Contents lists available at ScienceDirect

Expert Systems With Applications

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

A pseudo-label supervised graph fusion attention network for drug–target interaction prediction

Yining Xie [a,*], Xiaodong Wang [b], Pengda Wang [c], Xueyan Bi [d]

[a] College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin, 150040, China
[b] College of Computer and Control Engineering, Northeast Forestry University, Harbin, 150040, China
[c] University of Science and Technology of China, Hefei, 230026, China
[d] Heilongjiang Institute for Drug Control, Harbin, 150090, China

ARTICLE INFO

Keywords:
Drug–target interaction prediction
Pseudo-label supervision
Feature representation learning
Graph convolutional network

ABSTRACT

Drug–target interaction (DTI) prediction can reveal new drug targets and assist in drug repositioning. It can also help identify the most potential candidate drugs for specific targets, advancing new drug discovery. Graph Convolutional Networks (GCNs) have been employed to explore potential relationships between drug–target pairs (DTPs) due to their strong learning capabilities. However, existing methods primarily rely on static graphs constructed from topological structures. These graphs may contain missing or meaningless edges, limiting the ability of GCNs to capture node embeddings. Furthermore, the lack of labels in practical DTI prediction presents a significant challenge. To address these issues, this paper introduces a pseudo-label supervised graph fusion attention network for DTI prediction (PSF-DTI). Specifically, we establish a far-neighbor graph to capture robust differential information between DTPs, compensating for the limitations of traditional topology graphs. Additionally, we create an adaptive graph to dynamically update edge information for more accurate graph structures. During training, we assign pseudo-labels to unlabeled data based on feature similarities between DTPs, mitigating the impact of label scarcity. Comparative experiments with seven state-of-the-art algorithms on public datasets demonstrate the superior performance of PSF-DTI. Extensive ablation experiments validate the effectiveness of the proposed approaches. Our findings suggest that PSF-DTI offers significant advantages in DTI prediction, providing innovative methods and perspectives for future drug discovery and repositioning.

1. Introduction

Drug development is a vital part of modern medical research. However, this process is very complex and time-consuming, often requiring more than a decade of effort to successfully launch a new drug (Ding, Tang, & Guo, 2020). Traditional computational simulation and biochemical experimental methods often require a lot of time and resources (Shi et al., 2023). The prediction of drug–target relationships is the key to drug development. With the accumulation of biological data and the development of computational methods, many methods for predicting drug–target relationships have emerged. These methods can reveal the interactions between ligands and specific targets, thereby accelerating the drug discovery and design process and reducing research and development costs (Mostafa, Alhossary, Salem, & Mohamed, 2022). Therefore, the need for more accurate and efficient predictions of drug–target relationships is particularly urgent.

There are two main categories of the relationship between drugs and targets: drug–target affinity (DTA) and DTI (Zhang, Hu et al., 2023). DTA prediction quantifies the binding strength or affinity between drugs and targets through machine learning or deep learning models. For example, Öztürk, Özgür, and Ozkirimli (2018) only used the sequence information of proteins and drugs, extracting features through CNNs to predict DTA. Zhu, Yao, Qi et al. (2023) proposed a DTA prediction method (MT-DTA) that combined variational autoencoders and attention mechanisms to enhance feature expression through interactive learning and autoencoding. Zhu, Yao, Zheng et al. (2023) combined Transformer and diffusion models to construct TD-GraphDTA, improving prediction accuracy and model interpretability through multi-scale information interaction and graph optimization. However, DTA prediction has high requirements for the sequence information of the data. When it is difficult to obtain sequence information or the data is limited, its prediction accuracy may be restricted.

DTI prediction focuses on predicting whether there is an interaction between a specific drug and a specific target, and can comprehensively

* Corresponding author.
E-mail addresses: yiningxie@nefu.edu.cn (Y. Xie), ahrisdream@nefu.edu.cn (X. Wang), wangpengda@cetccloud.com (P. Wang), yewubu@hljidc.org.cn (X. Bi).

https://doi.org/10.1016/j.eswa.2024.125264
Received 14 May 2024; Received in revised form 4 July 2024; Accepted 27 August 2024
Available online 30 August 2024

consider more interaction network information, providing higher accuracy (Nikraftar & Keyvanpour, 2023). Additionally, DTI prediction also provides important support for drug repositioning and prediction of drug side effects (He et al., 2024). Currently, DTI prediction methods are mainly divided into the following three categories: ligand-based methods, docking-based methods, and machine learning-based methods (Mei, Kwoh, Yang, Li, & Zheng, 2013).

Ligand-based methods quantitatively group and associate proteins based on the chemical similarity of their ligands. One common approach is using existing active small molecule structures to infer potential interactions. González-Díaz et al. (2011) introduced a multi-target quantitative structure–activity relationship classifier for comprehensive DTI predictions. However, when only a few ligands bind to the target, this method is difficult to capture enough information.

Docking-based methods are widely used, and their core idea is to predict DTI by simulating and analyzing the binding process between them. These methods often use molecular docking techniques, such as molecular dynamics simulations, to evaluate the binding mode between a drug and its target (Morris et al., 2009). This approach requires the three-dimensional structure of the molecule for docking simulations (Li, An, & Jones, 2011). Shaikh, Sharma, and Garg (2016) successfully predicted and validated DTI for anticancer drugs using enhanced proteochemometric (PCM) modeling and molecular docking. However, if the protein's spatial structure is unknown or unavailable, this approach cannot be utilized (Xuan, Fan, Cui, Zhang, & Nakaguchi, 2022).

In recent years, machine learning has become a powerful approach for predicting DTI, including both traditional and deep learning methods. Traditional methods often assume that drugs with similar properties interact with similar targets, and vice versa. Perlman, Gottlieb, Atias, Ruppin, and Sharan (2011) introduced SITAR, which successfully predicts DTI by combining multiple drug–drug and gene–gene similarity measures and adopting a new scoring system. He, Heidemeyer, Ban, Cherkasov, and Ester (2017) proposed SimBoost to predict drug–target binding affinity, and proposed SimBoostQuant to evaluate the confidence of affinity by calculating the prediction interval. Additionally, eigenvector-based methods, such as the approach by Fu et al. (2016) that utilized meta paths and random forest models, were used for DTI prediction. However, traditional machine learning methods may struggle to fully capture the feature information of drugs and targets, limiting their ability to capture deep interactions.

Deep learning is increasingly popular for DTI prediction. Wen et al. (2017) used unsupervised pre-training to extract drug and target representations and constructed a deep learning model called DeepDTIs. Wang et al. (2018) applied a stacked autoencoder to analyze raw data from drug structures and protein sequences. Peng, Li, and Shang (2020) introduced DTI-CNN, which extracts key features from similarity networks using a denoising autoencoder to reduce dimensionality and identify essential features. Wang, You et al. (2020) used sparse principal component analysis (SPCA) to compress features by associating evolutionary protein features with drug substructure fingerprints. Zhang, Wei, Che, and Jin (2022) combined a Transformer with CNNs to capture drug molecular structure information. Chen et al. (2021) used the tree based XGBoost model for feature extraction and SMOTE for handling data imbalance. However, previous deep learning algorithms lack the extraction of topological structure information of DTI networks. GCN is widely applied in bioinformatics due to its excellent capability in processing graph data (Du, Yao, Tang, Zhao, & Gou, 2024). Therefore, it can effectively solve this problem. Zhao, Hu, Valsdottir, Zang, and Peng (2021) built a drug–target pair (DTP) network and used GCN to learn DTP features. Li, Wang, Lv, Zhang, and Wang (2021) developed the IMCHGAN model, incorporating a two-level neural attention mechanism to learn potential drug and target features, using inductive matrix completion. Cheng, Yan, Wu, and Wang (2021) combined graph attention networks with a multi-head self-attention mechanism to capture context in amino acid sequences. Li, Qiao, Gao, and Wang (2022) proposed the SGCL-DTI method, which uses supervised contrastive learning to group drug–target nodes with similar labels in embedded space. Wu, Gao, Zeng, Zhang, and Li (2022) introduced virtual nodes in BridgeDPI to create a learnable DTP association network for information transfer. Zhang, Wang, Wang, Meng, and Cui (2023) utilized meta-paths, hierarchical transformers, and graph attention networks to enhance sensitivity to complex topology and relationships between multi-node types. Li, Cai, Xu, and Ji (2023) developed MHGNN, a model that uses meta path aggregation to model high-order relationships and builds a DTP correlation graph based on node feature similarity.

However, existing GNN-based DTI prediction methods still face many challenges (Li, Wu et al., 2023). DTP networks typically have complex topological structures and integrate multi-source heterogeneous data. Current GNN-based methods assume that topology-based static graphs are complete and accurate, which may not hold true in real-world tasks (Wang et al., 2021). Therefore, we construct a far-neighbor graph and a dynamic adaptive graph to address the limitations of the topology graph. We use an attention mechanism to integrate features learned from three different graph representation methods, fully exploring the potential associative information between DTPs. Additionally, in DTI prediction tasks, the data often have highly complex structures and limited labeled data in practical applications, making it difficult for traditional supervised learning methods to achieve satisfactory performance (Xue, Li, Xie, & Wang, 2018). To address this, we propose a pseudo-label supervision strategy that uses unlabeled data to assist model training, thereby enhancing the model's robustness.

The main contributions of this paper are as follows:

- We introduce an innovative graph representation method for the DTP network that leverages the distance between node features to construct a far-neighbor graph. This approach connects distant nodes, effectively capturing their differentiation and enriching the information obtained from the topology graph for a more comprehensive representation of the DTP network.
- We design an adaptive graph with an optimizable adjacency matrix in PSF-DTI. This dynamic graph structure overcomes the limitations of static graphs in classification tasks. Additionally, we integrate an attention mechanism to fuse features from the three distinct graph representations, yielding more precise DTP feature information.
- We first introduce the pseudo-label assignment mechanism into the DTI prediction task and propose a unique pseudo-label assignment method. Using unlabeled data to assist model training significantly enhances the model's generalization ability when facing scarce DTI labeled data.
- The graph representation method and model structure of PSF-DTI is a general method that can be easily extended to other practical applications.

2. Method

In this section, we will introduce our model PSF-DTI in detail. It is mainly divided into three parts: construction of DTP network graph representation, graph fusion attention network, and pseudo-label supervision strategy. Fig. 1 illustrates the complete process of PSF-DTI.

2.1. Construction of graph representation for DTP

To leverage GCN to capture the deep and comprehensive relationship between DTPs, we first need to establish a graph representation of the DTP network. The topology graph has been well studied in the area of data mining for structural information capturing (Wang, Zhu et al., 2020). To this end, we build a topology graph to capture the structural information of the DTP network. However, the topology graph only aggregates topology neighbor information, which is difficult to fully mine the feature relationship information of the whole DTP network. To address this limitation, we create a far-neighbor graph that connects distant nodes to complement the topology graph. We focus on two key aspects: node feature representation learning and adjacency matrix construction.

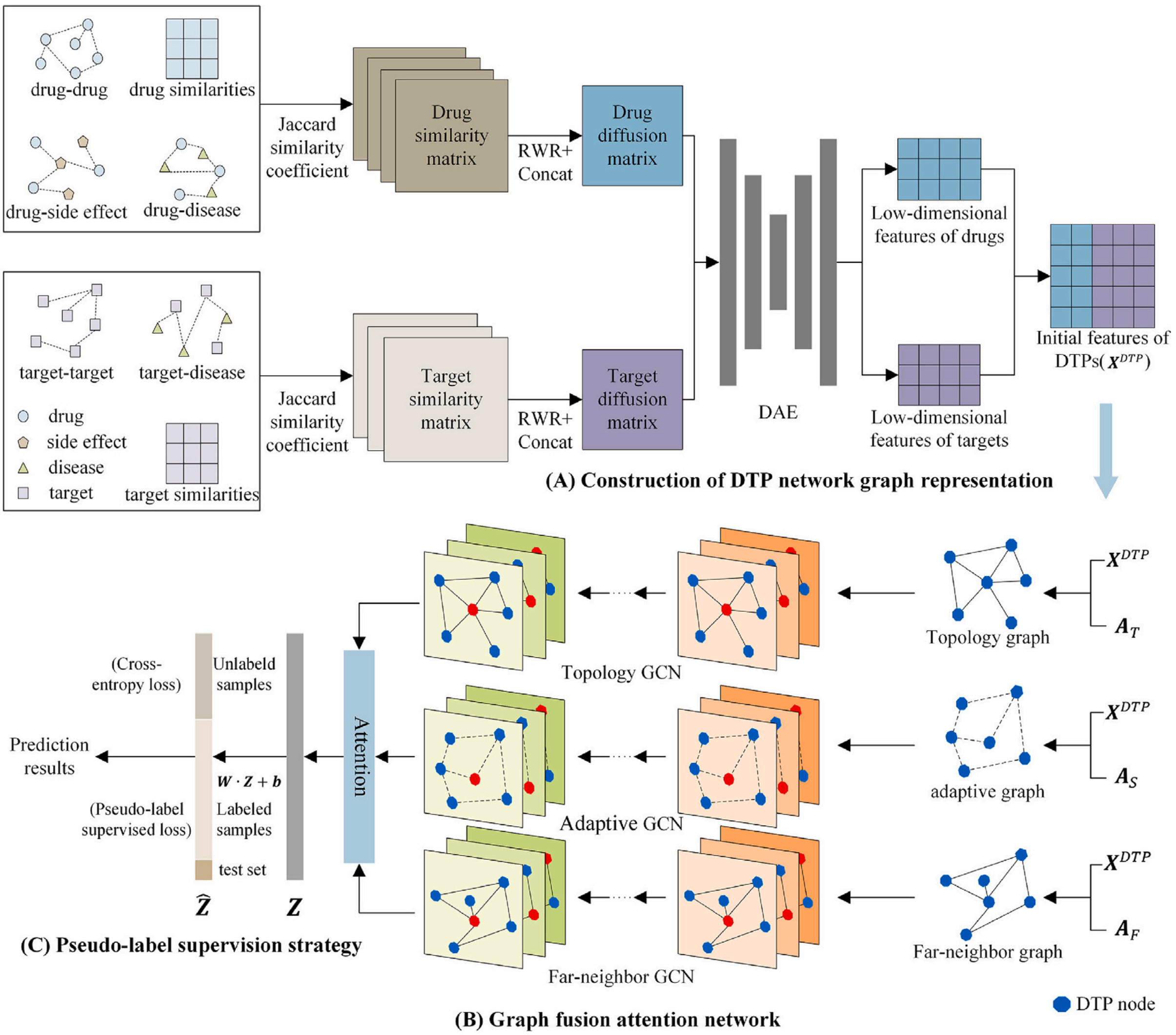

Fig. 1. Overview of the PSF-DTI process. (A) The random walk with restart (RWR) method and denoising autoencoder (DAE) are used to extract low-dimensional features from the similarity matrices of drugs and targets. By combining these features, initial representations $\mathbf{X}^{\mathrm{DTP}}$ are generated, thereby constructing the topology graph, far-neighbor graph, and adaptive graph of the DTP network. (B) Three GCNs are used to process topology graph, far-neighbor graph, and adaptive graph respectively, and attention mechanism is used to integrate information to obtain the final representation $\mathbf{Z}$ of the DTP node. (C) Pseudo-labels are assigned to unlabeled DTP nodes based on feature similarities between DTPs. The loss function considers both the loss of labeled samples and the loss from pseudo-label supervision.

2.1.1. *Node feature representation learning*

First, we collect multi-source information from multiple heterogeneous networks, including four drug-related networks (drug–drug interaction, drug–disease relation, drug-side effect relation, and drug similarity based on chemical structure) and three target-related networks (target–target interaction, target–disease relation, and target similarity based on primary protein sequences). We calculate the similarity matrices for drugs and targets using Jaccard similarity. The drug-related networks result in matrices $\mathbf{H}_{\text{drug-drug}}$, $\mathbf{H}_{\text{drug-disease}}$, $\mathbf{H}_{\text{drug-side effect}}$, $\mathbf{H}_{\text{drug similarity}}$ while the target-related networks result in matrices $\mathbf{H}_{\text{target-target}}$, $\mathbf{H}_{\text{target-disease}}$, $\mathbf{H}_{\text{target similarity}}$. For instance, the target-disease relation network is represented by $\mathbf{M}_{\text{target-disease}} \in \mathbb{R}^{N_{\text{target}} \times N_{\text{disease}}}$, where N_{target} and N_{disease} denote the numbers of targets and diseases, respectively. Each element $m_{i,j}$ represents the association between target i and disease j. Therefore, each row can be regarded as the eigenvector of each target. The calculation method of elements $h_{i,j}$ of matrix $\mathbf{H}_{\text{target-disease}}$ is as follows:

$$h_{i,j} = \frac{\left|\mathbf{M}_i \cap \mathbf{M}_j\right|}{\left|\mathbf{M}_i \cup \mathbf{M}_j\right|} \tag{1}$$

where $\mathbf{M}_i$ denotes the set of diseases associated with target i in $\mathbf{M}_{\text{target-disease}}$, and $\mathbf{M}_j$ is similar. $\left|\mathbf{M}_i \cap \mathbf{M}_j\right|$ represents the intersection of the two sets, $\left|\mathbf{M}_i \cup \mathbf{M}_j\right|$ represents their union.

For each drug similarity matrix, we use RWR to obtain diffusion state and capture global network information. After acquiring all diffusion state matrices, we concatenate them to form the overall diffusion state matrix $\mathbf{V}_{\mathrm{d}}$ for drugs. The same approach is used to obtain the overall diffusion state matrix $\mathbf{V}_{\mathrm{t}}$ for targets.

The overall diffusion matrices of drugs and targets are usually high-dimensional and noisy. High-dimensional features increase computational complexity, potential overfitting, and hinder effective GCN feature aggregation. Additionally, noise in the data further obscures the true underlying relationships and negatively affects the performance of downstream tasks. To address these issues, we refer to Peng et al. (2020) and apply a denoising autoencoder (DAE) to reduce feature dimensionality and denoise the data, extracting the most critical features from the original input and obtaining a more robust representation. Fig. 2 shows the overall structure of the DAE. We add noise to the original input $\mathbf{V}_{\mathrm{d}}$ and $\mathbf{V}_{\mathrm{t}}$, then input them into the DAE. The DAE restores the noisy input data to the original form by learning the abstract, low-dimensional features $\mathbf{X}_{\mathrm{d}}$ and $\mathbf{X}_{\mathrm{t}}$. Finally, we combine the

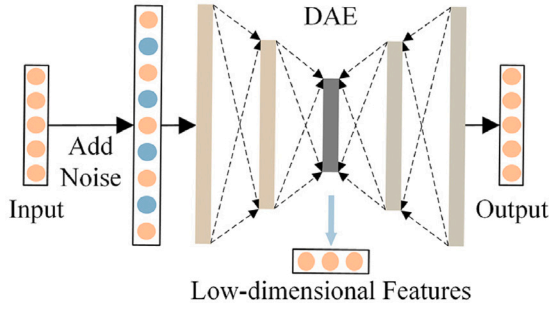

Fig. 2. DAE Structure Diagram. Including the process of adding noise to the input data, the encoding and decoding steps within the DAE, and the extraction of low-dimensional features.

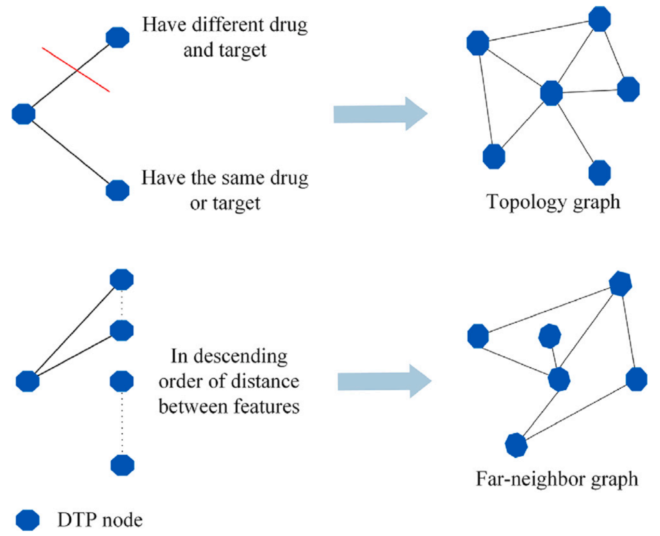
Have different drug and target
Have the same drug or target
Topology graph
In descending order of distance between features
Far-neighbor graph
DTP node

Fig. 3. Construction of adjacency matrices for topology graph and far-neighbor graph.

low-dimensional features $\mathbf{X}_d$ and $\mathbf{X}_t$ of the drug and target to obtain the initial feature representation $\mathbf{X}^{\text{DTP}}$ of the DTP node.

2.1.2. *Adjacency matrix construction*

As shown in Fig. 3, we construct the adjacency matrix of the topology graph according to the following rule: if two DTP nodes contain the same drug or target, they are considered neighbors. Let $\mathbf{A}_\text{T} \in \mathbb{R}^{N_\text{DTP} \times N_\text{DTP}}$ be the adjacency matrix of the topology graph, where N_DTP represents the number of DTP nodes. When the element $a_{i,j}^{T} = 1$ in the adjacency matrix $\mathbf{A}_\text{T}$, it indicates that the DTP nodes represented by the ith row and jth column are adjacent, otherwise $a_{i,j}^{\text{T}} = 0$.

The rule for constructing the adjacency matrix of the far-neighbor graph is as follows: For each DTP node, calculate the distance between its features and those of other nodes. Then, select the top K DTP nodes with the largest distances as its neighbors. In this paper, we choose the Manhattan distance, and compare it with other distance metrics that are discussed in the ablation experiments. Let $\mathbf{A}_\text{F} \in \mathbb{R}^{N_\text{DTP} \times N_\text{DTP}}$ be the adjacency matrix of the far-neighbor graph. If the DTP node represented by the ith row is a neighbor of the DTP node represented by the jth column, then element $a_{i,j}^{\text{F}} = 1$, otherwise it is 0.

To sum up, we can construct the topology graph $\text{G}_\text{t} = (\mathbf{A}_\text{T}, \mathbf{X}^{\text{DTP}})$ and far-neighbor graph $\text{G}_\text{f} = (\mathbf{A}_\text{F}, \mathbf{X}^{\text{DTP}})$ of the DTP network.

2.2. *Graph fusion attention network*

The topology graph and far-neighbor graph of the DTP network are both static graphs, which may contain missing or meaningless edges and cannot be adjusted or updated based on downstream information. To address this limitation, we also construct an adaptive graph. By treating the adjacency matrix as a parameter and optimizing it through backpropagation, the graph structure can be dynamically adjusted. Three separate GCNs are employed to extract features of DTP nodes, and an attention mechanism is used to fuse these features.

2.2.1. *Convolution module of topology graph and far-neighbor graph*

Given the topology graph $\text{G}_\text{t} = (\mathbf{A}_\text{T}, \mathbf{X}^{\text{DTP}})$ as input, GCN can effectively capture the information transmission between nodes and their neighbors according to the adjacency matrix $\mathbf{A}_\text{T}$, thereby updating the initial feature $\mathbf{X}^{\text{DTP}}$ of nodes. In the lth layer of the GCN, the output is represented as:

$$\mathbf{Z}_\text{T}^{(l)} = \sigma(\tilde{\mathbf{D}}_\text{T}^{-\frac{1}{2}} \tilde{\mathbf{A}}_\text{T} \tilde{\mathbf{D}}_\text{T}^{-\frac{1}{2}} \mathbf{Z}_\text{T}^{(l-1)} \mathbf{W}_\text{T}^{(l)}) \tag{2}$$

where $\mathbf{Z}_\text{T}^{(0)} = \mathbf{X}^{\text{DTP}}$, $\mathbf{Z}_\text{T}^{(l)}$ represents the embedding of DTP nodes in the lth layer. $\tilde{\mathbf{D}}_\text{T}^{-\frac{1}{2}} \tilde{\mathbf{A}}_\text{T} \tilde{\mathbf{D}}_\text{T}^{-\frac{1}{2}}$ is used for normalization, where $\tilde{\mathbf{A}}_\text{T}$ is the adjacency matrix plus identity matrix, and $\tilde{\mathbf{D}}_\text{T}$ is the diagonal degree matrix of $\tilde{\mathbf{A}}_\text{T}$. $\mathbf{W}_\text{T}^{(l)}$ is the shared weight matrix of layer l, σ indicates the activation function. The final layer's output, $\mathbf{Z}_\text{T}$, is the topology graph representation matrix of DTP nodes.

Similarly, for the far-neighbor graph $\text{G}_\text{f} = (\mathbf{A}_\text{F}, \mathbf{X}^{\text{DTP}})$ as input, the output in the lth layer of the GCN is represented as:

$$\mathbf{Z}_\text{F}^{(l)} = \sigma(\tilde{\mathbf{D}}_\text{F}^{-\frac{1}{2}} \tilde{\mathbf{A}}_\text{F} \tilde{\mathbf{D}}_\text{F}^{-\frac{1}{2}} \mathbf{Z}_\text{F}^{(l-1)} \mathbf{W}_\text{F}^{(l)}) \tag{3}$$

The output of the final layer is represented as $\mathbf{Z}_\text{F}$. Through these two GCN models, the representation matrices $\mathbf{Z}_\text{T}$ and $\mathbf{Z}_\text{F}$ of DTP nodes in the topology graph and far-neighbor graph can be obtained.

2.2.2. *Convolution module of adaptive graph*

The topology and far-neighbor graphs' adjacency matrices may not fully capture the actual information (Jin et al., 2020). Thus, combining $\mathbf{Z}_\text{T}$ and $\mathbf{Z}_\text{F}$ may not be used as the final representation of DTP nodes. We construct an adaptive graph $\text{G}_\text{s} = (\mathbf{A}_\text{S}, \mathbf{X}^{\text{DTP}})$ for the DTP network, where $\mathbf{A}_\text{S}$ is an adaptive adjacency matrix. Unlike the topology graph or far-neighbor graph, the elements of the adaptive graph adjacency matrix are not binary values (0 or 1), but real numbers. During the backpropagation process, the values in the adjacency matrix can be updated by calculating the gradient of each element. These values represent the strength or weight of the connection between nodes, thereby capturing more complex relationships between nodes, rather than just the presence or absence of a connection (0 or 1). Therefore, the adaptive adjacency matrix can better reflect the connection between nodes, thereby improving the representation ability and performance of the model. In order to ensure that the adaptive graph can fully represent the DTP network information, we initialize the adaptive adjacency matrix as follows:

$$\mathbf{A}_\text{S} = \frac{\mathbf{A}_\text{T} + \mathbf{A}_\text{F}}{2} \tag{4}$$

Similarly, when the adaptive graph $\text{G}_\text{s} = (\mathbf{A}_\text{S}, \mathbf{X}^{\text{DTP}})$ is input into the adaptive GCN, the output in the lth layer is expressed as:

$$\mathbf{Z}_\text{S}^{(l)} = \sigma(\tilde{\mathbf{D}}_\text{S}^{-\frac{1}{2}} \tilde{\mathbf{A}}_\text{S} \tilde{\mathbf{D}}_\text{S}^{-\frac{1}{2}} \mathbf{Z}_\text{S}^{(l-1)} \mathbf{W}_\text{S}^{(l)}) \tag{5}$$

The final layer's output, $\mathbf{Z}_\text{S}$, is the adaptive graph representation matrix of DTP nodes.

2.2.3. *Feature fusion based on attention mechanism*

The three GCNs yield three representation matrices for DTP nodes: $\mathbf{Z}_\text{T}$ from the topology graph, $\mathbf{Z}_\text{F}$ from the far-neighbor graph, and $\mathbf{Z}_\text{S}$ from the adaptive graph. To achieve the optimal combination, we use an attention mechanism to learn the relative importance of each representation. The attention vector $\boldsymbol{\alpha}$ is then applied for feature fusion. The process is outlined as follows:

$$\boldsymbol{\alpha}(\boldsymbol{\alpha}_\text{t}, \boldsymbol{\alpha}_\text{f}, \boldsymbol{\alpha}_\text{s}) = \text{Att}(\mathbf{Z}_\text{T}, \mathbf{Z}_\text{F}, \mathbf{Z}_\text{S}) \tag{6}$$

where $\boldsymbol{\alpha}_\mathrm{t}, \boldsymbol{\alpha}_\mathrm{f}, \boldsymbol{\alpha}_\mathrm{s} \in \mathbb{R}^{N^{\mathrm{DTP}} \times 1}$ represent the attention scores of $\mathbf{Z}_\mathrm{T}$, $\mathbf{Z}_\mathrm{F}$, and $\mathbf{Z}_\mathrm{S}$, respectively. N^{DTP} represents the number of DTP nodes, and $\mathrm{Att}(\cdot)$ represents the attention function.

Take node i as an example to introduce the function implementation of $\mathrm{Att}(\cdot)$. Assume that the representation of node i in $\mathbf{Z}_\mathrm{T}$ is $\mathbf{z}_i^\mathrm{T} \in \mathbb{R}^{1\times h}$. By applying a nonlinear transformation and a shared attention vector $q \in \mathbb{R}^{h' \times 1}$, we calculate the attention value ω_i^T as follows:

$$\omega_i^\mathrm{T} = \mathbf{q}^T \cdot \tanh(\mathbf{W} \cdot (\mathbf{z}_i^\mathrm{T})^T + \mathbf{b}) \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{h' \times h}$ is the weight matrix, $\mathbf{b} \in \mathbb{R}^{h' \times 1}$ is the bias vector, h' is the attention layer dimension, and h is the feature dimension of the DTP node in the three representation matrices. Similarly, we compute the attention values ω_i^F and ω_i^S for node i in the representation matrices $\mathbf{Z}_\mathrm{F}$ and $\mathbf{Z}_\mathrm{S}$. Next, we use the SoftMax function to normalize the attention values ω_i^T, ω_i^F and ω_i^S to obtain the final weights α_i^T, α_i^F, and α_i^S:

$$\alpha_i^\mathrm{T} = \mathrm{SoftMax}(\omega_i^\mathrm{T}) = \frac{\exp(\omega_i^\mathrm{T})}{\exp(\omega_i^\mathrm{T}) + \exp(\omega_i^\mathrm{F}) + \exp(\omega_i^\mathrm{S})} \tag{8}$$

Similar, we calculate α_i^F and α_i^S. For all N^{DTP} nodes, we obtain the learned weights, $\boldsymbol{\alpha}_\mathrm{t} = [\alpha_i^\mathrm{T}], \boldsymbol{\alpha}_\mathrm{f} = [\alpha_i^\mathrm{F}], \boldsymbol{\alpha}_\mathrm{s} = [\alpha_i^\mathrm{S}] \in \mathbb{R}^{N^{\mathrm{DTP}} \times 1}$, and define the diagonal matrices $\boldsymbol{\alpha}_\mathrm{T} = \mathrm{diag}(\boldsymbol{\alpha}_\mathrm{t})$, $\boldsymbol{\alpha}_\mathrm{F} = \mathrm{diag}(\boldsymbol{\alpha}_\mathrm{f})$, and $\boldsymbol{\alpha}_\mathrm{S} = \mathrm{diag}(\boldsymbol{\alpha}_\mathrm{s})$. Finally, the final representation $\mathbf{Z}$ of the DTP node is obtained by weighting the three representation matrices:

$$\mathbf{Z} = \boldsymbol{\alpha}_\mathrm{T} \cdot \mathbf{Z}_\mathrm{T} + \boldsymbol{\alpha}_\mathrm{F} \cdot \mathbf{Z}_\mathrm{F} + \boldsymbol{\alpha}_\mathrm{S} \cdot \mathbf{Z}_\mathrm{S} \tag{9}$$

Finally, final representation $\mathbf{Z}$ is used for node classification by inputting it into a linear transformation layer and a SoftMax activation function:

$$\hat{\mathbf{Z}} = \mathbf{W} \cdot \mathbf{Z} + \mathbf{b} \tag{10}$$

$$\mathbf{Z}' = \mathrm{SoftMax}(\hat{\mathbf{Z}}) \tag{11}$$

where $\mathbf{W}$ is the weight matrix and $\mathbf{b}$ is the bias vector. The linear transformation result is converted into a probability distribution using the SoftMax function.

2.3. Pseudo-label supervision strategy

The pseudo-label strategy is a common approach for mitigating the scarcity of labeled data in semi-supervised classification tasks. In typical graph network node classification problems, pseudo-labels are usually generated based on topology graphs, neglecting valuable information about node features (Yang et al., 2023). Since the features of nodes in the same category are usually similar, we allocate pseudo-labels to unlabeled data based on the cosine similarity. Let $\mathbf{M} \in \mathbb{R}^{N\times N}$ represents the cosine similarity matrix, with element $m_{i,j}$ calculated as follows:

$$m_{i,j} = \frac{\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle}{\|\hat{\mathbf{z}}_i\| \cdot \|\hat{\mathbf{z}}_j\|} \tag{12}$$

where $\hat{\mathbf{z}}_i$, $\hat{\mathbf{z}}_j$ represent the final representation of node i and node j respectively, $\langle\cdot\rangle$ represents the inner product of two vectors, and $\|\cdot\|$ represents the norm of the vector. Then $m_{i,j}$ can represent the similarity between node i and node j. Let $\mathrm{V_N}$ unlabeled training set and $\mathrm{V_L}$ represents the labeled training set. Given an unlabeled sample $\mathrm{v}_i \in \mathrm{V_N}$, compute $m_{i,j}$ between v_i and each $\mathrm{v}_j \in \mathrm{V_L}$. The higher $m_{i,j}$, the greater the possibility that v_i and v_j have the same label. Therefore, we assign the label of sample v_j corresponding to the maximum value of $m_{i,j}$ to v_i. It is shown as follows:

$$\mathrm{y}_i = \arg\max\nolimits_j m_{i,j}, \quad \mathrm{s.t.}\ \mathrm{v}_i \in \mathrm{V_N},\ \mathrm{v}_j \in \mathrm{V_L} \tag{13}$$

To ensure the quality of pseudo-labels, we set a threshold value β to control which similarity is considered high enough for selectively assigning pseudo-labels. The index t_i is calculated as follows:

$$t_i = \begin{cases} 1, & \text{if} \max_j m_{i,j} \geq \beta \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

Algorithm 1 calculation process of PSF-DTI

```
Require: Initial feature matrix $\mathbf{X}^{\mathrm{DTP}}$, topology graph adjacency matrix $\mathbf{A}_\mathrm{T}$, far-neighbor graph adjacency matrix $\mathbf{A}_\mathrm{F}$, label matrix $\mathbf{Y}$, number of layers $L$, learning rate $\eta$, number of epochs $T$, hyper-parameter $\beta$
Ensure: Classification results $\mathbf{Z}'$
Initialize the PSF-DTI model parameters $\Theta$, $\mathbf{A}_\mathrm{S}$, $\boldsymbol{\alpha}$
for each epoch $t = 1, 2, \ldots, T$ do
    Forward Propagation:
    $\mathbf{Z}_\mathrm{T}^{(0)} = \mathbf{Z}_\mathrm{F}^{(0)} = \mathbf{Z}_\mathrm{S}^{(0)} = \mathbf{X}^{\mathrm{DTP}}$
    for each layer $l = 1, 2, \ldots, L$ do
        Calculate $\mathbf{Z}_\mathrm{T}^{(l)}$ and $\mathbf{Z}_\mathrm{F}^{(l)}$ by Eqs. (2) and (3)
        Calculate $\mathbf{Z}_\mathrm{S}^{(l)}$ by Eq. (5)
    end for
    Calculate attention vector $\boldsymbol{\alpha}$ and $\mathbf{Z}$ by Eqs. (6) and (9)
    Calculate cosine similarity matrix $\mathbf{M}$ by Eq. (12)
    Calculate the predicted result $\mathbf{Z}'$ by Eq. (11)
    Compute Loss:
    Calculate loss $\mathrm{L}_1$ and $\mathrm{L}_2$ by Eqs. (15) and (16)
    Calculate the overall loss L by Eq. (17)
    Backward Propagation:
    Update $\mathbf{A}_\mathrm{S} \leftarrow \mathbf{A}_\mathrm{S} - \eta \frac{\partial \mathrm{L}(\mathbf{X}^{\mathrm{DTP}}, \mathbf{A}_\mathrm{T}, \mathbf{A}_\mathrm{F}, \boldsymbol{\alpha}, \Theta)}{\partial \mathbf{A}_\mathrm{S}}$
    Update $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \frac{\partial \mathrm{L}(\mathbf{X}^{\mathrm{DTP}}, \mathbf{A}_\mathrm{T}, \mathbf{A}_\mathrm{F}, \mathbf{A}_\mathrm{S}, \Theta)}{\partial \boldsymbol{\alpha}}$
    Update $\Theta \leftarrow \Theta - \eta \frac{\partial \mathrm{L}(\mathbf{X}^{\mathrm{DTP}}, \mathbf{A}_\mathrm{T}, \mathbf{A}_\mathrm{F}, \mathbf{A}_\mathrm{S}, \boldsymbol{\alpha})}{\partial \Theta}$
end for
return $\mathbf{Z}'$
```

When $t_i = 1$, it indicates that the maximum value of $m_{i,j}$ can be considered sufficiently similar. Therefore, the label of the corresponding sample v_j can be assigned to v_i as a pseudo-label. Then the loss function of unlabeled samples supervised by pseudo-labels can be expressed as:

$$\min\nolimits_{\theta, \mathbf{A}_\mathrm{S}} \mathrm{L}_1 = \sum_{\mathrm{v}_i \in \mathrm{V_N}} t_i \cdot \varphi(\mathbf{z}_i', \mathbf{y}_i) \tag{15}$$

where θ represents the model parameters, $\mathbf{z}_i'$ denotes the model's predictions for the sample v_i, $\mathbf{y}_i$ denotes the pseudo-label of the sample v_i, $\varphi(\cdot)$ represents the cross-entropy loss function.

Additionally, the loss function for labeled samples is as follows:

$$\min\nolimits_{\theta, \mathbf{A}_\mathrm{S}} \mathrm{L}_2 = \sum_{\mathrm{v}_i \in \mathrm{V_L}} \varphi(\mathbf{z}_i', \mathbf{y}_i) \tag{16}$$

In summary, our model's loss function is expressed as:

$$\begin{aligned} \min\nolimits_{\theta, \mathbf{A}_\mathrm{S}} \mathrm{L} &= \mathrm{L}_1 + \mathrm{L}_2 \\ &= \sum_{\mathrm{v}_i \in \mathrm{V_N}} t_i \cdot \varphi(\mathbf{z}_i', \mathbf{y}_i) + \sum_{\mathrm{v}_i \in \mathrm{V_L}} \varphi(\mathbf{z}_i', \mathbf{y}_i) \end{aligned} \tag{17}$$

Algorithm 1 shows the calculation process of model training after obtaining the initial features of DTPs.

3. Experiments

3.1. Datasets

We build a dataset based on Luo et al. (2017). First, we obtain relevant information about drugs, including drug–drug interactions, from the DrugBank database (Knox et al., 2010). Information about target proteins, including target–target interactions, is obtained from the HPRD database (Prasad et al., 2009). Information about diseases, including drug–disease relations and target–disease relations, is obtained from the CTD database (Davis et al., 2021). Information about side effect, including drug-side effect relations, is obtained from the SIDER database (Kuhn, Letunic, Jensen, & Bork, 2016). In addition, we

Table 1
Statistics of the datasets.

Node type	Numbers	Resources	Edge type	Numbers	Resources
Drug	708	DrugBank	Drug-Drug	10 036	DrugBank
Target	1512	HPRD	Target-Target	7363	HPRD
Disease	5603	CTD	Drug-Disease	199 214	CTD
			Target-Disease	1 596 745	CTD
Side effect	4192	SIDER	Drug-Side effect	80 164	SIDER
			Drug-Target	1923	DrugBank

obtain two similarity networks based on the chemical structure of the drug and the primary sequence information of the target protein.

By amalgamating this information, we form a heterogeneous network consisting of four node types, five edge types, and two similarity networks. The network contains 12,015 nodes and 1,895,445 edges. Meanwhile, we use the known DTIs obtained from the DrugBank database as positive sample labels in the dataset and randomly select the same number of DTPs without interactions as negative samples. Table 1 summarizes the data details.

3.2. Experimental settings

3.2.1. Evaluation metric

We utilize three widely accepted evaluation metrics: Area Under the ROC Curve (AUC), Area Under the Precision–Recall Curve (AUPR), and Matthews Correlation Coefficient (MCC). AUC assesses the model's capability to differentiate between positive and negative samples across various threshold settings, while AUPR emphasizes the balance between precision and recall. MCC provides a single comprehensive indicator that considers true positives, true negatives, false positives, and false negatives, offering a holistic evaluation of the PSF-DTI model in DTI prediction tasks. The combination of these three indicators comprehensively evaluates the performance of the PSF-DTI model.

3.2.2. Implementation detail

We conduct 10-fold cross-validation on the dataset described above. The dataset is randomly divided into 10 non-overlapping subsets. In each iteration, one subset served as the test set while the remaining 9 subsets formed the training set. This process is repeated 10 times to compute the average result.

Our proposed PSF-DTI model incorporates three parallel GCNs to construct the overall framework. All GCNs have the same structure and are composed of two hidden layers, with dimensions of 256 and 64 respectively. We utilized the Adam optimizer for model optimization, setting the learning rate and weight decay to 1e-3, and applying a dropout rate of 0.2 to address overfitting. It is important to note that the threshold β in the pseudo-label assignment process should ensure that the unlabeled data can obtain sufficient high-quality pseudo-labels, and is set to 0.99 in all experiments in this paper. In the process of utilizing DAE to extract low-dimensional features, we configured the dimensions of $\mathbf{X}_d$ and $\mathbf{X}_t$ as 100 and 400, respectively.

3.2.3. Baseline

To assess our model's performance, we compared it with six classic or state-of-the-art methods:

DTINet (Luo et al., 2017) accurately predicts new drug target interactions from the constructed heterogeneous network by learning low-dimensional vector representations and vector space projection, achieving significant performance improvement.

NeoDTI (Wan, Hong, Xiao, Jiang, & Zeng, 2019) proposes a nonlinear end-to-end learning model, which integrates diverse information of heterogeneous network data and automatically learns the feature representations of drugs and targets.

IMCHGAN (Li et al., 2021) combines inductive matrix completion and heterogeneous graph attention networks to learn the latent feature representations of drugs and targets, and uses a prediction score model to calculate the optimal drug-to-target projection.

DTI-MGNN (Li, Qiao, Wang, & Wang, 2022) is based on multichannel GCNs and graph attention mechanisms, combining topological structure and semantic information to improve representation learning capabilities in DTI prediction.

SGCL-DTI (Li et al., 2022) generates contrastive losses to guide supervised optimization of the model by comparing the topological structure and semantic characteristics of the DTP network, as well as a new selection strategy for positive and negative samples.

MHGNN (Li et al., 2023) achieves DTI prediction by utilizing metapathway aggregation to capture complex structures and rich semantics in biologically heterogeneous graphs.

AMGDTI (Su et al., 2024) automatically aggregates semantic information from heterogeneous networks for DTI prediction, overcoming the limitations of manually designed meta-graphs.

3.3. Performance comparison

We divide the data in the training set into labeled data and unlabeled data. According to the proportion of unlabeled data in the total data in the training set, starting from 10%, setting a level for every 10% increase until 90%, used to verify robustness of PSF-DTI model in the face of lack of labeled data. We test the model at these nine proportions, as shown in Table 2, and compare it with the best performance of the aforementioned baseline methods, as shown in Fig. 4. When the proportion of unlabeled data is 60%, our results are similar to the best AUC and AUPR of the baseline methods. Therefore, in Table 3, we compare the performance of all baseline methods and our method at this proportion. The training set data of the baseline method are all labeled.

As can be seen from Table 2 and Fig. 4, before the proportion of unlabeled data reaches 70%, although the performance of the model has decreased, it still maintains a relatively stable level. This is because our proposed pseudo-label supervision strategy can generate reliable pseudo-labels for unlabeled data during training, indicating that our model still performs well when faced with a small amount of labeled data. After 70%, although the decrease becomes larger, which may be because the amount of labeled data is too small and the threshold β limits the allocation of pseudo-labels, AUC, AUPR and MCC still maintain a high level. When the proportion of unlabeled data is low, PSF-DTI outperforms all baseline methods in terms of AUC and AUPR. When the amount of our unlabeled data increases to 60%, our AUC and AUPR are still higher than the best method in the baseline. Regarding MCC, as evident from Tables 2 and 3, our method surpasses all methods except MHGNN, and is only slightly lower than MHGNN at 10%. In the DTI prediction task, negative samples may contain potential true DTI, which affects the model's ability to correctly identify negative samples. Since MCC is very sensitive to negative sample classification, this may cause our method to be slightly inferior to MHGNN in MCC score. Therefore, the above results show that our model performs well in terms of robustness and generalization ability.

In addition, by comparing the results of different methods in Table 3, we can draw the following conclusions: (i) DTINet only extracts features from heterogeneous networks through a computational pipeline for DTI prediction, without employing methods such as GCN for feature learning. Consequently, its performance is lower compared to other GCN-based methods. (ii) NeoDTI and IMCHGAN use GCN to independently extract features of drugs and targets but do not consider the high-order connections between DTP nodes, which limits their performance. (iii) DTI-MGNN and SGCL-DTI construct topology graphs and feature graphs to aggregate features of DTP nodes. However, the feature graphs constructed using the k-nearest neighbor method overlap significantly with the topology graphs, and both are static, leading to poor performance. Additionally, SGCL-DTI includes the label information of the test set when constructing the meta-path, so its actual performance

Table 2
Model performance under different proportions of unlabeled data.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
AUC	**0.9875**	0.9873	0.9845	0.9841	0.9817	0.9781	0.9727	0.9531	0.9267
AUPR	**0.9880**	0.9879	0.9848	0.9846	0.9838	0.9791	0.9722	0.9553	0.9319
MCC	**0.9089**	0.9030	0.8912	0.8887	0.8826	0.8648	0.8428	0.7955	0.7245

Table 3
Comparison results of the proposed models and baselines.

Methods	AUC	AUPR	MCC
DTINet	0.9058	0.9194	0.6676
NeoDTI	0.9210	0.9291	0.6985
IMCHGAN	0.9072	0.9210	0.6914
DTI-MGNN	0.9229	0.9190	0.7450
SGCL-DTI	0.8542	0.8900	0.6053
MHGNN	0.9727	0.9552	**0.9174**
AMGDTI	0.9779	0.9781	0.8131
PSF-DTI (60%)	**0.9781**	**0.9791**	0.8648

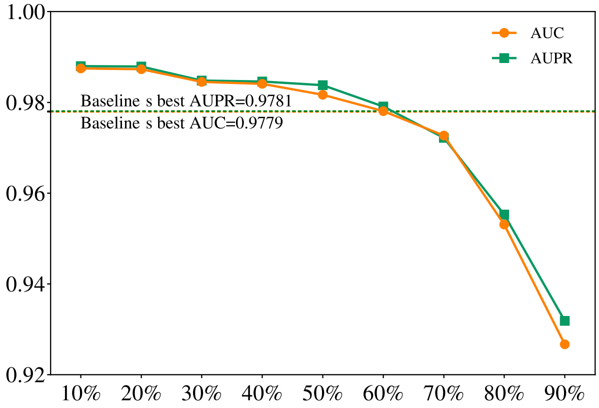

Fig. 4. Comparison of model performance under different proportions of unlabeled data. The two dashed lines represent the best AUC and AUPR in the baseline.

is the lowest. (iv) MHGNN assigns weights to the edges between DTP nodes instead of simple 0 or 1, achieving better performance and the highest MCC, but the graph structure it constructs is still static and cannot adjust the edge information. (v) AMGDTI uses adaptive metagraphs to flexibly extract features of drugs and targets but only uses the inner product of drug and protein feature representations for prediction, lacking feature aggregation between DTP nodes. (vi) Our model, PSF-DTI, constructs a far-neighbor graph and a topology graph to capture the difference information and neighbor relationships between DTP nodes, respectively, focusing more comprehensively on the overall structure of the DTP network. It compensates for the shortcomings of static graphs through a dynamic adaptive graph, thus achieving the best performance.

3.4. Ablation study

To validate the effectiveness of each component of the proposed PSF-DTI model, we perform ablation experiments on the key aspects of the model when the proportion of unlabeled data in the training set is 20%, 50%, and 80%.

3.4.1. Ablation study on graph fusion attention network

To validate the effectiveness of the proposed graph fusion attention network, we conduct ablation experiments with four variants.

- **PSF-T:** Serves as a baseline, considering only the topology graph and utilizes a single GCN for DTI prediction.
- **PSF-TS:** Combines the topology graph and the adaptive graph while excluding the far-neighbor graph.
- **PSF-TF:** Combines the topology graph and the far-neighbor graph but does not utilize the adaptive graph.
- **PSF-NA:** Aggregates by directly summing the feature matrices $\mathbf{Z}_{\mathrm{T}}$, $\mathbf{Z}_{\mathrm{F}}$, and $\mathbf{Z}_{\mathrm{S}}$ without employing the attention mechanism for fusion.

All experimental results are presented in Table 4, from which we can draw the following conclusions: (i) using only the topology graph results in the poorest performance because relying solely on the topological structure fails to explore the underlying relationships between DTP nodes. (ii) The use of the far-neighbor graph compensates for the limitations of topology graph modeling and effectively enhances model performance. (iii) Regardless of whether the far-neighbor graph is present, the adaptive graph can further improve the model's performance. (iv) The attention mechanism can effectively fuse features to enhance model performance.

The complete model achieves the best results at unlabeled data proportions of 20% and 50%. Compared to the baseline using only the topology graph, it shows improvements in AUC by 12.75% and 13.34%, AUPR by 13.67% and 13.44%, and MCC by 30.14% and 29.68%, respectively. However, when the proportion is 80%, the model without the attention mechanism performs better. This may be due to higher proportions of unlabeled data introducing more noise, making it difficult for the attention mechanism to effectively identify key information.

Therefore, we supplement the experiments with a comparison between using and not using the attention mechanism across nine proportions of unlabeled data. The experimental results are shown in Fig. 5. Only when the proportion of unlabeled data is high, the effect without attention mechanism will be better. When the proportion is low, the attention mechanism improves the performance of the model.

3.4.2. Ablation study on pseudo-label supervision strategy

To assess the effectiveness of our proposed pseudo-label supervision strategy, we conduct experiments by removing the pseudo-label supervision loss from the loss function, resulting in a variant SF-DTI. The experimental results are shown in Fig. 6. It can be seen that when the proportion of unlabeled data in the training set is 20%, the AUC, AUPR, and MCC decrease by 1.57%, 1.47%, and 7.96% respectively. When the proportion increase to 80%, the decreases are 1.63%, 1.51%, and 6.05% respectively. It shows that our method can alleviate the problem of insufficient label data in practical applications.

3.4.3. Ablation study on far-neighbor graph

In PSF-DTI, we use Manhattan distance to construct the far-neighbor graph. To verify the effectiveness of constructing the far-neighbor graph based on the farthest distance, we construct a near-neighbor graph test model based on the nearest distance. Additional experiments are conducted using Euclidean distance to construct graphs, and the results are shown in Table 5. It can be seen from the table that among the two methods of calculating distance, the effect of the far-neighbor graph is better than that of the near-neighbor graph. In addition, Manhattan distance achieves better results for both near-neighbor graphs and far-neighbor graphs. We believe one of the main reasons is the high overlap between the adjacency matrices of the near-neighbor graph and the topology graph. We use Manhattan distance to check for edges that exist in the near-neighbor graph but not in the topology graph, and the proportion is only 5.25%. In contrast, the far-neighbor graph and the topology graph exhibit almost no overlapping edges. This complementarity effectively avoids information redundancy. Compared to the near-neighbor graph, the far-neighbor graph provides a more comprehensive perspective, revealing deeper patterns and associations. When combined with the topology graph, it helps the model understand the overall structure and distribution of the DTP network, capturing more global features and thus improving overall performance. Furthermore, relying on more dispersed connections, the far-neighbor graph is less susceptible to local disturbances, thereby enhancing the model's robustness to noise and outliers.

Table 4
Results of ablation study on graph fusion attention network.

Variants	20%			50%			80%		
	AUC	AUPR	MCC	AUC	AUPR	MCC	AUC	AUPR	MCC
PSF-T	0.8598	0.8512	0.6016	0.8483	0.8494	0.5858	0.8095	0.7997	0.5367
PSF-TS	0.9707	0.9679	0.8555	0.9619	0.9604	0.8347	0.9371	0.9362	0.7639
PSF-TF	0.9817	0.9815	0.8771	0.9715	0.9730	0.8490	0.9472	0.9495	0.7720
PSF-DTI	**0.9873**	**0.9879**	**0.9030**	**0.9817**	**0.9838**	**0.8826**	0.9531	0.9553	0.7955
PSF-NA	0.9835	0.9839	0.8856	0.9783	0.9801	0.8701	**0.9580**	**0.9612**	**0.8114**

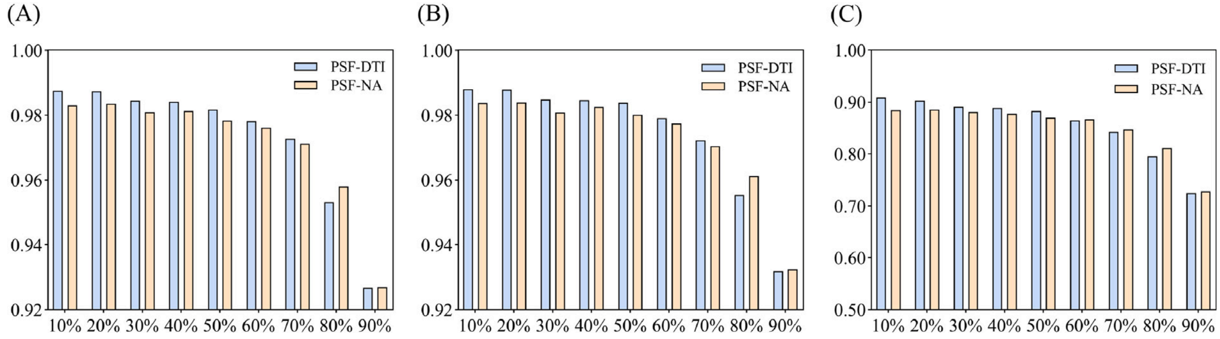

Fig. 5. Performance comparison of attention mechanism models. (A), (B), and (C) respectively represent the AUC, AUPR, and MCC results under different proportions.

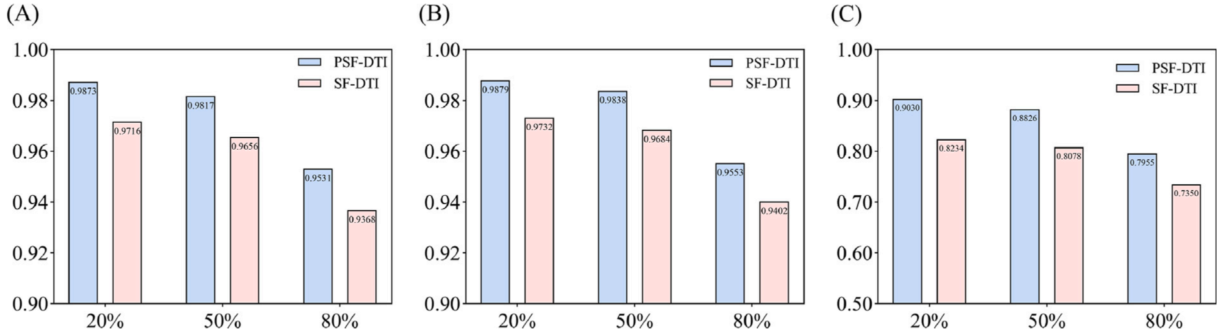

Fig. 6. Results of ablation study on pseudo-label supervision strategy. (A), (B), and (C) respectively represent the AUC, AUPR, and MCC results under different proportions.

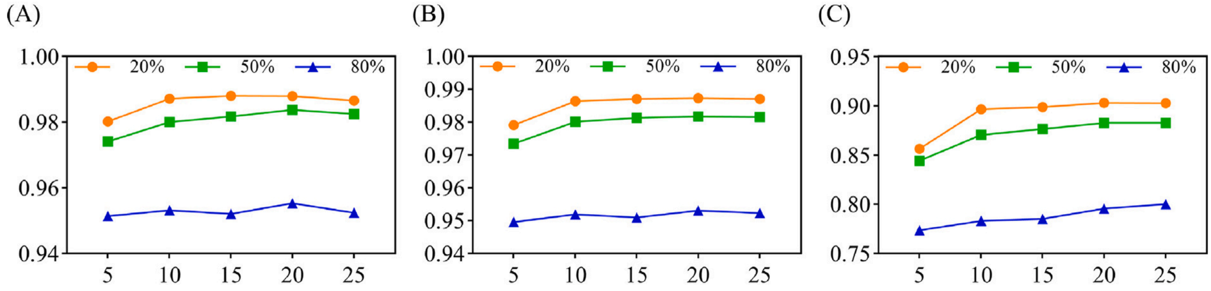

Fig. 7. Results of parameter study. (A), (B), and (C) respectively represent the AUC, AUPR, and MCC results under different K values.

3.5. Parameter study

In our method, an important hyperparameter is the number of neighbors K selected when constructing the far-neighbor graph. In order to study the impact of different K values on model performance, we selected the following set of K values for experiments: {5, 10, 15, 20, 25}. The experimental results are shown in Fig. 7. It can be seen that when the K value is low, increasing K can significantly improve the model performance because a larger K value enhances the ability of nodes to aggregate differential information. However, once K reaches 10, the performance improvement slows down, and beyond K = 20, the performance begins to decline. This is because a larger K value will generate more noisy edges, which will affect the model performance. Therefore, K = 20 is selected in PSF-DTI.

3.6. Model robustness test

In our method, DTP nodes contain both drug and target protein information. Therefore, the DTP nodes in the test set and the DTP nodes in the training set may contain the same drug and the targets are homologous, resulting in overperformance of the model. At the same time, if the training set contains homologous samples, the model will over-rely on specific features or patterns during training and fail to generalize to a wider data distribution. To this end, we conducted a robustness test on PSF-DTI. Since the DTI prediction task mainly focuses

Table 5
Results of ablation study on far-neighbor graph.

Settings	20%			50%			80%		
	AUC	AUPR	MCC	AUC	AUPR	MCC	AUC	AUPR	MCC
Euclidean Near-neighbor	0.9424	0.9445	0.7778	0.9302	0.9360	0.7506	0.8964	0.9103	0.6840
Euclidean Far-neighbor	0.9847	0.9828	0.8897	0.9780	0.9776	0.8706	0.9502	0.9524	0.7868
Manhattan Near-neighbor	0.9439	0.9508	0.7865	0.9309	0.9413	0.7603	0.9010	0.9164	0.7007
Manhattan Far-neighbor	**0.9873**	**0.9879**	**0.9030**	**0.9817**	**0.9838**	**0.8826**	**0.9580**	**0.9612**	**0.7955**

Table 6
Results of robustness test.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
AUC	**0.9828**	0.9809	0.9791	0.9764	0.9724	0.9676	0.9560	0.9391	0.8981
AUPR	**0.9832**	0.9816	0.9796	0.9780	0.9741	0.9695	0.9594	0.9401	0.9112
MCC	**0.8802**	0.8751	0.8709	0.8571	0.8499	0.8319	0.7985	0.7645	0.6683

Table 7
Prediction results of drugs.

Drug ID: Name	Target ID: Name	Result
DB00162: Vitamin A	P10745: RBP3	True
	P12271: RLBP1	True
	O95237: LRAT	True
	P50440: GATM	True
	P49419: ALDH7A1	True
	P04156: PRNP	True
Accuracy		100%
DB01200: Bromocriptine	P08908: HTR1A	True
	P35348: ADRA1A	True
	P34969: HTR7	True
	Q13621: SLC12A1	True
	P35499: SCN4A	True
	Q03403: TFF2	True
Accuracy		100%

Table 8
Prediction results of targets.

Target ID: Name	Drug ID: Name	Result
P08173: CHRM4	DB00540: Nortriptyline	True
	DB00809: Tropicamide	True
	DB01142: Doxepin	True
	DB01394: Colchicine	True
	DB01393: Bezafibrate	True
	DB01059: Norfloxacin	True
Accuracy		100%
P35372: OPRM1	DB00193: Tramadol	True
	DB00295: Morphine	True
	DB00704: Naltrexone	True
	DB00222: Glimepiride	True
	DB01250: Olsalazine	True
	DB04896: Milnacipran	True
Accuracy		100%

on the prediction of positive DTI, we consider target proteins with sequence similarity scores exceeding 40% as homologous (Luo et al., 2017), and remove positive DTIs containing homologous proteins from the data for experiments. Table 6 presents the experimental results, demonstrating that our model maintains strong performance even after eliminating homologous data.

3.7. Case study

Finally, we select two common drugs. For each drug, we randomly select 6 DTP nodes containing that drug. We use the data excluding these nodes to train the model and make predictions. The prediction results are shown in Table 7. Additionally, we select two common target proteins and randomly select 6 DTP nodes containing each target from the dataset. We again use the data excluding these nodes to train the model and make predictions. The prediction results are shown in Table 8. These results demonstrate the robust DTI prediction capability of our model.

4. Conclusion

This paper introduces a pseudo-label supervised graph fusion attention network for DTI prediction (PSF-DTI). We extract low-dimensional feature representations of DTP from diverse heterogeneous data sources. We construct a topology graph and a far-neighbor graph with DTP as nodes, and introduce an adaptive graph capable of dynamically updating the adjacency matrix to address static graph limitations. Feature fusion is performed through the attention mechanism to fully explore the potential correlation information between DTPs. During the model training phase, we introduce a pseudo-label supervision strategy to enhance the robustness of the model when faced with insufficient labeled data. Comparative experiments with baseline methods show that PSF-DTI still has better performance when facing a small amount of label data. Ablation experiments reveal that our proposed graph representation method and pseudo-label supervision strategy can improve the performance of the model. In addition, we conduct additional experiments to verify that the model is robust when removing homologous data from the data.

However, our study also has limitations. The two-stage nature of PSF-DTI increases the difficulty of optimization and may lead to the loss of important information in the initial feature extraction stage. To address this, future work could focus on developing an end-to-end learning framework that integrates feature extraction and prediction to minimize information loss and simplify optimization. Additionally, such models can improve computational efficiency when processing large-scale data through pruning and optimization techniques. Furthermore, our GCN-based approach may perform poorly when there are few connections between DTP nodes, especially in the presence of isolated nodes. We believe that combining drug and target protein information from more heterogeneous data sources can alleviate this problem.

CRediT authorship contribution statement

Yining Xie: Conceptualization, Writing – review & editing, Supervision, Project administration. **Xiaodong Wang:** Methodology, Software, Writing – original draft, Writing – review & editing. **Pengda Wang:** Validation, Visualization. **Xueyan Bi:** Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by Harbin Science and Technology Bureau Manufacturing Innovation Talent Project, China (CXRC20221110393), Heilongjiang Science and Technology Department Provincial Key R and D Program Applied Research Project, China (SC2022ZX06C0025), and Heilongjiang Science and Technology Department Provincial Key R and D Program Guidance Project, China (GZ20220088).

References

Chen, C., Shi, H., Jiang, Z., Salhi, A., Chen, R., Cui, X., et al. (2021). DNN-dtis: Improved drug-target interactions prediction using xgboost feature selection and deep neural network. *Computers in Biology and Medicine*, *136*, Article 104676.

Cheng, Z., Yan, C., Wu, F.-X., & Wang, J. (2021). Drug-target interaction prediction using multi-head self-attention and graph attention network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(4), 2208–2218.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., et al. (2021). Comparative toxicogenomics database (CTD): update 2021. *Nucleic acids research*, *49*(D1), D1138–D1143.

Ding, Y., Tang, J., & Guo, F. (2020). Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion. *Knowledge-Based Systems*, *204*, Article 106254.

Du, Y., Yao, Y., Tang, J., Zhao, Z., & Gou, Z. (2024). Drug-target interactions prediction via graph isomorphic network and cyclic training method. *Expert Systems with Applications*, *249*, Article 123730.

Fu, G., Ding, Y., Seal, A., Chen, B., Sun, Y., & Bolton, E. (2016). Predicting drug target interactions using meta-path-based semantic network analysis. *BMC bioinformatics*, *17*, 1–10.

González-Díaz, H., Prado-Prado, F., García-Mera, X., Alonso, N., Abeijón, P., Caamano, O., et al. (2011). MIND-best: Web server for drugs and target discovery; design, synthesis, and assay of MAO-b inhibitors and theoretical- experimental study of G3PDH protein from trichomonas gallinae. *Journal of proteome research*, *10*(4), 1698–1718.

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, *9*, 1–14.

He, Y., Sun, C., Meng, L., Zhang, Y., Mao, R., & Yang, F. (2024). Flexible drug-target interaction prediction with interactive information extraction and trade-off. *Expert Systems with Applications*, *249*, Article 123821.

Jin, W., Ma, Y., Liu, X., Tang, X., Wang, S., & Tang, J. (2020). Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 66–74).

Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database—2009 update. *Nucleic acids research*, *37*(suppl_1), D767–D772.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2010). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, *39*(suppl_1), D1035–D1041.

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic acids research*, *44*(D1), D1075–D1079.

Li, Y. Y., An, J., & Jones, S. J. (2011). A computational approach to finding novel targets for existing drugs. *PLoS Computational Biology*, *7*(9), Article e1002139.

Li, M., Cai, X., Xu, S., & Ji, H. (2023). Metapath-aggregated heterogeneous graph neural network for drug–target interaction prediction. *Briefings in Bioinformatics*, *24*(1), bbac578.

Li, Y., Qiao, G., Gao, X., & Wang, G. (2022). Supervised graph co-contrastive learning for drug–target interaction prediction. *Bioinformatics*, *38*(10), 2847–2854.

Li, Y., Qiao, G., Wang, K., & Wang, G. (2022). Drug–target interaction predication via multi-channel graph neural networks. *Briefings in Bioinformatics*, *23*(1), bbab346.

Li, J., Wang, J., Lv, H., Zhang, Z., & Wang, Z. (2021). IMCHGAN: inductive matrix completion with heterogeneous graph attention networks for drug-target interactions prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 655–665.

Li, H., Wu, B., Sun, M., Ye, Y., Zhu, Z., & Chen, K. (2023). Multi-view graph neural network with cascaded attention for lncrna-mirna interaction prediction. *Knowledge-Based Systems*, *268*, Article 110492.

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications*, *8*(1), 573.

Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., & Zheng, J. (2013). Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, *29*(2), 238–245.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, *30*(16), 2785–2791.

Mostafa, A. A., Alhossary, A. A., Salem, S. A., & Mohamed, A. E. (2022). GBO-kNN a new framework for enhancing the performance of ligand-based virtual screening for drug discovery. *Expert Systems with Applications*, *197*, Article 116723.

Nikraftar, Z., & Keyvanpour, M. R. (2023). A comparative analytical review on machine learning methods in drugtarget interactions prediction. *Current Computer-Aided Drug Design*, *19*(5), 325–355.

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, *34*(17), i821–i829.

Peng, J., Li, J., & Shang, X. (2020). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC bioinformatics*, *21*(Suppl 13), 394.

Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., & Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. *Journal of Computational Biology*, *18*(2), 133–145.

Shaikh, N., Sharma, M., & Garg, P. (2016). An improved approach for predicting drug–target interaction: proteochemometrics to molecular docking. *Molecular Biosystems*, *12*(3), 1006–1014.

Shi, W., Peng, D., Luo, J., Chen, G., Yang, H., Xie, L., et al. (2023). A review on predicting drug target interactions based on machine learning. In *International conference on health information science* (pp. 283–295). Springer.

Su, Y., Hu, Z., Wang, F., Bin, Y., Zheng, C., Li, H., et al. (2024). AMGDTI: drug–target interaction prediction based on adaptive meta-graph learning in heterogeneous network. *Briefings in Bioinformatics*, *25*(1), bbad474.

Wan, F., Hong, L., Xiao, A., Jiang, T., & Zeng, J. (2019). Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics*, *35*(1), 104–111.

Wang, R., Mou, S., Wang, X., Xiao, W., Ju, Q., Shi, C., et al. (2021). Graph structure estimation neural networks. In *Proceedings of the web conference 2021* (pp. 342–353).

Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., et al. (2018). A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *Journal of Computational Biology*, *25*(3), 361–373.

Wang, Y.-B., You, Z.-H., Yang, S., Yi, H.-C., Chen, Z.-H., & Zheng, K. (2020). A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC medical informatics and decision making*, *20*, 1–9.

Wang, X., Zhu, M., Bo, D., Cui, P., Shi, C., & Pei, J. (2020). Am-gcn: Adaptive multi-channel graph convolutional networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1243–1253).

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., et al. (2017). Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, *16*(4), 1401–1409.

Wu, Y., Gao, M., Zeng, M., Zhang, J., & Li, M. (2022). Bridgedpi: a novel graph neural network for predicting drug–protein interactions. *Bioinformatics*, *38*(9), 2571–2578.

Xuan, P., Fan, M., Cui, H., Zhang, T., & Nakaguchi, T. (2022). GVDTI: graph convolutional and variational autoencoders with attribute-level attention for drug–protein interaction prediction. *Briefings in bioinformatics*, *23*(1), bbab453.

Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of drug repositioning approaches and resources. *International journal of biological sciences*, *14*(10), 1232.

Yang, Y., Sun, Y., Ju, F., Wang, S., Gao, J., & Yin, B. (2023). Multi-graph fusion graph convolutional networks with pseudo-label supervision. *Neural Networks*, *158*, 305–317.

Zhang, Y., Hu, Y., Han, N., Yang, A., Liu, X., & Cai, H. (2023). A survey of drug-target interaction and affinity prediction methods via graph neural networks. *Computers in Biology and Medicine*, *163*, Article 107136.

Zhang, R., Wang, Z., Wang, X., Meng, Z., & Cui, W. (2023). Mhtan-dti: Metapath-based hierarchical transformer and attention network for drug–target interaction prediction. *Briefings in Bioinformatics*, *24*(2), bbad079.

Zhang, P., Wei, Z., Che, C., & Jin, B. (2022). Deepmgt-DTI: Transformer network incorporating multilayer graph information for drug–target interaction prediction. *Computers in biology and medicine*, *142*, Article 105214.

Zhao, T., Hu, Y., Valsdottir, L. R., Zang, T., & Peng, J. (2021). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Briefings in bioinformatics*, *22*(2), 2141–2150.

Zhu, Z., Yao, Z., Qi, G., Mazur, N., Yang, P., & Cong, B. (2023). Associative learning mechanism for drug-target interaction prediction. *CAAI Transactions on Intelligence Technology*, *8*(4), 1558–1577.

Zhu, Z., Yao, Z., Zheng, X., Qi, G., Li, Y., Mazur, N., et al. (2023). Drug–target affinity prediction method based on multi-scale information interaction and graph optimization. *Computers in Biology and Medicine*, *167*, Article 107621.