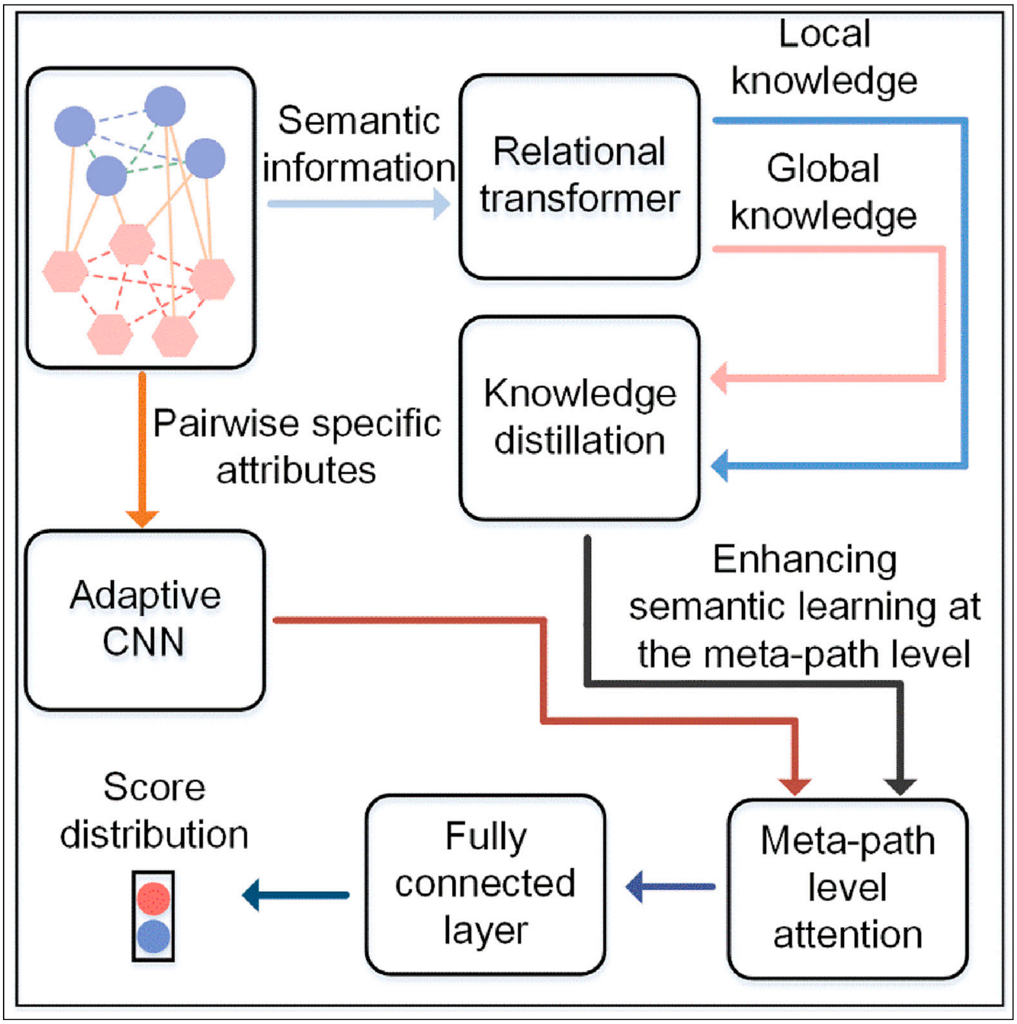
iScience

CellPress
OPEN ACCESS

Article

Graph reasoning method enhanced by relational transformers and knowledge distillation for drug-related side effect prediction



Honglei Bai, Siyuan Lu, Tiangang Zhang, Hui Cui, Toshiya Nakaguchi, Ping Xuan

pxuan@stu.edu.cn

Highlights

Two heterogeneous graphs were constructed based on two drug similarities

We established the relational transformer module to enhance the semantic information

Two knowledge distillation strategies were designed to extract knowledge

Adaptive convolutional neural networks were constructed to distinguish features

Bai et al., iScience *27*, 109571
June 21, 2024 © 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.109571

iScience

CellPress
OPEN ACCESS

Article

Graph reasoning method enhanced by relational transformers and knowledge distillation for drug-related side effect prediction

Honglei Bai,[1] Siyuan Lu,[1] Tiangang Zhang,[1,2] Hui Cui,[3] Toshiya Nakaguchi,[4] and Ping Xuan[5,6,*]

SUMMARY

Identifying the side effects related to drugs is beneficial for reducing the risk of drug development failure and saving the drug development cost. We proposed a graph reasoning method, RKDSP, to fuse the semantics of multiple connection relationships, the local knowledge within each meta-path, the global knowledge among multiple meta-paths, and the attributes of the drug and side effect node pairs. We constructed drug-side effect heterogeneous graphs consisting of the drugs, side effects, and their similarity and association connections. Multiple relational transformers were established to learn node features from diverse meta-path semantic perspectives. A knowledge distillation module was constructed to learn local and global knowledge of multiple meta-paths. Finally, an adaptive convolutional neural network-based strategy was presented to adaptively encode the attributes of each drug-side effect node pair. The experimental results demonstrated that RKDSP outperforms the compared state-of-the-art prediction approaches.

INTRODUCTION

Failure of drug clinical trials due to adverse reactions following normal drug use can pose health risks to participants and lead to significant financial losses.[1–3] Therefore, exploring the potential association between drugs and side effects can improve our understanding of drugs and reduce the cost of their development.[4,5] New computational predictions of drug side effect associations can screen drugs for potential side effects for biologists to conduct further experiments.[6–8] Earlier researchers proposed molecular docking-based approaches to estimate association score for drug side effect. These methods[9,10] use 3D structural information of drug-related proteins to screen potential drug candidates for side effects. However, the docking-based approach cannot be applied to those drugs of interest for which protein structure information is incomplete.[11]

Machine learning-based models can achieve better performance in predicting drug side effect associations. Zhao[12] used a minimum redundancy maximum correlation algorithm to prioritize all drug side effect node pairs. A triple matrix decomposition and a non-negative matrix decomposition model were used, respectively, by Guo et al.[13] and Galeano et al.[14] to predict new potential side effects. Other models based on machine learning methods, such as multi-label learning,[15] random walk,[16] graph regularization matrix decomposition models,[17] and multi-core learning,[18] also achieved excellent prediction performance. Nguyen et al. reviewed the studies on adverse drug reactions and evaluated the performances of multiple methods for predicting the adverse drug reactions.[19] Liu et al. constructed a prediction model based on support vector machine.[20] Alpay et al. assessed the ability of the previous methods in inferring the side effects which were contained by the pharmaceutical literature.[21] Galeano[22] constructed a geometric self-expressive model (GSEM) to learn a global optimization representation of drugs and side effects and identify the potential candidate side effects for the drugs. Several models were constructed based on the network-based inference (NBI) method,[23] multiple evidence fusion,[24] and multi-kernel learning[18] for predicting the potential side effects for drugs. These methods integrate surface-level data regarding drug side effect connections, but they are unable to disclose the complex associations between drug and side effect nodes.

More deep learning-based methods can extract deep features to predict possible side effects of drugs as deep learning techniques advance. Zhao et al.[25] suggested a multilayer perceptron-based model for predicting novel side effects by combining multi-source similarity data of the drugs and side effects. Li[26,27] developed a graph neural network-based model to calculate the possibility that a drug might have side effects. Xuan et al.[28] embedded the extensive similarity of drugs into a matrix and proposed a method based on graph convolutional

[1]School of Computer Science and Technology, Heilongjiang University, Harbin, China
[2]School of Mathematical Science, Heilongjiang University, Harbin, China
[3]Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC, Australia
[4]Center for Frontier Medical Engineering, Chiba University, Chiba, Japan
[5]Department of Computer Science and Technology, Shantou University, Shantou, China
[6]Lead contact
*Correspondence: pxuan@stu.edu.cn

https://doi.org/10.1016/j.isci.2024.109571

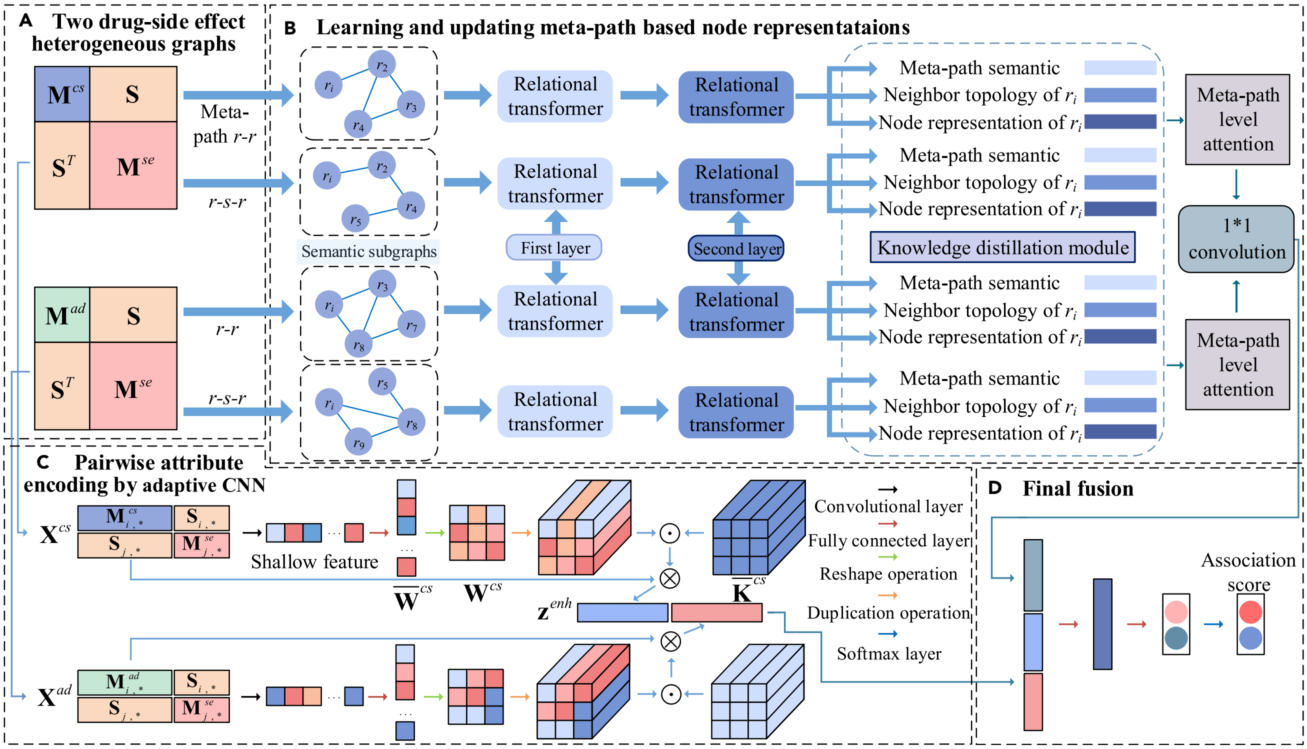


Figure 1. Framework of the constructed RKDSP model for predicting drug-related side effects

(A) construct two drug-side effect heterogeneous graphs based on two types of drug similarities (B) learn and update the representations of the drug and side effect nodes based on relational transformer and knowledge distillation on the semantic subgraphs (C) encode the attribute representation of each drug-side effect node pair by the adaptive convolutional neural networks (D) fuse multiple representations to estimate the association score for each pair of drug and side effect nodes.

auto-encoders. Zitnik et al. utilized the graph convolutional networks to model polypharmacy side effects.[29] Lin et al. proposed the twin network[30] and comparative learning-based methods to predict the drug-drug interactions,[31] respectively. Wang et al. designed a deep neural network model to utilize the chemical information of drugs, the biological one as well as their biomedical information.[32] Traditional transformer has been utilized in the link prediction, and it demonstrated the decent prediction performance.[33–37] Recently, several methods were proposed based on network analysis,[38] feature analysis,[39] and knowledge graph embedding.[40] However, the traditional transformer ignored the diverse connection relationships. The aforementioned approach ignores the importance of local knowledge in the connection relationships and global knowledge between the connection relationships.

A method referred to the relational transformers and knowledge distillation for drug-related side effect prediction (RKDSP) is presented to predict the potential side effects for the interested drugs (Figure 1). The unique contributions of this model are summarized as follows,

(1) The drug similarities based on their chemical substructures and the ones based on their associated diseases formed the specific topological structures. Most previous approaches ignored the importance of integrating drug similarity based on associated diseases for the prediction of potential drug side effects. Therefore, we constructed two heterogeneous graphs according to these two types of drug similarities (Figure 2). Each of heterogeneous graph is composed of the drug and side effect nodes, and the similarity connections and the association ones among these nodes.
(2) Most previous meta-path-based methods only aggregate the neighbors of the target node based on the current meta-path but ignore the attribute information of the connecting relationships in the meta-path. Therefore, we establish multiple meta-paths to embed the semantics of the similarity and association connections. The relational transformer-based module is constructed to enhance the semantic information of drug and side effect nodes within multiple meta-paths and to learn target node features based on meta-path instances.
(3) Multiple nodes within the same meta-path have locally tight connections between them, and individual nodes between multiple different meta-paths contain deep global relationships. Most prior meta-path learning approaches did not fully utilize this local and global knowledge. Thus, the knowledge distillation strategy is designed to update the node feature representation derived from these meta-paths (Figure 3). As the semantic learning from multiple meta-paths has various importance for the drug-side effect association prediction, the meta-path-level attention is presented to distinguish their importance.

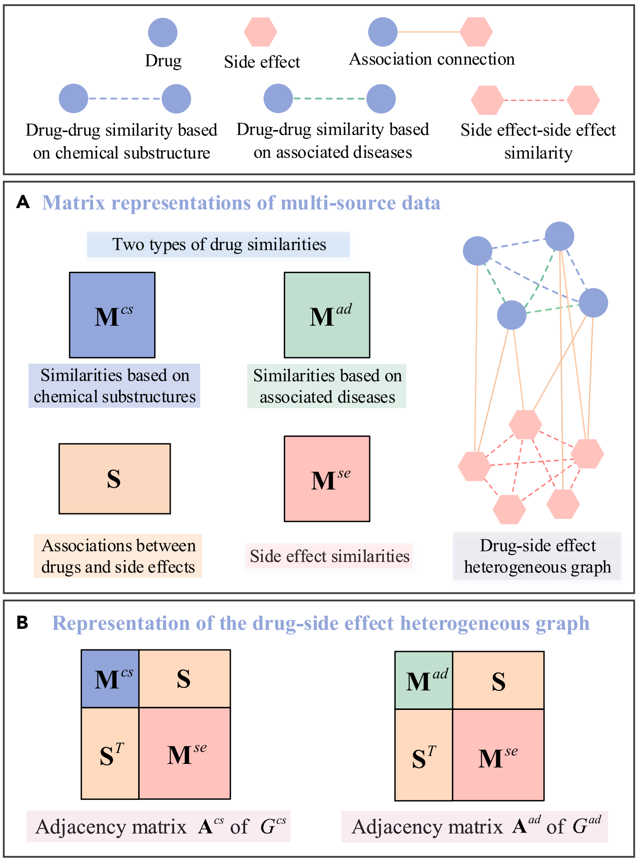

Drug
Side effect
Association connection
Drug-drug similarity based on chemical substructure
Drug-drug similarity based on associated diseases
Side effect-side effect similarity
A
Matrix representations of multi-source data
Two types of drug similarities
$\mathbf{M}^{cs}$
$\mathbf{M}^{ad}$
Similarities based on chemical substructures
Similarities based on associated diseases
$\mathbf{S}$
$\mathbf{M}^{se}$
Associations between drugs and side effects
Side effect similarities
Drug-side effect heterogeneous graph
B
Representation of the drug-side effect heterogeneous graph
$\mathbf{M}^{cs}$
$\mathbf{S}$
$\mathbf{S}^{T}$
$\mathbf{M}^{se}$
$\mathbf{M}^{ad}$
$\mathbf{S}$
$\mathbf{S}^{T}$
$\mathbf{M}^{se}$
Adjacency matrix $\mathbf{A}^{cs}$ of G^{cs}
Adjacency matrix $\mathbf{A}^{ad}$ of G^{ad}

Figure 2. Matrix notations of multi-source data and construction of an adjacency matrix for the entire drug-side effect heterogeneous graph
(A) Matrix notations of multi-source data, including two types of drug similarities, associations between drugs and side effects, and the side effect similarities.
(B) Construct the adjacency matrix of the drug-side effect heterogeneous graph.

(4) Most previously learned nodes' convolution operations on pairwise attributes focus on using standard convolution kernels, ignoring the importance of individual locations in the feature graph. To overcome this disadvantage, we constructed a module based on the convolutional neural networks with adaptive convolution kernels for learning the attributes of a pair of drug and side effect nodes. The adaptive convolution strategy enhances the ability in learning the local more important features for each drug-side effect pair.

RESULTS

Evaluation metrics

In order to assess the RKDSP's capacity to detect potential side effects of drugs, we conducted a 5-fold cross-validation. In particular, we separated all positive samples of drug-side effect associations known to exist into five random folds. In each fold, we trained the model using four known associations and an equal number of unknown associations, and the remaining positive and negative samples made up the test set. As an evaluation metric, we used the area under the receiver operating characteristic (ROC) curve (AUC).[41] The dataset has 2,887,722 unobserved drug-side effect node pairs compared to 80,164 known drug side effect associations. Therefore, as a further evaluation metric, we selected the area under the PR curve (AUPR).[42] AUPR can better assess the prediction performance when the unknown and known associations are heavily unbalanced. For each drug, we predicted its association score with 4,192 side effects. The recall of the top *k* also had to be calculated since biologists regularly select the top-ranked candidates for additional experimental validation.

Parameter settings

For the relational transformer-based module, we fine-tuned the features dimensions in the first and second layers with the values in {1600, 2400, 3200} and {800, 1200, 1600}. Finally, the dimension numbers are 2,400 and 1,200, respectively. The number of layers within the relational transformer was selected from {2, 3, 4}, and the best performance was achieved when it was 2. In the convolution layer, the number of adaptive convolution kernels was set to 16. We set the number of layers of the transformer of the meta-path-based representation learning module to 2. The output dimensions of the first-level and the second-level relation transformer are 2,400 and 1,200, respectively. The number of rounds,

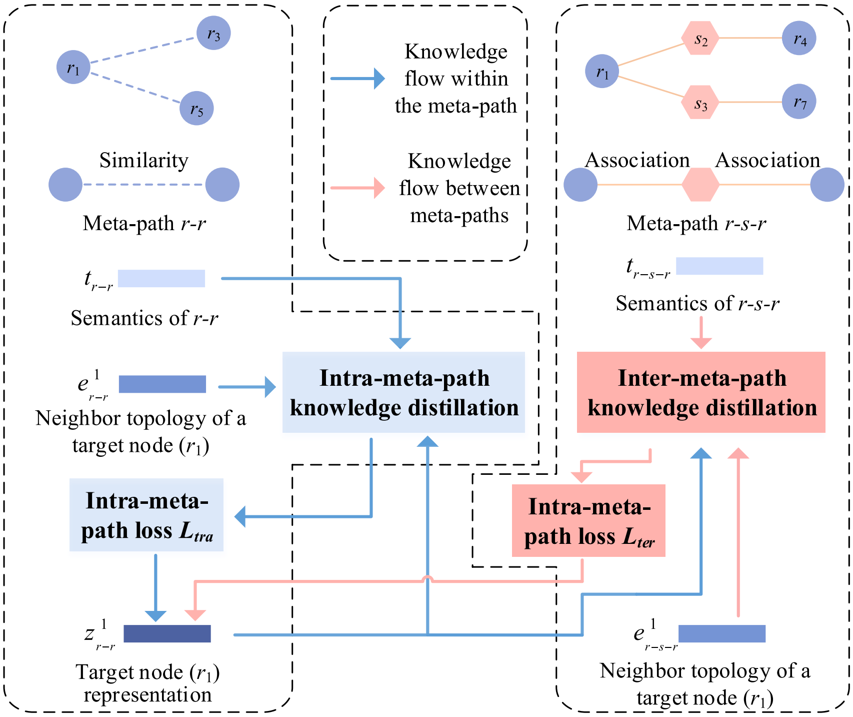

r_3
r_1
r_5
Similarity
Meta-path r-r
Knowledge flow within the meta-path
Knowledge flow between meta-paths
s_2
r_4
r_1
s_3
r_7
Association
Association
Meta-path r-s-r
t_{r-r}
Semantics of r-r
t_{r-s-r}
Semantics of r-s-r
e^1_{r-r}
Neighbor topology of a target node (r_1)
Intra-meta-path knowledge distillation
Inter-meta-path knowledge distillation
Intra-meta-path loss L_{tra}
Intra-meta-path loss L_{ter}
z^1_{r-r}
Target node (r_1) representation
e^1_{r-s-r}
Neighbor topology of a target node (r_1)

Figure 3. Illustration of the knowledge distillation strategy for capturing the local knowledge within meta-paths and the global knowledge among multiple meta-paths

learning rate, and batch size were 50, 0.0005, and 32, respectively. With an NVIDIA GeForce RTX 2080TI, we apply the suggested method based on the PyTorch framework.

Ablation studies

We verified the effectiveness of meta-path semantic learning, knowledge distillation (both intra-(LKD) and inter-meta-path knowledge distillation (GKD)), meta-path level attention (MPA), and pairwise dual-view attribute encoding (PDE) with ablation studies. The experimental results appear in Table 1. The best AUC and AUPR were achieved by the final models, which also included LKD, GKD, MPA, and PDE, with scores of 0.970 and 0.353, respectively. To evaluate the effectiveness of the meta-path semantic learning based on the relational transformer, a model without the learning (MSL) was constructed. Eliminating the learning caused 4.1% and 6.6% decrease in AUC and AUPR compared the complete prediction model. Therefore, learning the semantic information of multiple meta-paths through the relational transformer is necessary for improving the prediction performance. Without LKD (GKD), the AUC and AUPR decreased by 3.6% (3.2%) and 4.8% (5.4%), respectively, compared to the final model because drug and side effect nodes contain different semantic information in their features within and between meta-paths. The significance for enhancing local and global information when predicting drug-related side effects is further confirmed by this result. AUC and AUPR were reduced by 2.8% and 2.4%, respectively, when MPA was removed. This decline results from the fact that various meta-paths have unique semantic information. Assigning greater weights to meta-paths containing more important

Table 1. Ablation study results of our method

MSL	LKD	GKD	MPA	PDE	SCS	SAD	Average AUC	Average AUPR
×	✓	✓	✓	✓	✓	✓	0.929	0.287
✓	×	✓	✓	✓	✓	✓	0.934	0.305
✓	✓	×	✓	✓	✓	✓	0.938	0.299
✓	✓	✓	×	✓	✓	✓	0.942	0.329
✓	✓	✓	✓	×	✓	✓	0.911	0.285
✓	✓	✓	✓	✓	×	✓	0.913	0.299
✓	✓	✓	✓	✓	✓	×	0.928	0.304
✓	✓	✓	✓	✓	✓	✓	0.970	0.353

Table 2. Experimental results for comparing the proposed relational transformer strategy with the previous strategies

Strategy	$RKDSP_{drug}$	$RKDSP_{rtnet}$	$RKDSP_{reltr}$	$RKDSP_{know}$	$RKDSP$
Average AUC	0.956	0.958	0.967	0.964	0.970
Average AUPR	0.338	0.336	0.345	0.341	0.353

information while appropriately reducing those containing edge information is more effective than directly aggregating each meta-path. Our method improved the accuracy of AUC and AUPR by 5.9% and 6.8% compared to the model with PDE removed. This result indicates the importance of adaptively assigning weights to the convolution kernel to distinguish the importance of different positions in the feature map. In addition, the AUC and AUPR of the model eliminating the drug similarities based on the chemical substructures (SCS) decreased by 5.7% and 5.4%, respectively. The AUC and AUPR of the model without the drug similarities (SAD) calculated by their associated diseases dropped by 4.2% and 4.9%. The results showed that drug similarities based on the chemical substructures contributes the most to the improvement of prediction performance.

To assess the effectiveness of the proposed relational transformer for the drug-side effect association prediction, we compared the model with it and the models with other transformers (Table 2). Unlike our prediction method, the method[43] ignored the semantic information within multiple meta-paths about the drug nodes, and it constructed the separated modules for the drug nodes and the protein nodes for learning their topology information. To compare with the method, we constructed a prediction instrance by replacing the relational transformer with its transformer strategy. The instrance is referred to as $RKDSP_{drug}$. The researchers also constructed the prediction models based on the relation transformer with an attention,[44] the relational converter with the encoder and decoder framework,[45] and the position encoding strategy.[33] Thus, three prediction instances were constructed by replacing our designed relational transformer with these transformer strategies, respectively. These instances were named as $RKDSP_{rtnet}$, $RKDSP_{reltr}$, and $RKDSP_{know}$.

Comparison with other methods

Seven state-of-the-art methods were compared to the suggested method, including FGRMF,[17] RW-SHIN,[16] GraRep,[38] Galeano's method,[14] EEG-DTI,[46] DTI-MGNN,[26] GCRS,[28] GSEM,[22] KGDNN,[41] and SDPred.[25] We set the hyperparameters in each comparison model according to the best parameters reported in the corresponding literature, where $\lambda = 4$ were used for FGRMF, $\alpha = 0.05$ for Galeano's method, number of layers of the graph convolutional network $= 3$ for EEG-DTI, $L_{CE} = L_{CD} = 2$ for GCRS, and $lr = 1e - 4$ and $r = 32$ for SDPred.

The average ROC and PR curves for all 708 drugs are displayed in Figure 4 for every method (Table 3). RKDSP reached the highest mean AUC 0.970, and it is 5.1%, 7.8%, 4.3%, 5.8%, 3.7%, 3.1%, 1.3%, 2.6%, 4.5%, and 2.4% greater than that of FGRMF,[17] RW-SHIN,[16] GraRep,[38] Galeano's method,[14] EEG-DTI,[46] DTI-MGNN,[26] GCRS,[28] GSEM,[22] KGDNN,[40] and SDPred.[25] RKDSP obtained the highest average AUPR for all the drugs (AUPR = 0.353), and its AUPR is 21.2%, 25.4%, 17.4%, 22.2%, 16.4%, 15.5%, 8.1%, 9.3%, 12.7%, and 12.7% higher than FGRMF,[17] RW-SHIN,[16] GraRep,[38] Galeano's method,[14] EEG-DTI,[46] DTI-MGNN,[26] GCRS,[28] GSEM,[22] KGDNN,[41] and SDPred,[25] respectively.

On both the AUC and AUPR evaluations, RKDSP performs best. The specific topology of nodes is learned through the second-best GCRS. The SDPred model performs the third best and takes into account various kinds of similarity information. The graph neural network-based DTI-MGNN and the graph convolutional network-based EEG-DTI also learned the deep embedding of nodes to predict potential side effects of drugs and also achieved good performance (DTI-MGNN's AUPR = 0.198, EEG-DTI's AUC = 0.933). The AUC of FGRMF is slightly higher than Galeano's method (0.7% difference), but its AUPR is 4.8% higher than that method. The possible reason is that both are matrix decomposition-based methods, but FGRMF aggregates additional drug similarity information. The fact that RW-SHIN only learns the topological information of drug nodes might be the cause of its worst performance. We discovered that the predicting performance could be improved through the integration of multiple drug similarities (AUPR was 4.8% lower for the Galeano method than for FGRMF). In addition, the proposed method RKDSP builds multiple meta-paths to deeply integrate semantic information and extracts local knowledge within meta-paths and global knowledge between meta-paths in heterogeneous graphs for guiding the deep feature learning process. The machine learning-based methods GCRS, SDPred, DTI-MGNN, and EEG-DTI captured the topological features in the heterogeneous graph with higher performance than the machine learning-based methods FGRMF, Galeano's method, and RW-SHIN. In summary, integrating semantic information in different meta-paths and capturing the depth properties of the nodes are crucial to improve the prediction performance.

The recall rate of the top-k potential side effects is displayed in (Figure 5). More drug-related side effects were correctly screened when the recall was higher. RKDSP outperformed all other methods for different thresholds k, ranking 53.8% in the top 30, 74.4% in the top 90, and 78.8% in the top 120. With recall rates of 47.0%, 66.8%, and 71.9% in the top 30, 90, and 120, respectively, GCRS placed in second overall. In the top 30, 90, and 120, SDPred received rankings of 41.8%, 62.3%, and 67.4%, respectively. With recall rates of 36.6%, 57.5%, and 62.9% for DTI-MGNN and 34.7%, 54.6%, and 59.8% for EEG-DTI in the top 30, 90, and 120, respectively, they were very close to each other. The recall rates of GSEM are 44.2%, 65.1%, and 70.3% when k was 30, 60, 90. The corresponding recall rates of KDGNN are 41.6%, 58.7%, and 63.1%, and they are 6.1%, 5.2%, and 4.8% higher than that of GraRep. Although FGRMF and Galeano's method have nearly identical recall rates, the former model performs better in the top 30, 90, and 120 samples by 0.5%, 0.8%, and 1.3%, respectively. RW-SHIN ranks 23.7%, 41.3%, and 47.2%, and its performance is the worst.

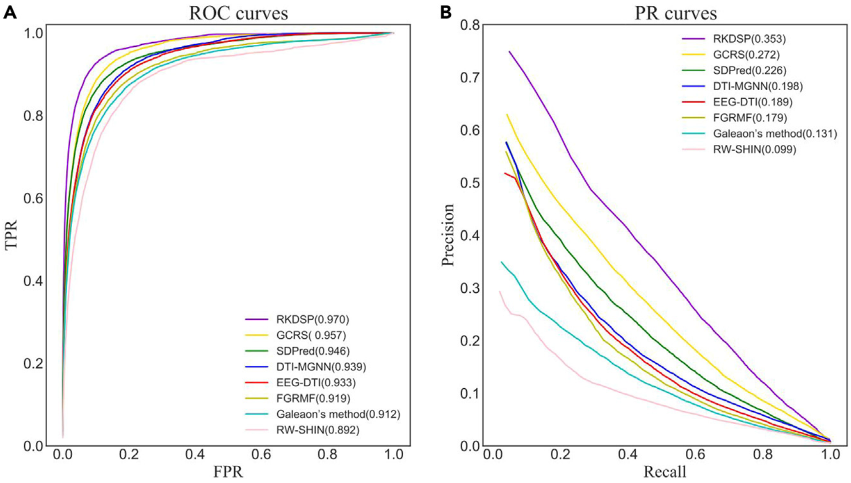


Figure 4. ROC curves (A) and PR curves (B) of all the methods in comparison of all drugs

For the top-ranked candidates, our method also obtained the highest precision rates and F1 scores as shown in the Table ST2. The precision rate (F1 score) of RKDSP was 4.6% (5.6%), 2.2% (3.4%), and 1.2% (2.1%) higher than that of the second best method (GCRS) for the top 30, 60, and 90 candidates. GSEM got the third highest precision rate (F1 score) which is 5.0% (6.6%), 2.9% (4.5%), and 2.1% (3.3%) lower than our model when k is 30, 60, and 90. GraRep performed slightly better than KGDNN, and the former's corresponding precision rates (F1 scores) are 1.0% (0.5%), 1.3% (1.0%), and 1.2% (1.5%) greater than the latter. RW-SHIN's performance was not as good as the other methods, its corresponding precision rates (F1 scores) are 14.3% (17.9%), 11.0% (16.6%), and 9.4% (15.3%).

In order to verify the validity of the proposed model for the prediction of potential side effects of new drugs, we conducted new drug effectiveness experiments. Since new drugs usually have no known associations with side effects. Therefore, we randomly selected 100 drugs, removed all known associations with side effects for these drugs, and considered these 100 drugs as new. Our method still achieved the highest performance compared to the other ten compared methods, and the experimental results are listed in Figures S1 and S2.

Case studies

We chose five medicines for the case study: loratadine, ibuprofen, oseltamivir, erlotinib, and ziprasidone in order to further evaluate the effectiveness of RKDSP in predicting side effects related to drug. We assigned a descending ranking to each drug's association scores for potential side effects. Table 4 shows the top fifteen candidate side effects for these five drugs.

744,709 drug-ADE associations are contained in the online database called MetaADEDB, which includes extensive data on adverse drug events (ADEs).[47] DrugCentral provides information on pharmaceutical products, drug mode of action, indications, and pharmacologic effects.[48] Rxlist is a clinically oriented searchable database of drug-related information, recording side effects, drug interactions, and other relevant information for over 5,000 drugs.[49] Of the 75 candidate side effects, MetaADEDB validated 50, DrugCentral included 39, and Rxlist covered 30. Thus, the results of the clinical trials demonstrated the presence of the listed side effects of the drugs. Three side effects were

Table 3. Performance comparison of RKDSP with state-of-the-art prediction methods

Methods	AUC	AUPR	Precision (top30)	Precision (top 90)	F1 score (top 30)	F1 score (top 90)
RKDSP	0.970	0.353	0.308	0.157	0.392	0.259
FGRMF	0.919	0.179	0.201	0.113	0.249	0.185
RW-SHIN	0.892	0.099	0.143	0.094	0.179	0.153
GraRep	0.927	0.187	0.212	0.117	0.276	0.193
Galeano's method	0.912	0.131	0.170	0.102	0.223	0.170
EEG-DTI	0.933	0.189	0.212	0.118	0.263	0.193
DTI-MGNN	0.939	0.198	0.214	0.121	0.270	0.200
GCRS	0.957	0.272	0.262	0.145	0.336	0.238
GSEM	0.944	0.260	0.258	0.136	0.326	0.226
KGDNN	0.925	0.226	0.202	0.105	0.271	0.178
SDPred	0.946	0.226	0.238	0.129	0.304	0.213

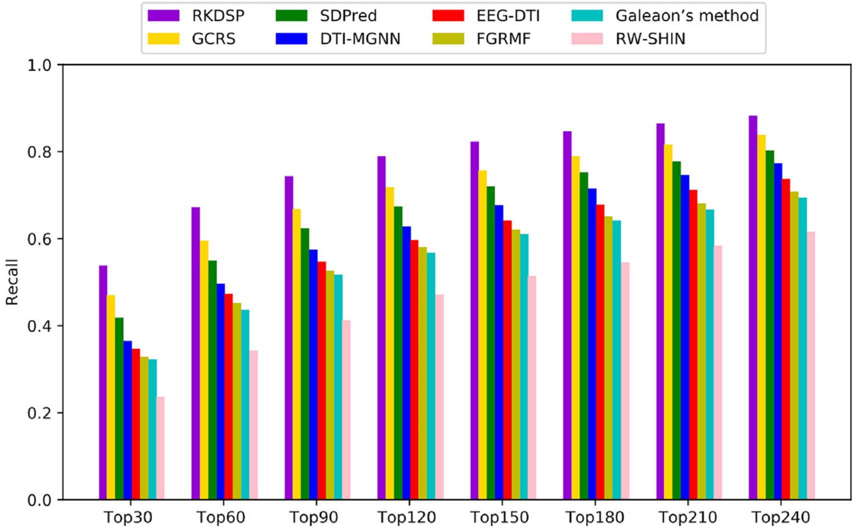


Figure 5. The average recalls over all the drugs at different top *k* values

supported by the published literature and marked as "literature".[50–52] We labeled eight unconfirmed candidate side effects as "unconfirmed." In summary, the five drug case studies demonstrate RKDSP's ability to identify potential side effects of drugs.

Prediction of new drug-associated side effects

Finally, our method employed all known associations to train the RKDSP model and predicted possible side effects of novel drugs. In Table ST3, we provide the top candidate side effects for all drugs to assist biologists in their quest to discover novel drug-associated side effects through further studies.

Limitations of the study

Recently, there has been an influx of rich information about long noncoding RNAs (lncRNAs) and diseases, including novel computational methods for lncRNA similarity and semantic similarity of diseases. In this context, one of the future research directions is to integrate data from multiple sources and multiple modalities to more comprehensively and multidimensional understand and predict the association between lncRNAs and diseases.

Conclusion

We proposed a method (RKDSP) to fuse the various semantics from multiple meta-paths within two heterogeneous graphs and to adaptively learn the pairwise attributes for predicting drug-related side effects. Two established drug-side effect heterogeneous graphs were helpful for the subsequent formulation of the attribute embeddings of the drug and side effect nodes. The constructed multiple meta-paths implied the diverse semantic information among the drug and side effect nodes. The node feature representations were formed by the constructed module based on relational transformers from multiple meta-paths. The knowledge distillation module was constructed to capture the local knowledge within meta-paths and the global knowledge among multiple meta-paths. The designed meta-path-level attention was able to assign higher importance for the informative meta-path semantics. The local and more important attributes for each drug-side effect pair were captured by the multi-layer convolutional neural networks with the adaptive convolution kernels. The comprehensive comparison results indicate RKDSP achieved superior performance than seven advanced methods in terms of both AUC and AUPR measure. RKDSP is also more attractive for the biologists since its top-ranked drug-related side effect candidates are more likely to contain the real drug-side effect associations. The case studies on five drugs further confirm RKDSP's ability in screen the potential candidate side effects for the interested drugs.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

Table 4. Top 15 candidate side effects of five drugs

Drug name	Rank	Side effect name	Evidence	Rank	Side effect name	Evidence
Loratadine	1	Nausea	MetaADEDB, Drugcentral	9	Confusion	MetaADEDB, Rxlist
	2	Dizziness	MetaADEDB, Drugcentral, Rxlist	10	Convulsions	Drugcentral
	3	Hypotension	MetaADEDB, Drugcentral	11	Arthralgia	MetaADEDB, Drugcentral
	4	Paresthesia	unconfirmed	12	Seizures	unconfirmed
	5	Dry mouth	Drugcentral, Rxlist	13	Anxiety	MetaADEDB, Drugcentral
	6	Agitation	MetaADEDB, Drugcentral	14	Arrhythmia	MetaADEDB
	7	Fever	Rxlist	15	Angina	MetaADEDB
	8	Sweating	Rxlist			
Ibuprofen	1	Nausea	Drugcentral, MetaADEDB, Rxlist	9	Pancreatitis	Drugcentral, MetaADEDB
	2	Urticaria	Drugcentral, MetaADEDB	10	Dyspepsia	Drugcentral, MetaADEDB
	3	Erythema	Drugcentral, MetaADEDB	11	Anxiety	Drugcentral, MetaADEDB
	4	Leukopenia	Drugcentral, MetaADEDB	12	Pancytopenia	Drugcentral, MetaADEDB
	5	Tachycardia	Drugcentral, MetaADEDB	13	Shock	Drugcentral, MetaADEDB, Rxlist
	6	Constipation	Drugcentral, MetaADEDB, Rxlist	14	Hyperglycemia	Unconfirmed
	7	Pain	Drugcentral, MetaADEDB, Rxlist	15	Myalgia	Drugcentral
	8	Paresthesia	unconfirmed			
Oseltamivir	1	Vomiting	Drugcentral, MetaADEDB, Rxlist	9	Fatigue	Drugcentral, MetaADEDB
	2	Pain	Drugcentral, MetaADEDB, Rxlist	10	Rhinitis	MetaADEDB
	3	Diarrhea	Drugcentral, Rxlist	11	Fever	Drugcentral, Rxlist
	4	Edema	Drugcentral, MetaADEDB	12	Sweating	Unconfirmed
	5	Palpitations	Drugcentral	13	Psychiatric Disorders	Drugcentral, MetaADEDB, Rxlist
	6	Myalgia	Drugcentral	14	Dry mouth	Unconfirmed
	7	Anorexia	Drugcentral	15	Malaise	Drugcentral
	8	Gastritis	unconfirmed			
Erlotinib	1	Rash	Drugcentral, MetaADEDB, Rxlist	9	Urticaria	Drugcentral
	2	Pain	Drugcentral, MetaADEDB, Rxlist	10	Anemia	Rxlist
	3	Hypersensitivity	Drugcentral, MetaADEDB	11	Pleural	Drugcentral, MetaADEDB
	4	Edema	Drugcentral, MetaADEDB	12	Myalgia	Drugcentral
	5	Thrombocytopenia	Drugcentral, Rxlist	13	Pneumonia	Drugcentral
	6	Chest pain	Drugcentral, Rxlist	14	Paresthesia	unconfirmed
	7	Asthenia	Drugcentral, MetaADEDB	15	Heart failure	Drugcentral
	8	Stomatitis	Drugcentral, MetaADEDB			
Ziprasidone	1	Dystonia	Drugcentral, MetaADEDB, Rxlist	9	Dysphagia	Rxlist
	2	Jaundice	Rxlist	10	Dysarthria	Literature[47]
	3	Amenorrhea	Rxlist	11	Hyperventilation	Literature[48]
	4	Confusion	Rxlist	12	Hypertension	MetaADEDB, Rxlist
	5	Erythema multiforme	Drugcentral, MetaADEDB	13	Manic reaction	Rxlist
	6	Somnolence	Drugcentral, MetaADEDB, Rxlist	14	Vasodilation	Rxlist
	7	Abdominal pain	MetaADEDB, Rxlist	15	Glaucoma	Literature[49]
	8	Insomnia	Drugcentral, MetaADEDB, Rxlist			

- METHOD DETAILS
  - ○ Dataset
  - ○ Construction of drug-side effect dual-view heterogeneous graphs
  - ○ Extracting semantic subgraphs based on meta-paths
  - ○ Meta-path-based node representation learning by relational transformer
  - ○ Meta-path-based node representation update by knowledge distillation

  - ○ Pairwise attribute encoding based on CNN with adaptive convolution kernels
  - ○ Final fusion and optimization
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109571.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (62172143, 62372282), STU Scientific Research Initiation grant (NTF22032), and the Natural Science Foundation of Heilongjiang Province (LH2023F044).

All authors contributed to the article and approved the submitted version.

AUTHOR CONTRIBUTIONS

H.B.: Designed the method and participated in manuscript writing. S.L.: Designed the experiments and participated in manuscript writing. T.Z.: Designed the experiments and edited the manuscript. H.C.: Participated in method design and manuscript writing. T.N.: Participated in experiment design. P.X.: Participated in method design and manuscript writing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 6, 2023
Revised: September 29, 2023
Accepted: March 22, 2024
Published: March 26, 2024

REFERENCES

1. Sachdev, K., and Gupta, M.K. (2020). A comprehensive review of computational techniques for the prediction of drug side effects. Drug Dev. Res. *81*, 650–670.
2. Zhang, F., Sun, B., Diao, X., Zhao, W., and Shu, T. (2021). Prediction of adverse drug reactions based on knowledge graph embedding. BMC Med. Inf. Decis. Making *21*, 38.
3. Cakir, A., Tuncer, M., Taymaz-Nikerel, H., and Ulucan, O. (2021). Side effect prediction based on drug-induced gene expression profiles and random forest with iterative feature selection. Pharmacogenomics J. *21*, 673–681.
4. Li, J., Zheng, S., Chen, B., Butte, A.J., Swamidass, S.J., and Lu, Z. (2016). A survey of current trends in computational drug repositioning. Briefings Bioinf. *17*, 2–12.
5. Jiang, H., Qiu, Y., Hou, W., Cheng, X., Yim, M.Y., and Ching, W.-K. (2020). Drug side-effect profiles prediction: From empirical to structural risk minimization. IEEE ACM Trans. Comput. Biol. Bioinf *17*, 402–410.
6. Zheng, Y., Peng, H., Ghosh, S., Lan, C., and Li, J. (2019). Inverse similarity and reliable negative samples for drug side-effect prediction. BMC Bioinf. *19*, 554.
7. Seo, S., Lee, T., Kim, M.-h., Yoon, Y., et al. (2020). Prediction of side effects using comprehensive similarity measures. BioMed Res. Int. *2020*, 1357630.
8. Lee, W.-P., Huang, J.-Y., Chang, H.-H., Lee, K.-T., and Lai, C.-T. (2017). Predicting drug side effects using data analytics and the integration of multiple data sources. IEEE Access *5*, 20449–20462.
9. Yang, L., Chen, J., and He, L. (2009). Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. PLoS Comput. Biol. *5*, e1000441.
10. Luo, H., Chen, J., Shi, L., Mikailov, M., Zhu, H., Wang, K., He, L., and Yang, L. (2011). Drar-cpi: a server for identifying drug repositioning potential and adverse drug reactions via the chemical–protein interactome. Nucleic Acids Res. *39*, W492–W498.
11. Mizutani, S., Pauwels, E., Stoven, V., Goto, S., and Yamanishi, Y. (2012). Relating drug–protein interaction network with drug side effects. Bioinformatics *28*, i522–i528.
12. Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. Math. Biosci. *306*, 136–144.
13. Guo, X., Zhou, W., Yu, Y., Ding, Y., Tang, J., and Guo, F. (2020). A novel triple matrix factorization method for detecting drug-side effect association based on kernel target alignment. BioMed Res. Int. *2020*, 4675395.
14. Galeano, D., Li, S., Gerstein, M., and Paccanaro, A. (2020). Predicting the frequencies of drug side effects. Nat. Commun. *11*, 4575.
15. Zhang, W., Liu, F., Luo, L., and Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinf. *16*, 365–411.
16. Hu, B., Wang, H., and Yu, Z. (2019). Drug side-effect prediction via random walk on the signed heterogeneous drug network. Molecules *24*, 3668.
17. Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., and Li, X. (2018). Feature-derived graph regularized matrix factorization for predicting drug side effects. Neurocomputing *287*, 154–162.
18. Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing *325*, 211–224.
19. Nguyen, D.A., Nguyen, C.H., and Mamitsuka, H. (2021). A survey on adverse drug reaction studies: data, tasks and machine learning methods. Briefings Bioinf. *22*, 164–177.
20. Liu, M., Wu, Y., Chen, Y., Sun, J., Zhao, Z., Chen, X.-w., Matheny, M.E., and Xu, H. (2012). Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J. Am. Med. Inf. Assoc. *19*, e28–e35.
21. Alpay, B.A., Gosink, M., and Aguiar, D. (2022). Evaluating molecular fingerprint-based models of drug side effects against a statistical control. Drug Discov. Today *27*, 103364.
22. Galeano, D., and Paccanaro, A. (2022). Machine learning prediction of side effects for drugs in clinical trials. Cell Rep. Methods *2*, 100358.
23. Yu, Z., Wu, Z., Li, W., Liu, G., and Tang, Y. (2022). Adenet: a novel network-based inference method for prediction of drug adverse events. Briefings Bioinf. *23*, bbab580.
24. Cao, D.-S., Xiao, N., Li, Y.-J., Zeng, W.-B., Liang, Y.-Z., Lu, A.-P., Xu, Q.-S., and Chen, A.F. (2015). Integrating multiple evidence sources to predict adverse drug reactions based on a systems pharmacology model. CPT Pharmacometrics Syst. Pharmacol. *4*, 498–506.
25. Zhao, H., Wang, S., Zheng, K., Zhao, Q., Zhu, F., and Wang, J. (2022). A similarity-based deep learning approach for determining the frequencies of drug side effects. Briefings Bioinf. *23*, bbab449.

26. Li, Y., Qiao, G., Wang, K., and Wang, G. (2022). Drug–target interaction predication via multi-channel graph neural networks. Briefings Bioinf. *23*, bbab346.
27. Bongini, P., Scarselli, F., Bianchini, M., Dimitri, G.M., Pancino, N., and Lió, P. (2023). Modular multi–source prediction of drug side–effects with drugnn. IEEE ACM Trans. Comput. Biol. Bioinf *20*, 1211–1220.
28. Xuan, P., Wang, M., Liu, Y., Wang, D., Zhang, T., and Nakaguchi, T. (2022). Integrating specific and common topologies of heterogeneous graphs and pairwise attributes for drug-related side effect prediction. Briefings Bioinf. *23*, bbac126.
29. Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics *34*, i457–i466.
30. Lin, S., Wang, Y., Zhang, L., Liu, Y., Fang, Y., Jiang, M., Wang, Q., Zhao, B., Xiong, Y., Wei, D.Q., and Chu, Y. (2022a). Mdf-sa-ddi: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. Briefings Bioinf. *23*, bbab421.
31. Lin, S., Chen, W., Chen, G., Zhou, S., Wei, D.-Q., and Xiong, Y. (2022b). Mddi-scl: predicting multi-type drug-drug interactions via supervised contrastive learning. J. Cheminf. *14*, 81.
32. Wang, C.-S., Lin, P.-J., Cheng, C.-L., Tai, S.-H., Kao Yang, Y.-H., and Chiang, J.-H. (2019). Detecting potential adverse drug reactions using a deep neural network model. J. Med. Internet Res. *21*, e11016.
33. Li, G., Sun, Z., Hu, W., Cheng, G., and Qu, Y. (2023). Position-aware relational transformer for knowledge graph embedding. IEEE Transact. Neural Networks Learn. Syst. 1–15.
34. Li, Y., Guo, Z., Wang, K., Gao, X., and Wang, G. (2023b). End-to-end interpretable disease–gene association prediction. Briefings Bioinf. *24*, bbad118.
35. Wang, G., Zhang, X., Pan, Z., Rodríguez Patón, A., Wang, S., Song, T., and Gu, Y. (2022). Multi-transdti: transformer for drug–target interaction prediction based on simple universal dictionaries with multi-view strategy. Biomolecules *12*, 644.
36. Kim, Y., and Kwon, J. (2023). Attsec: protein secondary structure prediction by capturing local patterns from attention map. BMC Bioinf. *24*, 183–216.
37. Lin, S., Zhang, G., Wei, D.-Q., and Xiong, Y. (2022). Deeppse: Prediction of polypharmacy side effects by fusing deep representation of drug pairs and attention mechanism. Comput. Biol. Med. *149*, 105984.
38. Xue, R., Liao, J., Shao, X., Han, K., Long, J., Shao, L., Ai, N., and Fan, X. (2020). Prediction of adverse drug reactions by combining biomedical tripartite network and graph representation model. Chem. Res. Toxicol. *33*, 202–210.
39. Dey, S., Luo, H., Fokoue, A., Hu, J., and Zhang, P. (2018). Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinf. *19*, 476–513.
40. Joshi, P., Masilamani, V., and Mukherjee, A. (2022). A knowledge graph embedding based approach to predict the adverse drug reactions using a deep neural network. J. Biomed. Inf. *132*, 104122.
41. Hajian-Tilaki, K. (2013). Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. Caspian J. Intern. Med. *4*, 627–635.
42. Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One *10*, e0118432.
43. Hu, J., Yu, W., Pang, C., Jin, J., Pham, N.T., Manavalan, B., and Wei, L. (2023). Drugormerdti: Drug graphormer for drug–target interaction prediction. Comput. Biol. Med. *161*, 106946.
44. Huang, S., Li, J., Xiao, Y., Shen, N., and Xu, T. (2022). Rtnet: relation transformer network for diabetic retinopathy multi-lesion segmentation. IEEE Trans. Med. Imag. *41*, 1596–1607.
45. Cong, Y., Yang, M.Y., and Rosenhahn, B. (2023). Reltr: Relation transformer for scene graph generation. IEEE Trans. Pattern Anal. Mach. Intell. *45*, 11169–11183.
46. Peng, J., Wang, Y., Guan, J., Li, J., Han, R., Hao, J., Wei, Z., and Shang, X. (2021). An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. Briefings Bioinf. *22*, bbaa430.
47. Ali, T., Sisay, M., Tariku, M., Mekuria, A.N., and Desalew, A. (2021). Antipsychotic-induced extrapyramidal side effects: A systematic review and meta-analysis of observational studies. PLoS One *16*, e0257129.
48. MS II, L., Franks MD, A.M., Rollyson, M.S.I.V.,W., Barbour MD, J.K., Curry MD, M.B., et al. (2021). Anti-n-methyl-d-aspartate receptor encephalitis: a diagnosis obscured by concomitant recreational drug use. Marshall J. Med. *7*, 18.
49. Ciobanu, A.M., Dionisie, V., Neagu, C., Bolog, O.M., Riga, S., and Popa-Velea, O. (2021). Psychopharmacological treatment, intraocular pressure and the risk of glaucoma: a review of literature. J. Clin. Med. *10*, 2947.
50. Yu, Z., Wu, Z., Li, W., Liu, G., and Tang, Y. (2021). Metaadedb 2.0: a comprehensive database on adverse drug events. Bioinformatics *37*, 2221–2222.
51. Avram, S., Bologa, C.G., Holmes, J., Bocci, G., Wilson, T.B., Nguyen, D.-T., Curpan, R., Halip, L., Bora, A., Yang, J.J., et al. (2021). Drugcentral 2021 supports drug discovery and repositioning. Nucleic Acids Res. *49*, D1160–D1169.
52. Steigerwalt, K. (2015). Online drug information resources. Choice *52*, 1601–1611.
53. Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., and Zeng, J. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Commun. *8*, 573.
54. Kuhn, M., Letunic, I., Jensen, L.J., and Bork, P. (2016). The sider database of drugs and side effects. Nucleic Acids Res. *44*, D1075–D1079.
55. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., Wiegers, J., Wiegers, T.C., and Mattingly, C.J. (2021). Comparative toxicogenomics database (ctd): update 2021. Nucleic Acids Res. *49*, D1138–D1143.
56. Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. Bioinformatics *26*, 1644–1650.
57. Nair, V., and Hinton, G.E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines, pp. 807–814.
58. Kingma, D.P., and Ba, J. (2015). Adam: A method for stochastic optimization. In Int. Conf. Learn. Representations.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited		
Drug-side effect associations	SIDER	http://sideeffects.embl.de/; RRID:SCR_00431
Drug-disease associations	Comparative Toxicogenomics Database (CTD)	https://ctdbase.org/; RRID:SCR_006530
Train data	Lou et al.	https://www.nature.com/articles/s41467-017-00680-8
Software and algorithms		
Functional similarity	Wang et al.	https://pubmed.ncbi.nlm.nih.gov/20439255/
Adam algorithm	Kingma et al.	https://arxiv.org/abs/1412.6980
ReLU function	Nair et al.	https://doi.org/10.5555/3104322.3104425

RESOURCE AVAILABILITY

Lead contact

Further requests for information should be directed and will be handled by the lead contact, Ping Xuan, email: pxuan@stu.edu.cn.

Materials availability

All materials reported in this paper will be shared by the lead contact upon request.

Data and code availability

- Data reported in this paper will be shared by the lead contact upon request.
- This paper does not report original code.
- Any additional information for reanalyzing this work is available from the lead contact upon request.

METHOD DETAILS

We developed a deep learning model called RKDSP, which combines two types of drug similarities and side effect similarities predict drug candidate side effects. First, based on drugs and side effect similarities, bi-perspective heterogeneous graphs are constructed. Then, bi-perspective fusion representations and adaptively enhanced pairwise attributes are learned separately. Finally, by integrating these two representations, the association scores of drugs and side effect nodes are computed.

Dataset

The chemical structures of drugs may provide the information of the chemical positions, the atomic arrangements and the functional groups. Two drugs with more similar chemical structures are usually more similar, thus the drug similarities are calculated based on their chemical structures. The drugs with similar functions are more likely to be involved the similar disease processes, the drug similarities were also calculated based on their association diseases. The dataset originally extracted the 80,164 pairs of drug and side effect nodes which cover 708 drugs and 4,192 side effects. The dataset for drug-related side effect association prediction was derived from the method[53] and it contains the drug-side effect associations, the drug-disease associations, and the drug-drug similarities based on chemical substructures. The dataset originally extracted the 80,164 pairs of drug and side effect nodes from the SIDER (RRID:SCR_004321)[54] which cover 7 drugs and 4,192 side effects. The method also obtained the 199,214 associations among the 708 drugs and 5,603 diseases from the Comparative Toxicogenomics Database (CTD) (RRID:SCR_006530).[55]

Construction of drug-side effect dual-view heterogeneous graphs

To deeply integrate multiple data sources to facilitate the study, we constructed the drug-side effect association matrix $\mathbf{S} \in R^{N_r \times N_s}$, chemical substructure-based matrix of drug-drug similarity $\mathbf{M}^{cs} \in R^{N_r \times N_r}$, associated disease-based matrix of drug-drug similarity $\mathbf{M}^{ad} \in R^{N_r \times N_r}$, and side effect similarity matrix $\mathbf{M}^{se} \in R^{N_s \times N_s}$. It is more likely that the side effects of the drugs r_i and r_j will be similar if their chemical substructures are more similar. We used this biological premise to construct the chemical substructure-based drug-drug similarity matrix $\mathbf{M}^{cs}$. In addition, the similarity between two drugs tends to be higher when they have been associated with other diseases that are comparable. After revising our estimates of drug similarity based on the diseases connected with each drug,[56] we built the matrix of drug-drug similarity based on these related diseases $\mathbf{M}^{ad}$. Similarly, using the directed acyclic network of drugs related to side effects, we computed the side effect similarity matrix $\mathbf{M}^{se}$. In addition, we built the drug-side effect association matrix $\mathbf{S}$. If biological research has shown an association between the use of drug r_i and the incidence of side effect s_j, then the value of $\mathbf{S}_{ij}$ is 1. Otherwise, it is 0.

Drug-side effect bi-perspective heterogeneous graphs were built to fully integrate data from various sources on drugs and side effects. Two heterogeneous graphs from varying perspectives may be represented as $G^{cs} = (V, E^{cs})$ and $G^{ad} = (V, E^{ad})$ given the node set V, which contains all drug nodes V^{dr} and side effect nodes V^{se}, and the edge set E consisting of edges E^{cs} on graph G^{cs} and edges E^{ad} on graph G^{ad}. The weights on the edges E^{cs} and E^{ad} are represented by adjacency matrices $\mathbf{A}^{cs}$ and $\mathbf{A}^{ad}$, respectively,

$$\mathbf{A}^{cs} = \begin{bmatrix} \mathbf{M}^{cs} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{M}^{se} \end{bmatrix}, \mathbf{A}^{ad} = \begin{bmatrix} \mathbf{M}^{ad} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{M}^{se} \end{bmatrix}, \qquad \text{(Equation 1)}$$

where $\mathbf{S}^T$ represents the transpose matrix of $\mathbf{S}$.

Extracting semantic subgraphs based on meta-paths

We constructed heterogeneous graphs G^{cs} and G^{ad} based on the drug's chemical substructure and the disease that it associated with, respectively. These two heterogeneous graphs are built based on drug similarities from different perspectives and, therefore, have specific topological information. Dual perspective heterogeneous graphs G^{cs} and G^{ad} contain drug (r) and side effect (s) nodes. The graphs also include multiple relationships, where $r - r$, $r - s$, and $s - s$ indicate drug-drug similarity, the association between drugs and side effects, and side effect-side effect similarity, respectively. A heterogeneous graph might have several nodes connected by pathways that indicate many connections; these paths are referred to as meta-paths. Various meta-paths contain varying semantic information. We extract the semantic subgraphs by projecting the bi-perspective heterogeneous graphs according to each meta-path $m \in M$. Given the heterogeneous graph G^{cs} and the meta-path $r - r$, we project G^{cs} into the semantic subgraph G^{cs}_{r-r}. The node feature matrix of G^{cs}_{r-r} is denoted as $\mathbf{X}_{r-r}$. If the meta-path is specified as $r - s - r$, the node feature matrix of the generated semantic subgraph is $\mathbf{X}_{r-s-r}$,

$$\mathbf{X}_{r-s-r} = \mathbf{AM} \cdot \mathbf{AM}^T. \qquad \text{(Equation 2)}$$

Meta-path-based node representation learning by relational transformer

We constructed a subgraph G^{cs}_m containing semantic information and topology specific to the meta-path m. Faced with multiple semantic subgraphs, we built multi-layer relational transformers learn the meta-path-based node representation of drug and side effect nodes.

Relational transformer

Inspired by the success of transformer in natural language processing, we propose the relational transformer for learning meta-path-based node representations. The query matrix $\mathbf{Q}^{n,l}_m$, key matrix $\mathbf{K}^{n,l}_m$, and value matrix $\mathbf{V}^{n,l}_m$ are obtained for the n-th attention head in the process described below,

$$\mathbf{Q}^{n,l}_m = \mathbf{W}^{n,l}_{m,\mathrm{Q}} \cdot \mathbf{X}^{l-1}_m, \qquad \text{(Equation 3)}$$

$$\mathbf{K}^{n,l}_m = \mathbf{W}^{n,l}_{m,\mathrm{K}} \cdot \mathbf{X}^{l-1}_m, \qquad \text{(Equation 4)}$$

$$\mathbf{V}^{n,l}_m = \mathbf{W}^{n,l}_{m,\mathrm{V}} \cdot \mathbf{X}^{l-1}_m, \qquad \text{(Equation 5)}$$

where m refers to the meta-path, l is the number of layers of the transformer, and $\mathbf{W}^{n,l}_{m,\mathrm{Q}}$, $\mathbf{W}^{n,l}_{m,\mathrm{K}}$, and $\mathbf{W}^{n,l}_{m,\mathrm{V}}$ are the weight matrices corresponding to the different meta-paths and attention heads. $\mathbf{X}^0_m$ is the node attribute matrix $\mathbf{X}_m$. Then, we computed the attention weights between nodes in the subgraph and multiplied by $\mathbf{V}^{n,l}_m$ to obtain the hidden matrix $\mathbf{H}^l_m$,

$$\mathbf{H}^l_m = \Vert^{N_{an}}_{n=1} \left\lceil \mathbf{Q}^{n,l}_m, \mathbf{K}^{n,l}_m \right\rceil \cdot \mathbf{V}^{n,l}_m, \qquad \text{(Equation 6)}$$

where $\left\lceil \mathbf{Q}^{n,l}_m, \mathbf{K}^{n,l}_m \right\rceil = \exp\left(\frac{\mathbf{Q}^T\mathbf{K}}{d}\right)$ is the exponential scale dot product function, d is the dimensionality of each attention head, and $\Vert$ is the concatenation operation. To ensure a smoother representation learning process, we apply a gating mechanism. We obtain the matching matrix $\mathbf{T}^l_m$ as follows,

$$\mathbf{T}^l_m = \mathrm{sig}\left(\mathbf{W}_t \cdot \left[\mathbf{H}^l_m, \mathbf{X}^{l-1}_m\right]\right), \qquad \text{(Equation 7)}$$

where $[\cdot, \cdot]$ is the splicing operation, and *sig* refers to the sigmoid activation function. Finally, we obtain the meta-path-based node representation matrix $\mathbf{Z}^l_m$,

$$\mathbf{Z}^l_m = \tanh\left(\mathbf{H}^l_m\right) \otimes \mathbf{T}^l_m + \mathbf{X}^{l-1}_m \otimes \left(\mathbf{I} - \mathbf{T}^l_m\right), \qquad \text{(Equation 8)}$$

where $\otimes$ is the Hadamard product operation. $\mathbf{Z}^l_m$ is the learned node representation matrix based on meta-path m at layer l.

Meta-path-based node representation update by knowledge distillation

Each node in the semantic subgraph G_m^{cs} has a specific connectivity relationship with its neighbor nodes, reflecting the local knowledge within the meta-path. The different connection relations in the subgraphs corresponding to different meta-paths contain the global knowledge among the meta-paths. Therefore, we need not only to extract the topology of the subgraph but also to learn the local knowledge within meta-paths and the global knowledge between meta-paths. We propose intra-meta-path and inter-meta-path knowledge distillation, inspired by knowledge distillation, to update the meta-path-based node representation.

Intra-meta-path knowledge distillation

Knowledge distillation in meta-paths is done to update the node representation by extracting the local knowledge from each meta-path. Given a semantic subgraph G_m^{cs}, we define the neighbor topology representation as e_m^i,

$$e_m^i = \mathrm{sig}\left(\frac{1}{K}\sum_{j=1}^{K} z_m^j\right), \qquad \text{(Equation 9)}$$

where K denotes the number of neighbor nodes, i refers to the target node, and j is the source node. Meta-path representation t_m is obtained through

$$t_m = \sigma\left(\frac{1}{A}\sum_{i=1}^{A} z_m^i\right), \qquad \text{(Equation 10)}$$

where A refers to the number of target nodes, and σ represents the ReLU function.[57] We measure distillation by mutual information based on meta-paths node embedding z_m^i, neighbor topology e_m^i, and meta-path representation t_m. The loss of knowledge distillation within a meta-path L_{tra} is defined as follows,

$$L_{tra} = -\sum_{m\in M}\left(\sum_{i}^{|A|}\left(\mathrm{MI}\left(z_m^i, e_m^i\right) + \mathrm{MI}\left(z_m^i, t_m\right)\right)\right). \qquad \text{(Equation 11)}$$

Inter-meta-path knowledge distillation

The purpose of knowledge distillation between meta-paths is to extract the global knowledge between different meta-paths. We distill with neighbor topology with different meta-paths, meta-path representations to update the node representations. The loss of knowledge distillation between metapaths L_{ter} is defined as follows,

$$L_{ter} = -\sum_{i}^{|A|}\left(\sum_{m\in M}\sum_{n\in M, n\neq m} \mathrm{MI}\left(z_m^i, e_n^i\right) + \mathrm{MI}\left(z_m^i, t_n\right)\right). \qquad \text{(Equation 12)}$$

Meta-path level attention

Considering that different meta-path node representations have varying significance for predicting drug-side effect associations, we developed a meta-path-level attention mechanism to learn the attention scores of each representation and adaptively combine them. Given the updated meta-path based node representation $\hat{z}_m^i$, the attention weight s_m^i is defined as follows,

$$s_m^i = q_{mpl}\cdot\tanh\left(W_{mpl}\hat{z}_m^i + b_{mpl}\right), \qquad \text{(Equation 13)}$$

where b_{mpl} represents the bias vector, and W_{mpl} refers to the weight matrix. The normalized score is α_m^i,

$$\alpha_m^i = \frac{\exp\left(s_m^i\right)}{\sum_{n\in M}\exp\left(s_n^i\right)}. \qquad \text{(Equation 14)}$$

The level of the attention score reveals the importance of the corresponding meta-path. After a meta-path-based attention mechanism, the semantic representation $\mathbf{z}_{i,cs}^{sem}$ can be obtained as follows,

$$\mathbf{z}_{i,cs}^{sem} = \sum_{m\in M}\alpha_m^i\hat{z}_m^i. \qquad \text{(Equation 15)}$$

From the dual-view heterogeneous graphs G^{cs} and G^{ad}, we obtain two semantic representations, $\mathbf{z}_{cs}^{sem}$ and $\mathbf{z}_{ad}^{sem}$. To include more valid information, we use 1*1 convolution to aggregate them and get a dual-view fusion representation $\mathbf{z}^{fus}$.

Pairwise attribute encoding based on CNN with adaptive convolution kernels

Attribute embedding of drug-side effect node pair

If drug node r_i and side effect node s_j are similar or related to more common drugs and side effects, then r_i and s_j are more likely to be related. Based on this biological premise and the similarity of the two drugs, we propose an embedding strategy to obtain the attribute embedding $\mathbf{X}^{cs}$ and $\mathbf{X}^{ad}$ of two drug-side effect node pairs.

Given the similarity matrix $\mathbf{M}^{cs}$ based on chemical substructures, side effect similarity matrix $\mathbf{M}^{se}$, and association matrix $\mathbf{S}$, we combine the i-th row of $\mathbf{M}^{cs}$ and the j-th column of $\mathbf{S}$ to form x_{left},

$$x_{left} = \left[\mathbf{M}^{cs}_{i,*}; \mathbf{S}_{j,*}\right], \quad \text{(Equation 16)}$$

where; is the stacking operation. $\mathbf{M}^{cs}_{i,*}$ denotes the similarity of r_i to all drugs. $\mathbf{S}_{j,*}$ refers to the association of s_j with all drugs. The association between r_i and all side effects are contained in $\mathbf{S}_{i,*}$, which is obtained from the i-th row of $\mathbf{S}$. $\mathbf{M}^{se}_{j,*}$ is the j-th row of $\mathbf{M}^{se}$ that covers the similarity of s_j and all side effects. They are stacked to form x_{right},

$$x_{right} = \left[\mathbf{S}_{i,*}; \mathbf{M}^{se}_{j,*}\right]. \quad \text{(Equation 17)}$$

We define property embedding based on drug chemical substructures as $\mathbf{X}^{cs}$,

$$\mathbf{X}^{cs} = \left[x_{left} \quad x_{right}\right]. \quad \text{(Equation 18)}$$

Likewise, given the drug similarity matrix $\mathbf{M}^{ad}$, the association matrix $\mathbf{S}$, and the side effect similarity matrix $\mathbf{M}^{se}$, we can obtain another perspective of the r_i-s_j pairwise attribute embedding $\mathbf{X}^{ad}$.

Dual-view attribute encoding

Traditional CNNs use uniform convolutional kernels for feature extraction at different locations in the feature map, even though these locations contain information of varying importance. With adaptive convolution, the convolution kernels can adjust their adaptive weights according to the local content. Convolutional kernels at locations containing important information receive greater weights, and those at locations containing edge information are assigned smaller weights. In addition, each of two drug-side effect heterogeneous graphs (G^{cs} and G^{ad}) has its own specific features. Thus, we propose a pairwise attribute encoding module to obtain and aggregate the dual-view attributes of drug-side effect node pairs from $\mathbf{X}^{cs}$ and $\mathbf{X}^{ad}$. The encoding process is similar, and for brevity, we will only discuss the case where the input is $\mathbf{X}^{cs}$.

First, we send pairs of attributes $\mathbf{X}^{cs}$ to the convolutional layer with the LeakyReLU activation function to extract their shallow features. Then, shallow features are passed through a fully connected layer with a sigmoid activation function to learn the weights $\overline{\mathbf{W}}^{cs}$. $\overline{\mathbf{W}}^{cs}$ can perceive the potential relationship between each position in the node pair attribute. Finally, we reshape $\overline{\mathbf{W}}^{cs}$ to $\mathbf{W}^{cs}$, which we use as a scaling factor for each convolution kernel. We denote the scaled convolution kernel as $\mathbf{K}^{cs}$,

$$\mathbf{K}^{cs} = \mathbf{W}^{cs} \odot \overline{\mathbf{K}}^{cs}, \quad \text{(Equation 19)}$$

where $\odot$ represents the dot product operation. The network can consider the local content inconsistency of the feature map based on the obtained local background adaptive kernel. We performed zero padding on $\mathbf{X}^{cs}$ to preserve edge information. The number of zero padding is 1. The feature map $\mathbf{z}^{cs}$ after convolution is defined as $\mathbf{z}^{cs}$,

$$\mathbf{z}^{cs} = \tilde{X}^{cs} * \mathbf{K}^{cs}, \quad \text{(Equation 20)}$$

where $*$ is the convolution operation, and $\tilde{X}^{cs}$ is the paired attribute embedding after performing zero padding. Similarly, we can obtain feature maps $\mathbf{z}^{ad}$ by node pair embedding $\mathbf{X}^{ad}$. $\mathbf{z}^{enh}$ is obtained by splicing $\mathbf{z}^{cs}$ and $\mathbf{z}^{ad}$, which indicates adaptive enhancement of pairwise attributes.

Final fusion and optimization

Given the dual-view fusion representation $\mathbf{z}^{fus}$ and the adaptive enhancement pairwise attribute $\mathbf{z}^{enh}$, we connect and flatten the two to form a combined representation $\mathbf{z}^{com}$. We obtain the probability distribution of whether r_i is associated with s_j by applying the fully connected layer to $\mathbf{z}^{com}$,

$$p = \text{softmax}\left(Wz^{com} + b\right), \quad \text{(Equation 21)}$$

where the bias vector and weight matrices, respectively, are denoted by b and W. The difference between the labels and the predicted score distribution is measured using the cross-entropy loss function as *loss*,

$$\text{loss} = \sum_{i=1}^{N} \sum_{j=1}^{D} R_j \log\left(p\right)_j + L_{\text{tra}} + L_{\text{ter}}, \quad \text{(Equation 22)}$$

where $D = 2$. R_j represents the true label, indicating whether there is a true association between a pair of nodes. The Adam algorithm[58] optimizes the loss function *loss*.

QUANTIFICATION AND STATISTICAL ANALYSIS

To verify whether RKDSP outperforms the comparison method, we performed the Paired Wilcoxon Test. In terms of each prediction method, its AUC and AUPR were obtained for each drug, and it have 708 AUCs and AUPRs. For RKDSP and one the compared methods, 708 AUC (AUPR) pairs went through the Paired Wilconxon Test. Compared to other approaches (Table ST1), RKDSP performs significantly better (p-value < 0.05).