identical training and testing

Table 7
Performance comparison between single-conformer and multi-conformer molecular image representations.

Method	ACC	AUPR	AUC	F1	PRE	REC
Single	0.967	0.992	0.999	0.936	0.945	0.935
Top-3	0.966	0.991	0.999	0.937	0.943	0.935

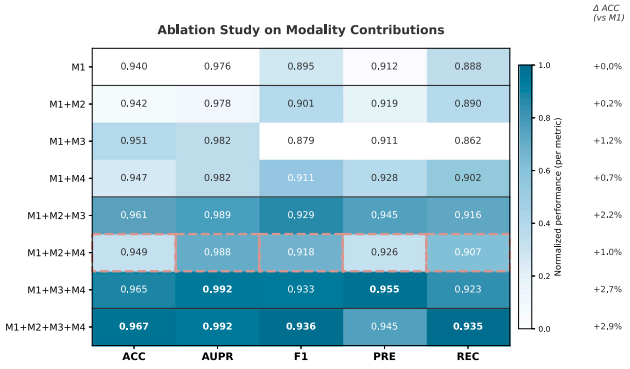

Fig. 6. Modality ablation under cold-start settings. M1, M2, M3, and M4 denote the multi-source knowledge graph structural, chemical substructure, dual-level semantic, and molecular visual modalities, respectively.

settings. Fig. 7 presents the performance of different modality combinations under the conventional setting (Task 1), Fig. 6 shows the modality ablation results under the two cold-start settings (Task 2 and Task 3), and Fig. 8 compares several pre-trained language models used as the semantic encoder.

The knowledge graph structural modality alone achieves 0.940 ACC, confirming the effectiveness of topology-based representations for DDI prediction. Among additional modalities, the semantic modality provides the largest performance gain, highlighting the importance of dual-level semantic enhancement for capturing pharmacological interaction information. The molecular image modality contributes the second-largest improvement, providing complementary stereochemical and spatial information beyond graph topology.

In contrast, the chemical substructure modality introduces relatively limited improvements, reflecting partial redundancy with the drug-substructure channel already incorporated in the knowledge graph structural encoding. Nevertheless, the full four-modality framework still achieves the best overall performance, demonstrating the effectiveness of multimodal feature integration.

To further evaluate the contributions of different modalities under cold-start settings, modality ablation experiments were additionally performed on Task 2 and Task 3, as shown in Fig. 6. The results show that the prediction performance consistently improves as additional modalities are incorporated under both tasks. Compared with the three-modality configuration, incorporating the molecular image modality (M4) further improves all evaluation metrics in both Task 2 and Task 3, demonstrating its consistent contribution under cold-start settings.

Since the semantic modality is identified as the most influential component, we further investigate the impact of different pre-trained language models. As shown in Fig. 8, BERT-base-uncased [37] achieves the best overall performance (ACC = 0.967), followed by BioBERT-base-cased [55] and RoBERTa-base [56], while Qwen1.5-1.8B performs comparatively worse. This result indicates that interaction-oriented fine-tuning effectively adapts general-domain language representations to DDI semantics, whereas generation-oriented large language models are less suited to discriminative interaction feature extraction.

To further validate the proposed fusion strategy, we compared it with a simple feature concatenation baseline, as shown in Table 8. The proposed strategy consistently outperforms simple concatenation across all evaluation metrics, demonstrating the effectiveness of the proposed multimodal fusion mechanism.

Table 8
Comparison of the proposed multimodal fusion strategy with simple feature concatenation.

Method	ACC	AUPR	AUC	F1	PRE	REC
Concat	0.963	0.991	0.999	0.933	0.942	0.921
Proposed	0.967	0.992	0.999	0.936	0.945	0.935

Overall, the ablation results validate the complementary contributions of the proposed multimodal representations, the effectiveness of the dual-level semantic enhancement strategy, and the superiority of the proposed multimodal fusion strategy.

4.9. Parameter sensitivity analysis

To evaluate the robustness of DSMV-DDI, parameter sensitivity analyses were conducted on two task-specific hyperparameters introduced by the proposed framework: semantic description token length and multimodal fusion projection dimension. Since the GNN architecture and the ResNet50 backbone were fixed throughout all experiments as predefined model components, they were not included in the sensitivity analysis. The corresponding results are shown in Fig. 9.

Token Length Sensitivity.
Fig. 9(a) shows that DSMV-DDI maintains relatively stable performance across different token lengths, indicating that the proposed semantic representation framework is not overly sensitive to textual granularity. The best performance is achieved at 100 tokens, suggesting that moderate-length descriptions provide an effective balance between semantic informativeness and redundancy. When the token length increases to 512, the performance slightly declines. Notably, the model remains highly competitive within the 64–256 range, demonstrating the robustness of the proposed dual-level semantic enhancement strategy.

Projection Dimension Sensitivity.
Fig. 9(b) presents the impact of different projection dimensions on multimodal fusion performance. Compared with direct feature concatenation, projection-based fusion significantly improves F1-score and recall, highlighting the importance of modality-specific feature alignment. The performance steadily improves as the projection dimension increases from 128 to 512, while larger dimensions provide diminishing returns and slightly degrade performance at 2048. Overall, the 512-dimensional projection achieves the best trade-off between representation capacity, computational efficiency, and generalization ability.

Overall, the parameter sensitivity analyses demonstrate that DSMV-DDI exhibits strong robustness and stable performance across different hyperparameter settings.

4.10. Case study on representative DDI types

To provide a more granular performance analysis beyond aggregate metrics, Table 9 reports the per-class prediction results on ten representative DDI types. All 65 categories were first sorted in descending order by sample count and assigned sequential class IDs from 0 to 64. Representative types were then selected at regular intervals from this ranked list, ensuring comprehensive coverage across the full spectrum of data frequencies, from high-prevalence to extremely rare interaction types.

The selected types cover a wide range of sample frequencies, from metabolism decrease (n = 9810) to hyperglycemic activities increase (n = 5). On high-frequency types, the model achieves excellent performance with F1 scores exceeding 0.97, demonstrating reliable prediction for common pharmacokinetic events such as metabolism alterations

Table 9
Performance of DSMV-DDI on representative DDIs in Task 1.

DDI Type	Precision	Recall	F1	Counts
The metabolism decrease	0.976	0.974	0.975	9810
The hypotensive activities increase	0.970	0.981	0.975	1086
The serotonergic activities increase	0.964	0.983	0.973	188
The sedative activities decrease	0.970	0.971	0.969	102
The risk or severity of hypotension increase	0.929	0.943	0.936	70
The anticoagulant activities decrease	0.964	0.981	0.972	55
The vasoconstricting activities decrease	0.950	0.950	0.951	40
The risk or severity of rhabdomyolysis increase	0.722	0.867	0.788	15
The hyponatremic activities increase	1.000	1.000	1.000	9
The hyperglycemic activities increase	0.833	1.000	0.909	5

and serum concentration changes. Performance remains robust on moderately frequent types, with F1 scores consistently above 0.93 across diverse pharmacological effects including hypotensive, serotonergic, sedative, and anticoagulant activities. For DDI types with limited training samples, the results exhibit greater variability. Some rare types achieve perfect precision and recall (hyponatremic activities increase, $n = 9$). Others show moderate performance (rhabdomyolysis increase, $n = 15$, F1 0.788), indicating that certain rare interactions share overlapping characteristics with more prevalent types, making them harder to distinguish without sufficient data.

The per-class results reveal consistent patterns across different pharmacological categories. High-frequency DDI types associated with metabolic alterations, such as metabolism decrease (F1 0.975), achieve the strongest predictive performance. For pharmacological activity modulation types, such as hypotensive activities increase (F1 0.975) and serotonergic activities increase (F1 0.973), strong performance is also maintained, reflecting the framework's capacity to characterize diverse pharmacological interaction mechanisms beyond metabolic pathways. This consistency across pharmacological categories confirms that the multimodal fusion strategy effectively integrates complementary molecular representations for robust DDI prediction across a broad spectrum of interaction types.

Despite the overall strong performance across representative DDI types, prediction on low-frequency categories remains more challenging because of the highly imbalanced class distribution.

To further investigate the impact of long-tail optimization on low-frequency DDI prediction, we evaluated several representative loss functions, including Focal Loss (FL), Tailed Focal Loss (TFL), and Class-Balanced Loss (CB). As shown in Fig. 10, the standard cross-entropy (CE) loss achieves the best overall performance under the current benchmark, while the alternative loss functions produce comparable but slightly lower results across the evaluated metrics.

4.11. Visual explanations for DSMV-DDI

To gain deeper insights into the learned representations, we performed both qualitative and quantitative analyses under different modality configurations. As shown in Fig. 11, panels (a)–(c) present the 2D PCA projections of the learned drug-pair representations for the M1-only, w/o M3, and full DSMV-DDI models, while panel (d) reports the average inter-class centroid distance and intra-class dispersion computed in the original embedding space. In the PCA visualizations, the top ten most frequent DDI types are highlighted with distinct colors, whereas all remaining classes are shown in gray.

As shown in Fig. 11(a)–(d), the M1-only model exhibits substantial overlap among different DDI types, while removing the semantic modality (w/o M3) results in improved but still mixed class distributions. In contrast, the full DSMV-DDI model presents more clearly separated clusters in the PCA visualization. The quantitative analysis further shows that the full model achieves a larger average inter-class

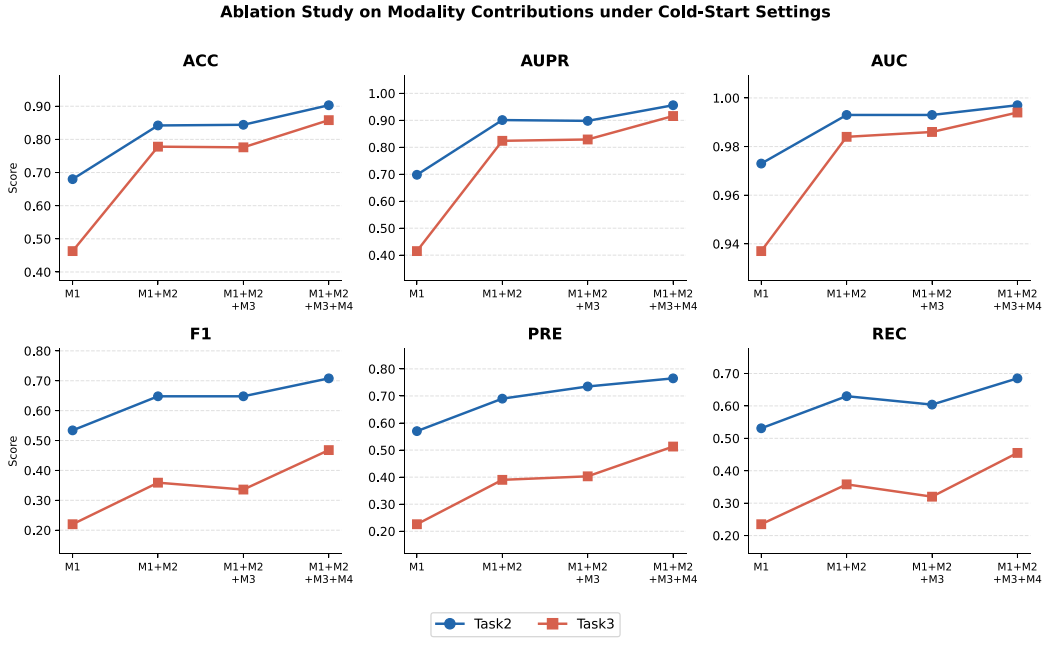

Fig. 7. Modality contribution analysis. M1, M2, M3, and M4 denote the multi-source knowledge graph structural, chemical substructure, dual-level semantic, and molecular visual modalities, respectively.

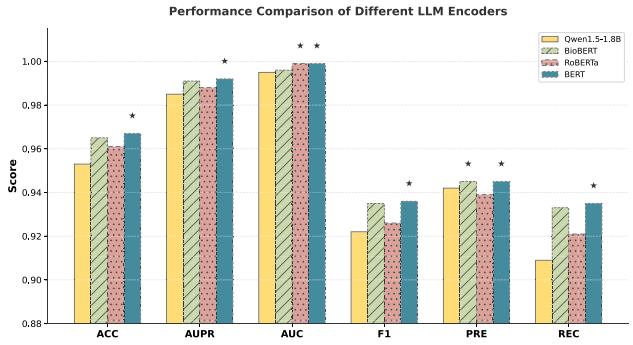

Fig. 8. Semantic encoder comparison. Performance of different pre-trained language models as semantic encoders under the full-modality setting.

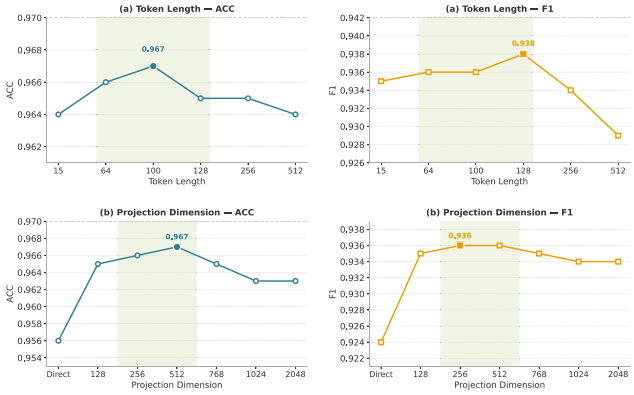

Fig. 9. Parameter sensitivity analysis. (a) Effect of description token length on ACC and F1. (b) Effect of projection dimension on ACC and F1.

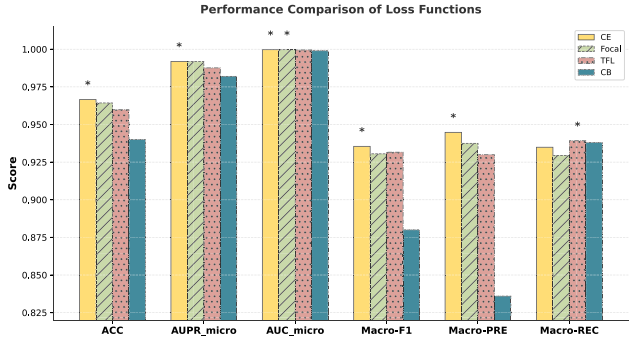

Fig. 10. Performance comparison of different loss functions for long-tail DDI prediction.

centroid distance together with a smaller average intra-class dispersion than the corresponding ablated configurations, providing complementary quantitative measurements of representation separability and compactness.

5. Conclusion and future work

DSMV-DDI demonstrates that integrating dual-level pharmacological semantics with stereochemical molecular visual representations substantially improves drug–drug interaction prediction under both conventional and cold-start settings. Experimental results demonstrate that the proposed multimodal framework effectively leverages complementary structural, chemical, semantic, and visual information to achieve consistently superior predictive performance. Extensive analyses further confirm the complementary contributions of different modalities and the effectiveness of the proposed multimodal representation learning framework. These findings highlight the importance of integrating complementary pharmacological and stereochemical information for reliable DDI prediction, particularly under realistic clinical scenarios involving previously unseen drugs.

Beyond improving DDI prediction performance, DSMV-DDI provides a unified multimodal representation learning paradigm that jointly models complementary pharmacological semantics and stereochemical molecular characteristics. This framework may also provide useful insights for developing multimodal learning approaches for other biomedical prediction tasks involving heterogeneous molecular information.

Although DSMV-DDI achieves strong performance under both conventional and cold-start settings, several challenges remain. Prediction performance on extremely low-frequency DDI types can still be improved. In addition, the current molecular image representation is based on a fixed conformer generation strategy and may not fully exploit the conformational diversity of flexible drug molecules.

Future work will investigate adaptive multi-conformer representation learning, more effective multimodal fusion and representation learning strategies, handling of long-tail and rare DDI types through imbalance-aware learning techniques, and evaluation on more diverse benchmark datasets and broader drug safety assessment tasks involving large-scale polypharmacy, thereby facilitating practical clinical decision support.

CRediT authorship contribution statement

Fangni Chen: Writing – original draft, Validation, Software, Methodology, Investigation, Data curation. **Tingting Jiang:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Shuai Yang:** Writing – review & editing, Visualization, Software, Resources. **Xiaohui Yuan:** Writing – review & editing, Supervision, Resources. **Chao Wang:** Writing – review & editing, Supervision, Resources. **Lichuan Gu:** Writing – review & editing, Supervision, Resources, Funding acquisition.

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

This work was partly supported by the National Natural Science Foundation of China (no 32501770, no 32472007, no 62306008).

Data availability

Codes and datasets of DSMV-DDI are available at https://github.com/CvChen-111/dsmv_ddi_model.

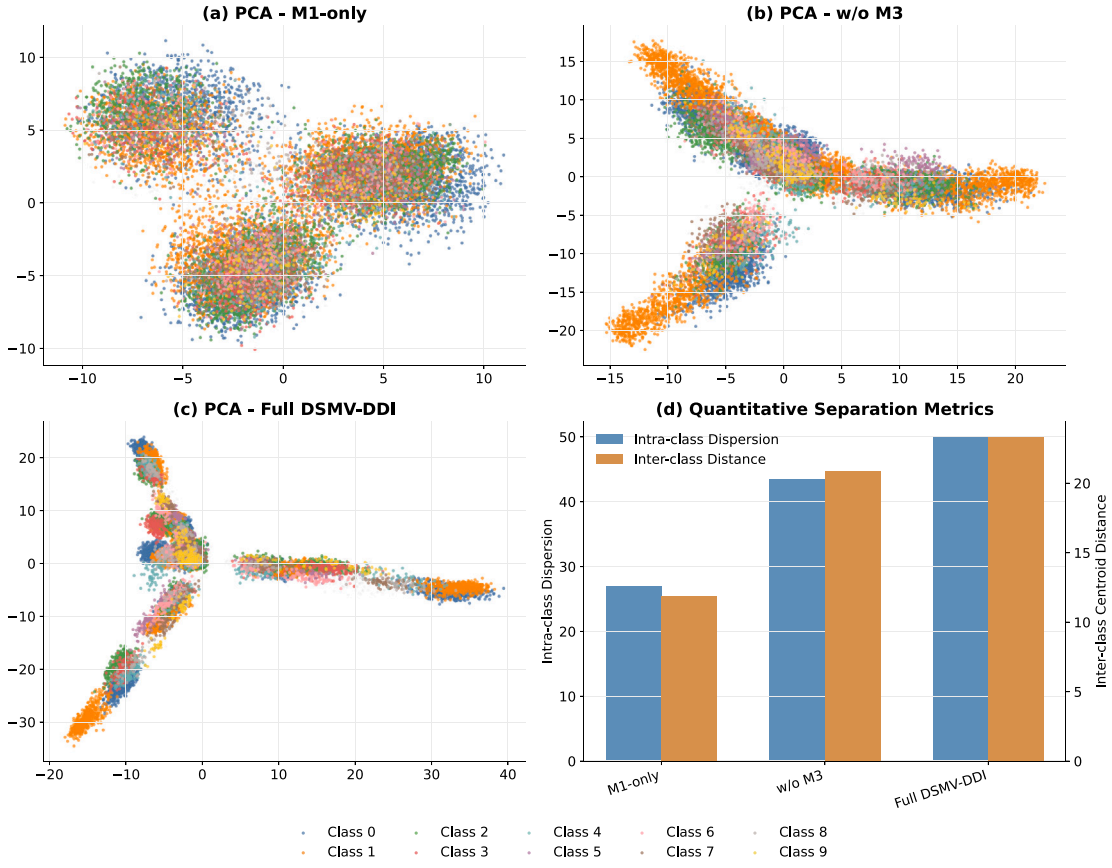

Fig. 11. Qualitative and quantitative representation analysis of different modality configurations. Panels (a)–(c) show PCA visualizations, while panel (d) reports the average inter-class centroid distance and intra-class dispersion computed in the original embedding space.

References

[1] B. Percha, R.B. Altman, Informatics confronts drug–drug interactions, Trends Pharmacol. Sci. 34 (3) (2013) 178–184, http://dx.doi.org/10.1016/j.tips.2013.01.006.

[2] M. Zitnik, M. Agrawal, J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 34 (13) (2018) i457–i466, http://dx.doi.org/10.1093/bioinformatics/bty294.

[3] S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman, N.P. Tatonetti, Similarity-based modeling in large-scale prediction of drug–drug interactions, Nat. Protoc. 9 (9) (2014) 2147–2163, http://dx.doi.org/10.1038/nprot.2014.151.

[4] S. Bangalore, G. Kamalakkannan, S. Parkar, F.H. Messerli, Fixed-dose combinations improve medication compliance: a meta-analysis, Am. J. Med. 120 (8) (2007) 713–719.

[5] D.N. Juurlink, M. Mamdani, A. Kopp, A. Laupacis, D.A. Redelmeier, Drug–drug interactions among elderly patients hospitalized for drug toxicity, Jama 289 (13) (2003) 1652–1658, http://dx.doi.org/10.1001/jama.289.13.1652.

[6] H. Jiang, Y. Lin, W. Ren, Z. Fang, Y. Liu, X. Tan, X. Lv, N. Zhang, Adverse drug reactions and correlations with drug–drug interactions: A retrospective study of reports from 2011 to 2020, Front. Pharmacol. 13 (2022) 923939, http://dx.doi.org/10.3389/fphar.2022.923939.

[7] Y. Qiu, Y. Zhang, Y. Deng, S. Liu, W. Zhang, A comprehensive review of computational methods for drug–drug interaction detection, IEEE/ACM Trans. Comput. Biology Bioinform. 19 (4) (2021) 1968–1985, http://dx.doi.org/10.1109/TCBB.2021.3081268.

[8] X. Lin, L. Dai, Y. Zhou, Z.-G. Yu, W. Zhang, J.-Y. Shi, D.-S. Cao, L. Zeng, H. Chen, B. Song, et al., Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction, Brief. Bioinform. 24 (4) (2023) bbad235, http://dx.doi.org/10.1093/bib/bbad235.

[9] A. Dehghan, K. Abbasi, M.R. Kazemi Najaf Abadi, Multi-DDA: drug–disease association prediction using a hybrid graph convolutional network with multi-modal drug representations, Bioinform. Adv. 6 (1) (2026) vbag034.

[10] R. Norouzi, K. Abbasi, P. Razzaghi, S. Gharaghani, MT-ConBiFormer-GPT: multi-target molecular generation for low-data drug discovery via a contrastive BiFormer-GPT architecture and curriculum learning with cross-domain generalization, Brief. Bioinform. 27 (3) (2026) bbag079.

[11] J.Y. Ryu, H.U. Kim, S.Y. Lee, Deep learning improves prediction of drug–drug and drug–food interactions, Proc. Natl. Acad. Sci. 115 (18) (2018) E4304–E4311, http://dx.doi.org/10.1073/pnas.1803294115.

[12] S. Lin, Y. Wang, L. Zhang, Y. Chu, Y. Liu, Y. Fang, M. Jiang, Q. Wang, B. Zhao, Y. Xiong, et al., MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism, Brief. Bioinform. 23 (1) (2022) bbab421, http://dx.doi.org/10.1093/bib/bbab421.

[13] S. Lin, W. Chen, G. Chen, S. Zhou, D.-Q. Wei, Y. Xiong, MDDI-SCL: predicting multi-type drug–drug interactions via supervised contrastive learning, J. Cheminformatics 14 (1) (2022) 81, http://dx.doi.org/10.1186/s13321-022-00659-8.

[14] X. Lin, Z. Quan, Z.-J. Wang, T. Ma, X. Zeng, KGNN: Knowledge graph neural network for drug–drug interaction prediction, in: IJCAI, vol. 380, 2020, pp. 2739–2745, http://dx.doi.org/10.24963/ijcai.2020/380.

[15] Y. Yu, K. Huang, C. Zhang, L.M. Glass, J. Sun, C. Xiao, SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization, Bioinformatics 37 (18) (2021) 2988–2995, http://dx.doi.org/10.1093/bioinformatics/btab207.

[16] D. Wu, W. Sun, Y. He, Z. Chen, X. Luo, Mkg-fenn: A multimodal knowledge graph fused end-to-end neural network for accurate drug–drug interaction prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (9) 2024, pp. 10216–10224, http://dx.doi.org/10.1609/AAAI.V38I9.28887.

[17] W. Li, Y. Zhou, W. Ma, W. Liang, X. Tang, MDG-DDI: multi-feature drug graph for drug–drug interaction prediction, BMC Bioinformatics 26 (1) (2025) 268, http://dx.doi.org/10.1186/s12859-025-06288-w.

[18] Y. Shi, M. He, J. Chen, F. Han, Y. Cai, SubGE-DDI: a new prediction model for drug–drug interaction established through biomedical texts and drug-pairs knowledge subgraph enhancement, PLoS Comput. Biol. 20 (4) (2024) e1011989, http://dx.doi.org/10.1371/journal.pcbi.1011989.

[19] Y. He, T. Ma, C. Li, P. Ma, H. Xiang, J. Wang, Y. Liu, B. Song, X. Zeng, ImageDDI: Image-enhanced molecular motif sequence representation for drug–drug interaction prediction, Inf. Fusion (2025) 103574, http://dx.doi.org/10.1016/j.inffus.2025.103574.
[20] Z. Pan, X. Yang, C. Wang, T. Han, Q. Zhao, Multimodal feature fusion for bone toxicity prediction and local platform, J. Chem. Inf. Model. 65 (23) (2025) 12799–12810.
[21] P. Li, S. Zhang, R. Liu, L. Shao, Q. Zhao, Dual-modal fusion based on peptide sequence and molecular graph with bidirectional cross-attention for precise peptide toxicity prediction, J. Chem. Inf. Model. 66 (6) (2026) 3091–3102.
[22] T. Han, Z. Pan, W. Ge, Q. Zhao, Multimodal multi-granularity fusion model with mamba architecture for ames mutagenicity prediction, J. Med. Chem. (2026).
[23] J. Hu, X. Yang, C. Yao, M. Zhang, S. Shen, L. Na, Q. Zhao, AMPred-MFG: Investigating the mutagenicity of compounds using motif-based graph combined with molecular fingerprints and graph attention mechanism, Interdiscip. Sci.: Comput. Life Sci. (2025) 1–17.
[24] B. Jin, J. Wang, X. Yang, L. Na, Q. Zhao, hERG-MFFGNN: An explainable deep learning model for predicting cardiotoxicity using multi-feature fusion and graph neural networks, Interdiscip. Sci.: Comput. Life Sci. (2025) 1–19.
[25] Q. Yang, X.-t. Ma, T.-f. Wang, X. Dou, M. Zhao, X.-w. Yang, DrugAgent: A multi-agent digital biosensor framework for interpretable virtual drug screening, J. Mol. Graph. Model. (2026) 109396.
[26] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, S. Liu, A multimodal deep learning framework for predicting drug–drug interaction events, Bioinformatics 36 (15) (2020) 4316–4322, http://dx.doi.org/10.1093/bioinformatics/btaa501.
[27] J.-Y. Shi, K.-T. Mao, H. Yu, S.-M. Yiu, Detecting drug communities and predicting comprehensive drug–drug interactions via balance regularized semi-nonnegative matrix factorization, J. Cheminformatics 11 (1) (2019) 28, http://dx.doi.org/10.1186/s13321-019-0352-9.
[28] L.F. Ribeiro, P.H. Saverese, D.R. Figueiredo, struc2vec: Learning node representations from structural identity, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 385–394, http://dx.doi.org/10.1145/3097983.3098061.
[29] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, Line: Large-scale information network embedding, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 1067–1077, http://dx.doi.org/10.1145/2736277.2741093.
[30] D. Wang, P. Cui, W. Zhu, Structural deep network embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1225–1234, http://dx.doi.org/10.1145/2939672.2939753.
[31] Q. Zhou, Y. Zhang, S. Wang, D. Wu, Drug–drug interaction prediction based on local substructure features and their complements, J. Mol. Graph. Model. 124 (2023) 108557.
[32] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (D1) (2018) D1074–D1082.
[33] G. Landrum, et al., RDKit: A software suite for cheminformatics, Comput. Chem. Predict. Model. (2019).
[34] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? 2018, arXiv preprint arXiv:1810.00826.
[35] T. Kipf, Semi-supervised classification with graph convolutional networks, 2016, http://dx.doi.org/10.48550/arXiv.1609.02907, arXiv preprint arXiv:1609.02907.
[36] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, Y. Sun, Masked label prediction: Unified message passing model for semi-supervised classification, 2020, arXiv preprint arXiv:2009.03509.
[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.
[38] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al., Deepseek-v3 technical report, 2024, arXiv preprint arXiv:2412.19437.
[39] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: large-scale self-supervised pretraining for molecular property prediction, 2020, arXiv preprint arXiv:2010.09885.
[40] T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, J. Comput. Chem. 17 (5–6) (1996) 490–519.
[41] J.D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95.
[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[43] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[44] M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, A.O. Hero, A deep learning architecture for metabolic pathway prediction, Bioinformatics 36 (8) (2020) 2547–2553, http://dx.doi.org/10.1093/bioinformatics/btz954.
[45] M. Asfand-E-Yar, Q. Hashir, A.A. Shah, H.A.M. Malik, A. Alourani, W. Khalil, Multimodal CNN-DDI: using multimodal CNN for drug to drug interaction associated events, Sci. Rep. 14 (1) (2024) 4076, http://dx.doi.org/10.1038/s41598-024-54409-x.
[46] T. Lyu, J. Gao, L. Tian, Z. Li, P. Zhang, J. Zhang, MDNN: A multimodal deep neural network for predicting drug–drug interaction events, in: Ijcai, vol. 3536, (3542.2021) 2021, http://dx.doi.org/10.24963/ijcai.2021/487.
[47] Z. Yang, K. Tong, S. Jin, S. Wang, C. Yang, F. Jiang, CNN-Siam: multimodal siamese CNN-based deep learning approach for drug–drug interaction prediction, BMC Bioinformatics 24 (1) (2023) 110, http://dx.doi.org/10.1186/s12859-023-05242-y.
[48] C. Zhang, Y. Lu, T. Zang, CNN-DDI: a learning-based method for predicting drug–drug interactions using convolution neural networks, BMC Bioinformatics 23 (Suppl 1) (2022) 88, http://dx.doi.org/10.1186/s12859-022-04612-2.
[49] L.-P. Kang, K.-B. Lin, P. Lu, F. Yang, J.-P. Chen, Multitype drug interaction prediction based on the deep fusion of drug features and topological relationships, Plos One 17 (8) (2022) e0273764, http://dx.doi.org/10.1371/journal.pone.0273764.
[50] X.-Y. Yan, P.-W. Yin, X.-M. Wu, J.-X. Han, Prediction of the drug–drug interaction types with the unified embedding features from drug similarity networks, Front. Pharmacol. 12 (2021) 794205, http://dx.doi.org/10.3389/fphar.2021.794205.
[51] Y. Li, L.-X. Hou, H.-C. Yi, Z.-H. You, S.-H. Chen, J. Zheng, Y. Yuan, C.-G. Mi, MOLGAECL: Molecular graph contrastive learning via graph auto-encoder pre-training and fine-tuning based on drug–drug interaction prediction, J. Chem. Inf. Model. 65 (6) (2025) 3104–3116, http://dx.doi.org/10.1021/acs.jcim.5c00043.
[52] K. Lin, L. Kang, F. Yang, P. Lu, J. Lu, MFDA: Multiview fusion based on dual-level attention for drug interaction prediction, Front. Pharmacol. 13 (2022) 1021329, http://dx.doi.org/10.3389/fphar.2022.1021329.
[53] G. Lee, C. Park, J. Ahn, Novel deep learning model for more accurate prediction of drug–drug interaction effects, BMC Bioinformatics 20 (1) (2019) 415, http://dx.doi.org/10.1186/s12859-019-3013-0.
[54] D. Pan, P. Lu, Y. Wu, L. Kang, F. Huang, K. Lin, F. Yang, Prediction of multiple types of drug interactions based on multi-scale fusion and dual-view fusion, Front. Pharmacol. 15 (2024) 1354540, http://dx.doi.org/10.3389/fphar.2024.1354540.
[55] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240, http://dx.doi.org/10.1093/bioinformatics/btz682.
[56] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, http://dx.doi.org/10.48550/arXiv.1907.11692, arXiv preprint arXiv:1907.11692.