Scientific Reports

https://doi.org/10.1038/s41598-026-36223-9

Article in Press

Enhanced drug disease association prediction through multimodal data integration and meta path guided global local feature fusion

Received: 6 March 2025

Accepted: 10 January 2026

Published online: 25 February 2026

Cite this article as: Wu S., Wang W., Jiao H. *et al.* Enhanced drug disease association prediction through multimodal data integration and meta path guided global local feature fusion. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-36223-9

Shengnan Wu, Wen Wang, Huizhi Jiao, Danhong Dong, Kexin Zhang & Xuechen Luo

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Enhanced Drug Disease Association Prediction through Multimodal Data Integration and Meta Path Guided Global Local Feature Fusion

Shengnan Wu[1,2*], Wen Wang[1,2], Huizhi Jiao[1,2], Danhong Dong[1,2], Kexin Zhang [1,2], Xuechen Luo[1]

(1. School of Management, Shanxi Medical University, Jinzhong, Shanxi 030600, China;

2. Shanxi Provincial Key Laboratory of Big Data for Clinical Decision-Making Research, Taiyuan, Shanxi 030001, China)

Corresponding. vivian19870220@163.com.

Abstract

Accurately predicting Drug-Disease Associations (DDAs) is of great significance for drug repurposing and new drug development. Although existing methods have promoted the development of this field to a certain extent, most of them are still limited to single-modal data and cannot fully characterize the complex features of drugs, diseases, and genes. At the same time, many methods only focus on either local neighborhoods or global structures during feature extraction, lacking the organic combination of the two, which limits the accuracy and generalization of predictions.

To address this, this paper proposes MedPathEx, a drug-disease association prediction method that combines multi-modal data integration and local-global feature learning. Specifically, we first construct a drug-gene-disease heterogeneous network and fuse multi-modal attributes such as drug chemical structures, ATC classifications, side effects, disease phenotypes and semantic information, as well as gene function annotations to generate more comprehensive node representations. Subsequently, we use graph convolutional networks to extract the attribute features of nodes themselves, capture local semantic relationships through meta-path modeling with a multi-head attention mechanism, and introduce a global attention mechanism to extract overall topological patterns, thereby achieving "micro-macro complementary" feature learning. Finally, by fusing node attributes and structural features, MedPathEx obtains a more discriminative comprehensive representation for the prediction of potential DDAs.

Experimental results show that MedPathEx outperforms existing methods in key indicators such as AUC, AP, and F1. Moreover, it successfully identifies new candidate drugs in cases of coronary artery disease and hypertension, demonstrating its great potential in practical applications.

Keywords: Drug-disease prediction; Heterogeneous network; Multi-modal data; Meta-path; Feature fusion

1. Introduction

With the rapid development of biomedical technology, the demand for data analysis and knowledge discovery in modern drug research and development is increasingly growing [1]. The successive opening of a large number of biomedical databases has provided rich resources covering information on biological entities such as drugs, diseases, and targets, laying a foundation for large-scale data mining [2]. However, in the face of massive datasets, traditional biological experimental methods are insufficient in terms of processing and analysis efficiency [3]. Against this backdrop, computational-based drug-disease association prediction methods have emerged. These methods can deeply analyze complex data, mine potential knowledge, reveal the associations between drugs and diseases, and screen candidate compounds with therapeutic potential, thereby significantly accelerating the process of new drug research and development [4.]

Computational-based drug-disease association prediction methods usually construct complex biological networks and extract rich features from them, enabling prediction models to effectively infer potential associations between drugs and diseases [5]. Early computational methods mainly relied on similarity measurements, such as drug chemical structure similarity, disease semantic similarity, or gene co-expression patterns, to infer potential associations [6]. Although such methods are intuitive and simple, due to the use of only single-modal information, they often have the problem of information missing, making it difficult to fully reflect the complexity of biological entities. Subsequently, researchers proposed matrix factorization and machine learning methods (such as non-negative matrix factorization, random walk, kernel methods, and ensemble learning)[7]. These methods can capture potential patterns from interaction matrices and improve prediction performance. However, they still need to be improved in terms of multi-source information integration and model generalization ability.

In order to better characterize the complex interaction relationships between drugs and diseases, heterogeneous network methods have gradually attracted attention[8]. This method can effectively reveal potential associations by simultaneously modeling different types of nodes such as drugs, diseases, genes, and their interrelationships. Early methods mostly relied on the topological structure of the network[9–11]. For example, Wu et al. [12] calculated the distance between nodes based on the static associations among drugs, diseases, genes, and side effects to predict potential associations. However, such methods usually ignore the attribute information of the nodes themselves (such as the chemical properties of drugs or the clinical phenotypes of diseases), resulting in limited node representation ability. To make up for this deficiency, some studies have tried to introduce node attributes into network modeling[13,14,15]. For example, FuHLDR [16] integrates drug molecular fingerprints, disease semantic similarity, and protein sequence features, and combines graph convolutional networks with meta-path methods to improve prediction effects. Nevertheless, most existing methods only consider single-modal attributes and fail to fully mine the multi-source information contained in drug-disease-gene relationships. In fact, different modal information is complementary: the chemical structure and ATC classification of drugs reveal molecular-level features, and side effect information reflects their clinical manifestations; the phenotypic and semantic information of diseases correspond to clinical manifestations and knowledge-level descriptions respectively; and gene function annotations describe molecular mechanism-level information. Therefore, fusing multi-modal data can provide more comprehensive and robust representations for nodes.

In terms of feature extraction, deep learning technologies and meta-path strategies are widely used in heterogeneous network analysis[17,18]. Researchers often use meta-paths to model the semantic relationships of local neighborhoods and combine attention mechanisms to enhance prediction performance. For example, Flam et al. [19] combine path embedding with graph neural networks to capture local substructures; the NEDD model proposed by Zhou et al. [20] uses meta-paths to model direct and high-order relationships between nodes; and the MAGNN model by Fu et al. [21] further combines node content conversion and multi-level aggregation mechanisms to effectively capture complex semantic patterns. Although these methods have made significant progress in local structure modeling, they still have limitations: most methods either over-rely on local neighborhood features or focus on global statistical information or embedded representations, failing to fully combine the advantages of both. Local features can meticulously describe neighborhood relationships such as drug-gene-disease, but lack perception of the overall network topology; global features can reflect long-range dependencies and macro laws between cross-modal entities, but often ignore local fine-grained semantics. Therefore, how to achieve complementary fusion of local and global features has become a key challenge in current Drug-Disease Association (DDA) prediction.

Based on the above research status and challenges, this paper proposes a drug-disease association prediction method combining multi-modal data integration and local-global feature extraction, named MedPathEx. Compared with existing methods, the main innovations of this study are reflected in the following three aspects:

(i)Multi-modal data integration: Integrate multi-modal attributes such as drug chemical structure, ATC classification, side effect information, disease phenotypic and semantic information, and gene function annotations under a unified framework to obtain more comprehensive node representations.

(ii)Local-global complementary feature learning: Capture local semantic relationships through meta-path modeling driven by a multi-head attention mechanism, and introduce a global attention mechanism to extract overall topological patterns, realizing complementary learning of micro-local features and macro-global features.

(iii)Deep feature fusion: Deeply fuse node attribute features with the learned local-global structural features to form a more discriminative comprehensive representation, which is ultimately used to accurately predict potential drug-disease associations.

2. Materials and Methods

2.1 Dataset

The data for this study were obtained from three public biomedical databases: Stanford Biomedical Network Dataset Collection (BIOSNAP24)[22], Comparative Toxicogenomics Database (CTD) [23], and Pharmacogenomics Knowledgebase (PharmGKB[24]). These databases provide extensive association data on drugs, genes, and diseases.

Specifically, from BIOSNAP, we acquired 9,761 drug-gene interaction records and 104,327 drug-disease interaction records, covering 4,349 drugs, 2,085 genes, and 565 diseases. From CTD, we obtained 36,321 disease-gene interaction records involving 3,969 genes and 561 diseases. From PharmGKB, we retrieved 3,422 disease-gene interaction records and 6,472 drug-gene interaction records, encompassing 1,614 drugs, 2,053 genes, and 730 diseases.

During data processing, we standardized the names of drugs, genes, and diseases and removed redundant and irrelevant entries, thereby ensuring the uniqueness of each association and the accuracy of the data. After meticulous filtering, we obtained a heterogeneous network with 12,661 nodes and 120,587 edges, including 1,148 diseases, 7,591 genes, 4,050 drugs, 69,034 disease-drug associations, 35,998 disease-gene associations, and 15,555 drug-gene associations (see Table 1).

Table 1.Summary of Entities and Associations in the Constructed Heterogeneous Network

Entity Type	Number of Entities	Association Type	Number of Associations
Drug	4050	Drug-Gene Association	15,555
Gene	7591	Disease-Gene Association	35,998
Disease	1020	Disease-Drug Association	69,034

2.2 Methods

The overall framework of the proposed approach is illustrated in Figure 1.

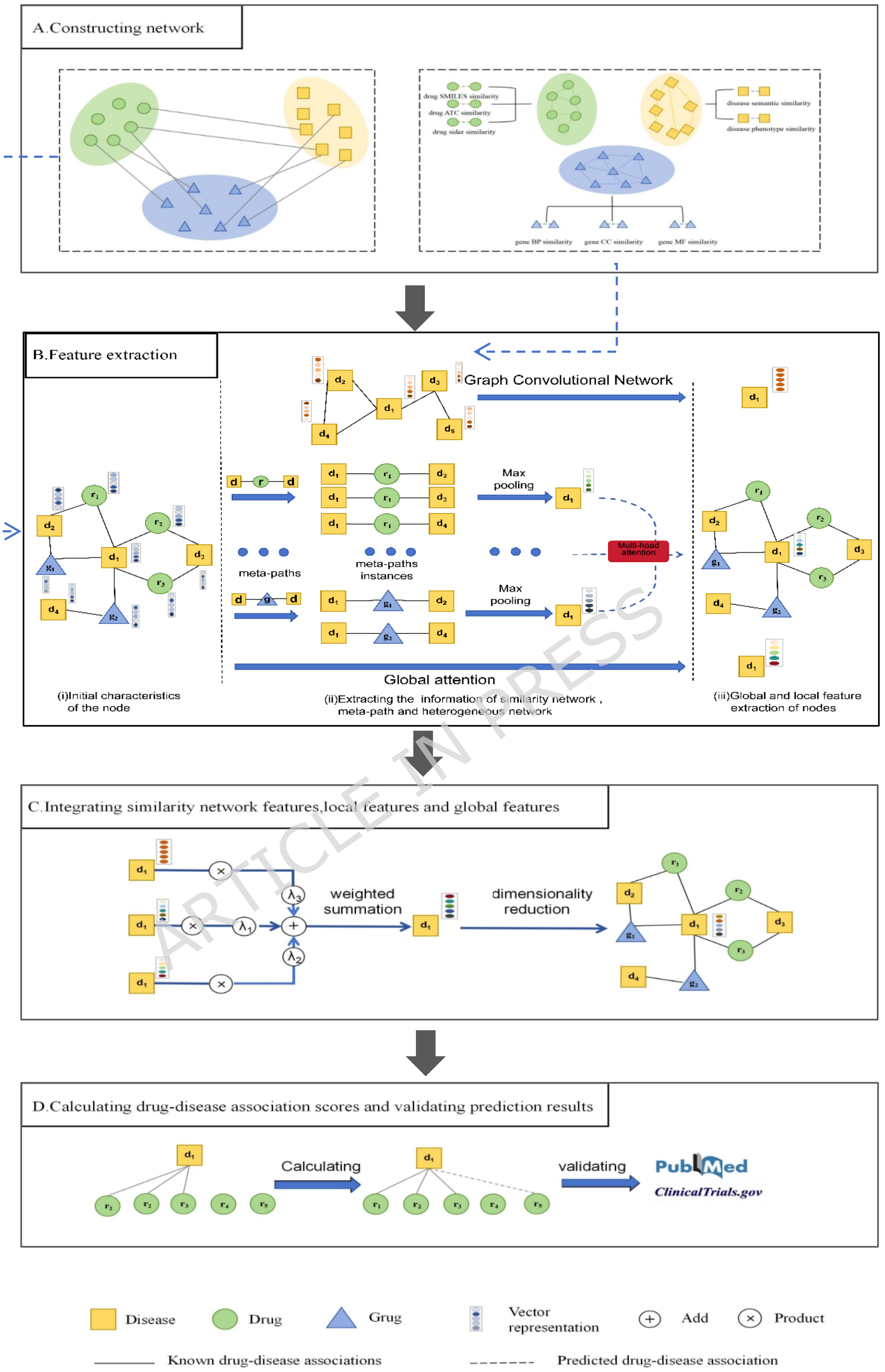
A.Constructing network
drug SMILES similarity
drug ATC similarity
drug sider similarity
disease semantic similarity
disease phenotype similarity
gene BP similarity
gene CC similarity
gene MF similarity
B.Feature extraction
Graph Convolutional Network
d r d
d g d
meta-paths
meta-paths instances
Max pooling
Multi-head attention
Global attention
(i)Initial characteristics of the node
(ii)Extracting the information of similarity network , meta-path and heterogeneous network
(iii)Global and local feature extraction of nodes
C.Integrating similarity network features,local features and global features
λ3
λ1
λ2
weighted summation
dimensionality reduction
D.Calculating drug-disease association scores and validating prediction results
Calculating
validating
PubMed
ClinicalTrials.gov
Disease
Drug
Grug
Vector representation
Add
Product
Known drug-disease associations
Predicted drug-disease association

Figure 1. Overall Framework of the Model.

2.2.1 Network Construction

To comprehensively represent the relationships among drugs, diseases, and genes, we construct two complementary structures: (i) per-type similarity graphs that capture intra-type relationships, and (ii) a heterogeneous network that encodes inter-type associations. Detailed methods are provided in Supplement~S1.

(i) Similarity graph construction

For each entity type, multiple features are integrated to quantify pairwise similarity.

- Drugs: Drug similarity is derived from three perspectives: (1) chemical structure similarity between drugs was calculated using SMILES fingerprints[25] obtained from DrugBank[26], with Tanimoto coefficients implemented via the RDKit toolkit; (2) clinical classification similarity was determined based on Anatomical Therapeutic Chemical (ATC) codes extracted from DrugBank, and quantified using the Jaccard coefficient[27]; and (3) side effect similarity from SIDER databases [28], also quantified with Jaccard scores.
- Diseases: We consider both phenotypic similarity and semantic similarity. Phenotypic similarity is calculated using OMIM descriptions via the MimMiner tool[29], while semantic similarity is measured from the MeSH directed acyclic graph (DAG) using MeSHSim[30].
- Genes: Gene similarity is evaluated from Gene Ontology (GO) annotations, including biological processes (BP), molecular functions (MF), and cellular components (CC). Using the GOSemSim algorithm[31], similarity is computed based on shared information content and the most informative common ancestor (MICA).

(ii) Similarity network fusion.

To integrate multi-source similarities, we employ Similarity Network Fusion (SNF). Given feature-specific similarity matrices S^i (e.g., S^{ATC}, S^{SMILES}, S^{SE}), an affinity matrix is constructed and iteratively updated:

$$P^i = S^i \cdot \left(\frac{\exp\left(-\frac{r^i}{\sigma} \right)}{\sum_j \exp\left(-\frac{r^i_{uj}}{\sigma} \right)} \right) \cdot S^{iT},$$

where r^i denotes the distance metric and σ is the affinity scaling parameter. The fused drug similarity matrix is obtained as:

$$R = \frac{1}{3}(P^{ATC} + P^{SMILES} + P^{SE}),$$

while disease and gene similarity matrices are fused by:

$$D = \mathrm{SNF}(S^{Phen}, S^{Sem}),$$

$$G = \mathrm{SNF}(S^{BP}, S^{MF}, S^{CC}).$$

(iii) Similarity graph representation

The final similarity networks are represented as $S_R \in R^{|V_R| \times |V_R|}$ for drugs, $S_D \in R^{|V_D| \times |V_D|}$ for diseases, and $S_G \in R^{|V_G| \times |V_G|}$ for genes. Each entry reflects the fused similarity between a pair of entities within the same type.

(iv) Heterogeneous network construction

Beyond intra-type similarities, we construct a heterogeneous network $G = (V, E)$ to encode inter-type associations. The node set V consists of drugs (V_R), diseases (V_D), and genes (V_G). Known associations are encoded in three bipartite adjacency matrices: the drug–disease matrix $A_{R,D} \in R^{|V_R| \times |V_D|}$, the disease–gene matrix $A_{D,G} \in R^{|V_D| \times |V_G|}$, and the drug–gene matrix $A_{R,G} \in R^{|V_R| \times |V_G|}$. An entry is set to 1 if an association exists and 0 otherwise. This heterogeneous graph provides a unified representation of cross-type interactions.

the similarity graphs preserve rich intra-type relational signals, while the heterogeneous graph complements them by encoding inter-type connectivity, together forming the foundation for subsequent feature learning.

To prevent information leakage, all drug–disease associations were split into five edge-level folds, and the test edges in each fold were removed from the heterogeneous network before feature extraction. MedPathEx was trained only on the remaining observed links. The detailed procedure and statistics are given in Supplementary Algorithm S1 and Supplement S2.

2.2.2 Feature Extraction

To obtain informative and complementary node representations, we design a three-level feature extraction strategy: (i) embeddings from similarity graphs, (ii) local semantics captured by meta-paths in the heterogeneous graph, and (iii) global structural features from attention over the entire network.

(i) Similarity-graph embeddings.

For each fused similarity network of drugs (S_R), diseases (S_D), and genes (S_G), the initial features are projected into a unified space:

$$h_S = W_S x_S, \ S \in \{R, D, G$$

where x_S denotes the original features and W_S is a learnable transformation matrix.

Subsequently, we employ Graph Convolutional Networks (GCNs) [32] to capture intra-type relationships. The propagation rule is defined as:

$$H^{(l+1)} = \mathrm{ReLU}\left(\grave{D}^{-\frac{1}{2}}\grave{A}\grave{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

where $\grave{A}$ is the similarity matrix with self-loops and $\grave{D}$ its degree matrix. After training, we obtain embeddings H_R, H_D, H_G that encode similarity-based structures within each entity type.

(ii) Heterogeneous local semantics via meta-paths.

In the heterogeneous drug–disease–gene graph, nodes are initialized from their incident associations and projected into a shared feature space:

$$h_v = W_v x_v,\ v \in \{R,D,G$$

where x_v is the binary association vector and W_v is trainable.

To capture typed semantics, we define symmetric meta-paths that reflect homophily and complementarity, such as RDR and RDGDR for drugs, DGD/DRD/DRRD for diseases, and GDRDG for genes(see Supplement~S1).

Each meta-path instance $P(v,u)$ connecting node v to u is summarized by max-pooling over its intermediate nodes:

$$h_{P(v,u),i} = \max\{h'_{t,i} | t \in P(v,u$$

and then aggregated using a graph attention mechanism:

$$\alpha_{vu}^{P} = \frac{\exp(e_{vu}^{P})}{\sum_{s \in N_v^P} \exp(e_{vs}^{P})},\ h_v^P = \sigma\left(\sum_{u \in N_v^P} \alpha_{vu}^{P} h_{P(v,u)}\right)$$

Multi-head attention is adopted to stabilize training:

$$h_v^P = \|_{k=1}^{K} \sigma\left(\sum_{u \in N_v^P} [\alpha_{vu}^{P}]_k \cdot h_{P(v,u)}\right)$$

Finally, path-level attention is used to weigh the contributions of different meta-paths:

$$h_v^{\Psi} = \sum_{P \in \Psi} \beta_P \cdot h_v^P$$

yielding semantic embeddings $h_r^{\Psi_R}, h_d^{\Psi_D}, h_g^{\Psi_G}$.

(iii) Global structural features

Although meta-paths capture local semantics, they cannot fully represent long-range dependencies. To complement this, we incorporate a global attention mechanism that directly models correlations across all nodes:

$$e_{i,j} = a^{\top} \mathrm{LeakyReLU}\,(W[h_i \parallel h_j]),\ \alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_k \exp(e_{i,k})},\ h_i^{'} = \sum_j \alpha_{i,j} h_j$$

Here, $h_i^{'}$ is the updated representation of node i, enriched by the weighted contributions of all other nodes. To ensure scalability, sparse neighborhoods and blockwise computation are applied (see Supplement~S1).

2.2.3 Feature Fusion

To integrate the complementary information extracted at different levels, we design a weighted fusion mechanism that combines local semantics, global structural patterns, and similarity-based embeddings. Taking disease nodes as an example, the feature fusion is formulated as:

$$h_d^{comb} = \lambda_1\, h_d^{\Psi_D} + \lambda_2\, h_d^{'} + \lambda_3\, [H_D]_d,$$

where $h_d^{\Psi_D}$ denotes the meta-path based representation, $h_d^{'}$ the global attention embedding, and $[H_D]_d$ the similarity-graph embedding. The coefficients $\lambda_1, \lambda_2, \lambda_3 \geq 0$ are learnable parameters that adaptively balance the contribution of each feature source.

The fused representation is further transformed into the final embedding through a linear mapping and non-linear activation:

$$h_d^{final} = \sigma\left(W_0 h_d^{comb}\right)$$

where W_0 is a trainable weight matrix and $\sigma(\cdot)$ is an activation function (e.g., ReLU). The same procedure is applied to drug and gene nodes, yielding h_r^{final} and h_g^{final}, respectively.

2.2.4 Association Score Calculation

Based on the final node embeddings, we predict the association probability between a drug–disease pair (i,j) using a sigmoid function applied to the dot product of their representations:

$$\grave{y}_{ij} = \sigma\left((h_i^{final})^{\top} h_j^{final}\right)$$

where h_i^{final} and h_j^{final} denote the fused feature vectors of drug i and disease j, respectively.
To optimize the model, we employ a binary cross-entropy loss with negative sampling:

$$L = - \sum_{(i,j)\in Y_P} \log \grave{y}_{ij} - \sum_{(i,j)\in Y_N} \log(1 - \grave{y}_{ij}$$

where Y_P and Y_N denote the sets of positive and negative drug–disease pairs, respectively. Minimizing this objective updates all trainable parameters of the framework, including GCN weights, attention modules, and fusion coefficients $\lambda_{1,2,3}$, in an end-to-end manner.

3. Results

3.1 Comparative Experiments

To evaluate the performance of the MedPathEx model, we conducted five-fold cross-validation and used the AUC, AP, and F1-score as evaluation metrics. The selected comparison models must have the ability to handle heterogeneous networks or utilize meta-paths for feature extraction, ensuring comparability in heterogeneous network analysis and node relationship capture. Therefore, we chose three categories of models: (1) traditional machine learning methods using the same multi-modal attributes as MedPathEx, (2) one comprises models that perform link prediction through meta-paths and (3) recent state-of-the-art heterogeneous graph learning models designed for drug–disease association prediction. Specifically, the models were as follows:

- **Random Forest (RF)** [33]models non-linear relationships among the integrated multi-modal attribute features using an ensemble of decision trees, serving as a classical machine learning baseline to assess the predictive contribution of attribute integration.
- **Support Vector Machine (SVM)** [34]trains a support vector machine classifier solely on the integrated multi-modal attributes of drugs and diseases, providing a non-graph baseline to evaluate predictive performance without network structural information.
- **Multi-Layer Perceptron (MLP)**[35] learns deep attribute representations through a fully connected neural network trained on the same integrated features, serving as a non-graph deep learning baseline for comparison with graph-based methods.
- **HAN** [36]leveraged heterogeneous graph neural networks to handle networks with varied node and relation types. It generated homogeneous subgraphs through meta-path random walks, used a hierarchical attention mechanism to aggregate neighboring node information, and learned structural and semantic information from multiple meta-paths.

- **MAGNN** employed a meta-path instance encoder on heterogeneous networks to integrate information from multiple meta-paths, capturing structural and semantic features of different node types to enhance node representations.
- **DRWBNCF**[37] integrated biological networks, such as drug-drug and disease-disease similarities. It used weighted bilinear graph convolution to capture node interactions and combined focal loss functions and graph regularization to model drug-disease associations.
- **FuHLDR**[16] A graph representation learning model based on biological heterogeneous information networks, FuHLDR fuses low-order representations from graph convolutional networks and high-order representations via meta-path strategies, using a random vector functional link network to predict drug-disease associations.
- **HDGAT**[38] integrates drug-disease similarities and associations in a heterogeneous graph, employing hierarchical and dynamic attention mechanisms to aggregate multi-level node information and residual connections to address over-smoothing, enabling drug-disease association prediction.

To ensure fairness, all models were evaluated under the same five-fold split. The comparative results are reported in Table 2. Experimental findings indicate that MedPathEx outperforms all baseline models, achieving superior AUC, AP, and F1 scores compared to other methods. Specifically, classical attribute-only baselines (SVM, RF, MLP), although not using graph structure, perform competitively on the fused multimodal attributes and in some cases surpass meta-path–only models (e.g., HAN), indicating that the integrated attribute space is highly discriminative. Nevertheless, methods that additionally exploit heterogeneous-graph semantics and stronger fusion mechanisms still yield higher overall accuracy.

Among meta-path–based heterogeneous GNNs, HAN and MAGNN capture semantic relations under different meta-paths and therefore improve over purely attribute-based models in some metrics; however, their limited use of diverse node attributes and global structure leaves a noticeable gap to MedPathEx.DRWBNCF enhances representation learning through weighted bilinear graph convolution and demonstrates improvements in capturing node interactions across biological networks. Nevertheless, its design lacks a mechanism to integrate global and local semantics comprehensively, resulting in relatively lower AP and F1 scores, which highlights the superior capacity of MedPathEx in multi-modal data integration. FuHLDR combines low-order GCN features with high-order meta-path information, showing stronger predictive power than DRWBNCF. Its advantage lies in capturing both local connectivity and higher-order semantics, but its fusion strategy is relatively simple, causing its performance to remain slightly inferior to MedPathEx.

HDGAT leverages hierarchical and dynamic attention to mitigate the over-smoothing problem and to strengthen multi-level feature aggregation, achieving results close to MedPathEx. However, it still exhibits small gaps in AUC and F1, indicating that its integration of global and local information is less effective.

Overall, MedPathEx demonstrates superior stability and scalability by jointly incorporating multi-modal similarities, meta-path-based local features, and global structural attention, achieving the best performance in the drug–disease association prediction task.

Table 2. Performance Comparison of Models

Model	AUC	AP	F1
RF	0.7952	0.7691	0.7868
SVM	0.7653	0.7451	0.7559
MLP	0.7852	0.7650	0.7706
HAN	0.7387	0.7043	0.7156
MAGNN	0.8014	0.7721	0.7836
DRWBNCF	0.8138	0.7966	0.8028
FuHLDR	0.8322	0.8265	0.8394
HDGAT	0.8499	0.8342	0.8536
MedPathEx	0.8559	0.8348	0.8668

3.2 Ablation Experiments

To evaluate the effectiveness of each component within the MedPathEx model, we conducted a series of ablation experiments to assess the impact of the different components on the performance of the model. The specific variations are as follows.

- **Experiment 1**: Removal of Similarity Network Features
  In this experiment, we removed the similarity network features from the model, ceasing to utilize the similarity information among drugs, diseases, and genes. The aim was to verify the impact of the inherent features of the node on the performance of the model.

- **Experiment 2**: Removal of Global Graph Features
  We removed the global heterogeneous graph learning component, relying solely on meta-paths and the similarity network to learn node representations. This experiment was designed to evaluate the contribution of global graph features to the model.

- **Experiment 3**: Removal of Meta-path Features
  Here removed the meta-path feature extraction component, relying instead on the original heterogeneous graph and similarity network. The objective was to assess the significance of the metapaths in the model.

- **Experiment 4**: Replacement of Attention Mechanism with Mean Aggregation
  In experiment, we replaced the attention mechanism in metapath aggregation with mean aggregation, no longer considering the importance of differences among different metapaths. The purpose of this study is to verify the effectiveness of the attention mechanism in enhancing the performance of the model.

The experimental results (see Table 3) indicate that each component significantly contributes to the final performance of the model. Removing meta-path features and global graph features leads to a substantial decline in performance, highlighting the importance of capturing high-order semantic relationships and the overall network structure. The removal of similarity network features also leads to performance reduction, although it is not as significant as in previous experiments, indicating that the incorporation of node attribute features aids in the prediction of drug-disease associations. The performance decreases the least when mean aggregation was used instead of the attention mechanism. This suggests that while the attention mechanism improves the role of metapaths, it does not have as much of an effect as other parts.

The full MedPathEx model performs the best on all tests (see Figure 2) owing to the important role that metapath features and global graph features play in capturing complex relationships and network structures. The similarity network features make better use of node attributes, and the attention mechanism makes meta-paths even more important during feature fusion. The synergy of all components enables MedPathEx to perform exceptionally well in drug–disease association prediction.

Table 3.Results of Ablation Experiment

	AUC	AP	F1
Experiment 1	0.8298	0.8008	0.8364
Experiment 2	0.8015	0.7984	0.8125
Experiment 3	0.7954	0.7891	0.8089
Experiment 4	0.8355	0.8204	0.8579
Complete Model	0.8559	0.8348	0.8668

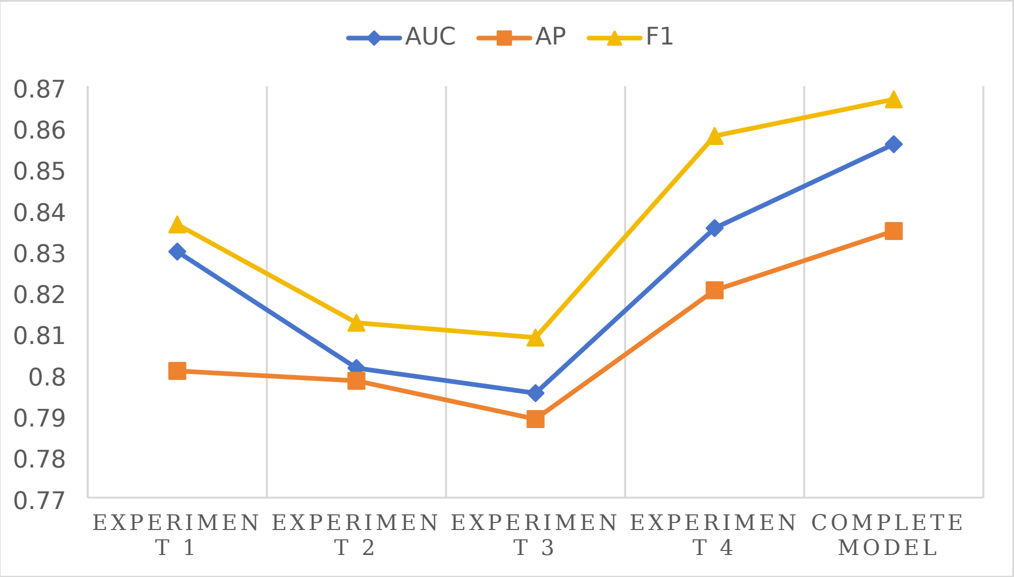

Figure 2.Performance Comparison of Ablation Experiments

3.3 Parameter Analysis

We evaluate the model's sensitivity to the parameters in this section. First, we analyze the impact of the fusion weights λ_1, λ_2, and λ_3 on the MedPathEx results, as shown in Figure 3. Because the fusion weights are constrained to sum to one, Figure 3 varies (λ_1, λ_2) on a grid while setting $\lambda_3=1-\lambda_1-\lambda_2$ at each point. The experimental results indicate that the meta-path features (λ_1) have the most significant impact, achieving the highest AUC of 0.8559 when $\lambda_1 = 0.7$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.1$. Global features (λ_2) provide complementary contributions, with moderate weights yielding better performance. Similarity network features (λ_3) have the least influence, and the model performs best when their weight is minimal. These results highlight the importance of prioritizing meta-path features while maintaining a balance between global and similarity network features.

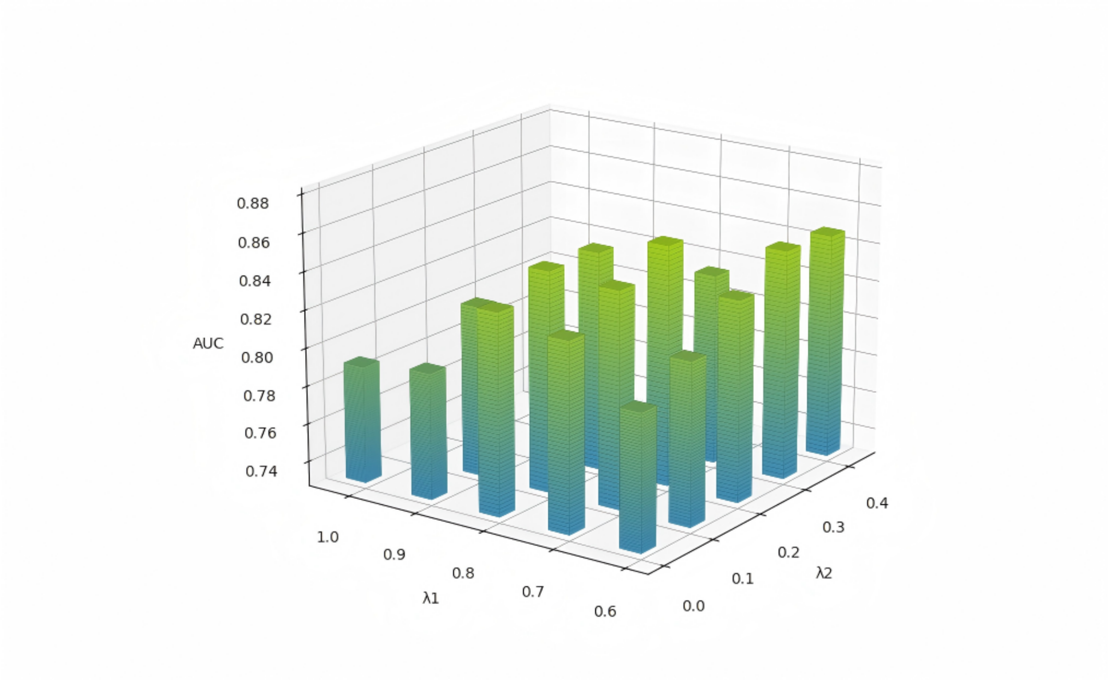

Figure 3. AUC as a function of λ_1, λ_2; $\lambda_3 = 1 - \lambda_1 - \lambda_2$.

Second, we analyze the effect of the number of heads K in the multi-head attention mechanism, as illustrated in Figure 4. The model achieves the best performance with $K = 8$. When $K = 1$, the model behaves as if no attention mechanism is applied, resulting in lower performance than other configurations. This demonstrates that the multi-head attention mechanism effectively distributes the weights across meta-paths and their instances. However, with $K = 16$, the performance decreases due to excessive parameters and potential overfitting. Therefore, we set $K = 8$ as the optimal value.

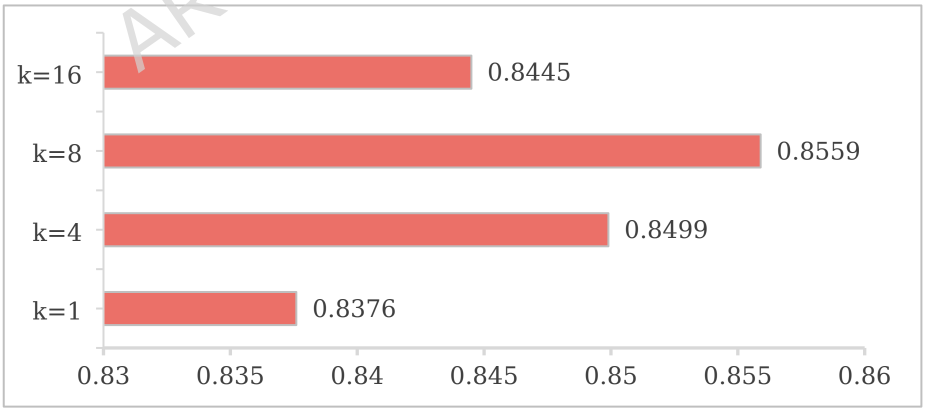

Figure 4. the perturbation of K on the AUC score

3.4 Case Studies

To validate the effectiveness of the model further, we verified the prediction results using publicly available literature from PubMed and ClinicalTrials.gov (https://clinicaltrials.gov/).

We selected coronary artery disease (CAD) and hypertension as validation cases owing to their high prevalence and mortality rates worldwide[39,40]. Evaluating the model's predictive performance for these two diseases demonstrated its ability to identify potential drugs and provide new research directions for their treatment.

CAD is a complex disease that involves inflammation, lipid metabolism, and vascular endothelial dysfunction [41]. Existing treatment methods for CAD are often ineffective and have significant side effects [41]. Therefore, the exploration of new therapeutic strategies is crucial. We predicted ten potential drugs associated with CAD (Table 4).

Table 4.The predicted Results for 10 Potential Drugs related to CAD

Disease Name	Drugbank ID	Drug Name	Evidence
Coronary Artery Disease	DB01076	Atorvastatin	ClinicalTrials/PMID:37852649/PMID：34926626
	DB01393	Bezafibrate	ClinicalTrials/PMID: 31272567
	DB01394	Colchicine	ClinicalTrials/PMID：37558377/PMID:34965168
	DB01427	Amrinone	PMID:7379283/PMID:6743422
	DB00758	Clopidogrel	ClinicalTrials/PMID: 18823340
	DB08162	fasudil	PMID: 30359818
	DB01001	Salbutamol	PMID:9651558
	DB01611	Hydroxychloroquine	ClinicalTrials/PMID:27372847
	DB06287	Temsirolimus	NA
	DB08910	Pomalidomide	NA

To elucidate the relationships between these drugs and CAD, we constructed a drug-gene-disease association network diagram, as shown in Figure 5. This diagram visualizes each drug and its potential target genes, offering an intuitive perspective on the mechanisms by which these drugs may be effective in treating CAD.

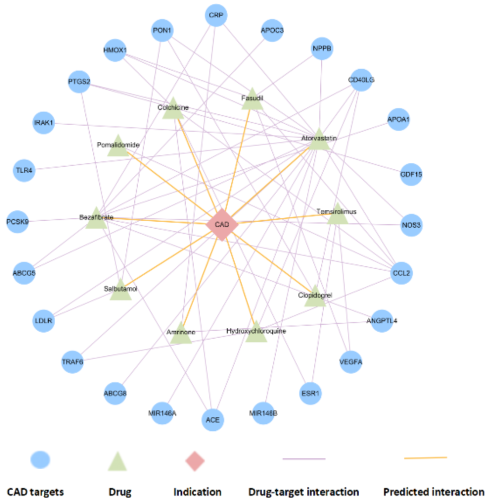

Figure 5.Drug-Gene-Disease Association Network of Potential CAD-Related Drugs

Among the predicted drugs, some have already been validated for the treatment of CAD and related cardiovascular diseases. For instance, Atorvastatin [42]is widely used to lower cholesterol levels and prevent cardiovascular events. Fasudil, a Rho kinase inhibitor, has not been extensively studied for CAD, but has a foundation in cardiovascular drug development. Fasudil shows potential for CAD treatment by inhibiting the Rho kinase pathway [43,44], particularly in cases involving coronary microvascular spasms, where it significantly improves myocardial ischemia [45].

Additionally, Temsirolimus and Pomalidomide, although relatively underexplored in CAD, have shown promise in modulating pathological calcification. Atherosclerosis is the key pathological basis of CAD, and vascular calcification is a critical factor in plaque maturation and instability. Therefore, targeting the calcification process may be an important strategy for CAD treatment [46]. Although relatively underexplored in CAD, Temsirolimus has shown promise in modulating pathological calcification by inhibiting the expression of key genes associated with calcification [47]. Similarly, Pomalidomide, which has not been extensively studied in CAD, mitigates plaque instability and calcification risk by reducing inflammatory responses [48,49].

We chose hypertension as the focus of our study because it is one of the most common cardiovascular diseases globally and is a major risk factor for cardiovascular conditions [40]. The ten potential drugs related to hypertension predicted by our model are listed in Table 5.

To visually represent the potential relationships between these drugs and hypertension, we constructed a drug-gene-disease association network diagram (Figure 6). This diagram illustrates the associations between each drug and its potential target genes, providing insights into the mechanisms by which these drugs may be effective in treating hypertension.

Table 5.Prediction Results of 10 Potential Drugs Related to Hypertension

Disease Name	Drugbank ID	Drugbank Name	Evidence
Hypertension	DB01359	Penbutolol	ClinicalTrials/PMID 7047173
	DB00275	Olmesartan medoxomil	PMID: 12076183
	DB00412	Rosiglitazone	ClinicalTrials/PMID 38411834
	DB01076	Atorvastatin	ClinicalTrials/PMID 37557013
	DB00783	Estradiol	ClinicalTrials/PMID 30896818
	DB01104	Sertraline	ClinicalTrials/PMID 29042710
	DB00482	Celecoxib	ClinicalTrials/PMID 30240679
	DB00310	chlorthalidone	ClinicalTrials/PMID 2258249
	DB01175	Escitalopram	NA
	DB00472	Fluoxetine	NA

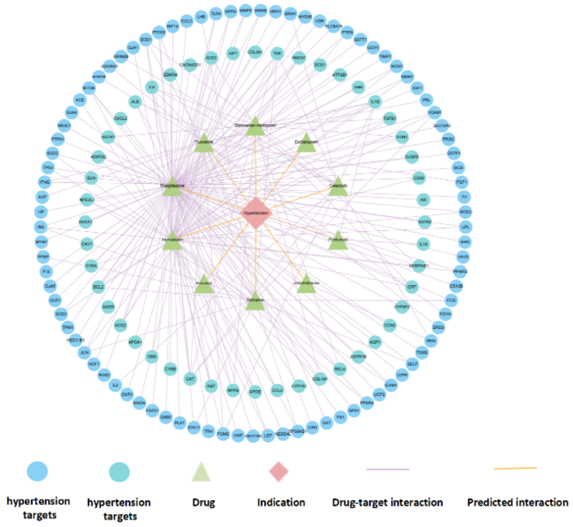

Figure 6.Drug-Gene-Disease Association Network of Potential hypertension-

Related Drugs

Among the predicted drugs, some have already been validated for the treatment of hypertension and related cardiovascular diseases. For example, Penbutolol [50], a β-blocker, is widely used to treat hypertension. Although less studied for hypertension, estradiol has shown potential in cardiovascular disease treatment, particularly in modulating vascular health and inflammatory responses [51]. Studies have indicated that estradiol protects against hypertension-related inflammation and tissue damage by regulating gap junction communication and reducing proinflammatory responses [52]. Additionally, a clinical trial (NCT00102141) suggested that estradiol combined with drospirenone may improve blood pressure control in postmenopausal women with hypertension.

Escitalopram and Fluoxetine, despite lacking direct evidence linking them to hypertension, offer new potential for hypertension treatment through the modulation of central nervous system function, which is associated with dysregulation of the central nervous system, particularly under conditions of stress and anxiety [53]. These drugs may indirectly influence blood pressure regulation by modulating serotonin levels in the central nervous system, thereby affecting the gene expression linked to blood pressure control. Further exploration of their potential applications in hypertension treatment could promote the development of multidimensional therapeutic strategies.

4. Discussion and Conclusion

Accurately predicting medication–disease associations is an essential step in drug development and has attracted substantial interest in recent years. In this study, we developed MedPathEx, a predictive framework that integrates multimodal biological information and combines meta-path–based local semantics with global structural features.

By incorporating diverse attributes—including drug chemical structures, ATC classifications, side-effect profiles, disease phenotypic and semantic information, and functional gene annotations—MedPathEx enriches the semantic representation of nodes within the heterogeneous network. In addition, by jointly modeling local meta-path semantics and global topological dependencies, the framework provides complementary perspectives that help improve the characterization of drug–disease relationships. Experimental results demonstrate that MedPathEx achieves improved predictive performance compared with representative baseline methods, indicating the effectiveness of integrating multimodal attributes and heterogeneous structural information for DDA prediction.

Despite these promising results, several limitations should be acknowledged. One important limitation is that the clinical applicability of the model has not yet been assessed using external real-world patient cohorts or prospective clinical studies. Moreover, although multiple data sources were integrated, the heterogeneous network still exhibits a degree of sparsity; incorporating additional biological entities and relationships—such as pathways, tissue-specific features, or molecular processes—may further enrich the representation space. In addition, the interpretability of the framework remains relatively limited. While meta-path attention offers some insight into semantic patterns, the overall representation learning process involves multiple neural components, making it difficult to fully trace how individual predictions are generated.

Future work will focus on incorporating more comprehensive biomedical entities, enhancing model interpretability through explainable learning mechanisms, and validating high-confidence predictions using external datasets or real-world clinical evidence. These efforts will help further improve the robustness and translational potential of MedPathEx within computational drug discovery.

Data availability

All datasets used in this study are publicly available from their respective official sources:
BioSNAP (https://snap.stanford.edu/biodata/),
CTD (https://ctdbase.org/),
PharmGKB (https://www.pharmgkb.org/),
DrugBank (https://go.drugbank.com/),
ChEMBL (https://www.ebi.ac.uk/chembl/),
SIDER (https://sideeffects.embl.de/),
OMIM (https://www.omim.org/),
MeSH (https://www.ncbi.nlm.nih.gov/mesh/),
Gene Ontology Consortium (http://geneontology.org/).

Code availability

The implementation of MedPathEx and the preprocessed data is available at https://github.com/wwiswwlucky/MedPathEx.

References

1. Wang, Y., Song, J., Dai, Q. & Duan, X. Hierarchical Negative Sampling Based Graph Contrastive Learning Approach for Drug-Disease Association

Prediction. *IEEE Journal of Biomedical and Health Informatics* **28**, 3146–3157 (2024).

2. Pasrija, P., Jha, P., Upadhyaya, P., Khan, M. S. & Chopra, M. Machine Learning and Artificial Intelligence: A Paradigm Shift in Big Data-Driven Drug Design and Discovery. *Curr Top Med Chem* **22**, 1692–1727 (2022).
3. Xiang, Z., Yunqiu, Z., Shaodan, S. & Liman, Z. Research on Drug Knowledge Discovery Method Fusing Meta-path Features of Heterogeneous Knowledge Network: Taking the Prediction of Drug-Target Relations as An Example. *Data Analysis and Knowledge Discovery %V %N* 1–19 (2023).
4. Gu, Y., Zheng, S., Yin, Q., Jiang, R. & Li, J. REDDA: Integrating multiple biological relations to heterogeneous graph neural network for drug-disease association prediction. *Computers in Biology and Medicine* **150**, 106127 (2022).
5. Jarada, T. N., Rokne, J. G. & Alhajj, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of Cheminformatics* **12**, 46 (2020).
6. Pan, X. *et al.* Deep learning for drug repurposing: Methods, databases, and applications. *WIREs Computational Molecular Science* **12**, e1597 (2022).
7. Luo, H. *et al.* Biomedical data and computational models for drug repositioning: a comprehensive review. *Briefings in Bioinformatics* **22**, 1604–1619 (2020).
8. Kim, Y., Jung, Y.-S., Park, J.-H., Kim, S.-J. & Cho, Y.-R. Drug-Disease

Association Prediction Using Heterogeneous Networks for Computational Drug Repositioning. *Biomolecules* **12**, (2022).

9. Liu, H., Song, Y., Guan, J., Luo, L. & Zhuang, Z. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinformatics* **17**, 539 (2016).
10. Wu, G., Liu, J. & Yue, X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC Bioinformatics* **20**, 134 (2019).
11. Jamali, A. A., Tan, Y., Kusalik, A. & Wu, F.-X. NTD-DR: Nonnegative tensor decomposition for drug repositioning. *PLOS ONE* **17**, 1–18 (2022).
12. Wu, P. *et al.* Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension. *Nature Communications* **13**, 46 (2022).
13. Huang, S., Wang, M., Zheng, X., Chen, J. & Tang, C. Hierarchical and Dynamic Graph Attention Network for Drug-Disease Association Prediction. *IEEE Journal of Biomedical and Health Informatics* **28**, 2416–2427 (2024).
14. Zhang, M.-L. *et al.* RLFDDA: a meta-path based graph representation learning model for drug–disease association prediction. *BMC bioinformatics* **23**, 516 (2022).
15. Sun, X., Jia, X., Lu, Z., Tang, J. & Li, M. Drug repositioning with adaptive graph convolutional networks. *Bioinformatics* **40**, btad748 (2024).
16. Zhao, B.-W. *et al.* Fusing Higher and Lower-Order Biological Information for

Drug Repositioning via Graph Representation Learning. *IEEE Transactions on Emerging Topics in Computing* **12**, 163–176 (2024).

17. Muzio, G., O'Bray, L. & Borgwardt, K. Biological network analysis with deep learning. *Briefings in Bioinformatics* **22**, 1515–1530 (2020).

18. Shi, C., Li, Y., Zhang, J., Sun, Y. & Yu, P. S. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* **29**, 17–37 (2017).

19. Jiao, Q., Jiang, Y., Zhang, Y., Wang, Y. & Li, J. Nsap: A neighborhood subgraph aggregation method for drug-disease association prediction. in *International Conference on Intelligent Computing* 79–91 (Springer, 2022).

20. Zhou, R. *et al.* NEDD: a network embedding based method for predicting drug-disease associations. *BMC Bioinformatics* **21**, 387 (2020).

21. Fu, X., Zhang, J., Meng, Z. & King, I. MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. in *Proceedings of The Web Conference 2020* 2331–2341 (Association for Computing Machinery, New York, NY, USA, 2020). doi:10.1145/3366423.3380297.

22. Zitnik, M., Sosic, R. & Leskovec, J. BioSNAP Datasets: Stanford biomedical network dataset collection. *Note: http://snap. stanford. edu/biodata Cited by* **5**, (2018).

23. Davis, A. P. *et al.* Comparative toxicogenomics database (CTD): update 2023. *Nucleic acids research* **51**, D1257–D1262 (2023).

24. Thorn, C. F., Klein, T. E. & Altman, R. B. PharmGKB: the pharmacogenomics

knowledge base. *Pharmacogenomics: Methods and Protocols* 311–320 (2013).

25. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research* **47**, D930–D940 (2019).
26. Knox, C. *et al.* DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic acids research* **52**, D1265–D1275 (2024).
27. Jaccard, P. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* **44**, 223–270 (1908).
28. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic acids research* **44**, D1075–D1079 (2016).
29. van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. & Leunissen, J. A. M. A text-mining analysis of the human phenome. *Eur J Hum Genet* **14**, 535–542 (2006).
30. Yu, G. Using meshes for MeSH term enrichment and semantic analyses. *Bioinformatics* **34**, 3766–3767 (2018).
31. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
32. Zhang, S., Tong, H., Xu, J. & Maciejewski, R. Graph convolutional networks: a comprehensive review. *Computational Social Networks* **6**, 1–23 (2019).
33. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
34. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their Applications* **13**, 18–28

(1998).

35. Kruse, R., Mostaghim, S., Borgelt, C., Braune, C. & Steinbrecher, M. Multi-layer Perceptrons. in *Computational Intelligence: A Methodological Introduction* 53–124 (Springer International Publishing, Cham, 2022). doi:10.1007/978-3-030-42227-1_5.
36. Wang, X. *et al.* Heterogeneous graph attention network. in *The world wide web conference* 2022–2032 (2019).
37. Meng, Y. *et al.* A weighted bilinear neural collaborative filtering approach for drug repositioning. *Briefings in Bioinformatics* **23**, bbab581 (2022).
38. Li, Y. *et al.* A comparative benchmarking and evaluation framework for heterogeneous network-based drug repositioning methods. *Brief Bioinform* **25**, (2024).
39. Brown, J. C., Gerhardt, T. E. & Kwon, E. Risk factors for coronary artery disease. (2020).
40. Tackling, G. & Borhade, M. B. Hypertensive heart disease. in *StatPearls [Internet]* (StatPearls publishing, 2023).
41. Alfaddagh, A. *et al.* Inflammation and cardiovascular disease: From mechanisms to therapeutics. *American journal of preventive cardiology* **4**, 100130 (2020).
42. McIver, L. A. & Siddique, M. S. Atorvastatin. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024).
43. Pireddu, R. *et al.* Pyridylthiazole-based ureas as inhibitors of Rho associated

protein kinases (ROCK1 and 2). *Medchemcomm* **3**, 699–709 (2012).

44. Nohria, A. *et al.* Rho kinase inhibition improves endothelial function in human subjects with coronary artery disease. *Circ Res* **99**, 1426–1432 (2006).

45. Mohri, M., Shimokawa, H., Hirakawa, Y., Masumoto, A. & Takeshita, A. Rho-kinase inhibition with intracoronary fasudil prevents myocardial ischemia in patients with coronary microvascular spasm. *J Am Coll Cardiol* **41**, 15–19 (2003).

46. Loscalzo, J. Molecular interaction networks and drug development: Novel approach to drug target identification and drug repositioning. *FASEB J* **37**, e22660 (2023).

47. Panwar, V. *et al.* Multifaceted role of mTOR (mammalian target of rapamycin) signaling pathway in human health and disease. *Signal Transduct Target Ther* **8**, 375 (2023).

48. Chanan-Khan, A. A. *et al.* Pomalidomide: the new immunomodulatory agent for the treatment of multiple myeloma. *Blood Cancer J* **3**, e143 (2013).

49. Chamberlain, P. P. *et al.* Evolution of Cereblon-Mediated Protein Degradation as a Therapeutic Modality. *ACS Med Chem Lett* **10**, 1592–1602 (2019).

50. Wiysonge, C. S., Volmink, J. & Opie, L. H. Beta-blockers and the treatment of hypertension: it is time to move on. *Cardiovasc J Afr* **18**, 351–352 (2007).

51. Knowlton, A. A. & Lee, A. R. Estrogen and the cardiovascular system.

Pharmacol Ther **135**, 54–70 (2012).

52. Ni, X. *et al.* β-estradiol alleviates hypertension- and concanavalin A-mediated inflammatory responses via modulation of connexins in peripheral blood lymphocytes. *Mol Med Rep* **19**, 3743–3755 (2019).

53. Qiu, T., Jiang, Z., Chen, X., Dai, Y. & Zhao, H. Comorbidity of Anxiety and Hypertension: Common Risk Factors and Potential Mechanisms. *Int J Hypertens* **2023**, 9619388 (2023).

Funding

This research received no external funding.

Acknowledgements

The authors thank the referees for suggestions that helped improve the paper substantially.

Author information

Contributions

Shengnan Wu and Wen Wang produced the main ideas, and did the modeling, computation and analysis and also wrote the manuscript. Huizhi Jiao,Danhong Dong ,Kexin Zhang and Xuechen Luo provided supervision and effective scientific advice and related ideas, research design guidance, and added value to the article through editing and contributing completions. All authors contributed to the article and approved the submitted version.

Corresponding authors

Correspondence to Shengnan Wu.

Ethics declarations

Competing interests

The authors declare no competing interests.