OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–21

https://doi.org/10.1093/bib/bbac434

Advance access publication date 14 October 2022

Problem Solving Protocol

MHADTI: predicting drug–target interactions via multiview heterogeneous information network embedding with hierarchical attention mechanisms

Zhen Tian†, Xiangyu Peng†, Haichuan Fang, Wenjie Zhang, Qiguo Dai and Yangdong Ye

Corresponding author: Zhen Tian, E-mail: ieztian@zzu.edu.cn; Yangdong Ye, E-mail: ieydye@zzu.edu.cn

†The first two authors contribute equally to the paper

Abstract

Motivation: Discovering the drug–target interactions (DTIs) is a crucial step in drug development such as the identification of drug side effects and drug repositioning. Since identifying DTIs by web-biological experiments is time-consuming and costly, many computational-based approaches have been proposed and have become an efficient manner to infer the potential interactions. Although extensive effort is invested to solve this task, the prediction accuracy still needs to be improved. More especially, heterogeneous network-based approaches do not fully consider the complex structure and rich semantic information in these heterogeneous networks. Therefore, it is still a challenge to predict DTIs efficiently. **Results**: In this study, we develop a novel method via **M**ultiview heterogeneous information network embedding with **H**ierarchical **A**ttention mechanisms to discover potential **D**rug–**T**arget **I**nteractions (**MHADTI**). Firstly, MHADTI constructs different similarity networks for drugs and targets by utilizing their multisource information. Combined with the known DTI network, three drug–target heterogeneous information networks (HINs) with different views are established. Secondly, MHADTI learns embeddings of drugs and targets from multiview HINs with hierarchical attention mechanisms, which include the node-level, semantic-level and graph-level attentions. Lastly, MHADTI employs the multilayer perceptron to predict DTIs with the learned deep feature representations. The hierarchical attention mechanisms could fully consider the importance of nodes, meta-paths and graphs in learning the feature representations of drugs and targets, which makes their embeddings more comprehensively. Extensive experimental results demonstrate that MHADTI performs better than other SOTA prediction models. Moreover, analysis of prediction results for some interested drugs and targets further indicates that MHADTI has advantages in discovering DTIs. **Availability and implementation:** https://github.com/pxystudy/MHADTI

Keywords: data fusion, multiview heterogeneous information network embedding, hierarchical attention mechanisms, drug–target interaction prediction

Introduction

Discovering drug–target interactions (DTIs) is an important step in the drug discovery pipeline, which could facilitate the understanding of drug action mechanisms, disease pathology and drug side effects [1]. Identifying DTIs by traditional biochemical experiments is extremely costly and time-consuming [2]. For example, the estimated cost of developing one new drug is about $1.8 billion, and takes approximately 13 years [3, 4]. Therefore, researchers have sought to employ computational-based approaches to discover DTIs, which could significantly reduce the high cost and narrow down the long period for developing new drugs [5].

Generally speaking, DTIs prediction approaches can be summarized into three categories, which are ligand-similarity-based [6], structure-based [7] and hybrid approaches [8]. Ligand-similarity-based and structure-based methods are two types of traditional computational prediction models. Specifically, ligand-similarity-based methods usually need a large number of known binding ligands for interested targets, while structure-based methods always require sufficient three-dimensional structures of target proteins for promoting their predictive power. However, at present both the known binding ligands and the three-dimensional structures of proteins are limited, requiring the performance of these two types of prediction models to be improved [9].

Zhen Tian received BS, MS and PhD degrees in computer science and technology from Harbin Engineering University in 2011, Harbin Institute of Technology in 2013 and 2017, respectively. Currently, he works at the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China. His current research interests include computational biology, complex network analysis and data mining.

Xiangyu Peng is studying for her master's degree in Zhengzhou University of School of Computer and Artificial Intelligence. Her research interests include data mining and computational biology.

Haichuan Fang is currently working toward the Master Degree of Engineering in Zhengzhou University, Zhengzhou, China. His research interests include knowledge graph embedding, bioinformatics and deep learning.

Wenjie Zhang is a lecturer at School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China. His current research interests lie in signal processing, deep learning and biomedical information analysis.

Qiguo Dai received BS, MS and PhD degrees in computer science and technology from Hubei University of Automotive Technology in 2006, Beijing University of Technology in 2010 and Harbin Institute of Technology in 2015, respectively. Currently, he is an associate professor at the School of Computer Science and Engineering, Dalian Minzu University. His research interests include bioinformatics and data mining.

Yangdong Ye (Member, IEEE) received the PhD degree from China Academy of Railway Sciences, Beijing, China, in 2002. He is currently a Professor with the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China. He has wide research interests, mainly including machine learning, pattern recognition, knowledge engineering and intelligent systems.

Received: June 5, 2022. **Revised:** August 19, 2022. **Accepted:** September 8, 2022

Currently, hybrid methods are believed to be another promising and effective way to predict DTIs. Hybrid methods mainly include two subcategories: similarity-based and network-based methods [10]. The assumption for similarity-based models is that *compounds with similar structures may have similar properties* [11]. Under this assumption, drugs and targets are always represented as feature vectors by utilizing their own information (without network information), and DTI prediction tasks can be formulated as one binary classification problem. Meanwhile, some effective similarity-based approaches such as SimBoost [12], DeepDTA [13], ML-DTI [14], GraphDTA [15], MGraphDTA [16], Mol2Context-vec [17] and DeepConv-DTI [18] are proposed and show advantages in the accuracy of DTI predictions. For example, SimBoost predicted continuous binding affinity values for compounds and proteins based on their different types of feature vectors [12]. However, features of drugs and targets need to be defined in advance, which affects their generalization ability to a certain extent. For bridging the gap between drug and target encoder, ML-DTI designed mutual learning layers, which were achieved by multi-head attention and position-aware attention to predict DTIs [14]. The features of drugs and proteins were learned from sequence information with one encoder. GraphDTA was a neural network architecture that models drugs as molecular graphs and learns the comprehensive representations of drugs and proteins with GNN-based method models [15]. However, it almost paid all attention to the representation of drugs and ignores the features of proteins. DTI-CDF generated multiple similarity-based features for drugs and the target proteins to improve the prediction performance of DTIs [19]. DTI-CDF only employed the traditional machine-learning models and neglected the use of advanced machine-learning methods such as graph neural network (GNN).

Network-based methods usually construct networks of drugs and targets and employ graph-based technology to learn the embedding of nodes and then discover their potential links in the networks. These methods usually follow the 'guilt-by-association' assumption that *drugs tend to bind to similar targets, and vice versa* [20]. Early network-based methods only take the DTI information to build the bipartite network and then predict DTIs [21, 22]. For example, Bleakley firstly proposed a supervised inference method to predict DTIs based on the bipartite local model [21]. However, these approaches only take the interactions between drugs and targets into consideration. Afterward, establishing heterogeneous networks which incorporate multiple information related with drugs and targets has gained much attention [23–25]. For instance, NRWRH established one integrated heterogeneous network by applying prior information of drugs and targets and then implemented the random walk with restart algorithm on the heterogeneous network to predict DTIs [26]. DTINet applied an unsupervised method to learn low-dimensional feature representations of drugs and target proteins from heterogeneous data and predicted DTI using inductive matrix completion [27]. NeoDTI constructed a heterogeneous network based on eight individual networks of drugs and targets and then learned their topology-preserving representations to predict DTIs [28]. However, for these methods, establishing the heterogeneous networks with the multisource information of drugs and targets in depth is a challenging task. Besides, these approaches mainly learn the feature representations for nodes in the network and it is difficult for them to achieve target prediction for drugs outside the heterogeneous network. More recently, biological knowledge graph (KG) based methods are also proposed to deal with DTI prediction [29, 30]. For instance, TriModel first utilized biomedical knowledge datasets to create a knowledge graph of entities connected to both drugs and targets and learned their comprehensive embedding representations [31]. Ye *et al.* [32] developed a unified framework called KGE_NFM for DTIs prediction, which could obtain the low-dimensional representations for various entities in the graphs. However, it is still a crucial and challenging task to establish the biomedical knowledge graphs systematically with the multi-omics data of biological entities.

Meanwhile, graph attention networks (GATs) [33] show great potential in modeling complex graph data, which has gained considerable interest in different research areas such as graph classifications [34] and recommender systems [35]. More importantly, GATs have been successfully utilized in some link prediction tasks in bioinformatics areas [36]. For instance, Long [37] put forward the hierarchical attention mechanism for microbe–drug interaction. Wang [38] proposed a heterogeneous graph attention network (HAN) model to learn node deep embedding from both node-level and semantic-level attentions. More recently, MLA-GNN first constructed multi-level graphs and then employs the attention mechanism on each graph to learn the comprehensive features of nodes [39]. Therefore, GAT has fully demonstrated its effective ability in fusing graph topological structure and learning node embeddings [40]. However, most of these approaches only apply the GAT on the node level and fewer studies pay much attention to the hierarchical attention mechanism on the graph level.

Multiview-based feature learning models have become another mainstream manner in dealing with the DTI prediction problem [41]. These approaches usually employ multitype information of drugs and targets such as chemical structures, fingerprints, side-effects to construct their similarity networks and then establish multi-view networks. Then embeddings of drugs and targets could be obtained based on these multiview similarity networks through various feature learning approaches [42, 43]. For example, MVGCN constructed multi-heterogeneous networks with the similarity networks of drugs and targets and learned their embeddings by aggregating the representations from inter-domain and intra-domain neighbors in multiview networks [44]. Afterward, Yuan [45] put forward a knowledge-enhanced multiview framework model which learned the comprehensive representations of nodes via multi-view attention mechanism and predicted DTIs on a large scale. Wei [46] proposed a multiview-based deep learning model called MDL-CPI which extracted the features of proteins and compounds and then obtained their interactive information for compound-protein interaction prediction. However, most of these approaches only take the homogeneous networks or biological features of entities themselves as multiview inputs.

Multiview HINs related to drugs and targets usually contain complex structural information and rich semantic information. How to fully capture the structural and semantic information in HINs for learnings embeddings of drugs and targets simultaneously is a challenging task [47]. Since meta-paths could capture complex relationships that effectively reveal structural and semantic information in HINs, various meta-paths will imply different semantic meanings well. Meanwhile, meta-path-based neighbors have different importance to the representations of drugs and targets. Therefore, learning their embeddings with hierarchical attention mechanisms has become an effective strategy.

In this study, we construct the multiview heterogeneous information networks (HINs) and expand the hierarchical attention mechanisms with three level attentions. A novel method is proposed employing **M**ultiview heterogeneous information networks with **H**ierarchical **A**ttention mechanisms to predict **D**rug-**T**arget **I**nteractions (**MHADTI**). An overview of MHADTI is shown in

Figure 1 and MHADTI mainly contains four steps. Firstly, multiple types of data for drugs and targets are collected from different datasets. Then we build different similarity networks for drugs and targets and construct multiview HINs. After that, MHADTI learns the deep embedding representations based on HINs with hierarchical attention mechanisms, which include the node-level, semantic-level and graph-level attentions, respectively. Lastly, we concentrate representations of drugs and targets and feed them into the multilayer perceptron (MLP) to predict DTIs. Our contributions can be summarized as follows:

(1) We establish multiview HINs of drugs and targets with their multisource information.
(2) MHADTI could capture the complex structure and rich semantic information in HINs with the hierarchical attention mechanisms.
(3) MHADTI could fully consider the importance of nodes, meta-paths and graphs with the hierarchical attention mechanisms which include node-level, semantic-level and graph-level attentions, respectively.
(4) Experimental results demonstrate that MHADTI is superior to other SOTA approaches in DTI prediction.

Materials and methods

In this section, we will first describe the datasets used in MHADTI. Then, multitype similarity networks of drugs or targets are calculated based on various similarity measurement models and multiview HINs are constructed based on these similarity networks combined with the DTI network. After that, learning the embeddings of drugs and targets based on multiview HINs with hierarchical attention mechanisms is well described. In the end, some implementation details are introduced.

Data collection

In this study, various types of information about drugs and targets are collected for MHADTI. Specifically, DTI data are mainly downloaded from the DrugBank database(v5.1) [48]. To measure the similarity between drugs and targets comprehensively, we also obtained their other types of information. Specifically, we adopted SMILES (Simplified Molecular Input Line Entry System) and side effect information of drugs, which are collected from PubChem [49] and Uniprot database [50], respectively. Meanwhile, functional annotation, protein domain and sequence information of targets are also downloaded from the Uniprot database [50]. A brief statistics about drugs and targets annotation information used is shown in Table 1. Here, we only select drugs and targets that all the annotation information is known. Consequently, a total of 15 252 DTIs involving 4358 drugs (compounds) and 2407 targets (proteins) are employed for establishing their HINs.

Table 1. A brief statistics about drugs and targets information.

	Data types	Number
Drugs	Side effects	747
	SMILES	4358
	DTIs	15 252
Targets	Protein Sequence	2407
	Protein Domain	2348
	BP	29 380
	CC	11 113
	MF	4181

Table 2. Main notations used in this study.

Notations	Descriptions
Φ	Meta-path
$\mathbf{h}$	Initial node feature
$\mathbf{h}'$	Projected node feature
$\mathbf{M}_\phi$	Type-specific transformation matrix
$\boldsymbol{q}$	Semantic-level attention vector
$\mathbf{c}$	Graph-level attention vector
$e_{v_i v_j}^{(\Phi,\mathcal{G})}$	Importance of node v_j to v_i under Φ in graph $\mathcal{G}$
$\alpha_{v_i v_j}^{(\Phi,\mathcal{G})}$	Weight of node v_j to v_i under Φ in graph $\mathcal{G}$
$\mathcal{N}^{(\Phi,\mathcal{G})}$	Meta-path based neighbors for Φ in graph $\mathcal{G}$
$w^{(\Phi,\mathcal{G})}$	Importance of meta-path Φ in graph $\mathcal{G}$
$\beta^{(\Phi,\mathcal{G})}$	Weight of meta-path Φ in graph $\mathcal{G}$
$\gamma^{\mathcal{G}}$	Weight of graph $\mathcal{G}$
$\mathbf{Z}^{(\Phi,\mathcal{G})}$	Node-level embedding under Φ in graph $\mathcal{G}$
$\mathbf{Z}^{\mathcal{G}}$	Semantic-level node embedding in graph $\mathcal{G}$
$\mathbf{Z}$	The final node embedding
(i, j)	The node pair of drug d_i and target t_j
y_{ij}	The ground truth of drug d_i and target t_j
$\hat{y}_{ij}$	The prediction interaction score of drug d_i and target t_j
$\mathbf{Y}^+$	The positive samples in the training set
$\mathbf{Y}^-$	The negative samples in the training set
IC	The information content
BP	Biological process
MF	Molecular function
CC	Cell component

The construction of multiview HINs

In this subsection, we will first describe the different similarity calculation models for drugs and targets (see **Appendix** A section) and then establish the multiview HIN.

Construction of Multiview HINs

Here, we can calculate similarities between all drug pairs with their side effect, fingerprint and Gaussian interaction profile kernel information. Meanwhile, the similarities of all target pairs can be evaluated by their functional annotation, protein domain and protein sequence information. Therefore, three different similarity networks for drugs and targets can be established. The side-effect-similarity network, fingerprint-based similarity network and Gaussian-interaction-profile-kernel-based similarity network of drugs are denoted as $Net_{sideeffect}$, $Net_{fingerprint}$ and Net_{GIP}. The annotation-based semantic similarity network, the domain-based similarity network and the sequence-based similarity network of targets are denoted as $Net_{integrated}$, Net_{domain} and $Net_{sequence}$, respectively. Combined with the known DTI network, three multiview HINs are constructed which are shown in Figure 1B.

Specifically, the three HINs in MHADTI model (Step 2 in Figure 1B) are represented as $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$, respectively. $\mathcal{G}_1$ is established based on annotation-based semantic similarity network of targets $Net_{integrated}$, side-effect-similarity network of drugs $Net_{sideeffect}$ and the known DTI network, while $\mathcal{G}_2$ is constructed by domain-based similarity network of targets Net_{domain}, fingerprint-based similarity network of drugs $Net_{fingerprint}$ and the known DTI network. Besides, $\mathcal{G}_3$ is built with sequence-based similarity network of targets $Net_{sequence}$, the Gaussian-interaction-profile-kernel-based similarity network of drugs Net_{GIP} and the known DTI network.

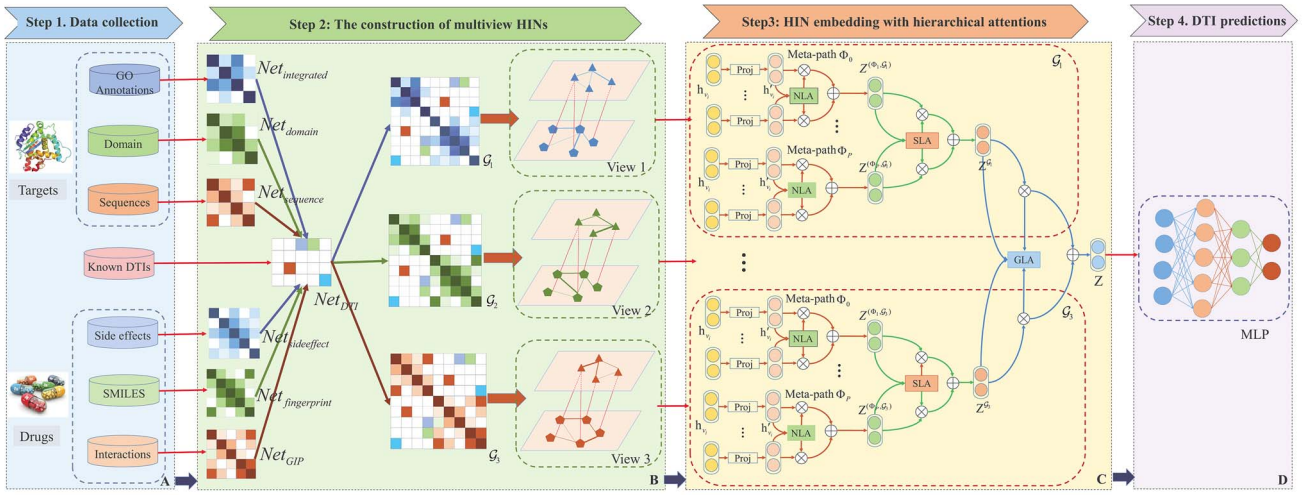

Figure 1. An overview of MHADTI for DTI prediction. MHADTI mainly contains four steps. **(A)** In the first step, we collect multisource information of drugs and targets from different datasets. **(B)** In the second step, various similarity calculation models are employed to measure the similarity for drugs and targets from different views and then MHADTI constructs the multiview heterogeneous information networks. **(C)** In the third step, MHADTI learns the embeddings of drugs and targets with the hierarchical attention mechanisms, which include the node-level, semantic-level and graph-level attentions. **(D)** We employ the multilayer perceptron to predict the potential DTIs based on the learned embeddings of drugs and targets in the third step. In the figure, $Net_{integrated}$, Net_{domain} and $Net_{sequence}$ denote the integrated semantic similarity network, domain-based similarity network and the sequence-based similarity network of targets, while $Net_{sideeffect}$, $Net_{fingerprint}$ and Net_{GIP} denote the side-effect-based similarity network, fingerprint-based similarity network and Gaussian-interaction-profile-kernel-based similarity network of drugs, respectively. Net_{DTI} represents the drug–target interaction network. NLA, SLA and GLA are the representations of node-level attention, semantic-level attention and graph-level attention. $\mathbf{Z}^{(\Phi,\mathcal{G})}$,$\mathbf{Z}^{\mathcal{G}}$ and $\mathbf{Z}$ denote the embedding representations at node-level, semantic-level and graph-level, respectively.

MHADTI model

In this subsection, we first define the basic concepts related to MHADTI model. Then learning embeddings of drugs and targets via hierarchical attention mechanisms is displayed in Figure 1C. After that, the decoder model and loss function are introduced. In the end, we will perform a description about the implementation details for MHADTI.

Related concepts

DTI prediction problem formulation.

Given a set of M drugs $D = (d_1, d_2, \cdots, d_M)$ and a set of N targets $T = (t_1, t_2, \cdots, t_N)$, and one drug–target pair $d_i \in D$ and $t_j \in T$ whose interaction information is unknown, the goal for MHADTI is to predict the interaction relationship $I(d_i, t_j)$ according to their final embeddings, which can be denoted as

$$I(d_i, t_j) = \begin{cases} 1, \ interaction \\ 0, \ no \ \ interaction \end{cases} \quad (1)$$

Definition 1. **Heterogeneous Information Network (HIN).**

The HIN (Graph) is one type of information network, which can be formulated as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \phi, \psi)$, where $\mathcal{V}$ denotes the node set and $\mathcal{E}$ represents the edge set. The node type mapping function and the edge mapping function are defined as $\phi : \mathcal{V} \to \mathcal{A}$ and $\psi : \mathcal{E} \to \mathcal{R}$. The $\mathcal{A}$ and $\mathcal{R}$ are node types and link types, respectively, and $|\mathcal{A}| + |R| > 2$.

Example. In the drug–target HIN (Figure 2B), there are two types of nodes which are drug and target, and two types of links which are node similarity and DTI. There are different relationships between nodes, i.e. DTI and drug–drug similarity.

Definition 2. **Meta-paths.**

A meta-path Φ can be described as the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, which is abbreviated as $A_1A_2 \cdots A_{l+1}$. The composition relation between node A_1 and A_{l+1} is formulated as $R = R_1 \circ R_2 \circ \cdots \circ R_l$, where $\circ$ denotes the composition operator on relations.

Example. In the drug–target HIN (Figure 2B), two drugs can be connected by different meta-paths (Figure 2C), such as Drug–Target–Drug (DTD) and Drug–Target–Target–Drug (DTTD). These meta-paths usually have different semantic meanings. For instance, DTD indicates that if two drugs interact with one common target, they will have a higher similarity. TDT indicates that if two targets interact with one common drug, they will also have a higher similarity.

Definition 3. **Meta-path-based neighbors.**

Suppose there is one node named v_i and one meta-path Φ in the HINs, the meta-path-based neighbors $\mathcal{N}_{v_i}^{\Phi}$ for node v_i can be defined as the nodes that connect with v_i based on the meta-path Φ. Note that $\mathcal{N}_{v_i}^{\Phi}$ is the set of nodes, which contains node v_i.

Example. As is shown in Figure 2D, for drug D_1, its DTD meta-path-based neighbors are D_1, D_4 and D_5. These meta-path-based neighbors have different importance to the embedding learning of drugs D_1.

Definition 4. **Heterogeneous Information Network Embedding.**

Suppose there is one HIN named $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \phi, \psi)$ and one node named v_i, the HIN embedding model is to learn a d-dimension representation from the original feature space via a mapping function $f : \mathcal{V} \to \mathbb{N}^d$, where $v_i \in \mathcal{V}$ and $d \ll |\mathcal{V}|$.

Example. In the MHADTI model, we employ the hierarchical attention mechanisms to learn the embedding representation of drugs and targets from their original feature space to the low-dimensional representation space.

Node-level attention

Node-level attention could effectively learn the importance of the neighbors for drugs and targets in HINs. And different meta-path-based neighbors of drugs and targets have diverse

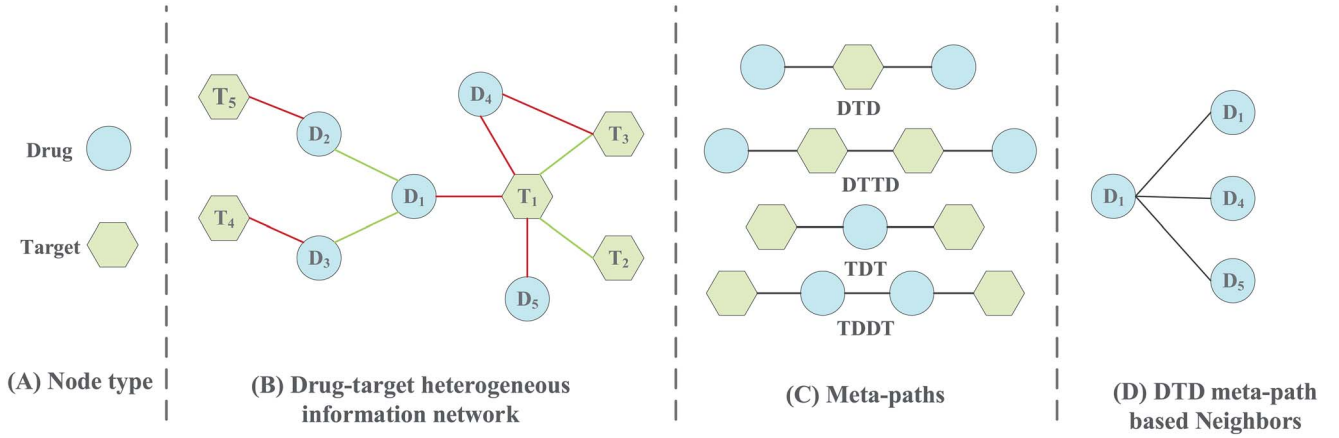

Figure 2. A toy example for MHADTI. **(A)** Two node types which are drug and target, **(B)** A heterogeneous information network contains two types of nodes: drug and target, and two types of links: similarity (green color) and interaction (red color), **(C)** Four meta-paths involved in MHADTI, which are DTD, DTTD, TDT and TDDT, **(D)** Drug D_1 and its DTD meta-path-based neighbors D_1, D_4 and D_5 based on the HIN (Figure 2B).

specific semantic meaning which is significant in learning their embeddings comprehensively. Therefore, the embeddings of drugs and targets can be generated by aggregating information from their meta-path-based neighbors with the node-level attention.

Since drugs and targets are two different types of nodes in HINs and they have different feature spaces, it is essential to transform their features into one same feature space firstly. The type-specific transformation matrix can be represented as $\mathbf{M}_{\phi}$ and the projection process can be defined as

$$\mathbf{h}'_{v_i} = \mathbf{M}_{\phi_{v_i}} \cdot \mathbf{h}_{v_i}, \phi_{v_i} \in \{drugs,\ targets\}, \tag{2}$$

where $\mathbf{M}_{\phi_{v_i}}$ denotes the transformation matrix for drugs or targets, and $\mathbf{h}_{v_i}$ and $\mathbf{h}'_{v_i}$ are the original and projection features for node v_i respectively. With this type-specific projection operation, drugs and targets with different features dimension can be transformed into one same space.

Next, we employ the self-attention mechanism to learn the weights for drugs and targets in HINs. Suppose there is one node-pair named (v_i, v_j), which is connected via the meta-path Φ in graph $\mathcal{G}$, the node-level attention $e^{(\Phi,\mathcal{G})}_{v_i v_j}$ indicates the importance of node v_j to node v_i. The importance of meta-path-based node pair (v_i, v_j) can be formulated as follows:

$$e^{(\Phi,\mathcal{G})}_{v_i v_j} = att_{node}(\mathbf{h}'_{v_i}, \mathbf{h}'_{v_j}; \Phi, \mathcal{G}), \tag{3}$$

where att_{node} represents the deep neural network which achieves the node-level attention. It could share all meta-path-based node pairs under the meta-path Φ in graph $\mathcal{G}$.

After that, we calculate $e^{(\Phi,\mathcal{G})}_{v_i v_j}$ for nodes $v_j \in \mathcal{N}^{(\Phi,\mathcal{G})}_{v_i}$, where $\mathcal{N}^{(\Phi,\mathcal{G})}_{v_i}$ denotes the meta-path-based neighbors based on Φ in graph $\mathcal{G}$ for node v_i including itself. Specifically, we normalize importance and get the weight coefficient $\alpha^{(\Phi,\mathcal{G})}_{v_i v_j}$ via softmax function:

$$\alpha^{(\Phi,\mathcal{G})}_{v_i v_j} = softmax\left(e^{(\Phi,\mathcal{G})}_{v_i v_j}\right) = \frac{exp\left(\sigma\left(\mathbf{a}^{T}_{(\Phi,\mathcal{G})} \cdot [\mathbf{h}'_{v_i} \| \mathbf{h}'_{v_j}]\right)\right)}{\sum_{v_k \in \mathcal{N}^{(\Phi,\mathcal{G})}_{v_i}} exp\left(\sigma\left(\mathbf{a}^{T}_{(\Phi,\mathcal{G})} \cdot [\mathbf{h}'_{v_i} \| \mathbf{h}'_{v_k}]\right)\right)}, \tag{4}$$

where σ denotes the activation function, $\|$ represents the concatenate operation and $\mathbf{a}^{T}_{(\Phi,\mathcal{G})}$ is the node level attention vector under meta-path Φ in graph $\mathcal{G}$. Please note that the weight coefficient is asymmetric, which means the importance of node v_i to v_j is not equal to the importance of v_j to v_i.

Then, the node-level embedding for v_i can be aggregated by its neighbors' projected features based on their corresponding coefficients weight defined as

$$\mathbf{z}^{(\Phi,\mathcal{G})}_{v_i} = \sigma\left(\sum_{v_j \in \mathcal{N}^{(\Phi,\mathcal{G})}} \alpha^{(\Phi,\mathcal{G})}_{v_i v_j} \cdot \mathbf{h}'_{v_j}\right), \tag{5}$$

where $\mathbf{z}^{(\Phi,\mathcal{G})}_{v_i}$ is the learned embedding of node v_i under the meta-path Φ in graph $\mathcal{G}$ and σ represents the activate function. Since the attention weight $\alpha^{(\Phi,\mathcal{G})}_{v_i v_j}$ is generated from a single meta-path, it has specific semantic meaning and is able to fully capture this kind of semantic information.

To make the training process more stable, we combine the node-level attention with the multihead attention together. The ultimate embedding for node v_i at the node-level attention is formulated as

$$\mathbf{z}^{(\Phi,\mathcal{G})}_{v_i} = \Big\|_{l=1}^{L} \sigma\left(\sum_{v_j \in \mathcal{N}^{(\Phi,\mathcal{G})}} \alpha^{(\Phi,\mathcal{G})}_{v_i v_j} \cdot \mathbf{h}'_{v_j}\right), \tag{6}$$

where $\|$ represents concatenation, L is the number of attention head and l is a variable ranging from 1 to L. In this study, MHADTI could learn the node-level embeddings of drugs and targets with the node-level attention, which aggregates their meta-path-based neighbors information. Given a set of meta-paths denoted as $\{\Phi_1, \ldots, \Phi_P\}$ in graph $\mathcal{G}$, after injecting features of all nodes into node-level attention, we can obtain P group of semantic-specific node embeddings, represented as $\{\mathbf{Z}^{(\Phi_1,\mathcal{G})}, \cdots, \mathbf{Z}^{(\Phi_P,\mathcal{G})}\}$.

Semantic-level attention

Generally speaking, drugs and targets in HINs may belong to multiple meta-paths and each meta-path has its specific semantic meaning. Further, one meta-path usually reflects only one aspect of semantic meaning of drugs and targets. Therefore, to obtain their more informative embeddings, we need to aggregate various semantics of different meta-paths. Here, four representative meta-paths for drugs and targets are selected, which are shown in Figure 2C.

To learn the importance of each meta-path automatically and fuse the different semantics of meta-paths of drugs and targets, we leverage the semantic-level attention into MHADTI. Here, still

taking the meta-path set $\{\mathbf{Z}^{(\Phi_1,\mathcal{G})},\cdots,\mathbf{Z}^{(\Phi_P,\mathcal{G})}\}$ as an example, these P groups of node-level embeddings are employed to learn the weight of each meta-path, which can be denoted as follows:

$$\left(\beta^{(\Phi_1,\mathcal{G})},\cdots,\beta^{(\Phi_P,\mathcal{G})}\right)=att_{sem}\left(\mathbf{Z}^{(\Phi_1,\mathcal{G})},\cdots,\mathbf{Z}^{(\Phi_P,\mathcal{G})}\right), \tag{7}$$

where att_{sem} is the deep neural network that performs the semantic-level attention.

Further, to learn the importance of each meta-path, we firstly transform node-level embeddings through a nonlinear transformation. Then, we evaluate the importance of the node-level embedding as the product between transformed embeddings and a semantic-level attention vector $\boldsymbol{q}$. In this way, we could average the importance of all semantic-specific node-level representations, which can be regarded as the importance of each meta-path. The importance of each meta-path is defined as

$$w^{(\Phi_p,\mathcal{G})}=\frac{1}{|\mathcal{V}|}\sum_{v_i\in\mathcal{V}}\boldsymbol{q}^T\cdot\tanh(\mathbf{W}\cdot\mathbf{z}_{v_i}^{(\Phi_p,\mathcal{G})}+\boldsymbol{b}), \tag{8}$$

where $\mathbf{W}$ is the weight matrix, $\boldsymbol{b}$ and $\boldsymbol{q}$ are the bias vector and semantic-level attention vector, respectively. Similarly, we operate the semantic-level importance and leverage the softmax function to normalize the weight coefficient of meta-path Φ_p, which can be expressed as follows

$$\beta^{(\Phi_p,\mathcal{G})}=\frac{exp\left(w^{(\Phi_p,\mathcal{G})}\right)}{\sum_{j=1}^{P}exp\left(w^{(\Phi_j,\mathcal{G})}\right)}, \tag{9}$$

where $\beta^{(\Phi_p,\mathcal{G})}$ denotes the normalized weight for meta-path Φ_p in graph $\mathcal{G}$. Finally, we can fuse the semantic-level embeddings to obtain the integrated embedding via the learned weight coefficients. It is calculated as follows

$$\mathbf{Z}^{\mathcal{G}}=\sum_{p=1}^{P}\beta^{(\Phi_p,\mathcal{G})}\cdot\mathbf{Z}^{(\Phi_p,\mathcal{G})} \tag{10}$$

In this way, the embedding representations for drugs and targets will be more comprehensive. Given multiview HINs $\{\mathcal{G}_1,\ldots,\mathcal{G}_K\}$, we can obtain K groups of semantic-level embeddings, which can be formulated as $\{\mathbf{Z}^{\mathcal{G}_1},\cdots,\mathbf{Z}^{\mathcal{G}_K}\}$, where $\mathbf{Z}^{\mathcal{G}_k}$ denotes the semantic-level node embeddings in graph $\mathcal{G}_k$.

Graph-level attention

When there are multiple attributes of drugs and targets such as drug fingerprints and protein annotations, we can adopt different attributes to construct their multiview HINs. Since various features of drugs and targets have different importance on their embedding learning, it is a great challenge to assign weights reasonably to their features which are learned from various heterogeneous graphs.

In this research, we construct the multiview HINs and propose a novel graph-level attention mechanism, which could automatically learn the importance of embeddings of drugs and targets from their different HINs. The weights for HINs can be formulated as

$$(\gamma^{\mathcal{G}_1},\gamma^{\mathcal{G}_2},\cdots,\gamma^{\mathcal{G}_K})=att_{graph}(\mathbf{Z}^{\mathcal{G}_1},\mathbf{Z}^{\mathcal{G}_2},\cdots,\mathbf{Z}^{\mathcal{G}_K}), \tag{11}$$

where att_{graph} is the deep neural network that performs the graph-level attention and γ^k is the weight coefficient for graph $\mathcal{G}_k$.

To learn the importance of graph $\mathcal{G}$, we first define a graph-level attention vector named $\mathbf{c}$. Then, we evaluate the weight coefficient in the k-th heterogeneous graph based on SoftMax activation function, which can be expressed as

$$\gamma^{\mathcal{G}_k}=\frac{exp(\mathbf{c}\cdot\mathbf{Z}^{\mathcal{G}_k})}{\sum_{j=1}^{K}exp(\mathbf{c}\cdot\mathbf{Z}^{\mathcal{G}_j})}, \tag{12}$$

where $\gamma^{\mathcal{G}_k}$ is the weight coefficient of $\mathcal{G}_k$ after being normalized. Obviously, the bigger the weight is, the more important the attribute is.

Finally, the learned weight coefficients are employed to fuse the node embeddings from different HINs and we can obtain the final embedding $\mathbf{Z}$, which can be defined as

$$\mathbf{Z}=\sum_{k=1}^{K}(\gamma^{\mathcal{G}_k}\cdot\mathbf{Z}^{\mathcal{G}_k}). \tag{13}$$

In brief, graph-level attention is actually to assign distinct weight coefficients to the same nodes, for the sake of evaluating the importance of embedding in each HIN reasonably. The whole process of hierarchical attention mechanisms is displayed in Figure 1C.

An example of learning the embedding representation for one node with MHADTI is shown in Figure 3. The hierarchical attention mechanism consists of three parts, which are node-level attention, semantic-level attention and graph-level attention. MHADTI learns the feature representations of nodes with the hierarchical attention mechanisms which mainly contain three steps. In the first step, MHADTI learns the features of nodes with their meta-path-based neighbors based on GATs and gets their node-level embeddings. In the second step, MHADTI first measures the coefficient weights of features of nodes that come from the node-level and then obtains the semantic-level embeddings of nodes with the semantic-level attention. In the third step, MHADTI learns the coefficient weights of features of nodes that come from the semantic-level and gets the graph-level embeddings of nodes with graph-level attention. The output of MHADTI at the node-level is the input of the MHADTI at the semantic level, and the output of MHADTI at the semantic-level is the input of the MHADTI at the graph-level.

Final decoder

In this study, we first feed the learned drug–target embedding representations into MLP and then perform the element-wise multiplication on the embeddings of drugs and targets. Finally, the interaction probability scores $\hat{y}_{ij}$ for the input drug d_i and target t_j can be evaluated.

$$\hat{y}_{ij}=\sigma(\mathbf{W}^T(\mathbf{z}_{d_i}\odot\mathbf{z}_{t_j})), \tag{14}$$

where $\mathbf{z}_{d_i}$ and $\mathbf{z}_{t_j}$ denote the final embedding for drug d_i and target t_j. The operation $\odot$ denotes the element-wise multiplication for drug $\mathbf{z}_{d_i}\in R^{F'}$ and target $\mathbf{z}_{t_j}\in R^{F'}$ and $\mathbf{W}^T$ is the transpose of matrix $\mathbf{W}\in R^{F'\times 1}$. Besides, σ denotes the activation functions including *ReLU* and *Sigmoid*.

Loss function

In this study, we adopt the binary cross-entropy as the loss function to train MHADTI. The objection is to minimize loss $\mathcal{L}$, which

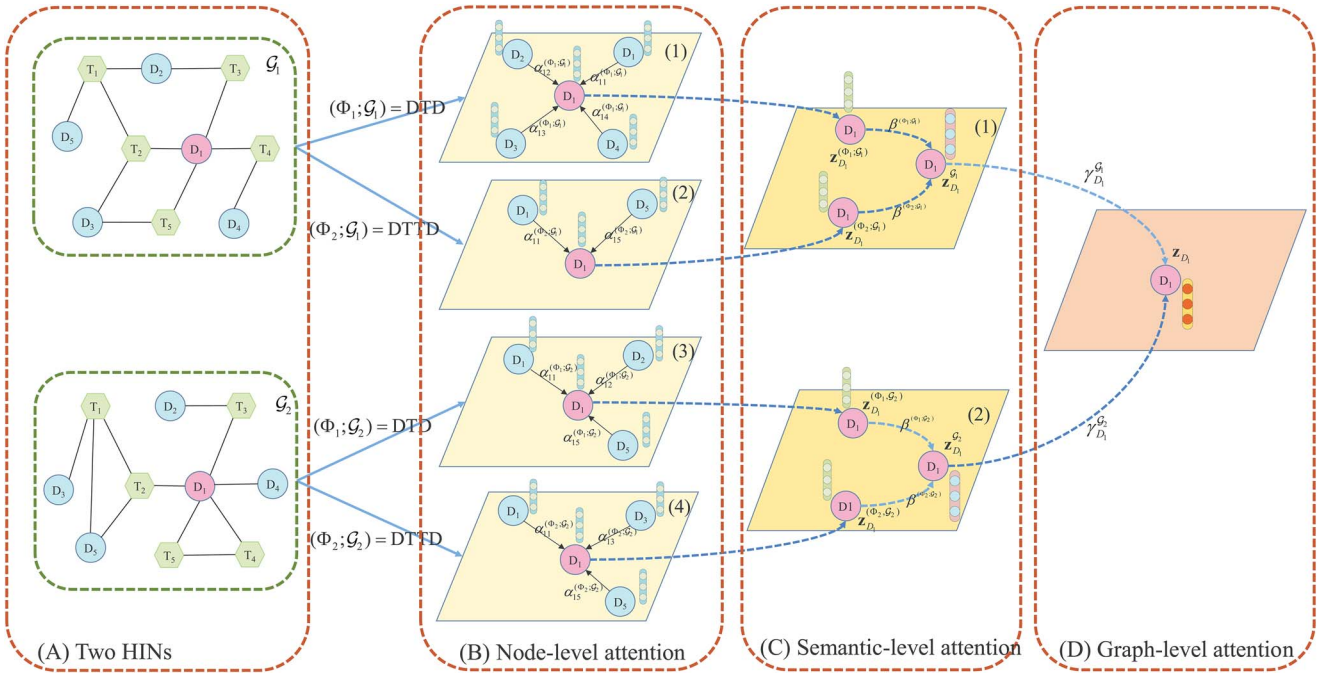

Figure 3. A toy example for learning the embedding of node D_1 with MHADTI. **(A)** Two HINs of D_1, which are constructed from different views. Φ_1 and Φ_2 denote two meta-paths, which are DTD and DTTD. **(B)** Embedding learning of D_1 at node-level attention. **(C)** Embedding learning of D_1 at semantic-level attention. **(D)** Embedding learning of D_1 at graph-level attention. Figure 3B denotes the neighbors of D_1 are generated under $(\Phi_1, \mathcal{G}_1)$, $(\Phi_2, \mathcal{G}_1)$, $(\Phi_1, \mathcal{G}_2)$ and $(\Phi_2, \mathcal{G}_2)$, respectively. Besides, $\alpha_{ij}^{(\Phi,\mathcal{G})}$ represents the weight of D_j to D_i under Φ in graph $\mathcal{G}$. $\beta^{(\Phi,\mathcal{G})}$ denotes the weight of Φ in graph $\mathcal{G}$. $\gamma_{D_1}^{\mathcal{G}}$ denotes the weight of D_1 in graph $\mathcal{G}$. $\mathbf{z}_{D_1}^{(\Phi,\mathcal{G})}$,$\mathbf{z}_{D_1}^{\mathcal{G}}$ and $\mathbf{z}_{D_1}$ denote the embeddings D_1 at node-level, semantic-level and graph-level, respectively. Some toys for learning the embeddings of some drugs and targets with different level attentions are displayed in Appendix B.

can be defined as

$$\mathcal{L} = -\sum_{(i,j)\in \mathbf{Y}^+ \cup \mathbf{Y}^-} y_{ij}\log \hat{y}_{ij} + (1-y_{ij})\log(1-\hat{y}_{ij}), \tag{15}$$

where (i,j) denotes the drug–target pair for drug d_i and target t_j, $\mathbf{Y}^+$ and $\mathbf{Y}^-$ represent the positive and negative drug–target pairs in the training set, respectively. If the drug–target pair $(i,j) \in \mathbf{Y}^+$, the ground truth y_{ij} is 1. If $(i,j) \in \mathbf{Y}^-$, the ground truth is 0. The prediction probability interaction score for drug d_i and target t_j is represented as $\hat{y}_{ij}$.

Complexity analysis of MHADTI

Given a HIN $\mathcal{G}$, MHADTI chooses one of the representative meta-paths named Φ for establishing the meta-path-based matrix $M_{\mathcal{G},\Phi}$. For one GAT attention head, its input feature dimension and out feature dimension are F and F', respectively, and its time complexity is $O(|V|EF' + |E|F')$, where $|V|$ and $|E|$ are the number of nodes in matrix $M_{\mathcal{G},\Phi}$ [38]. The number of attention heads in MHADTI is L, the time complexity will be $O(|V|EF'L + |E|F'L)$. Further, there are total K HINs and P meta-paths in MHADTI, the ultimate time complexity of Algorithm 1 is $O(KP|V|EF'L + KP|E|F'L)$.

In this study, there are three different HINs and four selected meta-paths. The number of the multi-heads L is 8. Therefore, the overall complexity is $O(|V|EF' + |E|F')$, which is linear to the number of nodes and edges in $M_{\mathcal{G},\Phi}$. The proposed model can be easily parallelized, because the node-level, semantic-level and graph-level attention can be parallelized. As a result, we can effectively achieve MHADTI.

Implementation details

To train MHADTI, we first assign a binary class label 0 or 1 to each drug–target pair in the training set. Specifically, drug–target pairs with known interactions will be labeled with 1, while the rest drug–target pairs in the training set are labeled with 0. The output for MHADTI is the interaction probability scores for drug–target pairs in the testing set.

In the construction process of multiview HINs (Figure 1B), there are 4358 drugs and 2407 targets in the DTI network. For measuring the fingerprint-based drug similarity, the fingerprint dimension for each drug is 167. For calculating the sequence-based target similarity, we set $\lambda = 1$ and each target is represented as one 40-dimensional vector. To fuse the three semantic similarity networks of targets, we adopt SNF method and the iteration number is six when SNF converges. In addition, different similarity thresholds are screened for the similarity networks of drugs and targets, respectively. The sparseness of these similarity networks ranges from 0.2% to 6%, which is consistent with the characteristics of biological networks [51].

In the embedding learning process of drugs and targets with hierarchical attention mechanisms (Figure 1C), MHADTI effectively implements multiple levels of attention components based on the PyTorch framework. On the whole, MHADTI contains two steps: one is the feature initialization step and the other is the hierarchical attention mechanism-based embedding learning step. The feature initialization step consists of two MLP layers, which are employed to encode the initial features of drugs and targets, respectively. The dropout value is 0.3. The initial dimensions for drugs and target are 881 and 40 and their output dimensions are both 128. In the hierarchical attention mechanism-based embedding learning step, the input dimension for node-level attention module is 128, the multi-head is set to

8 and the dropout value is also 0.3. The semantic-level attention vector q is set to 128, while the graph-level attention vector $\mathbf{c}$ is set to torch.tensor (128, 1).

The final decoder (Fig. 1D) contains three MLP layers. The input dimension is 128. The dropout value is 0.5. The final output of MHADTI is a scalar which indicates the prediction interaction probability score.

Besides, the parameters are initialized and optimized with Adam [52]. The learning rate is 0.005 and the regularization parameter is 0.0005. Furthermore, we adopt an early stop with a patience of 100. That is, MHADTI will stop training if the validation loss does not decrease in 100 consecutive epochs anymore. When compared with SOTA approaches, all methods are evaluated on the same training and test sets.

Algorithm 1 : The overall process of MHADTI

Input:

Heterogeneous information network set $\boldsymbol{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_K\}$,
Node feature $\{\mathbf{h}_{v_i}, \forall v_i \in V\}$,
Meta-path set $\{\Phi_0, \ldots, \Phi_p\}$,
Number of attention head L.

Output:

Final node embedding $\mathbf{Z}$,
Node-level attention weight α,
Semantic-level attention weight β,
Graph-level attention weight γ.

for $\mathcal{G}_k \in \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$ **do**
  Type-specific transformation $\mathbf{h}' \leftarrow \boldsymbol{M}_\phi \cdot \mathbf{h}$;
  for $\Phi_p \in \{\Phi_0, \ldots, \Phi_P\}$ **do**
    for $l = \{1, \ldots, L\}$ **do**
      for $v_i \in V$ **do**
        Find meta-path-based neighbors $\mathcal{N}_{v_i}^{(\Phi_p, \mathcal{G}_k)}$;
        for $v_j \in \mathcal{N}_{v_i}^{(\Phi_p, \mathcal{G}_k)}$ **do**
          Calculate weight coefficient $\alpha_{v_i v_j}^{(\Phi_p, \mathcal{G}_k)}$;
        end for
        Learn node-level embedding by $\mathbf{z}_{v_i}^{(\Phi_p, \mathcal{G}_k)} \leftarrow \sigma(\sum_{j \in \mathcal{N}_{v_i}^{(\Phi_p, \mathcal{G}_k)}} \alpha_{v_i v_j}^{(\Phi_p, \mathcal{G}_k)} \cdot \mathbf{h}_{v_j})$;
      end for
    end for
    Concatenate learned embeddings from all attention heads by $\mathbf{z}_{v_i}^{(\Phi_p, \mathcal{G}_k)} \leftarrow \Big\Vert_{l=1}^{L} \sigma(\sum_{j \in \mathcal{N}_i^{(\Phi_p, \mathcal{G}_k)}} \alpha_{v_i v_j}^{\Phi_p} \cdot \mathbf{h}_{v_j})$;
    Calculate weight $\beta^{(\Phi_p, \mathcal{G}_k)}$ of meta-path Φ_p in $\mathcal{G}_k$;
  end for
  Learn semantic-level node embedding by $\mathbf{Z}^{\mathcal{G}_k} \leftarrow \sum_{p=1}^{P} \beta^{(\Phi_p, \mathcal{G}_k)} \cdot \mathbf{Z}^{(\Phi_p, \mathcal{G}_k)}$;
  Calculate the graph-level attention weight vector $\gamma^{\mathcal{G}_k}$;
end for
Calculate the final embedding by $\mathbf{Z} \leftarrow \sum_{k=1}^{K} \gamma^{\mathcal{G}_k} \cdot \mathbf{Z}^{\mathcal{G}_k}$;
Calculate model losses by Cross-Entropy loss function;
Update model parameters via back propagation;
Return: $\mathbf{Z}, \alpha, \beta, \gamma$.

Results

In this section, we first give an introduction about the evaluation metrics used in this study. Then a comprehensive comparison with other competitive methods is displayed. After that, ablation experiments and parameters analysis experiments about MHADTI are invested. Lastly, we analyze the prediction results for some interested drugs and targets.

Experimental setup and evaluation metrics

In this study, we adopt the 5-fold-cross-validation (5-CV) strategy to evaluate the performance of MHADTI, which is similar to other approaches[53, 54]. Specifically, all the positive drug–target pairs are treated as the known positive samples and all the remained drug–target pairs are regarded as the negative samples. We randomly select the same number of negative samples as the positive samples. The positive and negative samples are together constructed as the experimental set.

In the 5-CV experiment, we divided the experimental set into five subsets with the same number. Each subset in turn is utilized as the testing subset, while the remaining four subsets are regarded as the training sample sets. The true positive (TP), false positive (FP), true negative (TN) as well as the false negative (FN) can be calculated, respectively. To reduce the data bias in the cross-validation experiments, we conduct five times on each prediction model to be evaluated and calculated their average values. All the comparison methods are performed with the 5-fold-cross-validation (5-CV), which is same to MHADTI. The detailed execution process of 5-CV strategy is shown in Figure 9 in Appendix C.

We mainly employ five widely used metrics, which are Accuracy (ACC), Area Under the Precision-Recall Curve (AUPRC), Area Under the Receiver Operating characteristic Curve (AUROC), Matthews Correlation Coefficient (MCC) and F1 score to evaluate the performance of the comparison methods as well as MHADTI. For a detailed introduction of these five metrics one can refer to reference [4] and here we do not repeat them anymore.

Comparison with other baseline methods

In this study, we select 12 state-of-the-art approaches and compare them with MHADTI model. These 12 approaches contain two molecular-feature-based methods: RF (Random Forests) [55], SVM(Support-vector machine) [56] ; two knowledge-graph-based models: TriModel [31], KGE_NFM [32] and eight GNN-based approaches: GCN [57], GAT [33], DTIGAT [58], DTICNN [59], DTIMGNN [23], EEGDTI [10], MK-TCMF [54] and SGCL-DTI [60].

To comprehensively evaluate the performance of MHADTI, we compare MHADTI with 12 baseline methods on our DrugBank dataset as well as other four commonly used datasets. These four datasets are Luo's dataset [27], Zheng's dataset [61], Yamanishi's dataset [62]and An's dataset [53]. The results of MHADTI and all the baseline approaches are presented in Tables 3–5, respectively.

- RF [55] is one of the ensemble learning methods for classification and its output is the class selected by most trees. We feed the features of drugs and targets for DTI predictions
- SVM [56] is a traditional supervised learning method. We feed the features of drugs and targets, respectively, into it directly for DTI predictions.
- TriModel [31] learns the embedding of drugs and proteins from the specific knowledge graph for DTI predictions.
- KGE_NFM [32] first obtains the low-dimensional representation for drugs and targets and then integrates other information via a neural factorization machine for DTI predictions.
- GCN [57] is a typical semi-supervised graph convolutional network. We feed the DTI network into GCNs and obtain the embeddings for drugs and targets to predict their interactions.
- GAT [33] is a semi-supervised neural network with the attention mechanism. We feed the DTI network into GATs and obtain the features of drugs and targets to complete the prediction task.

- DTIGAT [58] is the deep neural network method with the attention mechanisms, which could employ the interaction pattern and drug and target information.
- DTICNN [59] is a deep neural network-based method and it obtains essential features from the heterogeneous network to predict DTIs.
- DTIMGNN [23] constructs the topology graph and feature graph based on the drug–protein pair similarity and then predicts the DTIs via the multi-channel graph convolutional network with the graph attention mechanism.
- EEGDTI [10] is an end-to-end learning-based framework which could learn the comprehensive feature representations of drugs and targets based on graph convolutional networks.
- MK-TCMF [54] decomposes the original adjacency matrix into three matrices and then integrates the multiple kernel matrices to predict DTIs.
- SGCL-DTI [60] is a co-contrastive learning model which generates two different views and employs a contrastive loss to train the model for DTI predictions.

The results on our DrugBank dataset can be seen in Table 3; MHADTI achieves the best performance on all evaluation metrics. The results on ACC, AUC, AUPRC, MCC and F1 are 0.8801, 0.9522, 0.9352, 0.7631 and 0.8775, respectively. The AUC shows that MHADTI is effective in predicting a high number of positive samples and the value of AUPRC proves that MHADTI predicts a high percentage of correct DTIs. Besides, SGCL-DTI ranks second on ACC and AUC metrics, which are 0.8767 and 0.9356, while DTIGAT wins the second rank on AUPRC and F1 metrics, which are 0.9285 and 0.8731. Besides, TriModel wins the second rank on MCC and its score is 0.7365. Moreover, to get a more stable and fair comparison, we conduct the experiments five times for each prediction approach and adopt the mean values on each metric. The variance of the results on five-time experiments are also displayed. The results demonstrate that MHADTI is significantly superior to other baseline methods in predicting DTIs.

The results on Luo's dataset [27], Zheng's dataset [61] and Yamanishi's dataset [62] are presented in Table 4. For Luo's dataset, MHADTI achieves the best performance on both AUC and AUPRC metrics and the scores are 0.9655 and 0.9589, respectively. Besides, EEGDTI gets the second rank on AUC and GCN wins the second highest score on AUC and AUPRC, respectively. Their scores are 0.9550 and 9540, respectively. On Zheng's data, MHADTI model wins the highest scores on AUC and AUPRC metrics and the scores are 0.9469 and 0.9316, respectively. TriModel gets the second rank on AUC and AUPRC, of which sores are 0.9406 and 0.9256, respectively. Yamanishi's dataset mainly has four sub-datasets, which are GPCR, Enzyme, IC and NR, and the results are presented separately. Overall, MHADTI wins the highest scores on AUC of GPCR, AUC and AUPRC of IC and AUC of NR. Besides, DTIGAT wins the first rank on AUC and AUPRC on Enzyme dataset, while SGL-DTI obtains the highest scores on AUPRC of GPCR and AUC of NR. Other results about MHADTI as well as the comparison approaches have displayed in Table 4. Although MHADTI cannot get the first rank on each metric, the results in these datasets fully demonstrate that the performance of MHADTI outperforms other SOTA approaches. Further we analyze the generalization of MHADTI at the discussion section.

The results of An's dataset [53] are presented in Table 5. The results demonstrate that MHADTI wins the highest scores on ACC, AUC, MCC and F1 metrics, and their values are 0.9432, 0.9757, 0.9622 and 0.9464, respectively. MHADTI gets the second rank on MCC and the score is 0.9622. Besides, GCN gets a score of 0.9739 on AUPRC and wins the first rank. GAT gets the second rank on AUC and MCC, and their scores are 0.9710 and 0.8707, respectively. Meanwhile, SGCL-DTI ranks second on ACC and F1, and their scores are 0.9388 and 0.9343, respectively. The results on An's dataset demonstrate that MHADTI has an advantage over other comparison approaches in DTI predictions.

Experimental results of prediction approaches with different ratios between positive and negative samples

Different ratios between the number of positive and negative samples have much influence on the performance of MHADTI as well as the baseline approaches on our DrugBank dataset. Therefore, we conduct this experiment with different ratios (# positive samples: # negative samples) to investigate their effects on MHADTI and all the baseline methods. Specifically, we build three experimental datasets and each of them contains all the 15 252 positive samples. Meanwhile, these three experimental sets contain the 15 252 negative samples, 76 260 (15 252*5) negative samples and 152 520 (15 252*10) negative samples, respectively. Then, we perform the 5-CV experiment in terms of these three experimental datasets for all the prediction approaches in turn. The corresponding results are shown in Table 6.

From the results, we can find that MHADTI performs best on AUC metric under different ratios. The AUC values are 0.9522, 0.9463 and 0.9279 with 1:1, 1:5 and 1:10, respectively. MHADTI also wins the best performance on AUPRC metric when the ratio is 1:1 and the value is 0.9352. Meanwhile, KGN_NFM and DTICNN get the best performance on AUPRC metric when the ratios are 1:5 and 1:10, respectively, and the corresponding scores are 0.8045 and 0.6846, respectively. Besides, MHADTI wins the second rank on AUPRC with the 1:10 ratio, while SGCL-DTI gets the second highest scores on AUC metric with all rations and AUPRC metric under the 1:5 ratio. DITGAT also wins the second highest value on AUPRC under the 1:1 ratio. Overall, MHADTI achieves the best performance in this experiment.

Validation of the top-ranked prediction results for MHADTI

In the study, we employ the top-ranked prediction strategy [53] to further evaluate the performance of MHADTI. Specifically, we adopt the 5-CV experiment and the top 200 prediction samples in each folder are selected. Therefore, we could merge the results in each folder and get the top 1000 (200*5) prediction samples. Then these top 1000 predicted samples are filtered out and formed the validation set. The results are shown in Figure 4.

For MHADTI, the top 10 and 20 prediction results are all real DTIs. Besides, 49 out of 50, 99 out of 100, 195 out of 200, 487 out of 500 and 965 out of 1000 are real DTIs. The top-ranked prediction results fully demonstrate that MHADTI has desirable correctness and high credibility.

Ablation experiments

MHADTI learns the embeddings of drugs and targets via multi-view HINs with hierarchical attention mechanisms. Therefore, multiview HINs and hierarchical attention mechanisms have much effect on the performance of MHADTI. Here, two sets of ablation experiments are set up to investigate the effectiveness of each component from these two aspects.

The first ablation experiment is to verify the effectiveness of each HIN. A total of three HINs are constructed and employed

Table 3. The evaluation results of MHADTI and other baseline methods on our DrugBank dataset

Methods	Types	ACC	AUC	AUPRC	MCC	F1
RF [55]	MFP-based	0.8077 ± 0.0031	0.8281 ± 0.0028	0.8465 ± 0.0019	0.6881 ± 0.0010	0.7738 ± 0.0037
SVM [56]	MFP-based	0.7965 ± 0.0022	0.8032 ± 0.0045	0.8334 ± 0.0008	0.6701 ± 0.0024	0.7802 ± 0.0025
TriModel [31]	KGE-based	0.8523 ± 0.0026	0.9071 ± 0.0009	0.9142 ± 0.0025	0.7365 ± 0.0011	0.8367 ± 0.0018
KGE_NFM [32]	KGE-based	0.8686 ± 0.0032	0.9178 ± 0.0022	0.9052 ± 0.0014	0.7298 ± 0.0027	0.8222 ± 0.0040
GCN [57]	GNN-based	0.6318 ± 0.0109	0.8857 ± 0.0013	0.8961 ± 0.0021	0.2703 ± 0.0082	0.7789 ± 0.0034
GAT [33]	GNN-based	0.7872 ± 0.0076	0.8895 ± 0.0025	0.8611 ± 0.0045	0.6241 ± 0.0066	0.8426 ± 0.0058
DTIGAT [58]	GNN-based	0.8468 ± 0.0141	0.9254 ± 0.0015	0.9285 ± 0.0025	0.7144 ± 0.0035	0.8731 ± 0.0022
DTICNN [59]	GNN-based	0.8561 ± 0.0043	0.9124 ± 0.0037	0.9171 ± 0.0023	0.6801 ± 0.0042	0.8385 ± 0.0072
DTIMGNN [23]	GNN-based	0.8673 ± 0.0082	0.9221 ± 0.0008	0.9184 ± 0.0008	0.7021 ± 0.0041	0.8585 ± 0.0036
EEGDTI [10]	GNN-based	0.8352 ± 0.0098	0.8931 ± 0.0036	0.8821 ± 0.0051	0.6629 ± 0.0022	0.8512 ± 0.0017
MK_TCMF [54]	GNN-based	0.8324 ± 0.0017	0.8902 ± 0.0005	0.8723 ± 0.0029	0.6498 ± 0.0046	0.8497 ± 0.0011
SGCL-DTI [60]	GNN-based	0.8767± 0.0049	0.9356 ± 0.0011	0.9201 ± 0.0015	0.7288 ± 0.0019	0.8709 ± 0.0023
MHADTI(Ours)	GNN-based	**0.8801**± 0.0027	**0.9522**± 0.0031	**0.9352** ± 0.0041	**0.7631** ± 0.0009	**0.8775** ± 0.0062

Note: The best results are marked in bold and the second-best results are marked as underlined. MFR stands for molecular-fingerprint-based approaches and KGE denotes the knowledge-graph-embedding-based approaches.

Table 4. The evaluation results of MHADTI and other baseline methods on Luo's dataset, Zheng's dataset and Yamanishi's datasets.

Methods	Luo's data		Zheng's data		Yamanishi's data							
	AUC	AUPRC	AUC	AUPRC	GPCR		Enzyme		IC		NR	
					AUC	AUPRC	AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
RF [55]	0.8923	0.9391	0.8705	0.9133	0.8423	0.8502	0.8202	0.8351	0.8402	0.8229	0.8400	0.8323
SVM [56]	0.8874	0.9265	0.8805	0.9234	0.8009	0.8534	0.7886	0.8116	0.8200	0.8199	0.8378	0.8196
TriModel [31]	0.9342	0.9449	0.9406	0.9256	0.8734	0.8301	0.9088	0.9177	0.8434	0.8110	0.8455	0.8363
KGE_NFM [32]	0.9423	0.9228	0.9102	0.9006	0.8703	0.8403	0.8805	0.8733	0.9088	0.8534	0.8201	0.8455
GCN [57]	0.9533	0.9540	0.9087	0.9097	0.7658	0.7676	0.7594	0.7970	0.7947	0.8156	0.7049	0.7553
GAT [33]	0.9294	0.9108	0.8790	0.8625	0.7753	0.7680	0.8485	0.8335	0.9084	0.8816	0.8204	0.8082
DTIGAT [58]	0.9390	0.9232	0.7286	0.7213	0.7622	0.7649	0.9627	0.9613	0.7867	0.7778	0.9120	0.9032
DTICNN [59]	0.9077	0.9070	0.9199	0.9116	0.8543	0.8510	0.9335	0.9338	0.8918	0.8890	0.7333	0.7447
DTIMGNN [23]	0.9491	0.9325	0.9053	0.8734	0.8634	0.8552	0.9132	0.8488	0.8879	0.8320	0.8603	0.8429
EEGDTI [10]	0.9550	0.9339	0.9115	0.8883	0.8793	0.8432	0.9001	0.8436	0.8960	0.8553	0.8778	0.8655
MK_TCMF [54]	0.9077	0.8988	0.9122	0.8734	0.8043	0.8339	0.8634	0.8799	0.8311	0.8589	0.8737	0.8846
SGCL-DTI [60]	0.9496	0.9388	0.9388	0.9199	0.8787	0.8720	0.9315	0.9177	0.8989	0.8883	0.9323	0.8994
MHADTI(Ours)	0.9655	0.9589	0.9469	0.9316	0.8814	0.8596	0.9440	0.9373	0.9173	0.8948	0.9099	0.9150

The best results are marked in bold and the second-best results are marked as underlined.

Table 5. Experimental results on An's dataset for MHADTI and all the comparison methods

Methods	ACC	AUC	AUPRC	MCC	F1
RF [55]	0.8222	0.8762	0.9210	0.7633	0.8323
SVM [56]	0.8078	0.8633	0.9037	0.7321	0.7989
TriModel [31]	0.8634	0.9123	0.9043	0.8922	0.9273
KGE_NFM [32]	0.8529	0.9298	0.8834	0.8444	0.8935
GCN [57]	0.8031	0.9437	0.9739	0.6432	0.8683
GAT [33]	0.9345	0.9710	0.9583	0.8707	0.9306
DTIGAT [58]	0.7954	0.8603	0.8443	0.5980	0.8111
DTICNN [59]	0.9038	0.9449	0.9413	0.7683	0.8874
DTIMGNN [23]	0.8885	0.9104	0.9232	0.8183	0.8504
EEGDTI [10]	0.9338	0.9229	0.9499	0.8425	0.8993
MK_TCMF [54]	0.9218	0.9327	0.9316	0.8283	0.8662
SGCL-DTI [60]	0.9388	0.9544	0.9537	0.8638	0.9343
MHADTI(ours)	0.9432	0.9757	0.9622	0.8887	0.9464

Note: The best results are marked in bold and the second-best results are marked as underlined.

to learn the embeddings of drugs and targets in MHADTI. Without loss of generality, we call these three HINs as $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$, respectively. The corresponding definitions about these three HINs named $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ can be seen in the subsection of **Construction of Multiview HINs** and Figure 1B. Then we feed these three HINs into MHADTI with different combinations. Specifically, three HINs are divided into six different combinations, which are $\{\mathcal{G}_1\}$, $\{\mathcal{G}_2\}$, $\{\mathcal{G}_3\}$, $\{\mathcal{G}_1+\mathcal{G}_2\}$, $\{\mathcal{G}_1+\mathcal{G}_3\}$, $\{\mathcal{G}_2+\mathcal{G}_3\}$. These six HIN combinations are fed into the MHADTI in turn and we can obtain the prediction results which are displayed in Table 7. The ACC,

Table 6. The results of MHADTI as well as other baseline methods under different ratios on our DrugBank dataset (# positive samples: # negative samples=1:1,1:5 and 1:10)

Methods	1:1		1:5		1:10	
	AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
RF [55]	0.8281 ± 0.0028	0.8465 ± 0.0019	0.8206 ± 0.0018	0.7563 ± 0.0057	0.8019 ± 0.0017	0.5258 ± 0.0063
SVM [56]	0.8032 ± 0.0045	0.8334 ± 0.0008	0.8117 ± 0.0034	0.7504 ± 0.0077	0.7832 ± 0.0033	0.5016 ± 0.0102
TriModel [31]	0.9071 ± 0.0009	0.9142 ± 0.0025	0.9129 ± 0.0014	0.7763 ± 0.0066	0.8944 ± 0.0045	0.5863 ± 0.0086
KGE_NFM [32]	0.9178 ± 0.0022	0.9052 ± 0.0014	0.9209 ± 0.0024	**0.8045** ± 0.0027	0.9006 ± 0.0054	0.6614 ± 0.0063
GCN [57]	0.8857 ± 0.0013	0.8961 ± 0.0021	0.9294 ± 0.0012	0.7700 ± 0.0004	0.9105 ± 0.0024	0.6449 ± 0.0031
GAT [33]	0.8895 ± 0.0025	0.8611 ± 0.0045	0.8822 ± 0.0009	0.6906 ± 0.0017	0.8371 ± 0.0046	0.5621 ± 0.0022
DTIGAT [58]	0.9254 ± 0.0015	0.9285 ± 0.0025	0.8911 ± 0.0025	0.7472 ± 0.0008	0.8073 ± 0.0011	0.6598 ± 0.0047
DTICNN [59]	0.9124 ± 0.0037	0.9171 ± 0.0023	0.9099 ± 0.0024	0.7655 ± 0.0033	0.9109 ± 0.0040	**0.6846** ± 0.0055
DTIMGNN [23]	0.9221 ± 0.0008	0.9184 ± 0.0008	0.9301 ± 0.0015	0.7746 ± 0.0021	0.9015 ± 0.0057	0.6205 ± 0.0037
EEGDTI [10]	0.8931 ± 0.0036	0.8821 ± 0.0051	0.8806 ± 0.0035	0.7508 ± 0.0025	0.8534 ± 0.0014	0.6333 ± 0.0096
MK_TCMF [54]	0.8902 ± 0.0005	0.8723 ± 0.0029	0.8989 ± 0.0028	0.7795 ± 0.0102	0.8651 ± 0.0125	0.5302 ± 0.0108
SGCL-DTI [60]	0.9356 ± 0.0011	0.9201 ± 0.0015	0.9308 ± 0.0024	0.7936 ± 0.0066	0.9177 ± 0.0010	0.6135 ± 0.0100
MHADTI(ours)	**0.9522**± 0.0031	**0.9352** ± 0.0041	**0.9463**± 0.0012	0.7825 ± 0.0005	**0.9279** ± 0.0031	0.6779 ± 0.0087

The best results are marked in bold and the second-best results are marked as underlined.

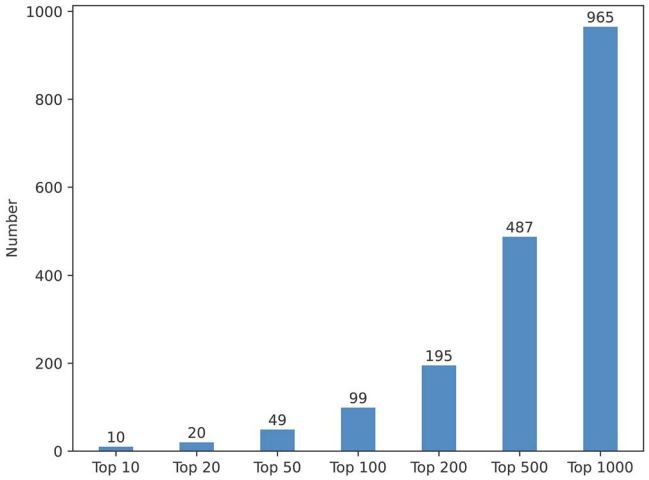

Figure 4. The number of DTIs verified in the top 1000 prediction results by MHADTI.

Table 7. The ablation experimental results on multi-view HINs for MHADTI

$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	ACC	AUC	AUPRC	MCC	F1
✓	✗	✗	0.7943	0.9212	0.9038	0.6138	0.7837
✗	✓	✗	0.7807	0.9138	0.8886	0.5850	0.7639
✗	✗	✓	0.7887	0.9160	0.8922	0.6000	0.7730
✓	✓	✗	0.7921	0.9277	0.9080	0.6124	0.7801
✓	✗	✓	0.8000	0.9191	0.8958	0.6184	0.7849
✗	✓	✓	0.8001	0.9292	0.9086	0.6236	0.7883
✓	✓	✓	**0.8801**	**0.9522**	**0.9352**	**0.7631**	**0.8775**

AUC, AUPRC, MCC and F1 are employed as the metrics to evaluate the performance. Besides, all the parameters in the experiment related MHADTI are consistent except for the input combination of HINs.

From the results in Table 7, we can see that MHADTI achieves the best performance in all the five metrics. The values for ACC, AUC, AUPRC, MCC and F1 are 0.8801,0.9522, 0.9352, 0.7631 and 0.8775, respectively. Besides, the prediction performance of MHADTI is improved with the increase of HIN number. The results

Table 8. The ablation experimental results on hierarchical attention mechanisms for MHADTI

NLA	SLA	GLA	ACC	AUC	AUPRC	MCC	F1
✗	✗	✗	0.6887	0.7011	0.7553	0.4255	0.6566
✓	✗	✗	0.8322	0.9451	0.9291	0.6842	0.8268
✓	✓	✗	0.8629	0.9501	0.9347	0.7356	0.8584
✓	✓	✓	**0.8801**	**0.9522**	**0.9352**	**0.7631**	**0.8775**

NLA, SLA and GLA denote the node-level attention, semantic-level attention and graph-level attention, respectively. The best results are in bold.

of taking two HINs as input is better than that of taking one HIN as input overall. The combination of $\{\mathcal{G}_2 + \mathcal{G}_3\}$ wins the second rank which are 0.8001, 0.9292, 0.9086, 0.6236 and 0.7883 on the corresponding evaluation metrics. However, comparing the combination $\{\mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3\}$, its performance is lower by 9.9%, 2.4%, 2.9%, 22.3% and 11.3% on ACC, AUC, AUPRC, MCC, and F1 respectively. In the results of taking one HIN as input $\mathcal{G}_1$ gets the best performance. The results in this set of experiments illustrate that three HINs are essential for MHADTI in predicting DTIs accurately.

The second ablation experiment is to evaluate the effectiveness of different level-attentions on MHADTI. The proposed hierarchical attention mechanisms contain node-level attention (NLA), semantic-level attention (SLA) and graph-level attention (GLA), respectively. Similar to the ablation experiment on HINs, we also split the attentions into three subsets, which are without all attentions, without SLA and GLA attentions and without GLA attention. Here, we still employ these five metrics to evaluate the performance of MHADTI.

For the results shown in Table 8 and Figure 5, with the addition of different level attentions, the performance of MHADTI is improved gradually. Specifically, the performance of MHADTI is worst when it does not employ any attention. In this case, the values on ACC, AUC, AUPRC, MCC and F1 are 0.6887, 0.7011, 0.7553, 0.4255 and 0.6566. Instead, MHADTI achieves the best performance when it adopts all the three level attentions. The ACC, AUC, AUPRC, MCC and F1 scores are 0.8801, 0.9522, 0.9352, 0.7631 and 0.8775, respectively. Considering more level attentions, the performance of MHADTI will achieve a better performance. Based on the results of this experiment, we can confirm that the

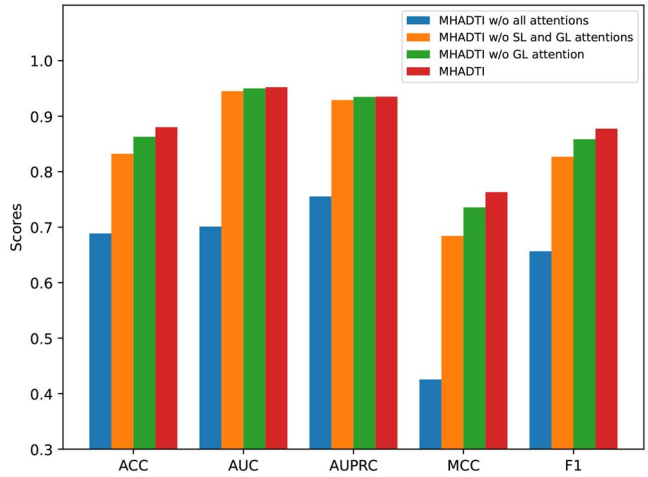

Figure 5. The ablation experimental results on hierarchical attention mechanisms for MHADTI.

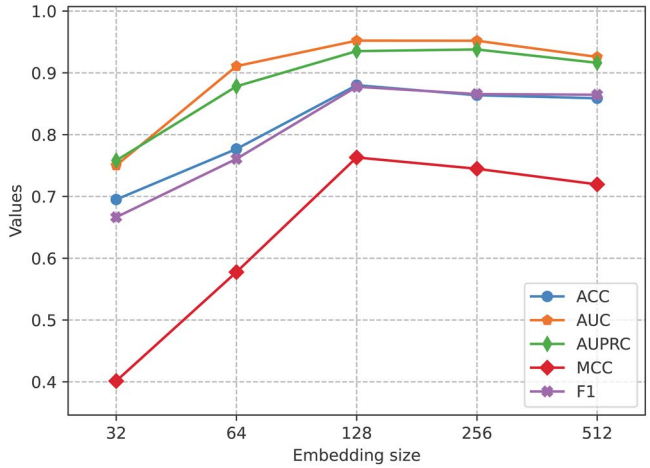

Figure 6. The values of ACC, AUC, APPRC, MCC and F1 under different embedding sizes. From the results, we can find out that MHADTI performs best when the embedding size is 128. Therefore, we adopt 128 as the embedding size for MHADTI.

hierarchical attention mechanism could improve the prediction accuracy of MHADTI effectively.

Node embedding size analysis

In this experiment, we investigate the effects of embedding size on MHADTI model and the results are shown in Figure 6. Here, the embedding size is 32, 64, 128, 256, and 512, respectively. MHADTI achieves the best performance when the embedding size is 128 overall. The values on ACC, AUC, MCC and F1 are highest when the embedding size is 128, which are 0.8801, 0.9522, 0.9352 and 0.8775, respectively. The value for AUPRC is highest when the embedding size is 256, which is 0.9379. It can be seen that the performance of MHADTI keeps getting better with the increase of embedding size from 32 to 128, while its performance decreases with the embedding size increasing from 128 to 512. As a result, we adopt the embedding size as 128 for MHADTI in this study.

Contributions of different meta-paths to MHADTI

In this study, MHADTI learns the semantic-level embeddings of drugs based on DTD and DTTD and learns the semantic-level embeddings of targets based on TDT and TDDT. These meta-paths have different semantic biological meanings. To fully evaluate the contribution of these meta-paths to MHADTI models in detail, we conduct an experiment on DrugBank dataset and the corresponding results are shown in Table 9. Particularly, MHADTI denotes that it contains all the four meta-paths.

Table 9. Results of the contribution of different meta-paths on MHADTI

	ACC	AUC	AUPRC	MCC	F1
MHADTI w/o DTTD	0.8376	0.9440	0.9283	0.6904	0.8305
MHADTI w/o DTD	0.8304	0.9292	0.9094	0.6737	0.8238
MHADTI w/o TDDT	0.8565	0.9476	0.9321	0.7204	0.8514
MHADTI w/o TDT	0.8137	0.9321	0.9141	0.6462	0.8050
MHDATI(ours)	**0.8801**	**0.9522**	**0.9352**	**0.7631**	**0.8775**

Note: MHADTI w/o DTTD, MHADTI w/o DTD, MHADTI w/o TDDT, MHADTI w/o TDT mean that MHADTI does not contain meta-path DTTD, DTD, TDDT and TDT, respectively.

From the results listed in the table, we can find that each meta-path plays an essential role on MHADTI. Specifically, meta-path DTD has a relatively great contribution to MHADTI than meta-path DTTD. The results for MHADTI w/o DTTD are 0.8376, 0.9440, 0.9283, 0.6904 and 0.8305, while the results for MHADTI w/o DTD are 0.8304, 0.9292, 0.9094,0.6737 and 0.8238 on ACC, AUC, AUPRC MCC and F1 metric, respectively. Besides, the result for meta-paths TDT and TDDT demonstrates that TDT has a relatively great contribution to MHADTI than meta-path TDDT. Based on the results, we can further confirm that each meta-paths have different contributions to the performance of MHADTI.

Results of similarity networks with different thresholds on MHADTI

In this study, we measure the similarity between drugs or targets from different aspects with their multisource information. Since many drugs or targets have higher similarities in the training and testing datasets, we adopt the similarity cutoff strategy on the training and testing sets to evaluate different similarity thresholds of drugs and targets on MHADTI [63]. The experiments are performed as follows:

First, we construct the similarity networks of drugs and targets. Then for the drug similarity networks, we calculate the different similarity cutoff thresholds as is mentioned in the reference [63]. Meanwhile, for the target similarity networks, we also measure the similarity cutoff thresholds in the same manner with drug similarity networks. After that, novel drug similarity networks and target similarity networks are both established under similarity cutoff thresholds. Lastly, we feed these novel drug similarity networks and target similarity networks into MHADTI to evaluate the performance of MHADTI.

Specifically, in the study [63], the similarity cutoff value is mean(**P**)+3×std(**P**), where **P** is the matrix representation of the drug (target) similarity networks, mean(**P**) denotes the mean value of matrix **P** and std(**P**) stands for the standard deviation of matrix **P**. Moreover, we conduct the experiments under four thresholds, which are mean(**P**), mean(**P**)+std(**P**), mean (**P**)+2×std(**P**) and mean(**P**)+3×std(**P**), respectively. A comparison report with similarity cutoff thresholds is shown in Table 10.

Results demonstrate that the different thresholds have effects on the performance of MHADTI. Specifically, when the threshold value is (mean(**P**)+std(**P**)), MHADTI performs best on ACC, MCC and F1. The corresponding scores are 0.9048, 0.8137 and 0.9103. When the threshold value is (mean(**P**)+2×std(**P**)), MHADTI performs best on AUC and AUPRC. The corresponding scores are 0.9689 and 0.9594, respectively. The last line in Table 10 is the

Table 10. Evaluation results of drug and target similarity networks with different thresholds on MHADTI

Cut-off thresholds	ACC	AUC	AUPRC	MCC	F1
mean(**P**)	0.8905	0.9520	0.9347	0.7851	0.8452
mean(**P**)+std(**P**)	**0.9048**	0.9565	0.9453	**0.8137**	**0.9103**
mean(**P**)+2×std(**P**)	0.8952	**0.9689**	**0.9594**	0.8015	0.9091
mean(**P**)+3×std(**P**)	0.8662	0.9530	0.9479	0.7156	0.8333
Without (ours)	0.8801	0.9522	0.9352	0.7631	0.8775

result of MHDATI without any similarity cutoff thresholds. We find that the experimental results of MHADTI have improved to a certain extent. The results of this experiment demonstrate and further confirm that cutoff thresholds of similarity networks of drugs and targets affect the performance of MHADTI.

Analysis of prediction results for common drugs and targets

Discovering the interactions accurately for some common drugs and targets is another effective manner to verify the effectiveness of DTI interaction prediction models [23]. In this subsection, we firstly select two representative drugs named Apomorphine and Fostamatinib and then analyzed the prediction results for these two drugs. Specifically, we employ all the known DTIs except for the interactions related to the candidate drug in the DTI network to train MHADTI and compare the prediction results with the ground-truth interactions. The corresponding results are shown in Table 11. For Apomorphine, the top-10 prediction DTIs are true interactions, while for Fostamatinib, nine of top-10 prediction DTIs are all true interactions.

Similarly, two common targets named ADRA1A and CYP19A1 are also employed to validate the effectiveness of MHADTI. The corresponding results are displayed in Table 12. The results demonstrate that 10 out of top-10 prediction interactions for ADRA1A and the nine out of top-10 prediction interactions for CYP19A1 are all true, which further illustrates the reliability of MHADTI.

Case study: Novel DTI predictions

Although MHADTI has a good performance on each benchmark dataset, it does not fully demonstrate that the proposed model is capable of effectively predicting novel DTIs. As a result, we execute the case study as Zhou [64] and Sameh [31] to further validate the ability of MHADTI in discovering novel DTIs. More importantly, case studies have been treated as an effective manner that employs MHADTI to solve practical problems.

Specifically, our case study experiment is performed on DrugBank dataset, which includes 15 252 DTIs involving 4358 drugs and 2407 targets. We firstly employ all the known DTIs to train MHADTI and then predict novel DTIs. After that, all the predicted DTIs are sorted according to the interaction probability values scored by MHADTI. We find that many predicted novel DTIs with high interaction probability scores are verified by the literature. Here, we select five novel predicted DTIs and the results are presented in Table 13.

The first one is the interaction between drug Pregabalin and gene CACNA2D1. Previous research identified CACNA2D1 as a gene modulator of intraocular pressure (IOP). In particular, drug pregabalin had a high affinity and selectivity with gene CACNA2D1, which played an important role in modulating IOP [65]. For gene CYP2D6, author Samer found that its enzyme product had an impact on oxycodone's metabolism and clinical efficacy. The activity of CYP2D6 was highly correlated with oxycodone experimental pain assessment [66]. Sunitinib was a drug which could interact with neurological diseases such as depression. Sunitinib was a predicted drug with gene prostaglandin-endoperoxide synthase 2 (PTGS2) by MHADTI model. It had been reported that gene PTGS2 was strongly associated with depression, which is supported by the published literature [67]. MHADTI uncovers a novel interaction between drug acetaminophen and gene POLE. A study found that acetaminophen could affect the expression of POLE mRNA by using hepatoma cells cultiv-atedinside a microfluidic biochip with or without acetaminophen [68]. Research showed that Gefitinib was the potential to be affected drug–drug interactions. For example, it was metabolized mainly by gene CYP3A5 and gene CYP3A5. CYP3A5 inhibitors or inducers may significantly alter their oral clearance and systemic or tumoral exposures[69]. All these five novel DTIs are identified by MHADTI and supported by the literature. These case studies fully could demonstrate the reliability of MHADTI in discovering interactions between drugs and targets.

Discussion

In this section, we will discuss two problems, one is the failure cases when establishing MHADTI model and the other is the generalization and applicability of MHADTI.

Failure cases for establishing MHADTI

In the process of building the MHADTI model, we encounter some failure cases and make an improvement on these aspects.

The first failure case is the number of layers at the hierarchical attention mechanisms. At first, MHADTI only employs node-level attention and semantic level attention like the HAN model. However, the performance of the proposed model is inferior to other baseline methods, such as SGCL-DTI [60]. Inspired by the multiview methods and coupled with the multisource of drugs and targets, we established three multiview drug–target HINs and employed the graph-level attention in MHADTI. With the hierarchical attention mechanisms, MHADTI could learn the embedding of drugs and targets from different views and achieve a relatively satisfactory result.

The second failure case is the selection of meta-paths. As we know, since meta-paths could capture complex relationships that effectively reveal structural and semantic information in HINs, various meta-paths will imply different semantic meanings well. At first, we only select two meta-paths, which are DTD and TDT for the proposed model. However, the result of MHADTI is the worst of all the baseline methods since MHADTI could not learn the embeddings with semantic-level attention. Hence, two other meta-paths which are DTTD and TDDT are applied to MHADTI. The performance of MHADT has great improvement. In the future, more meta-paths with rich semantic information can also be applied to the proposed model to further improve the performance of the model.

Besides, some parameters, such as the number of attention heads L and the feature dimension of drugs and targets, also need to tune based on the experimental results. For MHADTI, the number of attention head L is 8 and the dimension of drugs and targets is 128, ultimately.

Table 11. Prediction results for Apomorphine and Fostamatinib with MHADTI

Drug ID	Drug name	Target ID	Target name	Result
DB00714	Apomorphine	P02768	ALB	True
		P28335	HTR2C	True
		P28221	HTR1D	True
		P28222	HTR1B	True
		P49888	SULT1E1	True
		P08908	HTR1A	True
		P18825	ADRA2C	True
		O43704	SULT1B1	True
		Q05940	SLC18A2	True
		P50226	SULT1A2	True
DB12010	Fostamatinib	P17612	PRKACA	True
		Q9Y616	IRAK3	True
		P41240	CSK	True
		Q06418	TYRO3	True
		P53671	LIMK2	True
		Q9BYT3	STK33	True
		O43353	RIPK2	True
		Q06418	TYRO3	True
		P08684	CYP3A4	False
		P07332	FES	True

Table 12. Prediction results for ADRA1A and CYP19A1 with MHADTI

Target ID	Target name	Drug ID	Drug name	Result
P35348	ADRA1A	DB00502	Haloperidol	True
		DB00368	Norepinephrine	True
		DB01624	Zuclopenthixol	True
		DB00875	Flupentixol	True
		DB00777	Propiomazine	True
		DB01295	Bevantolol	True
		DB06144	Sertindole	True
		DB00610	Metaraminol	True
		DB00334	Olanzapine	True
		DB00935	Oxymetazoline	True
P11511	CYP19A1	DB00481	Raloxifene	True
		DB00655	Estrone	True
		DB00333	Methadone	True
		DB00856	Chlorphenesin	True
		DB00858	Drostanolone	True
		DB00624	Testosterone	True
		DB06147	Sulfathiazole	True
		DB14598	Anhydrous	True
		DB01406	Danazol	True
		DB00116	Tetrahydrofolate	False

The generalization and applicability of MHADTI

To fully demonstrate the generality of MHADTI, we perform the experiments on five different datasets and compare MHADTI with other SOTA approaches. Results demonstrate that MHADTI performs best on our DrugBank dataset, Luo's dataset [27], Zheng's dataset [61]. Meanwhile, on Yamanishi's dataset [62] and An's dataset [53], MHADTI does not get the highest scores on each evaluation metric. As a result, we would like to give a discussion about the generalization of MHADTI.

Firstly, the framework of MHADTI shows that it needs to construct multiview HINs with the multisource of drugs and targets. The information mainly contains drug fingerprints, drug side effects and target sequences and so on. Therefore, the rich information about drugs and targets is the data foundation of MHADTI. As we know, some of the drugs and targets need not always be annotated by all kinds of information. For example, in the Gene Ontology Annotation dataset, some targets do not have their annotation information, which means that we could not measure their GOA-based similarities. Besides, there are some drugs whose molecular structure does not exist. The reliability of the multiview HINs will be low, which may affect the performance of MHADTI. In practice, drugs and targets in our DrugBank datasets have rich annotation information than those in Yamanishi's dataset [62], which further verifies our assumption.

Secondly, the size of the heterogeneous network will affect the performance of MHADTI. MHADTI learns the embedding of drugs and targets from the meta-based neighbors with the node-level attention. If the size of the heterogeneous network is small, nodes

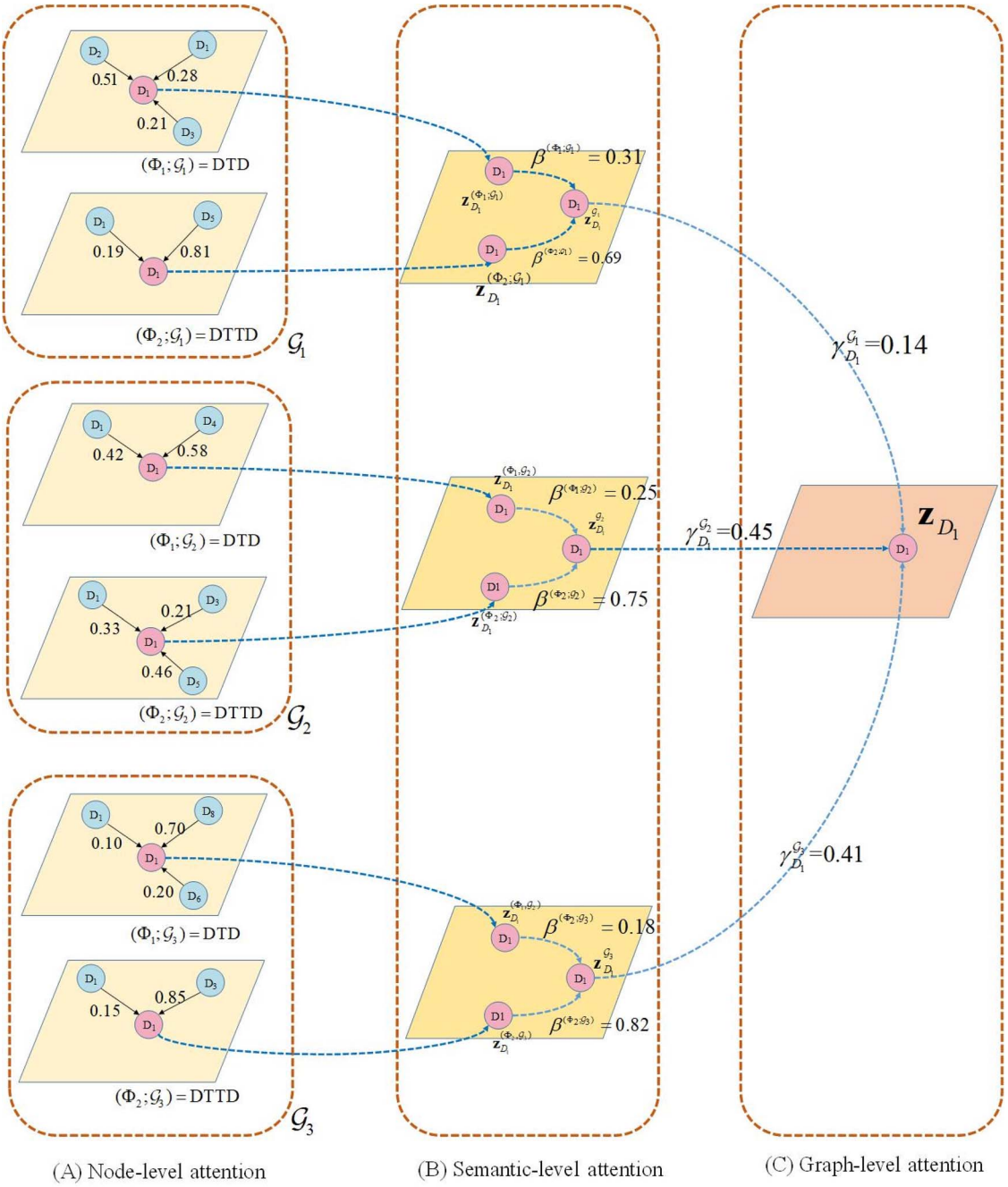

Figure 7. A toy example for learning the embedding representation of drug D1 (named Cariprazine) with MHADTI.

will have fewer meta-based-neighbors or even no meta-based-neighbor nodes. MHADTI could not learn the embedding at the node-level. For example, there are only 54 drugs and 26 targets in the NR of Yamanishi's dataset [62]. The performance of MHADTI on NR of Yamanishi's dataset is inferior to its performance on Luo's and Zheng's datasets (see Table 4).

In summary, MHADTI could achieve competitive results on higher quality and larger size multiview HINs.

Conclusion

Accurately identifying the DTIs is an essential step in drug discovery and drug repositioning. In this study, we propose a novel computational model called MHADTI to predict DTIs. Firstly, MHADTI adopts the multisource of drugs and targets to construct their similarity networks and establishes three drug–target HINs from different views. Then MHADTI learns the embeddings of drugs and targets from HINs with the hierarchical attention mechanisms which include the node-level, semantic-level and graph-level attentions. Lastly, MHADTI employs MLP to predict DTIs based on deep feature representations of drugs and targets.

To evaluate the performance of MHADTI, we conduct the 5-CV experiment and compare it with eight other state-of-the-art approaches. Experimental results demonstrate that MHADTI achieves the best performance on five evaluation metrics overall. Ablation and parameter sensitivity experiments are performed to obtain the best parameters of MHADTI. Analysis of prediction results for some interested drugs further exhibits the reliability of

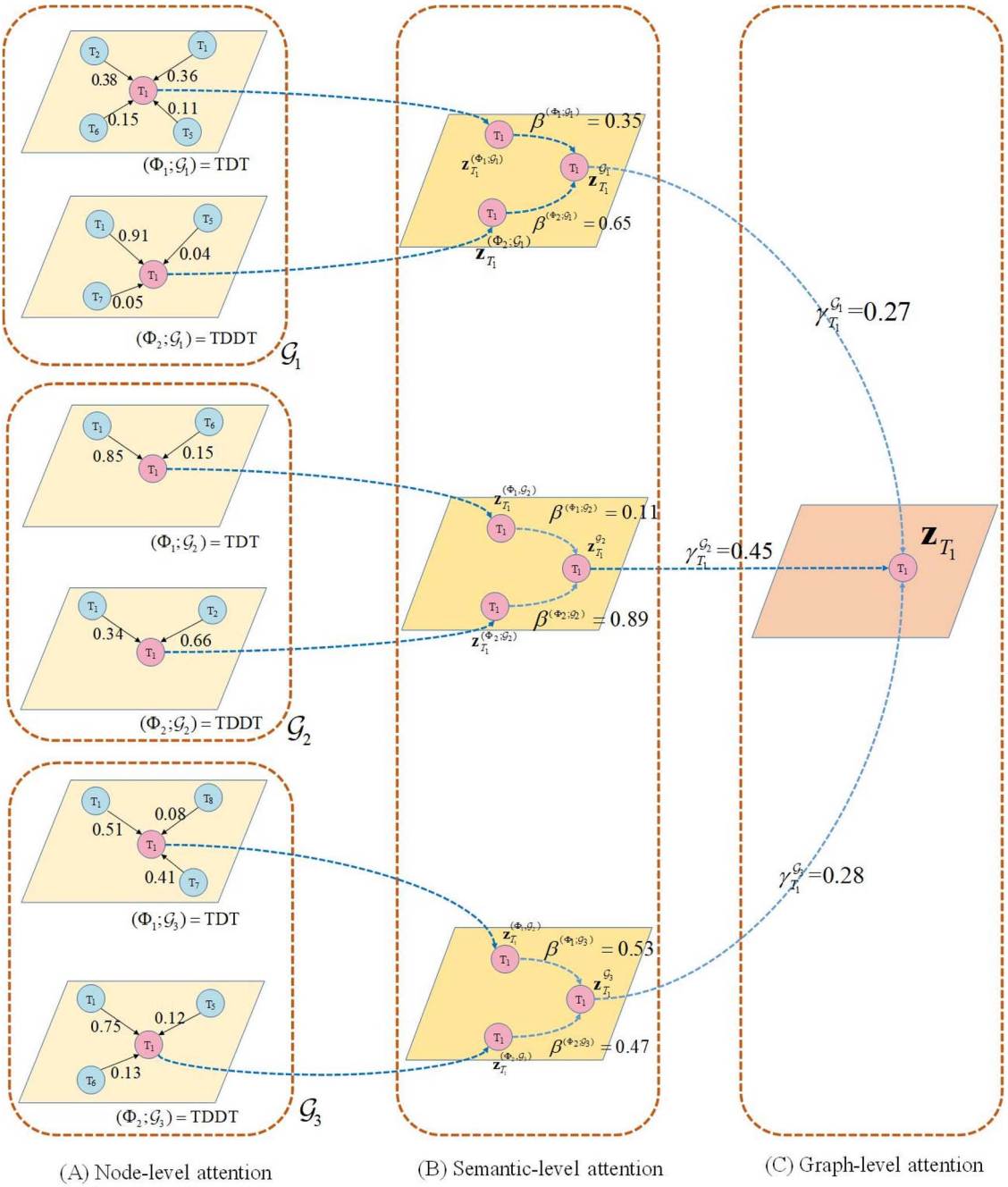

Figure 8. A toy example for learning the embedding representation of target T1 (named KLKB1) with MHADTI.

Table 13. Predicted novel DTIs supported by the literature

DrugID	DrugName	TargetID	TargetName	Scores	PMID
DB00230	Pregabalin	P54289	CACNA2D1	0.952	31714057
DB00497	Oxycodone	P10635	CYP2D6	0.934	20590588
DB01268	Sunitinib	P35354	PTGS2	0.973	31423209
DB00316	Acetaminophen	Q07864	POLE	0.977	22230336
DB00317	Gefitinib	P20815	CYP3A5	0.982	15900286

Note: Scores denote the predicted DTI interaction probability values by MHADTI.

MHADTI. Moreover, the code and data of MHADTI are uploaded on the Github, which is convenient to make a comparison and improvement.

In the future, we can do some work from the following aspects. First, we need to build a more reliable negative training set. In this study, we randomly selected the same number of negative

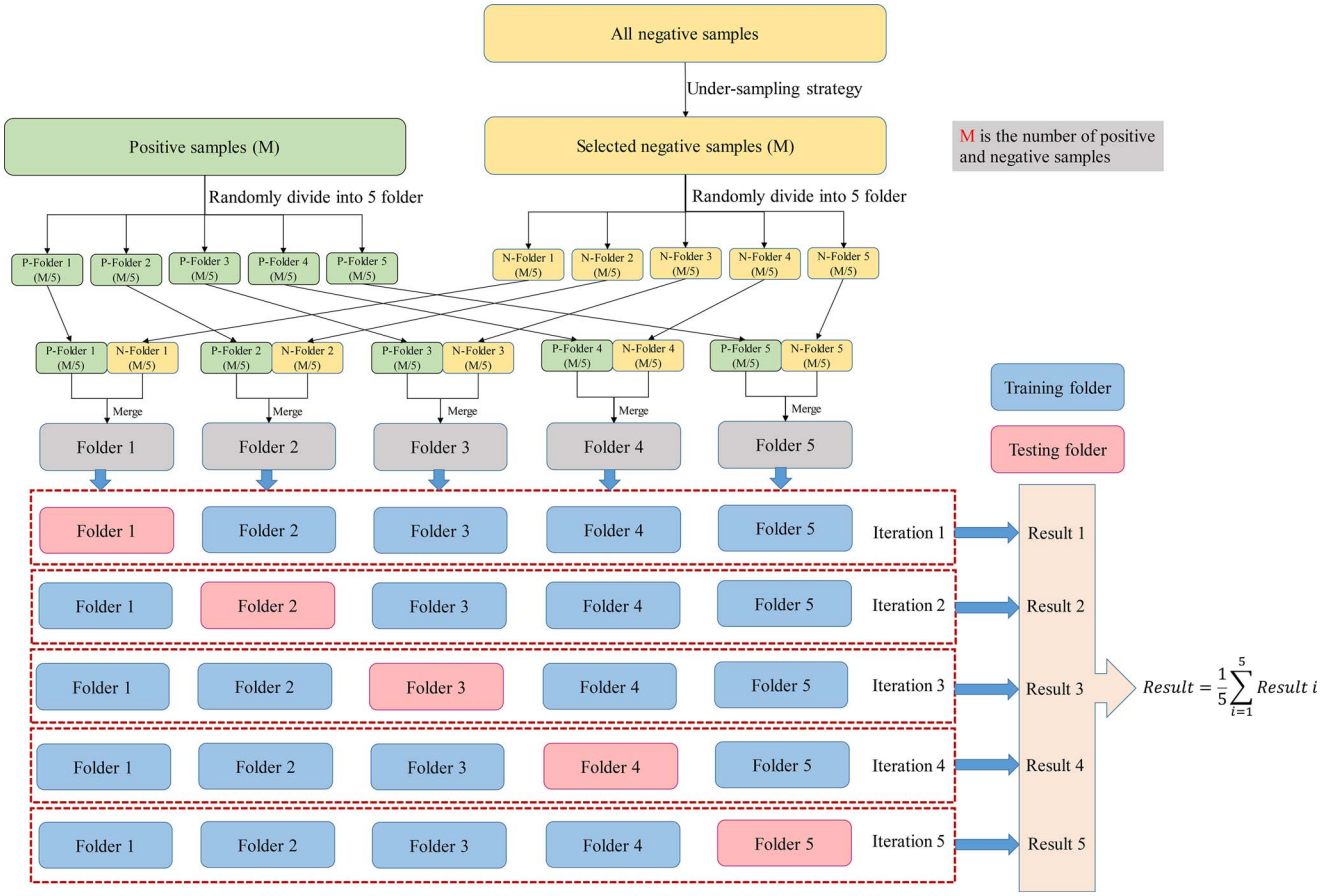

Figure 9. The 5-folder cross-validation strategy used in this study.

drug–target pairs as positive pairs. The negative drug–target pairs in the training set may be not reliable, which have a great influence on the performance of prediction models. Secondly, we select four representative meta-paths and learn the semantic-level embeddings for drugs and targets. More meta-paths with rich biological meaning should be selected for learning the embedding representations. Lastly, we can apply MHADTI to other link prediction problems such as miRNA–disease and drug–miRNA association prediction.

Key Points

- MHADTI evaluates the similarities of drugs and targets with their multisource information and constructs multiview HINs for learning their embeddings.
- MHADTI could learn the embeddings of drugs and targets with the hierarchical attention mechanisms which include node-level, semantic-level and graph-level attentions, respectively.
- MHADTI learns the optimal combination of meta-path-based neighbors, multi-meta-paths and multi-graphs in the hierarchical manner, which could better capture the complex structure and rich semantic information in different HINs.
- The evaluation results demonstrate that MHADTI outperforms other state-of-the-art approaches in DTI.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions.

Funding

This work is supported by funds from the National Science Foundation of China (NO.61801432, 62003308, 62176239 and 61701073).

Author contributions statement

Z.T. and X.P. developed the codes, conceived the experiment and drafted the whole manuscript together. T.Z. and H.F. set up the general idea of this study. W.Z. and Q.D revised this manuscript. Y.Y. gave some advice. All authors have read and approved the manuscript.

References

1. Yuan Q, Gao J, Dongliang W, *et al.* Druge-rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 2016;**32**(12): i18–27.
2. Whitebread S, Hamon J, Bojanic D, *et al.* Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov Today* 2005;**10**(21):1421–33.
3. Hopkins AL. Predicting promiscuity. *Nature* 2009;**462**(7270): 167–8.

4. Mahmud SMH, Chen W, Liu Y, *et al.* Predtis: prediction of drug–target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Brief Bioinform* 2021;**22**(5):bbab046.
5. Bagherian M, Sabeti E, Wang K, *et al.* Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Brief Bioinform* 2021;**22**(1):247–69.
6. Keiser MJ, Roth BL, Armbruster BN, *et al.* Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;**25**(2):197–206.
7. Cheng AC, Coleman RG, Smyth KT, *et al.* Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 2007;**25**(1):71–5.
8. Sydow D, Burggraaff L, Szengel A, *et al.* Advances and challenges in computational target prediction. *J Chem Inf Model* 2019;**59**(5):1728–42.
9. Xuan P, Fan M, Cui H, *et al.* Gvdti: graph convolutional and variational autoencoders with attribute-level attention for drug–protein interaction prediction. *Brief Bioinform* 2022;**23**(1):bbab453.
10. 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;**22**(5):bbaa430.
11. Zhou D, Zhijian X, Li WT, *et al.* Multidti: drug–target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network. *Bioinformatics* 2021;**37**(23):4485–92.
12. He T, Heidemeyer M, Ban F, *et al.* Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J Chem* 2017;**9**(1):1–14.
13. Öztürk H, Özgür A, Ozkirimli E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* 2018;**34**(17):i821–9.
14. Yang Z, Weihe Zhong L, Zhao, and Calvin Yu-Chian Chen. Mldti: mutual learning mechanism for interpretable drug–target interaction prediction. *The Journal of Physical Chemistry Letters* 2021;**12**(17):4247–61.
15. Nguyen T, Le H, Quinn TP, *et al.* Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics* 2021;**37**(8):1140–7.
16. Yang Z, Zhong W, Lu Z, *et al.* Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chem Sci* 2022;**13**(3):816–33.
17. Lv Q, Chen G, Lu Z, *et al.* Mol2context-vec: learning molecular representation from context awareness for drug discovery. *Brief Bioinform* 2021;**22**(6):bbab317.
18. Lee I, Keum J, Nam H. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput Biol* 2019;**15**(6):e1007129.
19. Chu Y, Kaushik AC, Wang X, *et al.* Dti-cdf: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief Bioinform* 2021;**22**(1):451–62.
20. Li Z-C, Huang M-H, Zhong W-Q, *et al.* Identification of drug–target interaction from interactome network with 'guilt-by-association'principle and topology features. *Bioinformatics* 2016;**32**(7):1057–64.
21. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009;**25**(18):2397–403.
22. Mei J-P, Kwoh C-K, Yang P, *et al.* Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 2013;**29**(2):238–45.
23. Li Y, Qiao G, Wang K, *et al.* Drug–target interaction predication via multi-channel graph neural networks. *Brief Bioinform* 2022;**23**(1):bbab346.
24. Kaimiao H, Cui H, Zhang T, *et al.* ALDPI: adaptively learning importance of multi-scale topologies and multi-modality similarities for drug–protein interaction prediction. *Brief Bioinform* 2022;**23**(2):bbab606.
25. Zeng X, Zhu S, Hou Y, *et al.* Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* 2020;**36**(9):2805–12.
26. Chen X, Liu M-X, Yan G-Y. Drug–target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst* 2012;**8**(7):1970–8.
27. Luo Y, Zhao X, Zhou J, *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017;**8**(1):1–13.
28. Wan F, Hong L, Xiao A, *et al.* Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* 2019;**35**(1):104–11.
29. Zheng S, Rao J, Song Y, *et al.* Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. *Brief Bioinform* 2021;**22**(4):bbaa344.
30. Wang X, Yang Y, Li K, *et al.* Bioerp: biomedical heterogeneous network-based self-supervised representation learning approach for entity relationship predictions. *Bioinformatics* 2021;**37**(24):4793–800.
31. Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* 2020;**36**(2):603–10.
32. Ye Q, Hsieh C-Y, Ziyi Yang Y, *et al.* A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nat Commun* 2021;**12**(1):1–12.
33. Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks[J]. arXiv preprintarXiv:1710.10903. 2017. https://arxiv.org/abs/1710.10903.
34. Gao J, Gao J, Ying X, *et al.* Higher-order interaction goes neural: A substructure assembling graph attention network for graph classification. *IEEE Transactions on Knowledge and Data Engineering* 2021, 1–1.
35. Wu Q, Zhang H, Gao X, *et al.* Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. In *The World Wide Web Conference, WWW'19*. Association for Computing Machinery, New York, NY, USA, 2019;2091–102.
36. Zhao Q, Zhao H, Zheng K, *et al.* Hyperattentiondti: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics* 2022;**38**(3):655–62.
37. Long Y, Wu M, Liu Y, *et al.* Ensembling graph attention networks for human microbe–drug association prediction. *Bioinformatics* 2020;**36**(Supplement_2):i779–86.
38. Wang X, Ji H, Shi C, *et al.* Heterogeneous graph attention network. In: *The World Wide Web Conference, WWW'19*. Association for Computing Machinery, New York, NY, USA, 2019, 2022–32.
39. Xing X, Yang F, Li H, *et al.* Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. *Bioinformatics* 2022;**38**(8):2178–86.
40. Zhao X, Zhao X, Yin M. Heterogeneous graph attention network based on meta-paths for lncrna-disease association prediction. *Brief Bioinform* 2022;**23**(1):bbab407.
41. Shang Y, Ye X, Futamura Y, *et al.* Multiview network embedding for drug-target interactions prediction by consistent and complementary information preserving. *Brief Bioinform* 2022;**23**(3):bbac059.

42. Li Y, Fang-Xiang W, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2018;**19**(2):325–40.
43. Tang X, Luo J, Shen C, *et al.* Multi-view multichannel attention graph convolutional network for mirna-disease association prediction. *Brief Bioinform* 2021;**22**(6):bbab174.
44. Haitao F, Huang F, Liu X, *et al.* Mvgcn: data integration through multi-view graph convolutional network for predicting links in biomedical bipartite networks. *Bioinformatics* 2022;**38**(2):426–34.
45. Yuan K, Zhang Y, Li Y, *et al.* A knowledge-enhanced multi-view framework for drug-target interaction prediction. *IEEE Transactions on Big Data* 2021;**8**(5):1387–98.
46. Wei L, Long W, Wei L. Mdl-cpi: Multi-view deep learning model for compound-protein interaction prediction. *Methods* 2022; **204**:418–27.
47. Zhao X, Zhao X, Yin M. Heterogeneous graph attention network based on meta-paths for lncRNA-disease association prediction. *Brief Bioinform* 2021; **23**(1):bbab407.
48. 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):D1074–82.
49. Kim S, Thiessen PA, Bolton EE, *et al.* Pubchem substance and compound databases. *Nucleic Acids Res* 2016;**44**(D1):D1202–13.
50. UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**(D1):D506–15.
51. Arita M. Scale-freeness and biological networks. *J Biochem* 2005; **138**(1):1–4.
52. Kingma DP, Ba J. Adam: A method for stochastic optimization-arXiv preprint arXiv:1412.6980. 2014.
53. An Q, Liang Y. A heterogeneous network embedding framework for predicting similarity-based drug-target interactions. *Brief Bioinform* 2021;**22**(6):bbab275.
54. Ding Y, Tang J, Guo F, *et al.* Identification of drug–target interactions via multiple kernel-based triple collaborative matrix factorization. *Brief Bioinform* 2022;**23**(2).
55. Cao D-S, Liang Y-Z, Deng Z, *et al.* Genome-scale screening of drug-target associations relevant to ki using a chemogenomics approach. *PloS one* 2013;**8**(4):e57680.
56. Hua Y, Chen J, Xue X, *et al.* A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PloS one* 2012;**7**(5):e37608.
57. Kipf TN, Wellin M. Semi-supervised classification with graph convolutional networksarXiv preprint arXiv:1609.02907. 2016.
58. Wang H, Zhou G, Liu S, *et al.* Drug-target interaction prediction with graph attention networksarXiv preprint arXiv:2107.06099. 2021.
59. 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 bioinformatics* 2020;**21**(13):1–13.
60. Yang L, Qiao G, Gao X, *et al.* Supervised graph co-contrastive learning for drug-target interaction prediction. *Bioinformatics* 2022;btac164.
61. Zheng Y, Peng H, Zhang X, *et al.* Predicting drug targets from heterogeneous spaces using anchor graph hashing and ensemble learning. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, 1–7.
62. Yamanishi Y, Araki M, Gutteridge A, *et al.* Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;**24**(13):i232–40.
63. Li S, Wan F, Shu H, *et al.* Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems* 2020;**10**(4):308–22.
64. Zhou M, Chen Y, Rong X. A drug-side effect context-sensitive network approach for drug target prediction. *Bioinformatics* 2019;**35**(12):2100–7.
65. Ibrahim MM, Maria DN, Mishra SR, *et al.* Once daily pregabalin eye drops for management of glaucoma. *ACS Nano* 2019; **13**(12):13728–44.
66. Samer CF, Daali Y, Wagner M, *et al.* Genetic polymorphisms and drug interactions modulating cyp2d6 and cyp3a activities have a major effect on oxycodone analgesic efficacy and safety. *Br J Pharmacol* 2010;**160**(4):919–30.
67. Luo L, Liang Y, Ding X, *et al.* Significance of cyclooxygenase-2, prostaglandin e2 and cd133 levels in sunitinib-resistant renal cell carcinoma. *Oncol Lett* 2019;**18**(2):1442–50.
68. Prot J-M, Bunescu A, Elena-Herrmann B, *et al.* Predictive toxicology using systemic biology and liver microfluidic 'on chip' approaches: application to acetaminophen injury. *Toxicol Appl Pharmacol* 2012;**259**(3):270–80.
69. Li J, Zhao M, He P, *et al.* Differential metabolism of gefitinib and erlotinib by human cytochrome p450 enzymes. *Clin Cancer Res* 2007;**13**(12):3731–7.
70. Landrum G. Rdkit documentation. *Release* 2013;**1**(1–79):4.
71. Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 2011;**27**(21):3036–43.
72. Chen X, Yan G-Y. Novel human lncrna–disease association inference based on lncrna expression profiles. *Bioinformatics* 2013;**29**(20):2617–24.
73. Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res* 2019;**47**(D1):D330–8.
74. Teng Z, Guo M, Liu X, *et al.* Measuring gene functional similarity based on group-wise comparison of go terms. *Bioinformatics* 2013;**29**(11):1424–32.
75. Tian Z, Fang H, Ye Y, *et al.* A novel gene functional similarity calculation model by utilizing the specificity of terms and relationships in gene ontology. *BMC bioinformatics* 2022;**23**(1): 1–14.
76. Wang B, Mezlini AM, Demir F, *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014; **11**(3):333–7.
77. Shi H, Liu S, Chen J, *et al.* Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure. *Genomics* 2019;**111**(6):1839–52.

APPENDIX A: The similarity calculation for drugs and targets

Drug similarity calculation

In this research, drug similarities are evaluated from three aspects which are side effects, molecular fingerprints and Gaussian Interaction Profile (GIP) kernel, respectively.

Each drug has different side effects which can formulate its drug side effect annotation set. Suppose there are two drugs named d_i and d_j, their corresponding drug side effect annotation sets are $\mathbf{S}_{d_i}$ and $\mathbf{S}_{d_j}$ and the drug side-effect-based similarity between d_i and d_j can be denoted as

$$Sim_{sideeffect}(d_i, d_j) = \frac{1}{2} \cdot \left(\frac{|\mathbf{S}_{d_i} \cap \mathbf{S}_{d_j}|}{|\mathbf{S}_{d_i}|} + \frac{|\mathbf{S}_{d_i} \cap \mathbf{S}_{d_j}|}{|\mathbf{S}_{d_j}|} \right) \tag{16}$$

For molecular fingerprint-based similarity, we firstly collect the SMILES of drugs and then obtain their 167 bits fingerprint vectors with RDKit software [70]. The name of the molecular fingerprint is MACCS key. Suppose there are two drugs named d_i and d_j, and their molecular fingerprint vectors be denoted as $\mathbf{V}_{d_i}$ and $\mathbf{V}_{d_j}$, respectively; the fingerprint-based cosine similarity can be formulated as

$$Sim_{fingerprint}(d_i, d_j) = \frac{\mathbf{V}_{d_i} \cdot \mathbf{V}_{d_j}}{\sqrt{(\mathbf{V}_{d_i})^2} \cdot \sqrt{(\mathbf{V}_{d_j})^2}} \tag{17}$$

Besides, the drug similarity can also be evaluated based on GIP kernel [71]. Suppose there is one DTI network; its matrix representation can be expressed as $\mathbf{M}$. The GIP kernel-based similarity between drugs can be formulated as

$$KD(d_i, d_j) = \exp\left(-\delta_d \| \boldsymbol{IP}(d_i) - \boldsymbol{IP}(d_j) \|^2\right), \tag{18}$$

where $\boldsymbol{IP}(d_i)$ and $\boldsymbol{IP}(d_j)$ represent the corresponding row for drug d_i and d_j in the matrix $\mathbf{M}$. The greater the difference between the rows belonging to drug d_i and d_j, the smaller the $KD(d_i, d_j)$, and vice versa. Parameter δ_d is widely employed to control the kernel bandwidth in the research [72], which can be expressed as

$$\delta_d = \delta'_d / \left(\frac{1}{n_d} \sum_{k=1}^{n_d} \| \boldsymbol{IP}(d_k) \|^2 \right), \tag{19}$$

where δ'_d equals to 1.0 and n_d denotes the number of drugs in the DTI network.

Target similarity calculation

For targets, we calculate their similarity based on functional annotation, protein domain and protein sequence information, respectively.

Gene Ontology is a controlled vocabulary and has three independent sub-ontologies, which are Biological Process (BP), Molecular Function (MF) and Cell Components (CC) [73]. Targets can be annotated with BP, MF and CC terms and their semantic similarity can be inferred from these three aspects [74]. With lost generality, we take BP as an example to calculate semantic similarity with STE approach [75]. First, the Information Content (IC) of one term t can be denoted as $IC(t)$, which can be expressed as

$$IC(t) = log(dp(t)) \cdot \left(log\left(\sum_{t_i \in Ac(t)} dp(t_i) \right) + 1 \right) \cdot \left(1 - \frac{log\,|(Ds(t))|}{log(N+1)} \right), \tag{20}$$

where $dp(t)$ denotes the depth of term t, $Ac(t)$ and $Ds(t)$ represent the ancestor term set and descendant term set in BP ontology. N is the number of total terms of term t in the whole GO structure. $|\cdot|$ denotes the number of elements in the set.

Then, we evaluate the weight for the semantic relationship between t_i and its parent term t_j, which can be formulated as

$$w_{t_i t_j} = \frac{\sum_{t_n \in Ds(t_j)} IC(t_n)}{\sum_{t_m \in Ds(t_i)} IC(t_m)}, \tag{21}$$

where t_m and t_n are the terms in $Ds(t_i)$ and $Ds(t_j)$, respectively.

Term IC can be divided into two parts: one is the inherited IC from its parents, and the other is the extended IC by itself [74]. The inherited IC of term t_i from its parent term t_j denoted as $IC_{Ih}(t_i)$ is measured as

$$IC_{Ih}(t_i) = w_{t_i t_j} \cdot IC(t_j). \tag{22}$$

The total inherited IC from its all parents terms can be expressed as

$$IC_{Ih}(t_i) = \sum_{t_k \in Pr(t_i)} w_{t_i t_k} \cdot IC(t_k), \tag{23}$$

where $Pr(t_i)$ denotes the parent term set of term t_i, and $w_{t_i t_k}$ can be calculated by Equation.6. The extended IC of term t_i is formulated as

$$IC_{extended}(t_i) = IC(t_i) - IC_{Ih}(t_i). \tag{24}$$

Suppose there is one term named t and its annotation set is S_t; its extended IC value for t is calculated as

$$IC_{extended}(S_t) = \sum_{t_i \in S_t} IC_{extended}(t_i). \tag{25}$$

Suppose there are two targets named n_i and n_j and their corresponding annotation term sets are $\mathbf{S}_{n_i}$ and $\mathbf{S}_{n_j}$, respectively; their semantic similarity can be expressed as

$$SimSWE(n_i, n_j) = \frac{IC_{extended}(\mathbf{S}_{n_i} \cap \mathbf{S}_{n_j})}{IC_{extended}(\mathbf{S}_{n_i} \cup \mathbf{S}_{n_j})} \tag{26}$$

We calculate the similarities for all target pairs and then construct the BP-based semantic similarity network that is fully connected. Similarly, MF-based and CC-based semantic similarity networks can also be constructed, which is the same to the BP-based semantic similarity network. These three semantic networks are represented as Net_{BP}, Net_{MF} and Net_{CC}, respectively. After that, we adopt the ***SNF*** (similarity network fusion) method [76], which is an effective and widely used way to fuse different similarity networks and ultimately get one integrated semantic similarity network named $Net_{integrated}$.

Protein domain-based similarity is also one common manner to measure the similarities between targets. Suppose there are two targets n_i and n_j and their corresponding domain annotation sets are $\boldsymbol{S}_{n_i}$ and $\boldsymbol{S}_{n_j}$. Their protein domain-based similarity can be formulated as

$$Sim_{domain}(n_i, n_j) = \frac{1}{2} \cdot \left(\frac{|\boldsymbol{S}_{n_i} \cap \boldsymbol{S}_{n_j}|}{|\boldsymbol{S}_{n_i}|} + \frac{|\boldsymbol{S}_{n_i} \cap \boldsymbol{S}_{n_j}|}{|\boldsymbol{S}_{n_j}|} \right). \quad (27)$$

Lastly, we measure the protein sequence similarity based on the pseudo-position specific scoring matrix (PsePSSM) [77]. Each protein sequence can be represented as a (20+20×λ)-dimensional vector. Suppose there are two targets denoted as n_i and n_j and their PsePSSM-based feature vectors can be denoted as $\boldsymbol{V}_{PP_{n_i}}$ and $\boldsymbol{V}_{PP_{n_j}}$, the sequence-based similarity can be formulated as

$$Sim_{sequence}(n_i, n_j) = \frac{\boldsymbol{V}_{PP_{n_i}} \cdot \boldsymbol{V}_{PP_{n_j}}}{\sqrt{(\boldsymbol{V}_{PP_{n_i}})^2} \cdot \sqrt{(\boldsymbol{V}_{PP_{n_j}})^2}} \quad (28)$$

APPENDIX B: Two examples of learning the embeddings of drugs and targets with the weights at different level attention

In this study, we employ MHADTI to learn to embeddings of drugs and targets with the hierarchical attention mechanisms. To show the learning process more concretely, we select one drug named Cariprazine (DB06016) and one target named KLKB1 (IDP03952) as the examples to display the weights at different level attention.

The process of learning the embeddings of drug Cariprazine with MHADTI is shown in Figure 7. There are totally two meta-paths named Φ_1 and Φ_2 and three views named $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$. Firstly, at the node-level, MHADTI learns the embedding of drug Cariprazine with their different meta-path-based neighbors under the meta-paths and views. Then, at the meta-path level, MHADTI learns the embedding of drugs Cariprazine with meta-path DTD and TDDT under each view. Finally, at the graph level, MHADTI learns the ultimate embedding of drugs from different graph-views. The weights at each step can be seen in Figure 7. The learning process of target KLKB1 is similar to drug Cariprazine and we do not repeat it anymore. The whole process has been presented in Figure 8.

APPENDIX C: 5-folder cross-validation strategy used in our study

We employ the 5-folder cross-validation (5-CV) strategy to evaluate the performance of MHADTI and the other comparison approaches. To demonstrate the 5-CV strategy more comprehensively, we depict its execution process by Figure 9. The flow of data and detailed implementation steps can be clearly presented.