BMC Bioinformatics

RESEARCH

Open Access

DTI-MHAPR: optimized drug-target interaction prediction via PCA-enhanced features and heterogeneous graph attention networks

Guang Yang[1], Yinbo Liu[1], Sijian Wen[1], Wenxi Chen[1], Xiaolei Zhu[1] and Yongmei Wang[1*]

*Correspondence:
wym0152@foxmail.com

[1] School of Information and Artificial Intelligence, Anhui Agricultural University, Changjiang West Road, Hefei 230036, Anhui, China

Abstract

Drug-target interactions (DTIs) are pivotal in drug discovery and development, and their accurate identification can significantly expedite the process. Numerous DTI prediction methods have emerged, yet many fail to fully harness the feature information of drugs and targets or address the issue of feature redundancy. We aim to refine DTI prediction accuracy by eliminating redundant features and capitalizing on the node topological structure to enhance feature extraction. To achieve this, we introduce a PCA-augmented multi-layer heterogeneous graph-based network that concentrates on key features throughout the encoding-decoding phase. Our approach initiates with the construction of a heterogeneous graph from various similarity metrics, which is then encoded via a graph neural network. We concatenate and integrate the resultant representation vectors to merge multi-level information. Subsequently, principal component analysis is applied to distill the most informative features, with the random forest algorithm employed for the final decoding of the integrated data. Our method outperforms six baseline models in terms of accuracy, as demonstrated by extensive experimentation. Comprehensive ablation studies, visualization of results, and in-depth case analyses further validate our framework's efficacy and interpretability, providing a novel tool for drug discovery that integrates multimodal features.

Keywords: DTIs prediction, Heterogeneous graph attention networks, Feature concatenation, PCA, Random forest

Introduction

Drug development is a lengthy, expensive, and high-risk complex process that encompasses several key stages, including discovery, development, clinical trials, and marketing. In recent years, with advancements in science and technology and improvements in data analysis methods, researchers have been able to predict associations between drugs and targets more accurately, significantly shortening development cycles and enhancing the efficiency of drug development. However, traditional methods for determining

drug-target interactions (DTIs) are often time-consuming, labor-intensive, and costly. Therefore, it is crucial to develop efficient computational methods for predicting drug-target associations.

In recent years, computational methods have proven to be effective in several tasks within the field of bioinformatics, such as predicting disease-related miRNAs [1, 2],predicting drug-target affinity [3],identifying potential disease-related genes [4] and disease-associated lncRNAs [5–7],predicting drug-drug interactions [8], protein-protein interactions [9], and specific protein localization within cells [10]. Some existing models in the field of drug discovery [11, 12] have been successfully applied to DTI prediction tasks through further expansion and optimization, demonstrating their effectiveness and efficiency. DTI computational prediction has enormous research potential and broad application prospects, offering advantages such as short time, low cost, high precision, and wide range in uncovering potential DTIs [13].

According to different methods, existing prediction models can mainly be divided into four categories: molecular docking-based methods, text mining-based methods, ligand-based methods, and chemogenomics-based methods. Molecular docking-based methods predict binding situations by simulating the physical and chemical interactions between drug molecules and target proteins [14], which have high biological interpretability. Text mining-based methods make predictions by extracting and analyzing relevant information from scientific literature, patents, and databases [15]. Ligand-based methods use the characteristics of known active ligands to predict new drug-target interactions through similarity searching and modeling [16], allowing for rapid predictions and suitable for high-throughput screening. Chemogenomics-based methods [17] integrate chemical and genomic data to predict drug-target interactions by constructing chemical-biological networks. Depending on the implementation, these methods can be further refined into similarity-based, pharmacological characteristic-based, and network-based methods [18, 19]. Similarity-based methods include the nearest neighbor method [20, 21], bipartite local model (BLM) [22, 23], and matrix factorization methods [24]. Network-based methods make associative predictions by constructing and analyzing network structures, which can be divided into binary network methods [25] and heterogeneous network methods [26, 27] according to different network topologies. Currently, many DTI prediction methods combine deep learning with heterogeneous networks [28–33]. Deep learning-based methods utilize the advantages of neural networks to automatically learn the features of nodes and the complex relationships between nodes.

Recently, a method combining metapath-based graph convolutional networks with large-scale heterogeneous networks to effectively learn representations of drugs and target proteins [34] has been proposed. Additionally, a metapath-based hierarchical transformer combined with an attention network for drug-target interaction prediction [35] has been proposed. This method applies a metapath instance-level transformer, single-semantics attention, and multi-semantics attention to generate low-dimensional vector representations of drugs and proteins. Additionally, the CFSSynergy method, which combines feature-based and similarity-based approaches, demonstrates outstanding performance in drug synergy prediction [36]. This method extracts discriminative features from drugs and cell lines, utilizes the Node2Vec

algorithm to compute protein similarities, and ultimately feeds these features into XGBoost for prediction. Furthermore, the methods of constructing heterogeneous knowledge graphs, employing heterogeneous graph transformer networks, and calculating relationship scores using fully connected networks have also proven to be highly effective [37], further highlighting the potential of heterogeneous graph neural networks in diverse applications. These studies still do not fully exploit the feature information of drugs and targets in DTI prediction, and the associations in heterogeneous biological information networks have not been fully utilized. At the same time, issues such as redundant features and low robustness of decoding methods have not been addressed. How to effectively utilize the topological structure between nodes to deeply mine feature information for DTI prediction remains an urgent problem to be solved. We aim to address the issue of redundant features and effectively leverage the topological structure among nodes to deeply mine feature information for DTI prediction.Therefore, we propose a framework MHAPR that constructs highly integrated multi-level information by concatenating feature vectors obtained from each convolutional layer in a multimodal heterogeneous graph attention network and then combines feature optimizer PCA [38] and random forest algorithm [39, 40].

More specifically, MHAPR first constructs a heterogeneous graph from various similarity information between drugs and targets, using the heterogeneous graph as input to the graph neural network for encoding. By aggregating heterogeneous nodes through the graph attention mechanism and setting up a multi-head self-attention mechanism and metapath weighting strategy, deep integration of multi-source similarity information in the heterogeneous graph is achieved. Then, the aggregated representation vectors of the heterogeneous graph neural networks with self-attention are concatenated, preserving multi-level representation information. Subsequently, we apply various effective feature selection methods to the representation encoding for feature selection and finally use Principal Component Analysis (PCA) to improve the quality of features. Compared to traditional models, MHAPR places greater emphasis on extracting key feature information and enhances the model's generalization ability through the robust ensemble learning mechanism of the random forest algorithm.In summary, our main contributions to this work are as follows:

- *Multimodal information fusion* By constructing a concatenated multi-layer heterogeneous graph attention network, we integrated various similarity information from drugs and targets (such as sequence and Gaussian similarity), enabling deep extraction and fusion of heterogeneous node representations. This method effectively captures the complex interactions between drugs and targets.
- *Feature selection and optimization* Through extensive experimentation, we demonstrated that different feature selection methods significantly enhance the quality of representation encoding. We explored their impact on improving multimodal biological information capabilities, providing theoretical support for feature optimization.
- *Selection of feature dimensionality reduction* By comparing four dimensionality reduction algorithms, we ultimately chose PCA as the feature optimizer, aiming to reduce feature redundancy and computational complexity while ensuring the

model focuses on key biological information, thereby enhancing prediction accuracy and reliability.

- *Enhancing model robustness* The reduced features were input into a random forest algorithm, leveraging its powerful ensemble learning mechanism to improve the model's tolerance to data noise, thereby enhancing overall robustness. This strategy ensures the model's stability and practicality in real-world applications.

Method and materials

Data collection

To validate the generalizability of our model, we downloaded two drug-target protein datasets for benchmarking from references [41, 42], both sourced from relevant databases. The first dataset, named the FSL dataset, is derived from the work of Tangmanussukum et al. [42] After removing duplicate relationships, it contains 862 drugs, 1,517 target proteins, and 3,583 known DTIs. The second dataset, named the DC dataset, is sourced from the work of Peng [41]. Drug nodes, known DTIs, and drug-drug interactions were extracted from the DrugBank database [43], while protein nodes and protein-protein interactions were extracted from the HPRD database [44]. This resulted in 708 drug samples, 1512 target protein samples, and 1923 known DTIs,Table 1 provides a detailed description of the datasets.

We found that the number of unknown association samples far exceeds the known ones, with a significant amount of noisy data. To mitigate noise and ensure dataset balance, we labeled all known drug-target protein associations as positive 1 and randomly selected an equal number of negative samples, labeling them as 0. We then used five-fold cross-validation, with 80% of the samples for training and 20% for testing.

Overview of processes

In bioinformatics research, biological information features that are noisy and high-dimensional often hinder models from capturing comprehensive and discriminative relationships within the data. Therefore, based on extensive comparative experiments and comprehensive consideration of the effectiveness of various feature combinations and their computational complexity, we propose a three-layer computational framework based on HAN-PCA-RF for predicting potential drug-target interactions. Figure 1 provides a detailed illustration of the DTI-MHAPR workflow, which specifically includes the following steps:

(1) Constructing a heterogeneous graph from target protein sequence similarity, Gaussian similarity, drug structure similarity, Gaussian similarity matrices, and

Table 1 Dataset details

Datasets	Proteins	Drugs	Interactions	Interaction proportion (%)
DC	708	1512	1923	0.17
FSL	1517	862	3583	0.27

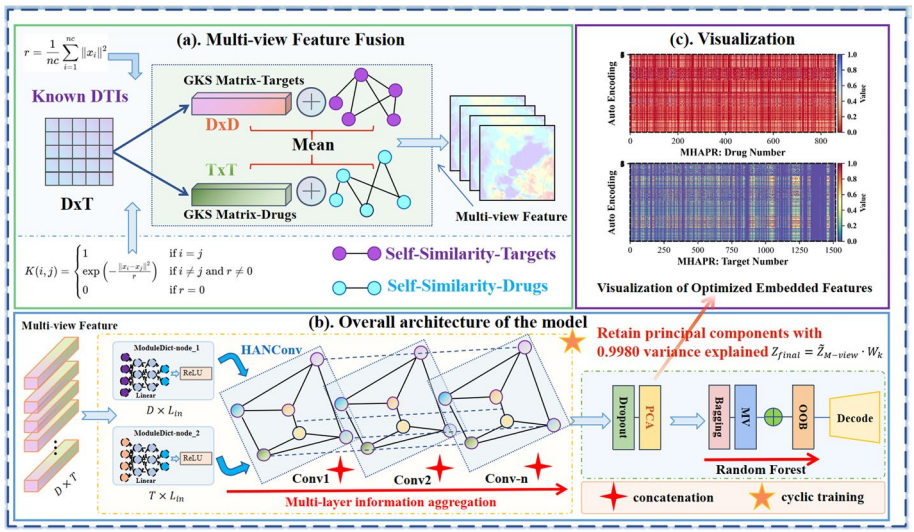

Fig. 1 Workflow diagram of DTI-MHAPR

known DTIs information. This heterogeneous graph effectively represents the multifaceted relationships between different types of nodes and edges.

(2) After applying a linear transformation and ReLU activation to the embeddings of different nodes, a graph neural network is used to encode the heterogeneous graph, incorporating multi-head self-attention mechanisms and meta-path weighting strategies to achieve deep integration of multi-source similarity information within the heterogeneous graph.

(3) Concatenating the representation vectors obtained from multiple layers of the heterogeneous attention network to retain and integrate multi-level representation information.

(4) PCA to project the original representation vectors onto the principal components, reducing the feature dimensions while focusing the model on key information.

(5) Using the random forest algorithm to predict the final interaction scores between drugs and target proteins, leveraging its ensemble learning capabilities to handle high-dimensional data and capture nonlinear relationships effectively.

Multi-view similarity feature integration

To gain a deeper understanding of the similarity information between drugs and targets, we employed a multi-modal feature extraction approach. Initially, we applied the Gaussian kernel function to the topological structure association network between bioinformatic nodes [45]. This method allows us to obtain more accurate Gaussian interaction profile kernel similarities by computing the Gaussian interaction profile kernel similarity. Similarly, we applied the Gaussian kernel function to the drug structure similarity matrix to derive the Gaussian similarity matrix for drugs.

Subsequently, we performed integration of multi-view similarity matrices to extract similarity features between drug-drug and target-target. This integration process

comprehensively considers the similarities from different data perspectives, providing a more holistic and precise similarity assessment.

$$M_t = \text{mean}\{S_t, G_t\} \tag{1}$$

$$M_d = \text{mean}\{S_d, G_d\} \tag{2}$$

where the drug structure similarity matrix $S_t \in \mathbb{R}^{T\times T}$, the Gaussian similarity matrix $G_t \in \mathbb{R}^{T\times T}$; the target sequence similarity matrix $S_d \in \mathbb{R}^{D\times D}$, and the Gaussian similarity matrix $G_d \in \mathbb{R}^{D\times D}$.

Heterogeneous graph attention network

HAN [46, 47] applies the popular multi-head self-attention mechanism to the metapath node feature representation of a heterogeneous graph, then uses semantic-level attention to weight and aggregate the metapath attention, obtaining the final semantic representation that includes node features and domain structure features. In a heterogeneous graph, drugs can reach other drugs or targets can reach other targets through different paths, which are called metapaths.In this paper, two types of meta-paths are considered: drug-edge-target and target-edge-drug.

In our in-depth study of the node-level attention framework, we focused on how to effectively process and integrate information from diverse metapath neighbors to construct richer and more meaningful embedding representations for each drug and target node. In this process, the node-level attention mechanism plays a crucial role, allowing the model to dynamically consider the importance of each neighbor node when aggregating neighbor information, thereby enhancing the model's expressiveness and adaptability. First, we use a specific type of transformation matrix M_{ϕ_i} for each type of node to project the features of different types of nodes into the same feature space. The projection process is as follows:

$$h_i' = M_{\phi_i} \cdot h_i \tag{3}$$

In the context of self-attention mechanism, h_i represents the original features of node i, and h_i' denotes the projected features. Using self-attention allows dynamic learning of relationship weights between different nodes, enhancing the model's understanding of relative importance among various node pairs. Within the self-attention framework, the importance of node pairs (i, j), especially those connected via specific metapaths $\emptyset_i$, can be represented by calculating their attention scores. This mechanism considers not only the nodes' own features but also their interactions, thereby more accurately capturing dependencies and information flow between nodes. The computational formula is as follows:

$$e_{ij}^{\emptyset} = \text{att}_{\text{node}}(h_i', h_j'; \emptyset) \tag{4}$$

Here,the node attention value $e_{ij}^{\emptyset}$ indicates the importance of node j to node i, while h_i' and h_j' represent the projected features of nodes i and j, respectively.att_{node} represents a deep neural network that executes node-level attention. Then, structural information is injected into the model through masked attention. After obtaining the importance

between node pairs based on meta-paths, these are normalized using the softmax function to obtain weight coefficients.)

$$\alpha_{ij}^{\emptyset} = \text{softmax}_j(e_{ij}^{\emptyset}) = \frac{\exp(\sigma(\mathbf{a}_{\emptyset}^T \cdot [\mathbf{h}_i' || \mathbf{h}_j']))}{\sum_{k \in N_i^{\emptyset}} \exp(\sigma(\mathbf{a}_{\emptyset}^T \cdot [\mathbf{h}_i' || \mathbf{h}_k']))} \tag{5}$$

The symbol $||$ represents the concatenation operation, $a_{\emptyset}^T$ is the node-level attention vector for the metapath $\emptyset$, and $\alpha_{ij}^{\emptyset}$ denotes the weight coefficient for each feature vector.The embedding $z_i^{\emptyset}$ based on the metapath $\emptyset$ for node i can be aggregated using the projected features of its neighboring nodes weighted by corresponding attention coefficients, and then passed through an activation function, as shown below:

$$z_i^{\emptyset} = \sigma\left(\sum_{k \in N_i^{\emptyset}} \alpha_{ij}^{\emptyset} \cdot \mathbf{h}_j'\right) \tag{6}$$

Due to the scale-free nature of heterogeneous graphs and their high variance in data, to further optimize the information aggregation process, we employ a multi-head attention mechanism. This mechanism runs multiple attention "heads" in parallel, allowing the model to capture information in different subspaces. Each "head" focuses on different types of information. Through this approach, the model can achieve more comprehensive and in-depth learning by integrating information from multiple dimensions. Ultimately, the attention scores computed by these different "heads" are combined to generate a comprehensive and informative new feature representation for each node i:

$$\mathbf{z}_i^{\emptyset} = \|_{k=1}^{K} \sigma\left(\sum_{j \in N_i^{\emptyset}} \alpha_{ij}^{\emptyset} \cdot \mathbf{h}_j'\right) \tag{7}$$

After concatenating embeddings, we obtain a node embedding set for specific semantics, denoted as $\mathbf{Z}_{(\emptyset_1)}, \ldots, \mathbf{Z}_{(\emptyset_p)}$. Each element of this specific semantic node embedding set serves as input. The learning weights for each metapath $(\omega_{(\emptyset_1)}, \ldots, \omega_{(\emptyset_p)})$ are computed as follows:

$$(\omega_{(\emptyset_1)}, \ldots, \omega_{(\emptyset_p)}) = \text{att}_{\text{sem}}(\{\mathbf{Z}_{(\emptyset_1)}, \ldots, \mathbf{Z}_{(\emptyset_p)}\}) \tag{8}$$

Here, att_{sem} represents a deep neural network performing semantic-level attention.Next, we calculate the importance $f_{(\emptyset_i)}$ of each metapath $(\emptyset_i)$. First, we use a single-layer MLP followed by an activation function for nonlinear transformation. Then, we measure the importance of specific semantic embeddings by their similarity to the semantic-level attention vector $\mathbf{q}$:

$$f_{(\emptyset_i)} = \frac{1}{|\nu|} \sum_{i \in \nu} \mathbf{q}^T \cdot \tanh(W \cdot \mathbf{z}_i^{(\emptyset_p)} + \mathbf{b}) \tag{9}$$

For different tasks, metapaths $(\emptyset_p)$ may have varying weights. By using learned weights as coefficients, we can integrate these specific semantic embeddings to obtain the final

embedding vector **Z**, which combines the features of drugs and targets incorporating information from neighboring nodes:

$$\mathbf{Z} = \sum_{P=1}^{P} \omega_{(\emptyset_p)} \cdot \mathbf{Z}_{(\emptyset_p)} \tag{10}$$

Here, $\omega_{(\emptyset_p)}$ represents the normalized importance of each metapath.

Feature optimiser

In our study, the multi-modal feature vectors extracted through the multi-layer heterogeneous graph attention network [46] contain rich information but also exhibit high dimensionality and a certain level of noise. To optimize these features and enable machine learning algorithms to more effectively learn discriminative information, we decided to use the PCA algorithm as a feature optimization tool within our framework.

Initially, we applied the heterogeneous graph attention network to extract a series of multi-modal feature vectors. Subsequently, we employed a strategy of concatenating feature vectors from various convolutional layers to integrate information from different levels. This process helps to maximize the retention of useful information while enhancing the diversity of features.

$$Z_{\text{M-view}} = \text{Concat}(Z_{\text{conv1}}, Z_{\text{conv2}}, Z_{\text{conv3}}) \tag{11}$$

The vectors $Z_{\text{conv1}}, Z_{\text{conv2}}, Z_{\text{conv3}}$ are feature vectors from different convolutional layers, and concat denotes the operation of concatenating these feature vectors from different layers. The vector $Z_{\text{M-view}}$ is the resulting concatenated representation.

During the feature optimization phase, we input the integrated feature vector $Z_{\text{M-view}}$ into the feature optimizer, compute the covariance matrix of the centered feature matrix, and then perform eigenvalue decomposition to select the principal components based on their eigenvalues. Through orthogonal transformation, PCA converts the originally potentially correlated feature vector variables into a set of linearly independent variables, thereby identifying the principal components of the data. This not only reduces the dimensionality of the data but also minimizes information loss to the greatest extent possible, effectively removing noise from the feature vectors.

$$\tilde{Z}_{\text{M-view}} = Z_{\text{M-view}} - \mu_{\text{col}} \tag{12}$$

$$C_{d-t} = \tilde{Z}^{T}_{\text{M-view}} \cdot \tilde{Z}_{\text{M-view}} \tag{13}$$

$$W_k = \text{Sort}\{\text{Decomp}[C_{d-t}]\} \tag{14}$$

$$Z_{\text{final}} = \tilde{Z}_{\text{M-view}} \cdot W_k \tag{15}$$

Here,the vector μ_{col} represents the mean vector for centering, and $\tilde{Z}_{\text{M-view}}$ denotes the centered feature vector, which is obtained by subtracting the mean of each column from the original feature vectors to ensure that the mean of the data in each dimension is zero. This is a necessary step for performing PCA.C_{d-t} represents the covariance matrix of

the centered feature matrix, and W_k is the transformation matrix. Decomp denotes the decomposition of the covariance matrix, and Sort refers to the process of sorting the eigenvectors by the magnitude of their corresponding eigenvalues. The transformation matrix is constructed by selecting the eigenvectors corresponding to the eigenvalues that collectively account for 99.80% of the cumulative variance contribution rate,the vector Z_{final} is obtained by multiplying the centered feature vector $\tilde{Z}_{\text{M-view}}$ with the transformation matrix W_k. This represents the dimensionality-reduced feature representation after PCA processing, which retains the most important information in the data.

In practical applications, we selected a dimensionality reduction index of 0.9980, which means that we retained 99.8% of the data's principal components. This choice was based on a trade-off between data simplification and information retention, ensuring that while significantly reducing data complexity, we preserved the vast majority of information critical to the model's predictive performance. This step is crucial for enhancing the overall performance and prediction accuracy of the model, providing a solid foundation for deeply mining and understanding complex bioinformatics data.

Classifier

After obtaining the optimized embedding vectors, we employ the random forest (RF) algorithm to predict the final drug-target interaction (DTI) associations. The RF algorithm introduces randomness during the training process to reduce the risk of overfitting, thereby exhibiting strong resistance to overfitting. Additionally, it consists of multiple trees, each trained on a different dataset, enabling the capture of various patterns in the data, which grants it a superior generalization capability compared to other machine learning algorithms.

The concatenated features resulting from the embedding process comprise $M \times N$ vectors, each with $2 \times L_{\text{in}}$ features. These features serve as the input to the random forest, which generates a series of decision trees. Through bootstrap aggregating, k samples are drawn with replacement from the dataset to form the training sets, and w decision trees are trained. Each training set contains duplicate samples. The final prediction value is obtained by aggregating the mean predictions of the w decision trees.

$$T_i \leftarrow \text{Train}(\mathbf{X}^{(i)}), \quad i = 1, 2, \ldots, w \tag{16}$$

where $\mathbf{X}^{(i)}$ denotes the i-th generated training set.

$$\hat{y}_{\text{DTI}} = \frac{1}{w} \sum_{i=1}^{w} T_i(\mathbf{x}) \tag{17}$$

where $\hat{y}_{\text{DTI}}$ is the predicted drug-target interaction value.

By utilizing this approach, the RF algorithm effectively aggregates the predictions from multiple decision trees, thereby enhancing the robustness and accuracy of the DTI association predictions.

Experiments and results

To evaluate the accuracy of the MHAPR computational framework in predicting drug-target interactions, we compared its performance on two benchmark datasets with seven state-of-the-art baseline models: MNGACDA [48], GATECDA [49], MINIMDA [50], HFHLMDA [51], CGHCN [52],MIDTI [53] and DTI-CNN [41].

MNGACDA [48] constructs a multimodal network using various information sources from drugs and circRNA, and then applies an inner product decoder based on the embedding representations of drugs and circRNA to predict their interaction scores. GATECDA [49] employs a graph attention autoencoder (GATE) to extract low-dimensional representations of drugs/circRNA, effectively preserving key information from sparse high-dimensional features and realizing effective integration of node neighborhood information. MINIMDA [50] constructs an integrated disease similarity network and miRNA similarity network using multiple information sources, then combines mixed high-order neighborhood information from multimodal networks to obtain disease and miRNA embeddings, and finally uses a multi-layer perceptron (MLP) for prediction. HFHLMDA [51] utilizes a hypergraph learning model to learn a projection matrix for calculating the association scores of uncertain diseases and miRNAs. CGHCN [52] combines graph convolutional networks and hypergraph convolutional networks, while DTI-CNN [41] leverages the Jaccard similarity coefficient and random walk model to obtain feature embeddings, and then uses convolutional neural networks to predict drug-target protein interactions after dimensionality reduction.The MIDTI method [53] first utilizes graph convolutional networks (GCNs) to simultaneously obtain the embedding representations of drugs and targets from multi-type networks. Then, it employs a deep interactive attention mechanism to further learn the discriminative embeddings of drugs and targets while fully considering the known DTI relationships.

To ensure fair and accurate results, we used the same similarity data for the baseline methods, specifically drug structure similarity, drug Gaussian similarity, target sequence similarity, and target Gaussian similarity. For the single-modal models CGHCN [52] and HFHLMDA [51], we only used the drug-target similarity matrix as training data.

To verify the generalizability of the model, we split the two benchmark datasets into training and testing sets with a 4:1 ratio, and applied 5-fold cross-validation (5-CV) on the training set to adjust the model parameters and structure. During the training process, for the FSL dataset, the embedding dimension was set to 1024, the number of attention mechanism heads to 4, and the number of layers in the heterogeneous attention network to 6. For the DC dataset, the embedding dimension was set to 512, the number of attention mechanism heads to 6, and the number of layers in the heterogeneous attention network to 4. The number of training epochs was set to 500, and the Adam optimizer was used. The optimal hyperparameter combination was a learning rate of 0.001 and a weight decay rate of 0.0002. Additionally, the dropout rate was set to 0.5 to randomly ignore some neurons, preventing overfitting.

Performance evaluation of the DC dataset

On the DC dataset, we trained the model using five-fold cross-validation. As shown in Table 2 and Fig. 2, compared to other methods, the DTI-MHAPR computational

Table 2 Comparison of performance using 5-CV on the DC dataset

Method	Auc	Aupr	F1-score	Acc	Recall	Spec	Precision
CGHCN	0.9001	0.9007	0.8370	0.8309	0.8664	0.7956	0.8107
DTI-CNN	0.9358	0.9423	0.8636	0.9412	0.8685	0.8726	0.8148
GATECDA	0.7918	0.8000	0.7501	0.7205	0.8270	0.6170	0.7011
HFHLMDA	0.8411	0.7529	0.8042	0.7938	0.8465	0.7407	0.7662
MINIMDA	0.9537	0.9359	0.9074	0.9030	0.9491	0.8569	0.8695
MNGACDA	0.8927	0.7500	0.9031	0.8928	1.0000	0.7854	0.8234
MIDTI	0.9769	0.9615	0.9609	0.9583	0.9938	0.9236	0.9285
Our model	0.9950	0.9941	0.9738	0.9735	0.9875	0.9595	0.9606

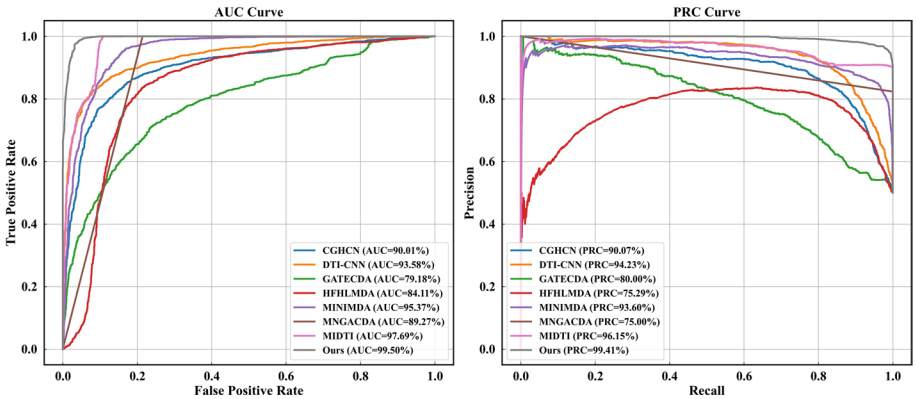

Fig. 2 Visualisation of AUC and AUPR values compared with the baseline model on the 5-CV

Table 3 Comparison of performance using 5-CV on the FSL dataset

Method	Auc	Aupr	F1-score	Acc	Recall	Spec	Precision
CGHCN	0.9058	0.9061	0.8441	0.8456	0.8375	0.8542	0.8524
GATECDA	0.9146	0.9276	0.8372	0.8396	0.8266	0.8520	0.8497
HFHLMDA	0.8875	0.8715	0.8217	0.8145	0.8547	0.7740	0.7922
MINIMDA	0.9408	0.9198	0.8898	0.8840	0.9356	0.8327	0.8489
MNGACDA	0.9674	0.9629	0.9086	0.9035	0.9585	0.8474	0.8640
MIDTI	0.9634	0.9578	0.9032	0.8997	0.9497	0.8676	0.8831
Our model	0.9890	0.9869	0.9544	0.9535	0.9778	0.9295	0.9324

framework achieved significant improvements across six evaluation metrics, except for Recall. Notably, it also attained a commendable score in Recall.

Performance evaluation of the FSL dataset

In addition to the aforementioned seven methods, we also compared our results with those obtained using the Heterogeneous Network that employs the Forward Similarity Integration (FSI) algorithm [42] on this dataset. We conducted five-fold cross-validation experiments and the results are presented in Table 3, showcasing the model's performance across seven evaluation metrics. Figure 3 visualizes the training curves. It can be

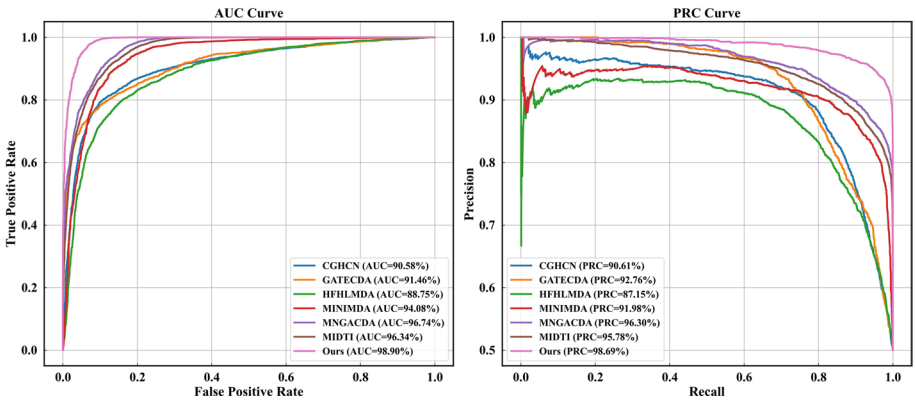

Fig. 3 Visualisation of AUC and AUPR values compared with the baseline model on the 5-CV

observed that the DTI-MHAPR computational framework achieved an AUC of 98.90% and an AUPR of 98.69%, demonstrating significant improvements.

In summary, these experiments conducted on two independent datasets indicate that the performance of DTI-MHAPR surpasses all other tested methods. This demonstrates the broad applicability of our approach, effectively enhancing the prediction of drug-target interactions (DTIs).

Dimensionality reduction analysis

The DC dataset contains 708 drugs and 1512 target proteins, while the FSL dataset includes 862 drugs and 1517 target proteins. Consequently, the dimensionality of the multi-modal feature vectors obtained after feature fusion is extremely high. Handling such high-dimensional data poses significant challenges for the model. However, much of this data is highly redundant. By employing dimensionality reduction algorithms, we can focus the network layers on key information, enabling the model to learn more discriminative features. Based on prior research, we experimented with various types of dimensionality reduction algorithms, including both linear and nonlinear methods. By comparing different algorithms and analyzing the performance of the model with varying dimensionality reduction parameters, we identified the optimal parameters for the model.

An exploration of dimensionality reduction algorithms

Feature engineering is an important preprocessing step that helps to extract transformational features from raw data to simplify machine learning models and improve the quality of machine learning algorithm results.Machine learning practitioners spend most of their time on data cleaning and feature engineering, so we are inspired to focus on the investigation of downscaling optimisation of the multimodal features mined by the HAN network, and the experimental steps are as follows:

- 1) Using segmented sampling method and dense sampling method to apply PCA [38], t-SNE [54], LLE (Locally Linear Embedding) [55], FastICA (Fast Independent Component Analysis) [56] algorithms are applied to the normalised dataset and the

Table 4 Comparison of PCA dimensionality reduction versus t-SNE dimensionality reduction on the DC dataset

DRA	Optimal AUC	Optimal AUPR	Best_n_components
PCA	0.9950	0.9941	0.9980
t-SNE	0.9927	0.9912	3
LLE	0.9943	0.9933	70
FastICA	0.9882	0.9875	32

Table 5 Comparison of PCA dimensionality reduction versus t-SNE dimensionality reduction on the FSL dataset

DRA	Optimal AUC	Optimal AUPR	Best_n_components
PCA	0.9890	0.9869	0.9980
t-SNE	0.9891	0.9867	3
LLE	0.9886	0.9860	24
FastICA	0.9872	0.9868	48

resulting dimensionality reduced dataset is experimented with using the ML algorithm.

- 2) Analyse the results obtained by the linear dimensionality reduction algorithms PCA, FastICA and the nonlinear dimensionality reduction algorithms LLE, t-SNE after dimensionality reduction using the ML algorithm, and investigate the effect of dimensionality reduction on the prediction performance of the ML algorithm with respect to DTIs.

After conducting the experiments, we obtained the results shown in Tables 4 and 5, with optimal values selected after intensive sampling. Among them, DRA represents different types of dimensionality reduction algorithms. The optimal AUC and AUPR values for PCA on the DC dataset are higher than those of the other three dimensionality reduction algorithms, reaching 99.50% and 99.41%, respectively. The two evaluation indexes of the PCA algorithm are also superior to the other algorithms on the FSL dataset, and only the AUC value of the t-SNE The AUC value of PCA algorithm is slightly higher than that of t-SNE algorithm.

We found that t-SNE algorithm needs to calculate the similarity between the samples in each iteration, and the time complexity required in processing a large amount of data is much larger than that of PCA. Therefore, considering the accuracy and time complexity of the prediction of DTIs, the use of PCA algorithm as the feature optimisation algorithm of our DTIs prediction framework is suitable for us. Therefore, considering the accuracy and time complexity of DTIs prediction, it is suitable to use PCA as our feature optimisation algorithm for DTIs prediction framework.

An investigation of the effect of PCA downscaling index on DTI prediction accuracy

After exploring two sets of advanced dimensionality reduction algorithms (linear and nonlinear), the result obtained is that PCA dimensionality reduction is superior to

Table 6 Dimensionality reduction analysis on the DC dataset

DC	AUC	PRC	F1_Score	Acc	Recall	Specificity	Precision
None	0.9808	0.9800	0.9225	0.9212	0.9396	0.9029	0.9072
0.9980	0.9950	0.9941	0.9738	0.9735	0.9875	0.9595	0.9606
0.9960	0.9941	0.9929	0.9722	0.9719	0.9845	0.9596	0.9603
0.9940	0.9939	0.9928	0.9714	0.9711	0.9823	0.9601	0.9607
0.9920	0.9936	0.9925	0.9689	0.9685	0.9833	0.9538	0.9550

Table 7 Dimensionality reduction analysis on the FSL dataset

DC	AUC	PRC	F1_Score	Acc	Recall	Specificity	Precision
None	0.9825	0.9822	0.9239	0.9228	0.9351	0.9105	0.9140
0.9980	0.9890	0.9869	0.9544	0.9535	0.9778	0.9295	0.9324
0.9960	0.9862	0.9835	0.9480	0.9468	0.9722	0.9220	0.9253
0.9940	0.9787	0.9724	0.9332	0.9305	0.9742	0.8872	0.8960
0.9920	0.9756	0.9682	0.9292	0.9266	0.9672	0.8855	0.8941

several other dimensionality reduction algorithms, next we will focus our attention on PCA dimensionality reduction (dense sampling after segmented sampling method).

PCA finds the main direction of the data by calculating the covariance matrix of the original data, and then solves for the eigenvalues and eigenvectors of this covariance matrix using eigenvalue decomposition. The eigenvalues represent the variance of the data in the direction of the eigenvectors, while the eigenvectors represent the main distribution of the data in each direction. It will select the eigenvectors with the largest eigenvalues as the principal components, and map the data onto the subspaces tensored by the principal components, thus achieving the purpose of data dimensionality reduction. For PCA dimensionality reduction index the experimental results are as follows Tables 6 and 7.Among them, DC denotes the percentage of principal components retained by the PCA algorithm.

It can be seen that on the DC dataset, the PCA dimensionality reduction index between 0.9920 and 0.9980 will play a role in eliminating redundant information in the data, improving the efficiency of data representation, and can reduce the cost of computation and storage, especially when the dimensionality reduction index is 0.9980 (i.e., 99.8% of the principal components are retained), the AUC value on this dataset reaches 99.50%, the The AUPR value reaches 99.41%, and the other five metrics are all improved in different magnitudes relative to no dimensionality reduction;

On the FSL dataset, the downscaling indices between 0.9960–0.9980 are better optimised for multimodal features, similarly, when the downscaling index is 0.9980, the AUC value reaches 98.90%, the AUPR value reaches 98.69%, and the other five evaluation metrics are also improved in different magnitudes compared to no downscaling.

Despite the fact that the heterogeneous graph network consists of nonlinear data, the feature matrix input to the PCA algorithm remains linear after feature extraction using HAN. It is important to note that PCA assumes the data follows a linear distribution, so it may fail when the data has a nonlinear structure. Additionally, the results of PCA are

Table 8 Performance of different machine learning algorithms for 5-CV on the FSL dataset

Classifier	AUC	PRC	F1_Score	Acc	Recall	Spec	Precision
KNN	0.9493	0.9377	0.8939	0.8859	0.9650	0.8062	0.8329
NB	0.8969	0.8414	0.8483	0.8256	0.9773	0.6730	0.7497
SVM	0.9724	0.9434	0.9489	0.9477	0.9758	0.9200	0.9241
LR	0.8941	0.8124	0.8587	0.8468	0.9339	0.7601	0.7967
RF	0.9890	0.9869	0.9544	0.9535	0.9778	0.9295	0.9324
DT	0.9423	0.8955	0.9438	0.9282	0.9251	0.9711	0.8800
XGB	0.9842	0.9787	0.9542	0.9534	0.9763	0.9305	0.9331

Table 9 Performance of different machine learning algorithms for 5-CV on the DC dataset

Classifier	AUC	PRC	F1_Score	Acc	Recall	Spec	Precision
KNN	0.9659	0.9570	0.9317	0.9269	0.9952	0.8583	0.8761
NB	0.9301	0.8844	0.9009	0.8900	0.9995	0.7804	0.8201
SVM	0.9850	0.9754	0.9706	0.9704	0.9807	0.9600	0.9608
LR	0.9402	0.8672	0.9262	0.9210	0.9921	0.8495	0.8686
RF	0.9950	0.9941	0.9738	0.9735	0.9875	0.9595	0.9606
DT	0.9569	0.9227	0.9512	0.9501	0.9760	0.9242	0.9278
XGB	0.9926	0.9898	0.9715	0.9711	0.9870	0.9554	0.9566

influenced by eigenvalue decomposition, and for high-dimensional data, approximate calculations of eigenvalue decomposition may be required, increasing computational complexity.

Notwithstanding these challenges, we achieved good results on both datasets, demonstrating that the PCA dimensionality reduction algorithm effectively optimizes features. Moreover, applying the PCA algorithm with a dimensionality reduction index of 0.9980 significantly improved the classification performance of multimodal data features in machine learning. Taking the DC dataset as an example, the embedding dimensions of drug and target features were 1536 before dimensionality reduction. Notably, since the data for each fold is reduced to different dimensions during the dimensionality reduction process, the final embedding dimensions of each fold vary. In five-fold cross-validation, the embedding dimensions for drug nodes were 8, 9, 10, 9, and 11, while for target nodes, they were 9, 8, 10, 9, and 12. For the FSL dataset, the embedding dimensions for drug nodes were 4, 6, 5, 6, and 4, while for target nodes, they were 3, 6, 5, 6, and 2.

Choice of classifier

We explore the classification effectiveness of seven machine learning algorithms on two datasets using grid point search. The evaluation metrics of the machine learning algorithms on five-fold cross-validation are shown in Tables 8 and 9, and their AUC vs. AUPR graphs are shown in Figs. 4 and 5. In the table, RF outperforms all other algorithms in five of the seven metrics and has the highest values of AUC vs. AUPR among the seven algorithms, indicating that the Random Forest algorithm outperforms other machine learning algorithms on both datasets. Therefore, we choose the random forest algorithm for supervised learning to perform DTI prediction. SVM

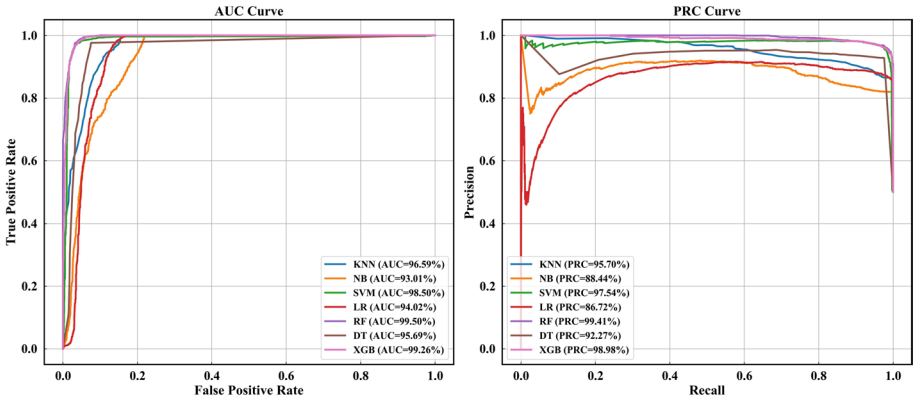

Fig. 4 AUC and AUPR curves for seven machine learning algorithms on the DC dataset

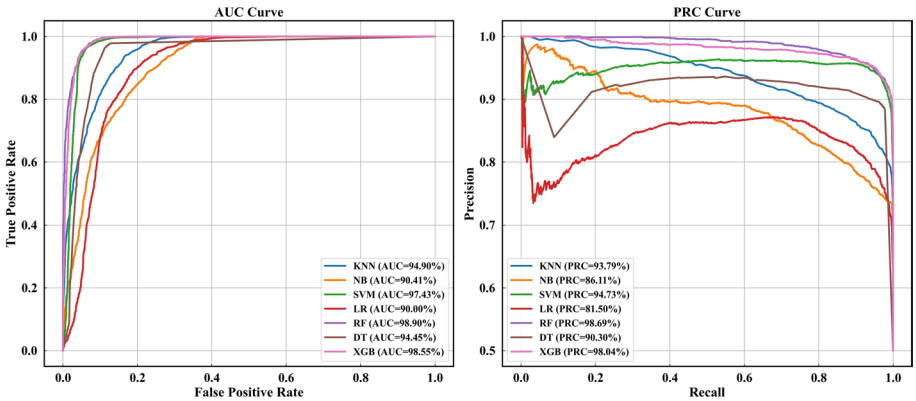

Fig. 5 AUC and AUPR curves for seven machine learning algorithms on the FSL dataset

refers to support vector machine, RF refers to random forest, XGB refers to eXtreme gradient boosting, KNN refers to K-nearest neighbors, DT refers to decision tree, LR refers to logistic regression, and NB refers to Naive Bayes.

The random forest algorithm has two critical parameters: n_estimators and max_depth. Generally, increasing n_estimators enhances the model's performance but also raises computational costs. The max_depth parameter controls the complexity of the trees, thereby preventing overfitting. We employed grid search to determine the optimal parameters for RF within the MHAPR framework.

As illustrated in Fig. 6, for the FSL dataset, the model achieves the highest AUC and AUPR values when max_depth is set to 20 and n_estimators is 300. For the DC dataset, shown in Fig. 7, the model reaches its optimal performance with a max_depth of 20. The AUC achieves its peak when n_estimators is 500, while the AUPR continues to show minor improvements up to n_estimators of 800. However, since the AUPR increases marginally beyond n_estimators of 500 and the model complexity significantly increases, we determined that the optimal parameters for the DC dataset are n_estimators of 500 and max_depth of 20.

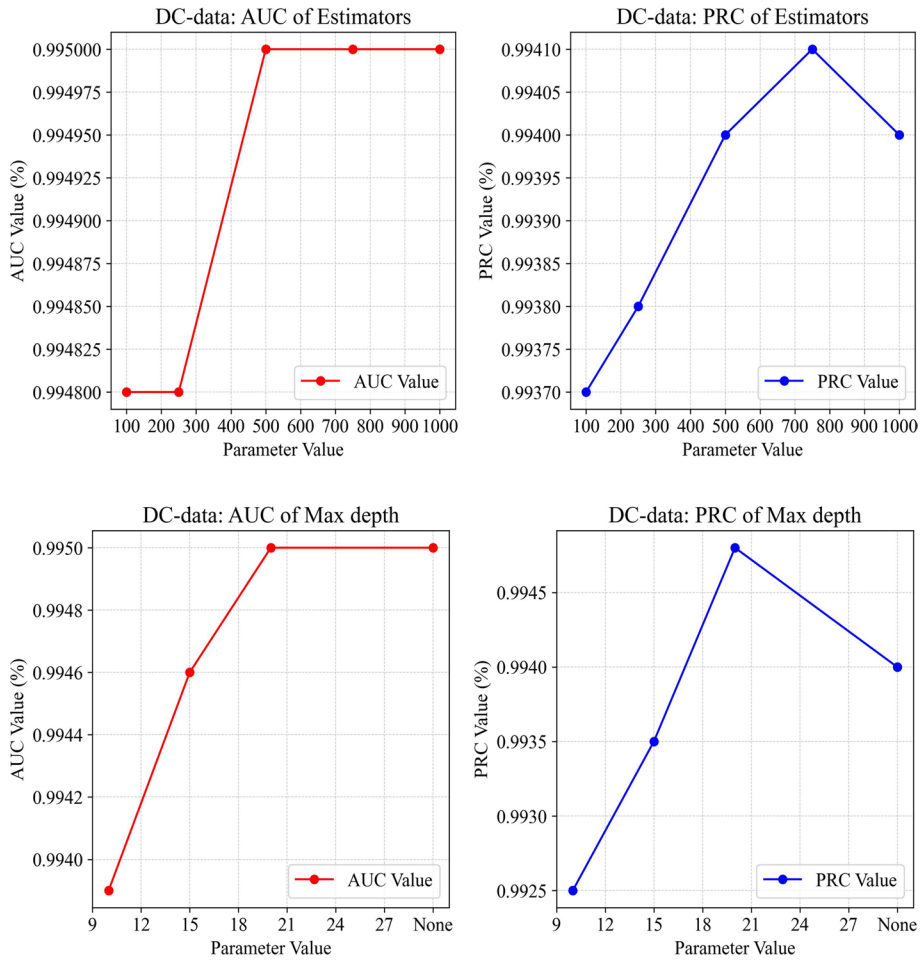

Fig. 6 Parameter grid search on the DC dataset

Ablation experiment

To evaluate the effectiveness of various modules within the MHAPR computational framework, we proposed three model variants: MHAPR-PCA, MHAPR-dropout, and MHAPR-concat. By controlling for major variables, we systematically excluded the PCA layer, dropout layer, and concat layer from the framework to understand the performance of each module within DTI-MHAPR. To assess the impact of the PCA layer in feature optimization, we first removed this layer from the model, naming this approach -PCA. Similarly, to evaluate the effect of the concat layer, we excluded the concatenation layer in the forward propagation, naming this approach -concat. The model with the dropout layer removed is referred to as -dropout.

All three methods are used in the DTI prediction task to compare the performance and the comparison results are shown in Fig. 8. The results show that the DTI-MHAPR framework can simultaneously obtain higher AUC, AUPR, F1-score, and ACC scores compared to the other three methods on both datasets, and the standard deviation of our model is very small with good generalisation ability, which indicates that the three modules of our model are well-designed.

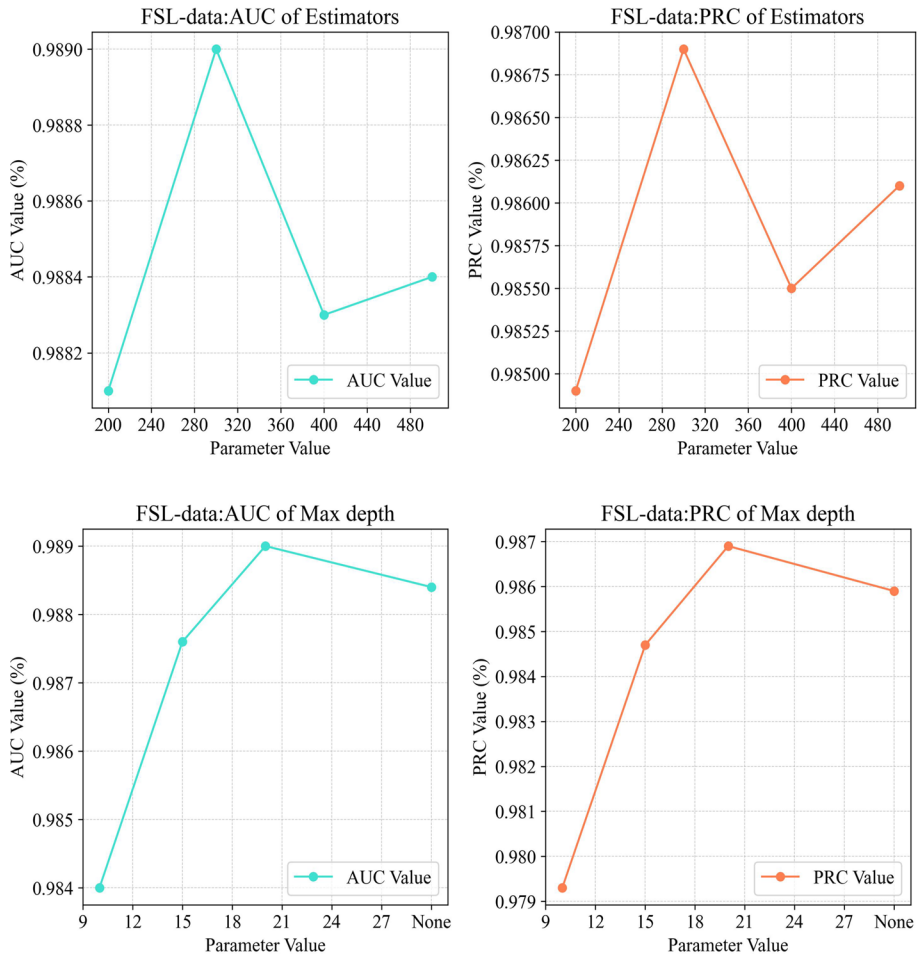

Fig. 7 Parameter grid search on the FSL dataset

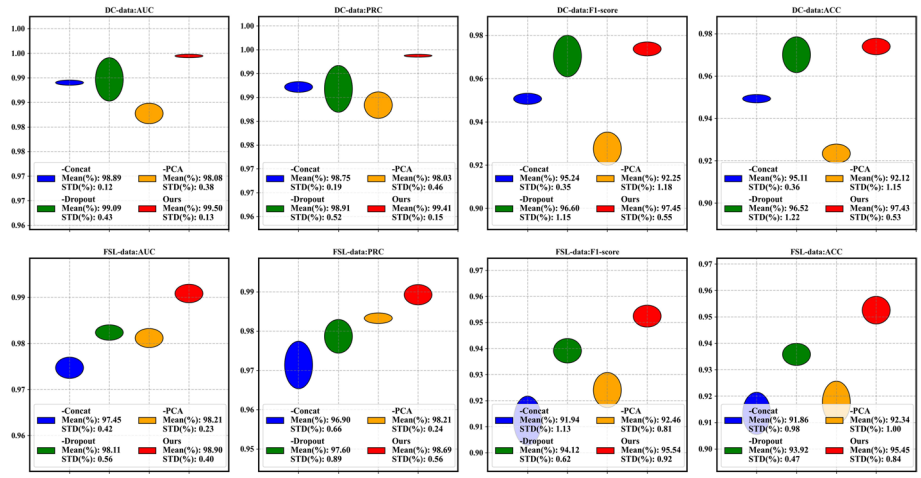

Fig. 8 Ablation experiment results on different network architectures. Mean represents the average, STD represents the standard deviation, and the vertical axis represents the evaluation indicators corresponding to each variant

Table 10 The top 20 targets associated with the drug Olanzapine

Ranking	Target	Evidence	Ranking	Target	Evidence
1	P21728	DrugBank	11	P28335	DrugBank
2	P20309	DrugBank	12	P34969	None
3	P21918	DrugBank	13	P28221	DrugBank
4	P08912	None	14	P28222	DrugBank
5	P08908	DrugBank	15	P08913	None
6	P18825	None	16	P08172	DrugBank
7	P21917	DrugBank	17	P18089	None
8	P08173	DrugBank	18	P35368	DrugBank
9	P28223	DrugBank	19	P35462	DrugBank
10	P35348	DrugBank	20	P14416	DrugBank

Table 11 The top 20 targets associated with the drug Quetiapine

Ranking	Target	Evidence	Ranking	Target	Evidence
1	P28223	DrugBank	11	P08173	DrugBank
2	P08913	DrugBank	12	P21728	DrugBank
3	P08172	DrugBank	13	P20309	DrugBank
4	P14416	DrugBank	14	P35462	DrugBank
5	P35368	DrugBank	15	P30536	None
6	P08908	DrugBank	16	P18825	DrugBank
7	P35348	DrugBank	17	P08912	DrugBank
8	P18089	DrugBank	18	P34969	DrugBank
9	P28222	DrugBank	19	P21918	DrugBank
10	P28221	DrugBank	20	P41595	None

Case study

We extracted the known DTIs from the DrugBank database and similarly selected the three drugs with high number of interactions among the known DTIs, which are Olanzapine, Quetiapine and Cabergoline.Then we found out the prediction results of the three drugs and obtained the target proteins with the score ranked in the top 20 for validation, and it was found that the The vast majority of DTIs were validated by the DrugBank database, as shown in Tables 10, 11,and 12 below, which shows the three drugs corresponding to the target proteins with the top 20 scores. In addition, in Fig. 9,we selected ten DTIs to be visualised using the knowledge graph, and the weights between the edges are the predicted scores for the extent to which the target is associated with the drug. These results indicate that the DTI-MHAPR method has good performance in drug-target interaction prediction.

Discussion

Motivated by the potential to harness the rich information encoded in the topological structure, our study introduces the DTI-MHAPR framework, which efficiently prioritizes and encodes the most salient features during the encoding-decoding process, thereby significantly enhancing the prediction of drug-target interactions.MHAPR

Table 12 The top 20 targets associated with the drug Cabergoline

Ranking	Target	Evidence	Ranking	Target	Evidence
1	P28223	DrugBank	11	P08173	DrugBank
2	P08913	DrugBank	12	P21728	DrugBank
3	P08172	None	13	P20309	None
4	P14416	DrugBank	14	P35462	DrugBank
5	P35368	DrugBank	15	P18825	DrugBank
6	P08908	DrugBank	16	P08912	None
7	P35348	DrugBank	17	P34969	DrugBank
8	P18089	DrugBank	18	P21918	DrugBank
9	P28222	DrugBank	19	P41595	DrugBank
10	P28221	DrugBank	20	P28335	DrugBank

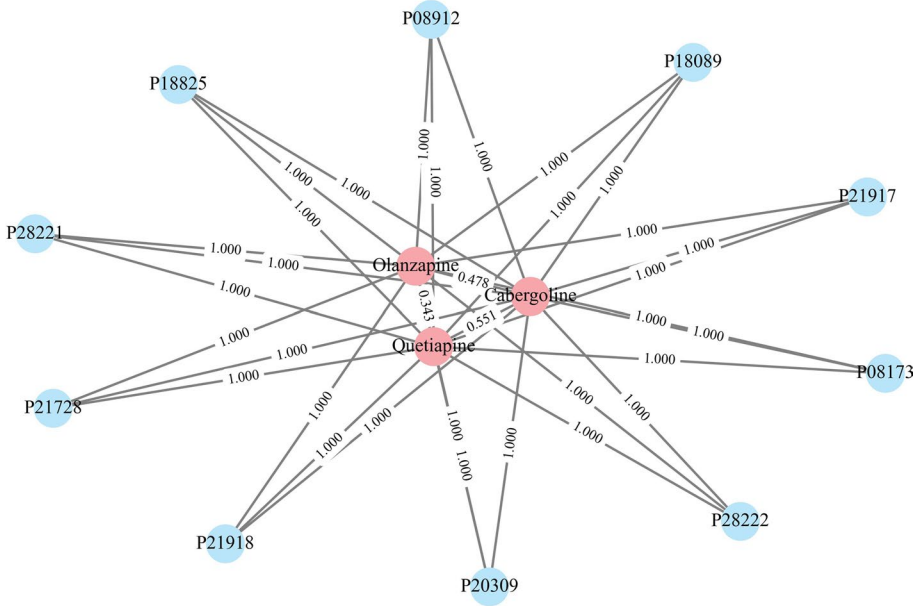

Fig. 9 Drug-target association subnetwork. The pink nodes represent the three drugs and the light blue nodes represent the top10 targets associated with the drugs

integrates diverse biological information of drugs and targets, providing the model with a rich biological information base. Additionally, by employing a hierarchical structure of multi-layer HAN, the framework effectively achieves deep representation extraction and fusion of multi-layer heterogeneous nodes. Furthermore, prior to feature decoding, MHAPR utilizes the PCA algorithm for feature optimization and employs the random forest algorithm as the classifier, enhancing the model's tolerance to data noise and improving overall robustness. Experimental results indicate that DTI-MHAPR outperforms existing methods in key evaluation metrics such as AUC, AUPR, and F1-Score.

However, despite the excellent performance of DTI-MHAPR, the framework still faces certain challenges and limitations. Specifically, in the application of the PCA

algorithm, the computational complexity of eigenvalue decomposition for high-dimensional data is relatively high, sometimes necessitating the use of approximate computation methods, which increases computational complexity.

To reduce the computational cost and complexity of the proposed method, we will implement model compression techniques in the future, utilizing pruning and quantization strategies to decrease the model's storage and computational demands. Additionally, we will optimize the data preprocessing pipeline by employing streaming processing techniques to filter and standardize data in real time, thereby enhancing the efficiency of subsequent training and prediction. Through these specific measures, we aim to improve the model's operability and efficiency, alleviating the high computational complexity associated with the eigenvalue decomposition of high-dimensional data from PCA, ensuring that it maintains good performance in larger-scale applications.

Moreover, the potential extensibility of the DTI-MHAPR framework opens new directions for future research. We plan to apply this framework to predict other types of interactions, such as drug-drug, drug-disease, and protein-disease interactions, to verify its applicability and extend its application scope.

Conclusion

In this study, we propose a novel computational method, DTI-MHAPR, designed to predict potential interactions between drugs and targets. Compared to existing models, DTI-MHAPR not only integrates sequence and Gaussian similarity information of drugs and targets but also constructs a multi-layer heterogeneous graph attention network (HAN) that effectively encodes and extracts representation vectors of drugs and targets. This approach can deeply explore the complex and subtle relationships between drugs and targets, enhancing the representation capability between nodes. To optimize model features, this study also investigates four different feature selection methods and ultimately adopts the PCA algorithm to improve the quality of feature representation and enhance the model's discriminative power. Extensive experimental results demonstrate that DTI-MHAPR provides an efficient new method for drug-target identification, contributing to the advancement of precision medicine and personalized treatment strategies.

In future research, we will focus on addressing the computational challenges of the PCA algorithm in handling high-dimensional data from the perspective of feature optimization. We plan to adopt advanced approximate eigenvalue decomposition methods, such as the Lanczos or Arnoldi algorithms, which can significantly reduce computational complexity, accelerate feature extraction speed, and enhance the model's responsiveness. Additionally, we will explore the introduction of ensemble learning methods to combine multiple feature selection algorithms, automatically identifying the features that contribute most to model performance, thereby reducing computational load while retaining important information.

Abbreviations

DTIs	Drug-target interactions
PCA	Principal component analysis
LLE	Locally linear embedding
t-SNE	T-distributed stochastic neighbor embedding
FastICA	Fast independent component analysis algorithms

HAN	Heterogeneous graph attention network
RF	Random forest algorithm
SVM	Support vector machine
XGB	EXtreme gradient boosting
DT	Decision tree
LR	Logistic regression
NB	Naive Bayes

Acknowledgements
We express our gratitude to the School of Information and Artificial Intelligence, Anhui Agricultural University, for providing the computational resources.

Author contributions
G.Y. was responsible for the main manuscript writing, conducting the experiments, and drawing the flowcharts. Y.B.L. was in charge of code modification and improvement. S.J.W. was responsible for the creation of Tables 1–4, W.X.C. for the creation of Tables 5–8, and X.L.Z. for the creation of Tables 9–11. Y.M.W. was responsible for revising the details of the manuscript.

Funding
This study was supported by the National Key Research and Development Program of China (Grant No. 2023YFC3205701) and also by the Anhui Beidou Precision Agriculture Information Engineering Research Center (Grant Nos. BDSY2023002 and 2022AH040122).

Availability of data and materials
The datasets and source code utilized in this study are publicly accessible via the following GitHub repository: https://github.com/Yang06092/MHAPR-DTI.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare no competing interests.

Received: 1 August 2024 Accepted: 20 December 2024
Published online: 13 January 2025

References

1. Zhang X, Zou Q, Rodriguez-Paton A, Zeng X. Meta-path methods for prioritizing candidate disease miRNAs. IEEE/ACM Trans Comput Biol Bioinf. 2019;16(1):283–91. https://doi.org/10.1109/TCBB.2017.2776280.
2. Zeng X, Liu LLL, Zou Q. Prediction of potential disease-associated microRNAs using structural perturbation method. Bioinformatics. 2018;34(14):2425–32. https://doi.org/10.1093/bioinformatics/bty112.
3. Hua Y, Song X, Feng Z, Wu X. MFR-DTA: a multi-functional and robust model for predicting drug-target binding affinity and region. Bioinformatics. 2023;39(2):056. https://doi.org/10.1093/bioinformatics/btad056 (https://academic.oup.com/bioinformatics/article-pdf/39/2/btad056/49096141/btad056.pdf).
4. Zeng X, Liao Y, Liu Y, Zou Q. Prediction and validation of disease genes using HeteSim scores. IEEE/ACM Trans Comput Biol Bioinf. 2017;14(3):687–95. https://doi.org/10.1109/TCBB.2016.2520947.
5. Sheng N, Huang L, Wang Y, Zhao J, Xuan P, Gao L, Cao Y. Multi-channel graph attention autoencoders for disease-related lncRNAs prediction. Brief Bioinform. 2022;23(2):604. https://doi.org/10.1093/bib/bbab604 (https://academic.oup.com/bib/article-pdf/23/2/bbab604/42805961/bbab604.pdf).
6. Sheng N, Cui H, Zhang T, Xuan P. Attentional multi-level representation encoding based on convolutional and variance autoencoders for lncRNA-disease association prediction. Brief Bioinform. 2020;22(3):067. https://doi.org/10.1093/bib/bbaa067 (https://academic.oup.com/bib/article-pdf/22/3/bbaa067/37965878/bbaa067.pdf).
7. Sheng N, Wang Y, Huang L, Gao L, Cao Y, Xie X, Fu Y. Multi-task prediction-based graph contrastive learning for inferring the relationship among lncRNAs, miRNAs and diseases. Brief Bioinform. 2023;24(5):276. https://doi.org/10.1093/bib/bbad276 (https://academic.oup.com/bib/article-pdf/24/5/bbad276/51711044/bbad276.pdf).
8. Kumar Shukla P, Kumar Shukla P, Sharma P, et al. Efficient prediction of drug-drug interaction using deep learning models. IET Syst Biol. 2020;14(4):211–6. https://doi.org/10.1049/iet-syb.2019.0116.
9. Hu L, Wang X, Huang YA, Hu P, You ZH. A survey on computational models for predicting protein-protein interactions. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbab036.
10. Emanuelsson O, Brunak S, Von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2(4):953–71. https://doi.org/10.1038/nprot.2007.131.
11. Zhao BW, Su XR, Hu PW, Ma YP, Zhou X, Hu L. A geometric deep learning framework for drug repositioning over heterogeneous information networks. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbac384.

12. Zhao BW, Hu L, You ZH, Wang L, Su XR. HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbab515.
13. Ezzat A, Wu M, Li XL, Kwoh CK. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform. 2019;20(4):1337–57. https://doi.org/10.1093/bib/bby002.
14. Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. IJMS. 2019;20(11):2783. https://doi.org/10.3390/ijms20112783.
15. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. Methods. 2015;74:97–106. https://doi.org/10.1016/j.ymeth.2015.01.015.
16. Acharya C, Coop A, Polli JE, MacKerell AD. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. CAD. 2011;7(1):10–22. https://doi.org/10.2174/157340911793743547.
17. Yamanishi Y. Chemogenomic approaches to infer drug–target interaction networks. In: Mamitsuka H, DeLisi C, Kanehisa M, editors. Data mining for systems biology, vol. 939. Totowa: Humana Press; 2013. p. 97–113. https://doi.org/10.1007/978-1-62703-107-3_9.
18. Zhang W, Chen Y, Liu F, Luo F, Tian G, Li X. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinform. 2017;18(1):18. https://doi.org/10.1186/s12859-016-1415-9.
19. Zhang W, Zou H, Luo L, Liu Q, Wu W, Xiao W. Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing. 2016;173:979–87. https://doi.org/10.1016/j.neucom.2015.08.054.
20. Van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PLoS ONE. 2013;8(6):66952. https://doi.org/10.1371/journal.pone.0066952.
21. Shi JY, Yiu SM. Srp: A concise non-parametric similarity-rank-based model for predicting drug-target interactions. In: 2015 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 1636–1641 2015. https://doi.org/10.1109/BIBM.2015.7359921
22. Mei JP, Kwoh CK, Yang P, Li XL, Zheng J. Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics. 2013;29(2):238–45. https://doi.org/10.1093/bioinformatics/bts670.
23. Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics. 2009;25(18):2397–403. https://doi.org/10.1093/bioinformatics/btp433.
24. Gönen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics. 2012;28(18):2304–10. https://doi.org/10.1093/bioinformatics/bts360.
25. Cheng F, Liu C, Jiang J, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol. 2012;8(5):1002503. https://doi.org/10.1371/journal.pcbi.1002503.
26. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. Mol BioSyst. 2012;8(7):1970. https://doi.org/10.1039/c2mb00002d.
27. Ba-alawi W, Soufan O, Essack M, Kalnis P, Bajic VB. Daspfind: new efficient method to predict drug-target interactions. J Cheminform. 2016;8(1):15. https://doi.org/10.1186/s13321-016-0128-4.
28. Wan F, Hong L, Xiao A, Jiang T, Zeng J. Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. Bioinformatics. 2019;35(1):104–11. https://doi.org/10.1093/bioinformatics/bty543.
29. Zeng X, Zhu S, Lu W, et al. Target identification among known drugs by deep learning from heterogeneous networks. Chem Sci. 2020;11(7):1775–97. https://doi.org/10.1039/C9SC04336E.
30. Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug-target interactions based on graph convolutional network and deep neural network. Brief Bioinform. 2021;22(2):2141–50. https://doi.org/10.1093/bib/bbaa044.
31. Peng J, Wang Y, Guan J, et al. An end-to-end heterogeneous graph representation learning-based framework for drug-target interaction prediction. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbaa430.
32. Zhou D, Xu Z, Li W, Xie X, Peng S. Multidti: drug-target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network Przytycka T, ed. Bioinformatics. 2021;37(23):4485–92. https://doi.org/10.1093/bioinformatics/btab473.
33. Li Y, Qiao G, Wang K, Wang G. Drug-target interaction predication via multi-channel graph neural networks. Brief Bioinform. 2021;23(1):346. https://doi.org/10.1093/bib/bbab346 (https://academic.oup.com/bib/article-pdf/23/1/bbab346/42258647/bbab346.pdf).
34. Qiao ZE, Wang G, Gsl-Dti LY. Graph structure learning network for drug-target interaction prediction. Methods. 2024;223:136–45. https://doi.org/10.1016/j.ymeth.2024.01.018.
35. Zhang R, Wang Z, Wang X, Meng Z, Cui W. Mhtan-dti: Metapath-based hierarchical transformer and attention network for drug-target interaction prediction. Brief Bioinform. 2023. https://doi.org/10.1093/bib/bbad079.
36. Rafiei F, Zeraati H, Abbasi K, Razzaghi P, Ghasemi JB, Parsaeian M, Masoudi-Nejad A. Cfssynergy: combining feature-based and similarity-based methods for drug synergy prediction. J Chem Inf Model. 2024;64(7):2577–85. https://doi.org/10.1021/acs.jcim.3c01486. (**PMID: 38514966**).
37. Gharizadeh A, Abbasi K, Ghareyazi A, Mofrad MRK, Rabiee HR. HGTDR: advancing drug repurposing with heterogeneous Graph Transf 2024. https://arxiv.org/abs/2405.08031
38. Mackiewicz A, Ratajczak W. Principal components analysis (PCA). Comput Geosci. 1993;19(3):303–42. https://doi.org/10.1016/0098-3004(93)90090-R.
39. Qi Y. Random forest for bioinformatics. In: Zhang C, Ma Y, editors. Ensemble machine learning. New York: Springer; 2012. p. 307–23. https://doi.org/10.1007/978-1-4419-9326-7_11.
40. Rigatti SJ. Random forest. J Insurance Med. 2017;47(1):31–9. https://doi.org/10.17849/insm-47-01-31-39.1.
41. Peng J, Li J, Shang X. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. BMC Bioinform. 2020;21(S13):394. https://doi.org/10.1186/s12859-020-03677-1.
42. Tangmanussukum P, Kawichai T, Suratanee A, Plaimas K. Heterogeneous network propagation with forward similarity integration to enhance drug-target association prediction. PeerJ Comput Sci. 2022;8:1124. https://doi.org/10.7717/peerj-cs.1124.

43. Wishart DS, Feunang YD, Guo AC, et al. Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res. 2018;46(D1):1074–82. https://doi.org/10.1093/nar/gkx1037.
44. Keshava Prasad TS, Goel R, Kandasamy K, et al. Human protein reference database-2009 update. Nucleic Acids Res. 2009;37(Database):767–72. https://doi.org/10.1093/nar/gkn892.
45. Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics. 2011;27(21):3036–43. https://doi.org/10.1093/bioinformatics/btr500 (https://academic.oup.com/bioinformatics/article-pdf/27/21/3036/48861183/bioinformatics_27_21_3036.pdf).
46. Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network. In: The World Wide Web conference. ACM, 2019; pp. 2022–2032. https://doi.org/10.1145/3308558.3313562
47. Li M, Cai X, Li L, Xu S, Ji H. Heterogeneous graph attention network for drug-target interaction prediction, 2022; pp. 1166–1176. https://doi.org/10.1145/3511808.3557346
48. Yang B, Chen H. Predicting circRNA-drug sensitivity associations by learning multimodal networks using graph auto-encoders and attention mechanism. Briefings Bioinf. 2023. https://doi.org/10.1093/bib/bbac596.
49. Deng L, Liu Z, Qian Y, Zhang J. Predicting circRNA-drug sensitivity associations via graph attention auto-encoder. BMC Bioinform. 2022;23(1):160. https://doi.org/10.1186/s12859-022-04694-y.
50. Lou Z, Cheng Z, Li H, Teng Z, Liu Y, Tian Z. Predicting miRNA-disease associations via learning multimodal networks and fusing mixed neighborhood information. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbac159.
51. Wang YT, Wu QW, Gao Z, Ni JC, Zheng CH. Mirna-disease association prediction via hypergraph learning based on high-dimensionality features. BMC Med Inform Decis Mak. 2021;21(S1):133. https://doi.org/10.1186/s12911-020-01320-w.
52. Liang X, Guo M, Jiang L, Fu Y, Zhang P, Chen Y. Predicting miRNA-disease associations by combining graph and hypergraph convolutional network. Interdiscip Sci Comput Life Sci. 2024. https://doi.org/10.1007/s12539-023-00599-3.
53. Song W, Xu L, Han C, Tian Z, Zou Q. Drug-target interaction predictions with multi-view similarity network fusion strategy and deep interactive attention mechanism. Bioinformatics. 2024;40(6):346. https://doi.org/10.1093/bioinformatics/btae346 (https://academic.oup.com/bioinformatics/article-pdf/40/6/btae346/58186433/btae346_supplementary_data.pdf).
54. Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. Distill. 2016;1(10):10. https://doi.org/10.23915/distill.00002.
55. Saul LK, Labs T, Ave P, Park F, Roweis ST. An introduction to locally linear embedding
56. Hyvarinen A, Oja E. A fast fixed-point algorithm for independent component analysis. Neural Comput. 1997;9(7):1483–92. https://doi.org/10.1162/neco.1997.9.7.1483.

Publisher's Note