OXFORD

Briefings in Bioinformatics, 2023, **24(6)**, 1–11
https://doi.org/10.1093/bib/bbad397
Problem Solving Protocol

DMFDDI: deep multimodal fusion for drug–drug interaction prediction

Yanglan Gan, Wenxiao Liu, Guangwei Xu, Cairong Yan and Guobing Zou

Corresponding author: Guobing Zou, School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai 200444, China.
E-mail: gbzou@shu.edu.cn

Abstract

Drug combination therapy has gradually become a promising treatment strategy for complex or co-existing diseases. As drug–drug interactions (DDIs) may cause unexpected adverse drug reactions, DDI prediction is an important task in pharmacology and clinical applications. Recently, researchers have proposed several deep learning methods to predict DDIs. However, these methods mainly exploit the chemical or biological features of drugs, which is insufficient and limits the performances of DDI prediction. Here, we propose a new **d**eep **m**ultimodal **f**eature fusion framework for **DDI** prediction, DMFDDI, which fuses drug molecular graph, DDI network and the biochemical similarity features of drugs to predict DDIs. To fully extract drug molecular structure, we introduce an attention-gated graph neural network for capturing the global features of the molecular graph and the local features of each atom. A sparse graph convolution network is introduced to learn the topological structure information of the DDI network. In the multimodal feature fusion module, an attention mechanism is used to efficiently fuse different features. To validate the performance of DMFDDI, we compare it with 10 state-of-the-art methods. The comparison results demonstrate that DMFDDI achieves better performance in DDI prediction. Our method DMFDDI is implemented in Python using the Pytorch machine-learning library, and it is freely available at https://github.com/DHUDEBLab/DMFDDI.git.

Keywords: drug–drug interaction; multimodal feature fusion; graphical neural network; attention mechanism

INTRODUCTION

Drug combination therapy is a well-established concept for the treatment of complex or co-existing diseases [1]. As drug–drug interactions (DDIs) usually occur when a drug is co-administered with another and multiple drugs, this treatment often can lead to an increased efficacy compared with single drug treatments. However, it also may increase the possibility of adverse drug reactions [2]. Therefore, to maximize synergistic benefits and minimize unexpected adverse drug reactions, it is crucial to accurately predict potential DDIs while treating complex diseases with drug combinations [3, 4].

The interactions between drugs currently are confirmed mainly through clinical trials. Although the clinical experiment has high reliability, it is usually labor-intensive, time-consuming and risky, which may cause patients to receive harmful treatments. Machine learning methods provide a new opportunity to efficiently predict DDIs. Feature similarity-based approaches assume that drugs with similar features have similar response patterns [5, 6], which mostly rely on the similarity of the drug characteristics, such as fingerprinting [7], chemical structure [8], pharmacological phenotype [9] and RNA [10]. Subsequently, multiple features are combined to improve model performance [11, 12, 13]. For example, NLLSS [3] combines three types of feature information and adopts a least squares classifier based on Laplace constraints to predict the combined efficacy of drugs. HNAI [14] analyzes four drug feature similarities and utilizes four classifiers to construct a prediction model. INDI [12] computes seven feature similarities and uses logistic regression models

Yanglan Gan is a professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. She received the Ph.D. degree in Computer Science from Tongji University in 2012, China. She has worked as a Visiting Scholar in the Department of Computer Science and Engineering at Washington University in St. Louis from 2009 to 2011, USA. Her research interests include bioinformatics, data mining, and Web services. She has published more than 40 papers on international journals and conferences, including Bioinformatics, IEEE/ACM TCBB, BMC Bioinformatics, BMC Genomics, Knowledge-based Systems, and Soft Computing. She served as a program committee member on BIBM 2021 and GIW 2018. She worked as a reviewer for a variety of international journals and conferences, such as BMC Bioinformatics, IEEE TCBB, Knowledge-based Systems.Yuhan Chen is currently a master student in the School of Computer Science and technology, Donghua University, China. Before that, he received a Bachelor degree in Henan Polytechnic University, 2021.
Wenxiao Liu is currently a master student in the School of Computer Science and technology, Donghua University, China. Before that, she received a Bachelor degree in Qufu Normal University, 2021. Her research interests include data mining and bioinformatics.
Guangwei Xu is a professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. He received the M.S. degree from Na njing University, Nanjing, China, in 2000, and the Ph.D. from Tongji University, Shanghai, China, in 2003. His research interests include data secure storage, data integrity verification and privacy protection, secure computing and sharing of outsourced data, QoS and routing of the wireless and sensor networks.
Cairong Yan is an assistant professor in the School of Computer Science and Technology at Donghua University, Shanghai, China. He received the Ph.D. degree from Xi'an Jiaotong University, Xi'an', China, in 2006. His research interests include data mining and Information Retrieval.
Guobing Zou is an Professor in the School of Computer Engineering and Science, Shanghai University, China. He received his PhD in Computer Science fromTongji University, Shanghai, China, 2012. His current research interests focus on data mining, intelligent algorithms and services computing. He has published around 70 papers on premier international journals and conferences, including Information Sciences, Expert Systems with Applications, Knowledge-based Systems, IEEE Transactions on Services Computing, AAAI, ICWS and ICSOC.
Received: March 27, 2023. **Revised:** September 28, 2023. **Accepted:** October 13, 2023

for classification. NDD [15] is a neural network-based method for predicting unknown DDIs using various information about drugs. DDI-IS-SL [16] predicts DDIs based on the integrated similarity and semi-supervised learning. Overall, these feature similarity-based methods have made great progress. However, these methods usually ignore the structural information of the drugs, and the selection of features relies on expert experience.

Recently, graph neural network (GNN) is introduced to learn the feature representation of drug chemical structures and predict DDIs. Current GNN-based approaches broadly fall into two categories. The first category of algorithms learns embedded features from drug molecular graphs, adopting a natural and effective approach to modeling graph structure data [17]. Atoms are represented as nodes in the drug molecular graph, and the edges are chemical bonds. Then the drug molecular graph is embedded by learning the atom features and the information passed over the chemical bonds. The other category of solutions takes advantage of the known interaction network of drugs. In DDI networks, by regarding drugs as nodes and interactions as edges, DDI prediction can be regarded as a link prediction task. MR-GNN [18] and GoGNN [19] exploit the powerful feature extraction capability of GNN to directly learn the embedding of the molecular structure of a drug. DPDDI [20] applies a deep graph autoencoder to learn potential representations of drugs from DDI network, and then feeds back the learned embedding to a deep feed-forward neural network for DDI prediction. CASTER [21] is an end-to-end model that predicts DDIs using substructure information extracted from drug SMILES strings to generate representations. KGNN [22] is based on a knowledge GNN to mine the relationships in knowledge graphs to solve the DDI prediction problem. Bi-GNN [23] utilizes Bi-layer GNN to solve the bio-link prediction task and learn drug feature representations, which is further used to predict DDIs. MIRACLE [24] predicts DDI through multi-view graph contrastive representation learning, which captures inter-view molecule structure and intra-view interactions between molecules simultaneously.

Despite the above models achieving remarkable results, they still have some limitations. Firstly, the existing methods are mainly based on biological, chemical or drug interaction features, seldom focusing on potential correlations between other multimodal features and DDI events. Second, information on distant atom pairs as well as chemical bonds in drug molecular graphs is rarely considered, even though they may exhibit important interactions in the molecule. Third, as drug features may contain redundant information, the traditional strategies of combining feature vectors neglect the different importance of multiple drug features to the drug representation.

To address the aforementioned limitations, we propose a new deep multimodal feature fusion framework for DDI prediction, named DMFDDI, which fuses drug molecular graphs, the DDI network and the biochemical similarity features of drugs to predict DDIs. The proposed framework is composed of three encoders for multimodal features, a feature fusion model and a predictor model. These three encoders, respectively, learn the low-dimensional embedded vectors for drug molecular graphs, the DDI network and biochemical similarity features. Specifically, based on the attention-gated graph neural network (AGRUNN), we extract the molecular structure from the given drug molecular graphs. A sparse graph convolution network (GCN) is introduced to learn the topological structure information of the DDI network. We extract the biochemical feature similarity by analyzing three feature similarity matrices. Then the learned different feature embedding of drugs are fused by a multimodal feature fusion module. Finally, we predict the scores of DDIs by Multi-Layer Perception (MLP). Extensive experiments on different scale datasets demonstrate that DMFDDI outperforms the 10 competing methods. Furthermore, detailed ablation experiments demonstrate that the AGRUNN and the multimodal feature fusion model play a key role in improving the prediction performance.

MATERIALS AND METHODS

Problem Definition

We formulate the DDI prediction task as a binary classification problem of determining whether two drugs interact with each other. Assuming that the drug interaction graph is represented by $N = (D, I)$, where D denotes the drug node set, and I denotes the edge set of DDIs represented by an adjacency matrix A. The drug d_i is further represented by the corresponding molecular structure graph $G = (V, E)$, where V denotes the atom nodes and E denotes the chemical bonds connecting pairs of atoms. We learn drug features from three perspectives, including drug molecular structure based on drug molecular graphs, drug relation features based on DDI network and biological features based on biological feature similarity matrices. Specifically, for the biological feature similarity, we choose three widely used features (target, enzyme and transporter) for similarity calculation [25]. The three biological features are, respectively, represented by drug–drug feature similarity matrices, target X_t, enzyme X_e and transporter X_p.

Given the adjacency matrix A, the molecular structure graph G and the three characteristic similarity matrices, DDI prediction problem is aimed to learn a prediction function $f\left(d_i \times d_j\right) \rightarrow [0, 1]$, to determine the probability of any two drugs interacting with each other.

Overview of DMFDDI

We propose DMFDDI, a new deep multimodal feature fusion framework to predict DDIs. As shown in Figure 1, DMFDDI consists of three main steps, including learning multimodal drug features based on three feature extraction modules, fusing learned feature embeddings and predicting the DDIs based on the predictor. The three feature extraction modules learn the low-dimensional embedding vectors of different drug features, including molecular structure-based feature extraction module, drug interaction network-based feature extraction module and biological similarity-based feature extraction module. Then the multimodal feature fusion module fuses the three aspects of drug feature embeddings. Finally, the fused feature vector is fed into the predictor to predict the scores of the DDIs.

Feature extraction modules

Feature extraction module based on drug molecular graph

The drug molecular graph is an important molecular structure representation of drugs. As illustrated in Figure 2, to learn the molecular structure of drugs, we construct a feature extraction module based on drug molecular graphs. Firstly, from SMILES sequences, the molecular graphs of drugs are obtained by RDKit tool [26]. Subsequently, local information of the drug molecular graph is extracted with updated node feature and chemical bond feature. Then the updated graph structure is fed into the AGRUNN. Finally, readout part obtains an embedding based on the molecular graph and the extracted key nodes.

Extracting local features of drug molecular structure. Firstly, the local feature information of the drug molecular graph is

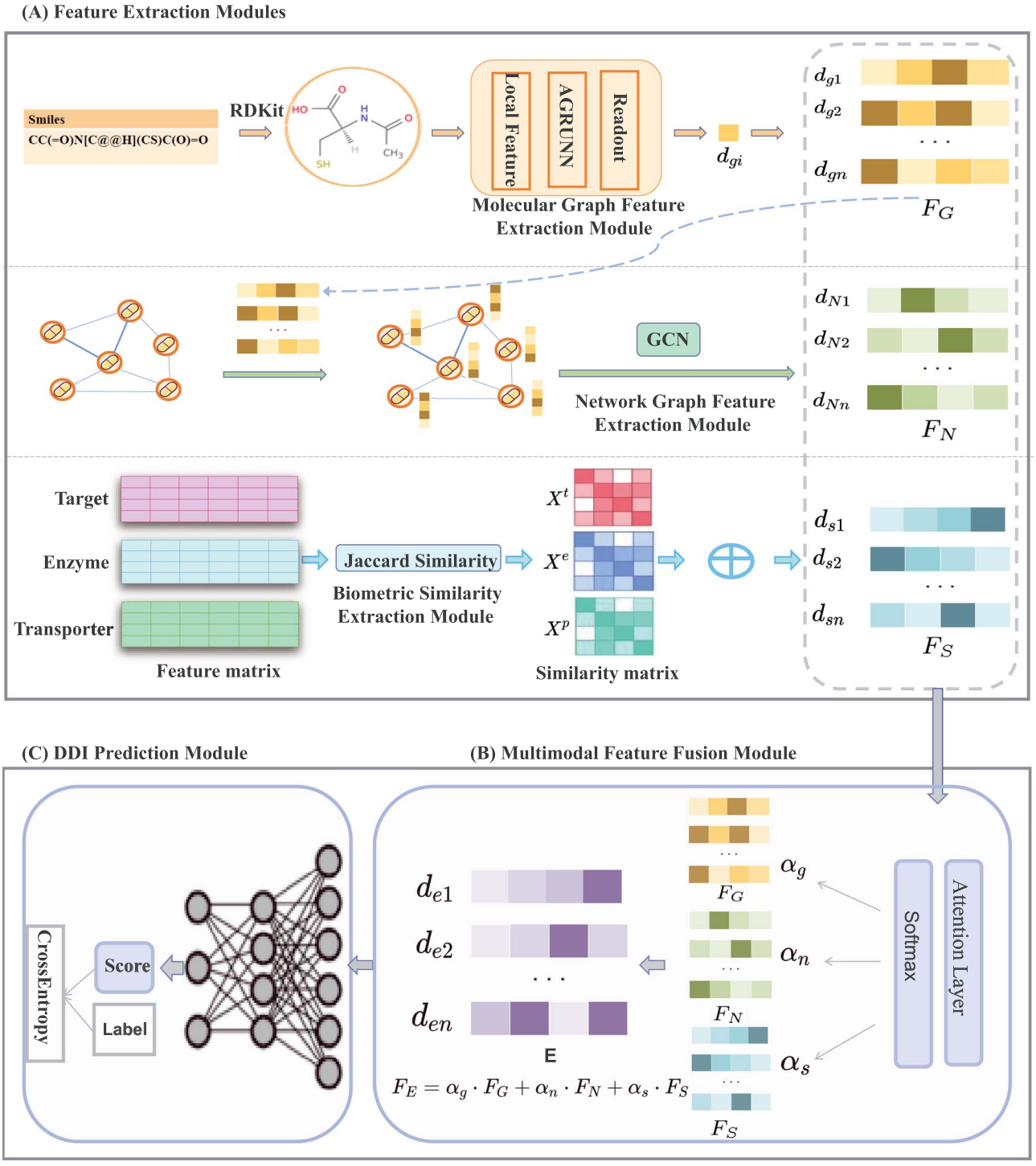

Figure 1. The framework of DMFDDI. DMFDDI consists of three major steps, including extracting three types of drug features, fusing multimodal drug features and predicting DDIs. (**A**) Three core feature extraction modules: feature extraction module based on the AGRUNN, extracting the molecular structure from drug molecular graph; feature extraction module based on sparse GCN, extracting drug relation features from drug interaction network graph; feature extraction module based on biochemical feature similarity matrices, extracting the biochemical feature similarities. (**B**) The multimodal fusion layer fuses the feature information of three feature extraction modules. (**C**) The predictor obtains the scores of drug interactions.

extracted. The initial feature $\bar{v}_i$ of each atom is obtained based on the atom symbol, formal charge, hybridization, chirality, etc. The initial feature $\bar{e}_{ij}$ of each bond is obtained based on the bond type, whether the bond is cyclic and conjugated, etc. To fully extract the features of atoms and chemical bonds, we first update the edges with the chemical bond features first and then the atom node features.

The features of any edge e_{ij} are updated according to the features of the nodes connected to the edge e_{ij}, and the updating process is defined as

$$e_{ij}^{(l)} = \mathrm{ReLU}\left[\left(\bar{e}_{ij}^{(l-1)} \left\| \bar{v}_i^{(l-1)} \right\| \bar{v}_j^{(l-1)}\right) w_e^{(l-1)} + b_e^{(l-1)}\right], \quad (1)$$

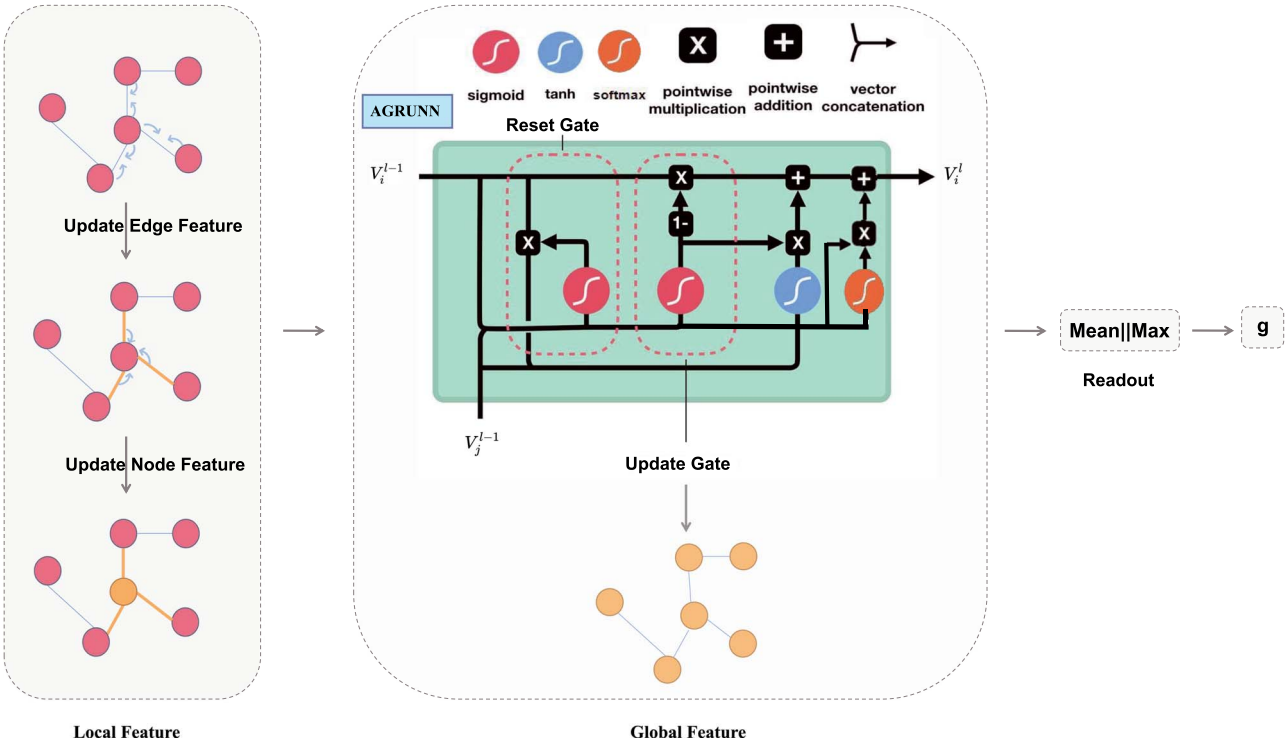

Figure 2. The process of learning the drug molecular structure representation from the drug molecule graph.

where ‖ is the splicing operation, $\tilde{v}_i$ is the initial characteristic of each atom, $W_e^{(l-1)}$ and $b_e^{(l-1)}$, respectively, represent the learnable weight matrix and the offset.

The feature of any node v_i is updated according to the features of connecting edges and the node itself, and the updating process is formalized as

$$\tilde{v}_i^{(l)} = \mathrm{ReLU}\left[\left(v_i^{(l-1)} \| \sum_{j \in M(i)} e_{ij}^{(l)}\right) W_v^{(l-1)} + b_v^{(l-1)}\right], \tag{2}$$

where $M(i)$ denotes the neighbors of node i. The atom features containing chemical bond information are updated to $\tilde{v}_i^{(l)}$. Then, the local information of drug molecular graph is extracted with the updated node feature and chemical bond feature.

Attention gated graph neural network module. Next, based on the drug molecular graph with updated node and chemical bond feature, we further construct an AGRUNN to aggregate and update feature information among different nodes through a gating mechanism.

In drug molecules, topologically distant pairs of atoms may also exhibit important interactions affecting the properties of the entire molecule, and the atoms may differently contribute to the representation of a drug molecule. A drug usually depends on a number of key substructures to interact with other drugs. Compared with the basic GNN model, the gated recurrent unit (GRU) model can capture the importance of substructures and higher order neighborhood information by adding additional control units (update gate, reset gate) and fusing them layer by layer [27]. Through effectively modeling the diffusion process of node information in the graph and improving long-term information propagation, the GRU model can extract local structural features and additional effective features, making the final drug feature embedding more accurate and representative. Meanwhile, to retain important feature information between nodes and eliminate redundant information, we add attention gates to automatically learn the flow of feature information between atomic nodes. The attention mechanism can improve the rationality and completeness of the node feature updating process. Specifically, the attention score determines the amount of feature information in each neighbor node that can be integrated into the current node. For each node, the attention score of its neighboring nodes is calculated as

$$att_i^l = \mathrm{softmax}\left(\tilde{v}_i^{(l-1)}\right) = \frac{\exp\left(\tilde{v}_i^{(l-1)} W\right)}{\sum_{j \in \mathcal{M}} \exp\left(\tilde{v}_j^{(l-1)} W\right)}, \tag{3}$$

where W denotes the linear transformation matrix, M denotes the set of current nodes and their first-order neighbors and the softmax function is used to normalize the attention vector.

Then, the feature information of each node is weighted and aggregated according to the attention scores of its neighbors:

$$\hat{v}_i^l = \sum_{j \in M} att_j^l \tilde{v}_j^{l-1}, \tag{4}$$

where M denotes the set of the current node and its first-order neighbors, and att_j^l is the attention score of the neighbor v_j.

With the attention gate, the AGRUNN is updated as

$$a^l = \hat{A}_g^{l-1} \tilde{v}_i^{l-1} W_a \tag{5}$$

$$Z = \sigma\left(W_z a^l + U_z \tilde{v}_i^{l-1} + b_z\right) \tag{6}$$

$$R = \sigma\left(W_r a^l + U_r \tilde{v}_i^{l-1} + b_r\right) \tag{7}$$

$$h^l = \tan\left(W_h a^l + U_h\left(R \odot \tilde{v}_i^{l-1}\right) + b_h\right) \tag{8}$$

$$v_i^l = h^l \odot Z + \tilde{v}_i^{l-1} \odot (1 - Z) + \hat{v}_i^l, \tag{9}$$

where Z and R, respectively, represent the update gate and reset gate, W and U are trainable weight matrices, b is the bias vector, σ represents the sigmoid activation function, $\hat{A}_g^{l-1}$ is the symmetric normalized adjacency matrix of the drug molecular graph and $\odot$ denotes the dot product operation.

The hidden state v_i^{l-1} of atom i is updated to v_i^l after AGRUNN. In this process, each node aggregates the feature information of neighboring nodes using update and reset gates, in addition to the semantic information associated with the current node through attention gates. Accordingly, the features of the atoms in the molecular graph are updated to v_i^l.

Readout. After updating the bond and atom information, the graph readout part is introduced to obtain an embedded representation of the molecular graph. To integrate the information of the whole drug molecular graph and highlight the features of important nodes, readout adopts the mean and the max function to integrate the feature representation of each key node as below:

$$g = \frac{1}{n}\sum_{i=1}^{n} v_i^l \| \max v_i^l, \tag{10}$$

where $\|$ denotes the splicing operation, and n is the number of atoms in the molecular graph. Then, the molecular graph-based feature matrix $F_G \in R^{N\times d_g}$ consisting of N drug molecular graph features is obtained.

Feature extraction module based on DDI networks

Based on the initial DDI network and the above-obtained drug molecular structure feature F_G, we construct a GCN to extract the drug relation feature. To handle the sparsity of DDI networks, we adopt sparse GCN [28] and sparse matrix multiplication (SpMM) in the module. Sparse GCN is a variant of graph convolution neural networks, which is suitable for the situations when there is a large amount of missing information in the graph. SpMM is a way of multiplying matrices that works well when one matrix is sparse and the other is dense. Sparse graph convolution can be efficiently computed by utilizing SpMM, which avoids dealing with the parts where no connections exist and thus improves computational efficiency.

Assuming that the adjacency matrix of the DDI network is represented by $A \in R^{N\times N}$, the number of drugs in the DDI network is denoted by N, and the drug molecular graph feature matrix $F_G \in R^{N\times dg}$ is input. Prior to the graph convolution operation, we normalize the adjacency matrix A as

$$\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}, \tag{11}$$

where $\tilde{D} = D + I$, $\tilde{A} = A + I$,I represents the identity matrix, and D represents the degree matrix.

Then we apply the GCN encoder framework as follows:

$$\mathbf{N}^{(1)} = U\left(A, G, \mathbf{W}_u^{(0)}, \mathbf{W}_u^{(1)}\right) = \widehat{A}\,\mathbf{ReLU}\left(\widehat{A}G\mathbf{W}_u^{(0)}\right)\mathbf{W}_u^{(1)}, \tag{12}$$

where $\mathbf{W}_u^{(0)}$ and $\mathbf{W}_u^{(1)}$, respectively, denote the two learnable weight parameters of layer 0 and layer 1 of the GCN encoder. After multiple layers of GCN, we finally obtain the drug relation feature matrix $F_N \in R^{N\times dg}$ from the DDI network.

For the GCN-based DDI feature extraction module, the weight parameters of the model are first initialized, and then the GCN model structure is defined, including the number of GCN layers and the dimensionality of each layer. During the training process, the drug molecular map features are input to the GCN model and are processed through multiple GCN layers to obtain the drug representation. To optimize the performance of the model, the binary cross entropy is chosen as the loss function to measure the difference between the predictions and the actual labels. The back propagation algorithm is adopted during training to calculate the gradient of the loss function, and the Adam optimizer is used to update the parameters of the model. To improve training stability and performance, an exponential decay method is used to adjust the learning rate. Together, these steps ensure that the GCN model can effectively extract features from the DDI network.

Feature extraction module based on biochemical feature similarities

We further extract biochemical features that may affect drug interactions, including targets, enzymes and transporters. A feature matrix of drugs (drug–enzyme, drug–target and drug–transporter) is constructed to represent these drug features. Each feature corresponds to a set of descriptors, and then the drug can be represented as a binary feature vector, where each entry indicates the presence or absence of the corresponding descriptor. However, the high sparsity and dimensionality of the feature vector may degrade the performance of the model. To reduce the sparsity and improve the accuracy of the feature vector, we utilize principal component analysis[29] for feature reduction. Then we calculate pairwise drug–drug similarity based on the Jaccard similarity metric:

$$J(f_{d_i}, f_{d_j}) = \frac{|f_{d_i} \cap f_{d_j}|}{|f_{d_i} \cup f_{d_j}|} = \frac{|f_{d_i} \cap f_{d_j}|}{|f_{d_i}| + |f_{d_j}| - |f_{d_i} \cap f_{d_j}|}, \tag{13}$$

where f_{d_i} and f_{d_j}, respectively, represent feature vectors drugs i and j, $|f_{d_i} \cap f_{d_j}|$ is the intersection of f_{d_i} and f_{d_j} and $|f_{d_i} \cup f_{d_j}|$ is the union.

Through the Jaccard similarity, we accordingly obtain the target similarity matrix $X^t \in R^{N\times dg}$, the enzyme similarity matrix $X^e \in R^{N\times dg}$ and the transporter similarity matrix $X^p \in R^{N\times dg}$, where N is the number of drugs and the dg is the dimensionality of the feature embedding.

Finally, to further explore the complementarity among different biochemical feature similarities, we concatenate these three vectors as the final heterogeneous feature embedding of d_i as

$$F_{S,d_i} = x_{d_i}^t \oplus x_{d_i}^e \oplus x_{d_i}^p, \tag{14}$$

where $x_{di}^t \in X^t$, x_{di}^t is the feature vector corresponding to drug d_i in row i of the target similarity matrix, and similarly, $x_{di}^e \in X^e$, $x_{di}^p \in X^p$, and the symbol $\oplus$ is the cascade operation of drug d_i features. The feature matrix $F_S \in R^{N\times dg}$ based on biometric similarity is finally obtained by feature dimensionality reduction operation, where the superscript d_g is the feature embedding dimension.

Multimodal feature fusion module

After learning the feature embedding of drug molecular graph, DDI network and biochemical feature similarities of drugs, we introduce a multimodal fusion neural layer to explore their complementarity and fuse the feature embedding. The multimodal

feature fusion layer fuses important features based on an attention mechanism. The critical information is fused by employing an attention mechanism to assign learnable weights.

Given the obtained F_G, F_N and F_S, the attention mechanism is calculated as follows:

$$(\alpha_g, \alpha_N, \alpha_S) = \mathrm{att}(F_G, F_N, F_S), \tag{15}$$

where α_g, α_N and α_S denote the attention coefficients of embedding F_G, F_N, F_S, respectively.

For these different types of features, their attentions $(\alpha_g, \alpha_N, \alpha_S)$ are learned during the training process of the model. The model automatically determines which features are more important, and adjusts the attention coefficients using the back propagation algorithm to optimize the loss function [30]. This learning process allows the model to better adapt to a given DDI prediction task and make full use of the information from each feature [31].

Supposed that node i in the embedding vector g is represented as g^i. We first apply the nonlinear transformation and then multiply it by the shared attention vector q to obtain its attention value Wg^i as follows:

$$Wg^i = q^T \cdot \tanh\left(w \cdot \left(g^i\right)^T + b\right), \tag{16}$$

where w is the weight matrix, b is a bias vectorand q is a shared attention vector. This step considers the contribution of each feature in the fusion and is modeled by a nonlinear transformation. Then, we use the softmax function to regularize the attention values and obtain their attention coefficients. These coefficients determine the weight of each feature in the final fusion. Softmax function ensures that the sum of these coefficients is equal to 1, so they indicate the relative importance of different features:

$$\alpha_g^i = \mathrm{soft}\left(w_g^i\right) = \frac{\exp\left(w_g^i\right)}{\exp\left(w_g^i\right) + \exp\left(w_n^i\right) + \exp\left(w_s^i\right)} \tag{17}$$

Similarly, $\alpha_n^i = \mathrm{soft}\left(w_n^i\right)$, $\alpha_s^i = \mathrm{soft}\left(w_s^i\right)$.

Finally, the final drug embedding E is obtained as

$$E = \alpha_g \cdot F_G + \alpha_n \cdot F_N + \alpha_s \cdot F_S \tag{18}$$

Then, the drug feature embedding E is obtained through the multimodal feature fusion layer, integrating information of the molecular structure, biological similarity and drug interaction network of drugs.

DDI prediction

Based on the multimodal feature fusion representation of drugs, we perform the prediction task using a fully connected deep learning network. First, we obtain an interaction-linked representation by multiplying the two drug representations, input it into the MLPand predict the probability score of the DDI as follows:

$$\hat{y}_{ij} = \sigma\left(\mathrm{MLP}\left(e_i \odot e_j\right)\right), \tag{19}$$

where σ denotes the element-wise product, and MLP consists of two fully connected layers.

Our learning goal is to minimize the distance between the prediction and the actual label. DMFDDI achieves the best DDI prediction performance by jointly training the feature extraction module and the feature fusion module, which learn together to minimize the loss function and learn effective feature representations from the different feature extraction modules. Since the prediction of DDI is a two-class problem, we use binary cross entropy as the loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_{ij} \log \hat{y}_{ij} + \left(1 - y_{ij}\right) \log\left(1 - \hat{y}_{ij}\right), \tag{20}$$

where $y_{ij} \in \{0, 1\}$ denotes the interaction label for the drug pair (d_i, d_j) in binary-classification tasks and $\hat{y}_{ij}$ is the predicted DDI probability.

EXPERIMENTS

Dataset

To validate the effectiveness of the proposed DMFDDI, we evaluate the model on small-, medium- and large-scale datasets. For the small-scale ZhangDDI dataset [32], it contains 572 drugs and 48 548 known edges. Although the number of drugs is relatively small, fingerprints of all drugs are available. In the medium-scale ChCh-Mine dataset[33], there are 1514 drugs and 48 514 known edges. It contains almost three times the drugs than the ZhangDDI dataset; however, only the same number of tagged DDI links are available. In the large-scale DeepDDI dataset [34], there are 1704 drugs and 192 284 known edges. In these datasets, drugs are associated with their SMILES string representations. Based on the open-source tool RDKit3, we convert the obtained SMILES string of each drug into the molecular graph, where atoms are represented as nodes, and chemical bonds as edges. Meanwhile, to integrate different biochemical feature similarities of drugs into DMFDDI, we obtain and combine the required biochemical feature data from well-known databases, including DrugBank [35] and KEGG [36]. As in the previous study [32], we select three features (targets, enzymes and transporters) for similarity calculation.

Evaluation metrics and Experimental setup

To estimate the performance of DMFDDI from different aspects, we adopt three widely used metrics, including area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC) and F1. These metrics have different emphasis. For classification problems, AUROC is suitable for class-balanced datasets, while AUPRC reflects the generalization ability of models on unbalanced datasets. In our experiments, we report the average values of these metrics in 10 replications.

We utilize the Adam optimizer to train the model [37] and Xavier to initialize the model [38]. The exponential decay method is adopted to set the learning rate, where the initial learning rate is 0.0001 and the multiplication factor is 0.96. The model applies a loss layer to the output of each intermediate layer [39], where the loss rate is 0.3. In the biochemical similarity-based feature extraction module, we utilize a three-layer network structure, and set the number of neurons in each layer as half of the previous layer. The number of neurons in the last hidden layer is fixed at 256. Also, the number of GCN layers is set to two for the drug molecular graph feature extraction module and DDI feature extraction module, and the dimensionality of the extracted drug representations is set to 256.

To evaluate the performance of different parameter settings and monitor the training process of the model, we randomly split the dataset into training, validation and testing sets with a ratio

of 6:2:2. The validation set provides an independent dataset for the model during training, which helps prevent overfitting and improve the model's generalization ability. As previous studies did [24], the drug pairs that are not known DDIs in the dataset are regarded as the negative samples. We randomly choose the same number of negative samples as the positive samples for our experiments. This balances the positive and negative samples, and prevents the model from being biased toward predicting positive samples.

Baselines Methods

In the comparative analysis, we compare our model against the following baselines, including the traditional and the recent state-of-the-art deep learning methods.

- NN [40] identifies new DDIs based on the molecular structural similarity of drugs involved in the established DDIs.
- GCN [41] uses GCNs for the semi-supervised node classification task. As a baseline, we apply GCN to encode the drug molecular graph, and then feed the embedded features into DNN for prediction.
- GIN [42] utilizes a graph isomorphic network (GIN) for learning molecular representations in various monomer property prediction tasks. As a baseline method, we use GIN to encode the drug molecular graph, and then feed the feature embedding into DNN for prediction.
- GAT [43] is based on graph attention network (GAT) to learn node embeddings by designing good attention mechanisms on the graph. We utilize GAT to obtain drug features based on DDI network, and then feed the extracted features into DNN for prediction as a baseline.
- SEAL-CI [44] applies a hierarchical graph representation learning framework to a semi-supervised graph classification task.
- DeepDDI [45] constructs a drug similarity matrix based on fingerprint, reduces the feature dimension and then utilizes DNN to predict DDIs.
- DPDDI [20] is a GCN-based method to predict DDIs. It utilizes GCN to capture the topological relationships of drugs in DDI networks.
- MIRACLE [24] learns both inter-view molecular structure and intra-view interaction information of the drug for DDI prediction.
- DM-DDI [31] combines drug features and topologies to learn representative drug embedding for DDI prediction. A deep neural network model is used on the drug feature matrix to extract feature information and a graph convolutional network model is used to capture structural information from the adjacency matrix.
- MFDA [46] employs a dual-level attention mechanism, including node-level and view-level, to obtain uniform drug embedding for drug interaction prediction.

Performance comparison with competing methods

We compare the DMFDDI with 10 competing DDI prediction models on three different scale datasets. We utilize three widely used metrics (AUROC, AUPRC and FI) to evaluate the prediction performance of each algorithm. Table 1 shows the performance of these DDI prediction methods. For each method, we show the average values of 10 replicate experiments. The best results are highlighted in bold. We observe that the proposed DMFDDI achieves better AUPRC and F1 values than the competitive methods on all the datasets with three different scales. For the evaluation metric AUROC, it achieves better performance on the ZhangDDI dataset and the ChCh-Miner dataset, while its performance is close with MIRACLE on the DeepDDI dataset. Specifically, the performance of the algorithm NN is relatively poor, which only uses the similarity-based fingerprints feature. In contrast, MIRACLE obtains better results because it utilizes different features, indicating the importance of integrating multiple features to predict DDIs. Also the graph-based methods have poor performance because they rely only on single-view graph information. GCN, GIN and GAT encode drug molecule graphs through different GNN frameworks and they make pairwise predictions of DDIs based on the obtained drug molecule representations. DPDDI learns drug representations directly from DDI relationships and utilizes the inner product of the target drug pair embedding results for prediction. The performances of DeepDDI and DPDDI are better than those of GCN, GIN and GAT only using drug molecular features. The two newly developed models DM-DDI and MFDA achieve better performance on the ZhangDDI dataset than on the medium and large datasets. The performance on different datasets may be affected by dataset differences. Further, the prediction models perform better when considering drug molecular maps and drug interaction network maps together, such as MIRACLE. Overall, DMFDDI achieves better performance on different datasets, which integrate three types of features, including molecular structure information, DDI information and biologically relevant feature similarities. The comparison results indicate that the fusion of multiple types of features can improve the performance of the model. Besides the ZhangDDI dataset, we also conduct experiments on the two large-scale datasets (the ChCh-Miner dataset and the DeepDDI dataset), and the results validate the scalability of DMFDDI.

Furthermore, we evaluate the ability of DMFDDI to predict DDIs between new drugs. First, the drugs rather than DDIs are randomly divided into five portions, where four portions are used as the training drugs and the rest as the test drugs. The models are trained on the DDIs between the training drugs and then are tested on the DDIs between the test drugs. Here, the two newly developed models, DM-DDI and MFDA, are also included in the comparison with DMFDDI on the ZhangDDI dataset. As shown in Table 2, DMFDDI achieves the best results compared with DM-DDI and MFDA. However, the performance of all models declines when new drugs are used for prediction. For DMFDDI, the AUROC, AUPR and F1 score are, respectively, decreased to 0.9314, 0.7241 and 0.4863. For prediction task on new drugs, positive samples are usually much less than negative samples, resulting in data imbalance, which might lead to the performance decrease.

Sensitivity analysis

Aiming to explore the impact of important parameters on prediction performance, we assign them with different values, and then evaluate the performance of DMFDDI on the ZhangDDI dataset. We analyze the dimensionality of the drug features d_g, the learning rate l_r and the number of GCN layers of the drug molecular graph feature extraction module L_m. By fixing other parameters, we investigate the effect of different key parameter settings on the performance of DMFDDI.

To determine the optimal setting of d_g, we vary the dimensionality from 2 to 1024. As shown in Figure 3(a), when d_g is set to 256, the three metrics are optimal and the model achieves the best performance. Specifically, as the dimensionality of drug features increases, DMFDDI can extract more useful information.

Table 1: Performance comparison of DMFDDI and 10 competing methods on three different scale datasets, measured by AUROC, AUPRC and F1

Dataset	Method	AUROC	AUPRC	F1
	NN	0.6781	0.5261	0.4984
	GCN	0.9193	0.8865	0.8161
	GIN	0.8145	0.7716	0.6415
	GAT	0.9149	0.9069	0.8121
ZhangDDI	SEAL-CI	0.9319	0.9285	0.8476
	DeepDDI	0.9201	0.8876	0.8025
	DPDDI	0.9516	0.9021	0.8423
	Miracle	0.9895	0.9817	0.9320
	DM-DDI	0.9876	0.9642	0.8521
	MFDA	**0.9989**	0.9627	0.8513
	DMFDDI	0.9927	**0.9836**	**0.9741**
	NN	0.7315	0.7723	0.5314
	GCN	0.8284	0.8427	0.7054
	GIN	0.7032	0.7241	0.6554
	GAT	0.8584	0.8814	0.7651
ChCh-Miner	SEAL-CI	0.9093	0.8938	0.8474
	DeepDDI	0.9212	0.9307	0.8541
	DPDDI	0.9583	0.9228	0.8452
	Miracle	0.9615	0.9557	0.9226
	DM-DDI	0.9587	0.7979	0.5476
	MFDA	**0.9795**	0.8468	0.7041
	DMFDDI	0.9712	**0.9851**	**0.9623**
	NN	0.8181	0.8082	0.7137
	GCN	0.8553	0.8327	0.7218
	GIN	0.7228	0.7027	0.6771
	GAT	0.8484	0.8114	0.7351
DeepDDI	SEAL-CI	0.9382	0.9044	0.8170
	DeepDDI	0.9226	0.9047	0.7922
	DPDDI	0.9285	0.9148	0.8531
	Miracle	**0.9551**	0.9234	0.8360
	DM-DDI	0.9065	0.4741	0.2298
	MFDA	0.9519	0.7247	0.4868
	DMFDDI	0.9514	**0.9633**	**0.9257**

Table 2: New Drug Predictive Analytical Experiment

Method	AUROC	AUPRC	F1
DM-DDI	0.8518	0.3646	0.1597
MFDA	0.9179	0.4745	0.2298
DMFDDI	**0.9314**	**0.7241**	**0.4863**

However, too high dimensionality may increase noise and lead to performance degradation. Similarly, we vary l_r to 0.01, 0.001, 0.0001, 0.00001 and fix the other parameter settings, and the performance of DMFDDI with different l_r is shown in Figure 3(b). We observe that the model performs best when l_r=0.0001. As shown in Figure 3(c), for the number of GCN layers of the drug molecular graph feature extraction module, the performance of DMFDDI improves as L_m increases. When L_m=2, these three metrics are optimal and the model achieves the best performance. However, too many layers may lead to over-smoothing and result in performance degradation.

Ablation study

To validate the effectiveness of each type of drug features, we use different feature combinations to predict DDIs on the ZhangDDI dataset. As shown in Figure 4, the combined features outperform the individual features, and the best performance is obtained when all three types of features are fused together, achieving AUROC, AUPRC and F1 values of 0.9921, 0.9815 and 0.9712, respectively. Each feature type contributes to DDI prediction to some extent. Specifically, the drug molecular graph features and the DDI features are more effective than the biochemical features. Among the pairwise feature combinations, the fusion of drug molecular graph features and DDI features leads to the best results, with AUROC, AUPRC and F1 values of 0.9815, 0.9806 and 0.9327, respectively. The model that integrates all three features improves the F1 value by 3.85% compared with the best pairwise combination, demonstrating that multi-modal feature fusion can better capture the multi-level characteristics of drugs. The results confirm that multi-modal feature fusion can enhance drug representation and DDI prediction performance.

In the drug molecular graph based feature extraction module, our model introduces an AGRUNN that captures the hidden key relationships and extracts the global features of the drug molecular graph. To validate its effectiveness, we compare the performance of DMFDDI with the AGRUNN and that of the model with the ordinary message passing network (MPNN) on the ChCh-Miner dataset. The results are shown in Table 2. We observe that DMFDDI with the AGRUNN model outperforms the model with the MPNN. The values of AUROC, AUPRC and F1 are improved

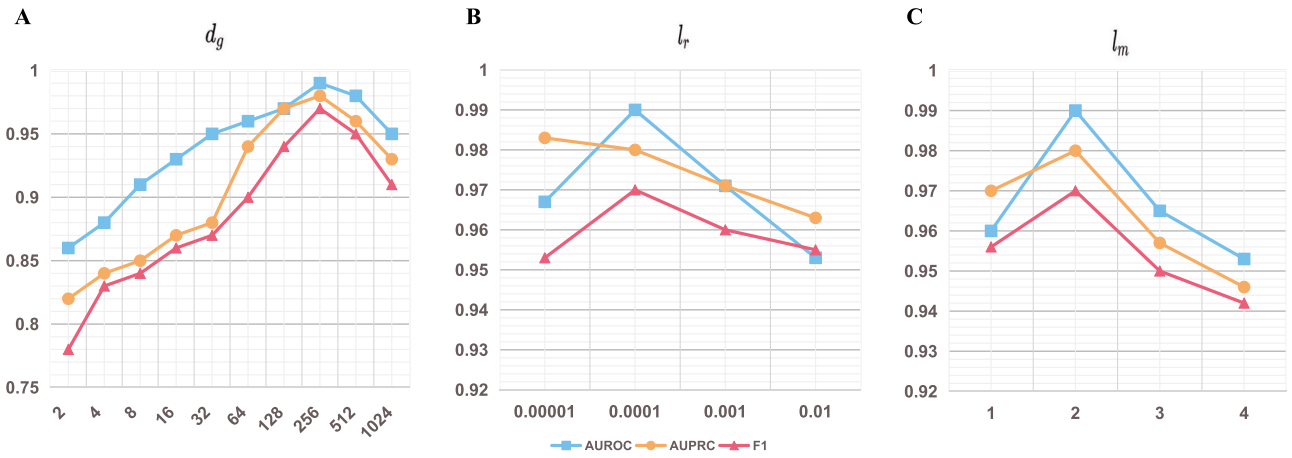

Figure 3. Performance comparison of the proposed DMFDDI with different drug feature dimensionality, learning rate and the number of GCN layers.

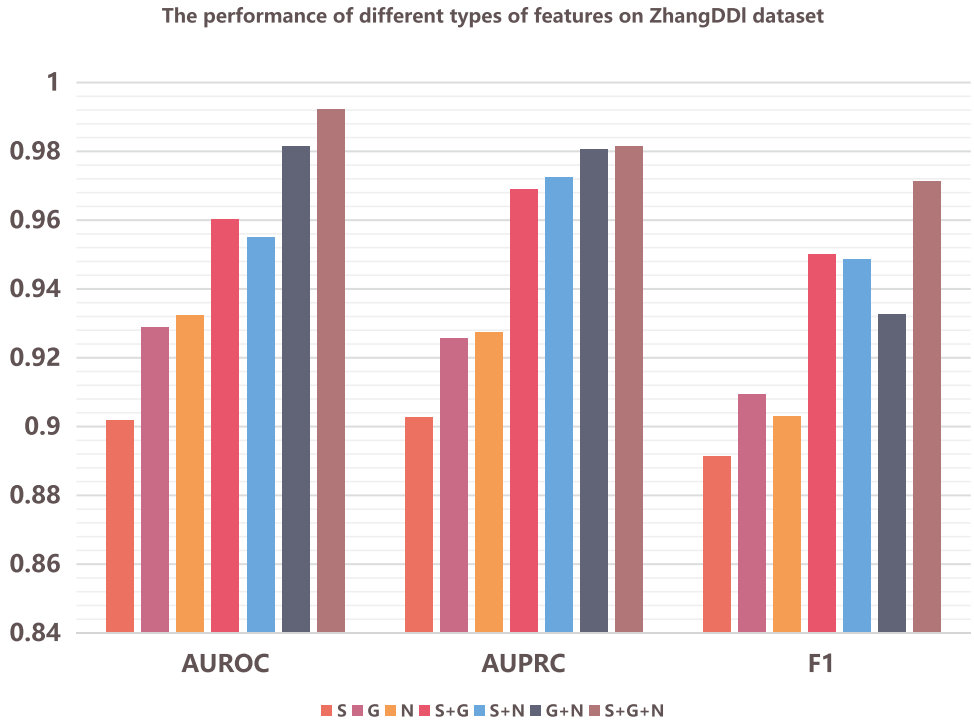

Figure 4. The performance comparison of DMFDDI with different feature combinations on ZhangDDI dataset. S denotes biochemical feature similarity, G denotes drug molecular graph features and N denotes DDI network features.

by 0.97%, 2.94% and 3.97%, respectively. The comparison results demonstrate that the AGRUNN can make full use of the neighborhood information of different orders, including the relationship between atom pairs separated by long topological distances, which further promotes accurate extraction of the features from the molecular graph and facilitates the DDI prediction.

To evaluate the effectiveness of the multimodal feature fusion model, we compare our model with traditional fusion strategies, including cascading and summation methods. We implement the traditional fusion approach in the fusion layer of the model. Table 3 shows the comparison results of different feature fusion strategies on the ZhangDDI dataset. DMFDDI outperforms DMFDDI_concat and DMFDDI_sum in all three evaluation metrics. Specifically, the performance is improved by 4.05%, 7.94% and 13.88% in AUROC, AUPRC and F1, respectively, over the second ranking DMFDDI_concat method, further validating the effectiveness of the proposed fusion strategy. Our feature fusion module fuses the feature information extracted from each module by exploring the complementarity between multimodal representations of drugs, resulting in better representation of the drug features.

Table 3: Ablation experiments of drug molecular graph feature extraction module

Method	AUROC	AUPRC	F1
MPNN	0.9615	0.9557	0.9226
AGRUNN	**0.9712**	**0.9851**	**0.9623**

Case study

Furthermore, we evaluate the performance of DMFDDI in predicting the unobserved DDIs. Specifically, we perform DDI predictions

Table 4: Ablation experiments of multimodal fusion layers

Method	AUROC	AUPRC	F1
DMFDDI_sum	0.9201	0.8653	0.8126
DMFDDI_concat	0.9516	0.9021	0.8324
DMFDDI	**0.9921**	**0.9815**	**0.9712**

Table 5: Top 10 DDIs predicted by DMFDDI

Number	Drug1	Drug2	Validation Source
1	Abemaciclib	Astemizole	Drugbank
2	Abiraterone	Fentanyl	Drugbank
3	Digoxin	Ergocalciferol	Drugbank
4	Acetylcholine	Cinchocaine	Drugbank
5	Naproxen	Clofarabine	Drugbank
6	Hydroxyzine	Cisplatin	N.A.
7	Anagrelide	Lomefloxacin	Drugbank
8	Alimemazine	Eprosartan	Drugbank
9	Aprepitant	Bivalirudin	N.A.
10	Bortezomib	Hydroxyzine	Drugbank

across 572 drugs and 37 264 pairs of drugs with a few of known interactions. We predict 289 920 unknown drug interactions to validate the capability of the model. Higher scores for unobserved drug pairs indicate a higher probability of interaction between these drugs. We validate the high ranking predictions in DrugBank [34]. The experimental results are shown in Table 4. Of the top 10 predicted DDI events, eight responses are confirmed. For example, the highest predicted response scores are for Abemaciclib and Astemizole, followed by higher scores for Abiraterone and Fentanyl, indicating an Abemaciclib–Astemizole interaction and an Abiraterone–Fentanyl interaction, respectively. This is also confirmed in Drugbank. The metabolism of Abemaciclib can be decreased when combined with Astemizole. The metabolism of Butyrfentanyl can be decreased when combined with Abiraterone. These case studies demonstrate that our DMFDDI can effectively detect the potential DDIs.

CONCLUSION

As adverse DDIs pose unexpected risk to patients, it is crucial to identify potential DDIs. As single drug feature is insufficient to comprehensively represent drug information. Therefore, in this study, we construct a new end-to-end DDI prediction learning framework, DMFDDI, which effectively fuses drug molecular structure features, DDI network and drug biochemical feature similarities. Specifically, based on the AGRUNN, we extract the molecular structure from the given drug molecular graphs. A sparse GCN is introduced to learn the topological structure information of the known DDI network. We extract the biochemical feature similarity by analyzing three feature similarity matrices. Then, in the multimodal feature fusion module, we explore the complementarity between multi-modal representations of drugs using an attention mechanism to fuse the feature information extracted from each module. Finally, the scores of DDIs are predicted by the predictor. Extensive experiments on different scale datasets demonstrate that DMFDDI outperforms the 10 competing methods. Furthermore, detailed ablation studies demonstrate that the AGRUNN and the multimodal feature fusion model play a key role in improving the prediction performance.

Key Points

- We propose DMFDDI, a multi-modal deep learning feature fusion framework for DDI prediction. The proposed model can effectively fuse multimodal features, including drug molecular structure features, the DDI network as well as biochemical feature similarities.
- To accurately extract the local and global structure of drug molecules, we construct a feature extraction module based on AGRUNN, which can aggregate and update feature information of different atom nodes and chemical bonds through a gating mechanism. A sparse GCN is introduced to learn the topological structure information of the DDI network.
- We introduce a multimodal fusion model to explore the complementarity among different types of drug features and fuse the feature embedding learned by each module, making the final drug feature representation more accurate.
- The extensive experiments demonstrate that DMFDDI achieves better performance compared with the 10 competitive methods. The predictions of DMFDDI are further validated by case studies, implying its capability in the predicting unknown DDIs.

ACKNOWLEDGEMENTS

We would like to thank MS. Huichun Zhu for fruitful discussions, and MS. Xingyu Huang for assistance in Python implementation.

FUNDING

This work were sponsored in part by the National Natural Science Foundation of China (62172088) and Shanghai Natural Science Foundation (21ZR1400400).

REFERENCES

1. Sun W, Sanderson PE, Zheng W. Drug combination therapy increases successful drug repositioning. *Drug Discov Today* 2016; **21**(7):1189–95.
2. Santana L, Lorberbaum T, Hripcsak G, *et al*. Similarity-based modeling in large-scale prediction of drug–drug interactions. *Nat Protoc* 2014;**9**(9):2147–63.
3. Chen M, Wang Q, Zhang L, *et al*. Nllss: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput Biol* 2016;**7**(12).
4. Ma L, Sutcliffe R, He F, *et al*. drug–drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy (Basel)* 2019;**21**(1).
5. Yan C, Duan G, Zhang Y, *et al*. Idnddi: An integrated drug similarity network method for predicting drug–drug interactions. In: Cai Z, Skums P, Li M (eds). *Bioinformatics Research and Applications*. Cham: Springer International Publishing, 2019, 89–99.
6. Chen X, Zhou C, Wang C-C, Zhao Y. Predicting potential small molecule–miRNA associations based on bounded nuclear norm regularization. *Brief Bioinform* 2021;**22**(6): bbab328.
7. Vilar S, Uriarte E, Santana L, *et al*. Detection of drug–drug interactions by modeling interaction profile fingerprints. *PloS One* 2013;**8**(3): e58321.

8. Takeda T, Hao M, Cheng T, *et al*. Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *J Chem* 2017;**9**:16.
9. Li P, Huang C, Yingxue F, *et al*. Large-scale exploration and analysis of drug combinations. *Bioinformatics* 2015;**31**(12):2007–16.
10. Wang C-C, Zhu C-C, Chen X. Ensemble of kernel ridge regression-based small molecule-mirna association prediction in human disease. *Brief Bioinform* 2022;**23**(1): bbab431.
11. Yan X-Y, Yin P-W, Wu XM, Han J. Prediction of the drug–drug interaction types with the unified embedding features from drug similarity networks. *Front Pharmacol* 2021;**12**.
12. Gottlieb A, Stein GY, Oron Y, *et al*. Indi: a computational framework for inferring drug interactions and their associated recommendations. *Mol Syst Biol* 2012;**8**:592.
13. Kastrin A, Ferk P, Leskosek B. Predicting potential drug–drug interactions on topological and semantic similarity features using statistical learning. *PloS One* 2018;**13**.
14. Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J Am Med Inform Assoc* 2014;**21**(e2): e278–86.
15. Rohani N, Eslahchi C. drug–drug interaction predicting by neural network using integrated similarity. *Sci Rep* 2019;**9**(1): 13645.
16. Yan C, Duan G, Zhang Y, *et al*. Predicting drug–drug interactions based on integrated similarity and semi-supervised learning. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**19**:168–79.
17. Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 1263–72. JMLR.org, 2017.
18. Nuo Xu, Pinghui Wang, Long Chen, Jing Tao, and Junzhou Zhao. Mr-gnn: Multi-resolution and dual graph neural network for predicting structured entity interactions. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI'19*, page 3968–74. AAAI Press, 2019.
19. Wang H, Lian D, Zhang Y, *et al*. Gognn: graph of graphs neural network for predicting structured entity interactions. *ArXiv, abs/200505537* 2020.
20. Feng Y-H, Zhang S-W, Shi J-Y. Dpddi: a deep predictor for drug–drug interactions. *BMC Bioinformatics* 2020;**21**(1): 419.
21. Huang K, Xiao C, Hoang TN, *et al*. Caster: predicting drug interactions with chemical substructure representation. *ArXiv, abs/191106446* 2019.
22. Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei MA, and Xiangxiang Zeng. Kgnn: knowledge graph neural network for drug–drug interaction prediction. In *International Joint Conference on Artificial Intelligence*, 2020.
23. Bai Y, Ken G, Sun Y, Wang W. Bi-level graph neural networks for drug–drug interaction prediction. *ArXiv, abs/200614002* 2020.
24. Wang Y, Min Y, Chen X, Ji W. Multi-view graph contrastive representation learning for drug–drug interaction prediction. 2021;2921–33.
25. Deng Y, Xinran X, Qiu Y, *et al*. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* 2020;**36**(15):4316–22.
26. Landrum G. *Rdkit: a software suite for cheminformatics, computational chemistry, and predictive modeling*. London: Academic Press, 2013.
27. Cho K, Merrienboer B, Gulcehre C, *et al*. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 2014.
28. Shi L, Wang L, Long C, *et al*. *Sgcn:sparse graph convolution network for pedestrian trajectory prediction*, 2021.
29. Shlens J. A tutorial on principal component analysis. *Educational* 2014;**51**.
30. Vaswani A, Shazeer N, Parmar N, *et al*. *Attention is all you need*. 2023.
31. Kang L-P, Lin K-B, Ping L, *et al*. Multitype drug interaction prediction based on the deep fusion of drug features and topological relationships. *PloS One* 2022;**17**:1–19.
32. Zhang W, Chen Y, Liu F, *et al*. Predicting potential drug–drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinformatics* 2017;**18**:18.
33. Ma T, Xiao C, Zhou J, Wang F. Drug similarity integration through attentive multi-view graph auto-encoders. 2018.
34. Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research*, pages 1263–72. PMLR, 2017.
35. Wishart D, Djoumbou Y, Guo AC, *et al*. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res* 2017;**46**.
36. Kanehisa M, Furumichi M, Tanabe M, *et al*. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**(D1): D353–61.
37. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.
38. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *J Mach Learn Res* 2010;**9**: 249–56.
39. Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
40. Vilar S, Harpaz R, Uriarte E, *et al*. drug–drug interaction through molecular structure similarity analysis. *J Am Med Inform Assoc* 2012;**19**:1066–74.
41. Kipf T, Welling M. Semi-supervised classification with graph convolutional networks. *ArXiv, abs/160902907* 2016.
42. Keyulu X, Weihua H, Leskovec J, Jegelka S. How powerful are graph neural networks? *ArXiv, abs/181000826* 2018.
43. Velickovic P, Cucurull G, Casanova A, *et al*. Graph attention networks. *ArXiv, abs/171010903* 2017.
44. Li J, Yu R, Cheng H, *et al*. Semi-supervised graph classification: a hierarchical graph perspective. 2019.
45. Ryu JY, Kim H, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. *Proc Natl Acad Sci* 2018;**115**:201803294.
46. Lin K, Kang L, Yang F, *et al*. Mfda: multiview fusion based on dual-level attention for drug interaction prediction. *Front Pharmacol* 2022;**13**.