如何使用异构图和基于异构图的算法来预测DTI？

如何构建异构生物信息网络？
节点类型：药物、蛋白质、疾病、药物副作用
边的类型：药物-药物，药物-蛋白质、药物-疾病、药物-副作用；蛋白质-蛋白质、蛋白质-疾病

OXFORD

Briefings in Bioinformatics, 2023, **24(2)**, 1–11

https://doi.org/10.1093/bib/bbad079

Advance access publication date 9 March 2023

Problem Solving Protocol

MHTAN-DTI：基于元路径的分层转换器和注意力网络的药物-靶标相互作用预测

MHTAN-DTI: Metapath-based hierarchical transformer and attention network for drug–target interaction prediction

metapath：元路径

Ran Zhang, Zhanjie Wang, Xuezhi Wang, Zhen Meng and Wenjuan Cui

Corresponding author: Xuezhi Wang, Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China; E-mail: wxz@cnic.cn

Abstract

Drug–target interaction (DTI) prediction can identify novel ligands for specific protein targets, and facilitate the rapid screening of effective new drug candidates to speed up the drug discovery process. However, the current methods are not sensitive enough to complex topological structures, and complicated relations between multiple node types are not fully captured yet. To address the above challenges, we construct a metapath-based heterogeneous bioinformatics network, and then propose a DTI prediction method with metapath-based hierarchical transformer and attention network for drug–target interaction prediction (MHTAN-DTI), applying metapath instance-level transformer, single-semantic attention and multi-semantic attention to generate low-dimensional vector representations of drugs and proteins. Metapath instance-level transformer performs internal aggregation on the metapath instances, and models global context information to capture long-range dependencies. Single-semantic attention learns the semantics of a certain metapath type, introduces the central node weight and assigns different weights to different metapath instances to obtain the semantic-specific node embedding. Multi-semantic attention captures the importance of different metapath types and performs weighted fusion to attain the final node embedding. The hierarchical transformer and attention network weakens the influence of noise data on the DTI prediction results, and enhances the robustness and generalization ability of MHTAN-DTI. Compared with the state-of-the-art DTI prediction methods, MHTAN-DTI achieves significant performance improvements. In addition, we also conduct sufficient ablation studies and visualize the experimental results. All the results demonstrate that MHTAN-DTI can offer a powerful and interpretable tool for integrating heterogeneous information to predict DTIs and provide new insights into drug discovery.

应用元路径实例级转换、单语义注意和多语义注意生成药物和蛋白质的低维向量表示。

Keywords: drug–target interaction prediction, metapath, transformer, graph attention network

Introduction

Drug discovery can be divided into early drug discovery stage and drug development stage. The early drug discovery process includes the selection and confirmation of drug targets and biomarkers, the determination of lead compounds, the study of structure-activity relationships, the screening of active compounds and the selection of drug candidates. After the drug candidates are determined, the new drug research enters the development stage, which includes preclinical toxicology studies and clinical studies [1].

The selection and confirmation of drug targets is the first and critical step in early drug discovery. Drug–target interaction (DTI) prediction can identify novel ligands for specific protein targets, which is important for drug repositioning, drug side effect prediction, polypharmacology and drug resistance studies [2, 3]. Since December 2019, coronavirus disease 2019 (COVID-19) pandemic with high infectivity and strong occult has spread rapidly worldwide, posing a serious threat to human health [4]. Drug development for the treatment of COVID-19 obviously should not follow the traditional new drug development process [5]. Drug repositioning discovers new clinical uses of strictly tested and verified drugs, which greatly shortens the drug development process and shows outstanding advantages in dealing with emergent diseases and treating rare diseases. In addition to acting on the main therapeutic targets, drugs may also interact with other proteins. Understanding drug–target information can help predict off-target toxicity and suboptimal efficacy, and also provide a new perspective for the study of molecular mechanisms of drug side effects.

The interactions between drugs and target proteins are measured using half maximal inhibitory concentration (IC50) or half maximal effective concentration (EC50), inhibition constant (Ki) and dissociation constant (Kd) values *in vitro* or *in vivo* experimental methods. Experiments to measure the affinity value for a large number of drug–target pairs are expensive and time-consuming, making the cost of developing new drugs high.

Ran Zhang is a master student in Computer Network Information Center, Chinese Academy of Sciences and University of Chinese Academy of Sciences. Her expertise is computational biology and artificial intelligence.

Zhanjie Wang is a master student in Computer Network Information Center, Chinese Academy of Sciences and University of Chinese Academy of Sciences. His expertise is data mining and artificial intelligence.

Xuezhi Wang is a researcher in Computer Network Information Center, Chinese Academy of Sciences. His expertise is biological data management and analysis.

Zhen Meng is a senior engineer in Computer Network Information Center, Chinese Academy of Sciences. Her expertise is biological data management and analysis.

Wenjuan Cui is an associate researcher in Computer Network Information Center, Chinese Academy of Sciences. Her expertise is computational biology and data mining.

Received: November 23, 2022. **Revised:** February 8, 2023. **Accepted:** February 13, 2023

为什么要使用异构图、元路径去实现节点特征的聚合！

In recent years, advances in computer processing abilities and continuous updating of computational algorithms have made virtual screening a popular tool in drug discovery. Computational methods have achieved good performance on multiple tasks in bioinformatics-related fields, such as disease-related miRNA prediction [6–8], disease genes prediction [9], drug–drug interaction prediction [10], protein–protein interaction prediction [11] and protein subcellular localization prediction [12]. Some models [13, 14] in drug discovery field can also be applied in DTI prediction through extension and improvement. DTI computational prediction has huge research potential and broad application prospects, which has advantages of short time, low cost, high accuracy and wide range in exploring potential DTIs [15]. Academia and industry have continuous and urgent needs for the development of technology.

Currently, computational methods for DTI prediction are mainly divided into four categories, namely molecular docking-based methods, ligand-based methods, text mining-based methods and chemogenomic-based methods.

Molecular docking-based methods model the three-dimensional (3D) structure of a protein and then simulate the docking process between a drug and a protein, requiring a large number of comprehensive sampling of all possible conformations to obtain the closest conformation to the real binding [16]. Ligand-based methods utilize the existing knowledge of pharmacology to analyze the drug structure that interacts with the target, revealing and predicting the relationship between the chemical structure and pharmacological activity of drugs [17]. Text-mining-based methods can automatically mine drug–protein association from scientific literature and discover candidate drug targets that can be used to intervene and treat diseases [18].

Chemogenomic-based methods [19] have recently achieved success in drug discovery and repositioning tasks, which integrate both the chemical space of compounds and the genomic space of proteins into a unified space: pharmacological space. These methods can be refined into different branches according to the implementation: similarity-based methods, pharmacological feature-based methods and network-based methods.

Similarity-based methods are based on the 'Guilt by Association' assumption that similar drugs tend to bind similar targets, and similar targets also tend to bind to similar drugs [20, 21]. Similarity-based methods contain the nearest neighbor methods [22, 23], bipartite local models (BLM) [24–26] and matrix factorization methods [27–30]. Pharmacological feature-based methods extract the pharmacological features of drugs and targets from different perspectives, and feed them into the machine learning model to identify DTIs [31–38]. Network-based methods construct a topological network for calculation and inference to predict DTIs. According to the different topology structure of the network, network-based methods can be divided into binary network-based methods [39] and heterogeneous network-based methods [40, 41]. At present, numerous DTI prediction methods combine deep learning with heterogeneous networks [42–49]. Deep-learning-based methods take advantages of neural networks to automatically learn the features of nodes and the complex relationships among them, which has been an effective tool for DTI prediction.

Recently, transformer [50] has also been introduced into the DTI prediction field [51–54]. Transformer applies self-attention mechanism to capture long-range contextual dependencies. In most transformer-based methods, transformer acts as a semantic feature extractor for drugs and proteins, respectively. According to the powerful ability of transformer to model both the long- and short-distance sequence features, we introduce transformer into metapath instance aggregation to realize feature passing between different types of pharmacological entities.

Many existing methods focus on the molecular structure level to obtain the representations of drugs and proteins with more features. Simultaneously, drug–drug interaction, drug–disease association, drug–side effect association, protein–protein interaction and protein–disease association also contain the potential and useful information related to DTIs. However, association mining is not thorough enough, and association relations in heterogeneous bioinformatics networks have not been fully exploited yet. How to effectively utilize the topological structure between nodes to accelerate DTI prediction is a difficult problem to be solved. We introduce association between multiple pharmacological entities, bringing both useful information and noise. Attention mechanism can capture global connections with long-range dependency information. In addition, attention mechanism highlights the importance of useful information and weakens the importance of noise or useless information, which has potentially good interpretability. Therefore, we apply the hierarchical attention mechanism to explain the DTI prediction process.

In this paper, we propose a DTI prediction method with metapath-based hierarchical transformer and attention network (MHTAN-DTI). First, we construct a heterogeneous bioinformatics network with 708 drugs, 1512 proteins, 5603 diseases and 4192 side effects. The nodes in the network represent four node types: drug, protein, disease and side effect, respectively. The edges represent drug–drug association, drug–protein association, drug–disease association, drug–side effect association, protein–protein association and protein–disease association, respectively. MHTAN-DTI applies metapath instance-level transformer, single-semantic attention and multi-semantic attention on the heterogeneous bioinformatics network to generate low-dimensional vector representations of drugs and proteins for accurate DTI prediction. 异构图中四种节点类型，六种边类型。

To summarize, we make the following major contributions in this work:

- We construct a metapath instance dataset based on multiple types of nodes and association to better describe the relations between drugs and targets.
- We propose MHTAN-DTI, an end-to-end DTI prediction framework, to effectively capture the structural and semantic information.
- MHTAN-DTI introduces transformer to message passing between the nodes of different types, and adopts two-level attention mechanism to learn the underlying pharmacological knowledge, which outperforms various state-of-the-art DTI prediction methods.

分层转换器是一个包含两个子结构的transformer编码器

Methodology

等级制的，层次的

As shown in Figure 1, we propose a hierarchical transformer and attention network representation learning method to predict DTIs. First, we construct a heterogeneous bioinformatics network containing four node types: drug, target, disease and side effect. Afterwards, we successively use transformer and attention mechanism to encode the intrinsic structural and semantic properties of the bioinformatics network to the low-dimensional latent embeddings of drugs and targets. Finally, we treat DTI prediction as a binary classification problem and minimize the loss function to update the weight coefficients of our model.

图1的整体情况：先构建了包含四种节点类型的异构图（生物信息学网络），然后成功使用transformer机制和attention机制去编码异构图（生物信息学网络）中固有的结构和语义属性，得到药物和靶标的低维嵌入，最后进行DTI预测，并根据损失更新权重参数。

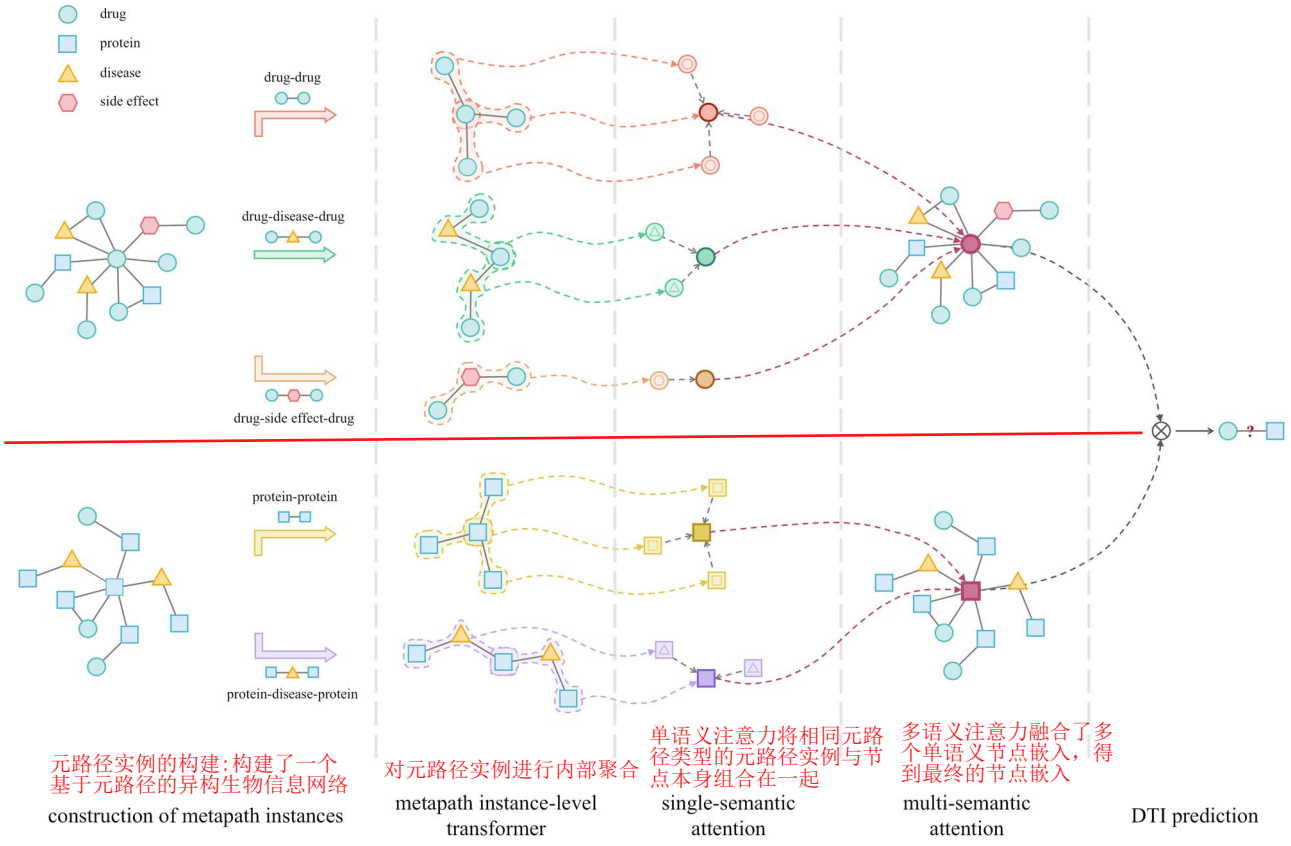


Figure 1. The overall architecture of MHTAN-DTI. MHTAN-DTI consists of five main steps: construction of metapath instances, metapath instance-level transformer, single-semantic attention, multi-semantic attention and DTI prediction. Construction of metapath instances constructs a heterogeneous bioinformatic network based on metapath. Metapath instance-level transformer conducts internal aggregation on metapath instances. Single-semantic attention combines metapath instances of the same metapath type with node itself, and multi-semantic attention fuses the multiple single-semantic node embeddings to get the final node embedding. Then we calculate the loss and optimize the model to identify DTIs.

Construction of metapath instances

We construct a bioinformatics network [55] with 708 drugs, 1512 proteins, 5603 diseases and 4192 side effects based on the Luo *et al.* dataset [34]. Luo *et al.* dataset contains six drug/protein-related networks: drug–drug interaction network [DrugBank (Version 3.0)] [56], protein–protein interaction network [HPRD database (Release 9)] [57], drug–protein interaction network [DrugBank (Version 3.0)] [56], drug–disease association network (Comparative Toxicogenomics Database) [58], protein–disease association network (Comparative Toxicogenomics Database) [58] and drug–side effect association network [sider database (version 2)] [59]. In addition, we also introduce drug chemical structure information as well as protein sequence information by creating two extra networks: the drug structure similarity network [60] and the protein sequence similarity network [61].

我们还建立两个额外的网络来引入药物化学结构信息和蛋白质序列信息：药物结构相似网络和蛋白质序列相似网络

使用Jaccard相似系数比较药物对之间的相似性

As exhibited in Figure 2, we first use Jaccard similarity coefficient [62] to calculate the similarity of each drug–drug pair through three association networks of drug–drug, drug–disease and drug–side effect, and obtain the similarity matrices based on interaction, disease and side effect, respectively. Referring to the existing research, we set thresholds to integrate drug chemical structure similarities, interaction-related drug similarities, disease-related drug similarities, side effect-related drug similarities and drug–drug interaction, so as to obtain drug–drug association. Similarly, protein–protein and protein–disease association networks are used to calculate the similarity of each pair of protein–protein to obtain the similarity matrices based on interaction and disease, respectively. Afterwards, we set thresholds to integrate protein sequence similarities, interaction-related protein similarities, disease-related protein similarities and protein–protein interaction to obtain protein–protein association. In this way, drug–drug similarity information and protein–protein similarity information are injected into the constructed bioinformatics network. Note that in order to avoid label leakage, we do not use the drug–protein interaction to calculate drug–drug similarity and protein–protein similarity, so that DTIs in the test set are always unknown in the training stage.

药物-药物的关联网络是什么，药物-疾病的关联网络是什么，药物和副作用的关联网络又是什么？并分别获得了它们的相似性矩阵，这个矩阵又是什么？

Definition 1. (heterogeneous graph). A heterogeneous graph is defined as $G = (V, E)$ associated with a node type mapping function $\phi : V \to \mathbb{A}$ and an edge type mapping function $\psi : E \to \mathbb{B}$. $\mathbb{A}$ and $\mathbb{B}$ denote the predefined sets of node types and edge types, satisfying $|\mathbb{A}| + |\mathbb{B}| > 2$. Heterogeneous graphs are also called heterogeneous information networks.

相似的药物往往以高概率结合相似的靶点

DTI prediction follows the 'Guilt by Association' assumption that similar drugs tend to bind similar targets with high probability. We introduce new node types such as disease and side effect into DTI prediction. The potential DTI information is probably hidden in the phenotypes of disease and the manifestations of side effects. Since the drugs that bind to the same target are similar, the drugs for treating the same disease are similar, the target positions related to the same disease are similar and

我们首先使用Jaccard相似系数，通过药物-药物、药物-疾病和药物-副作用三个关联网络，来计算每个（药物-药物对）的相似度，分别得到基于相互作用、基于疾病和基于副作用的相似矩阵。

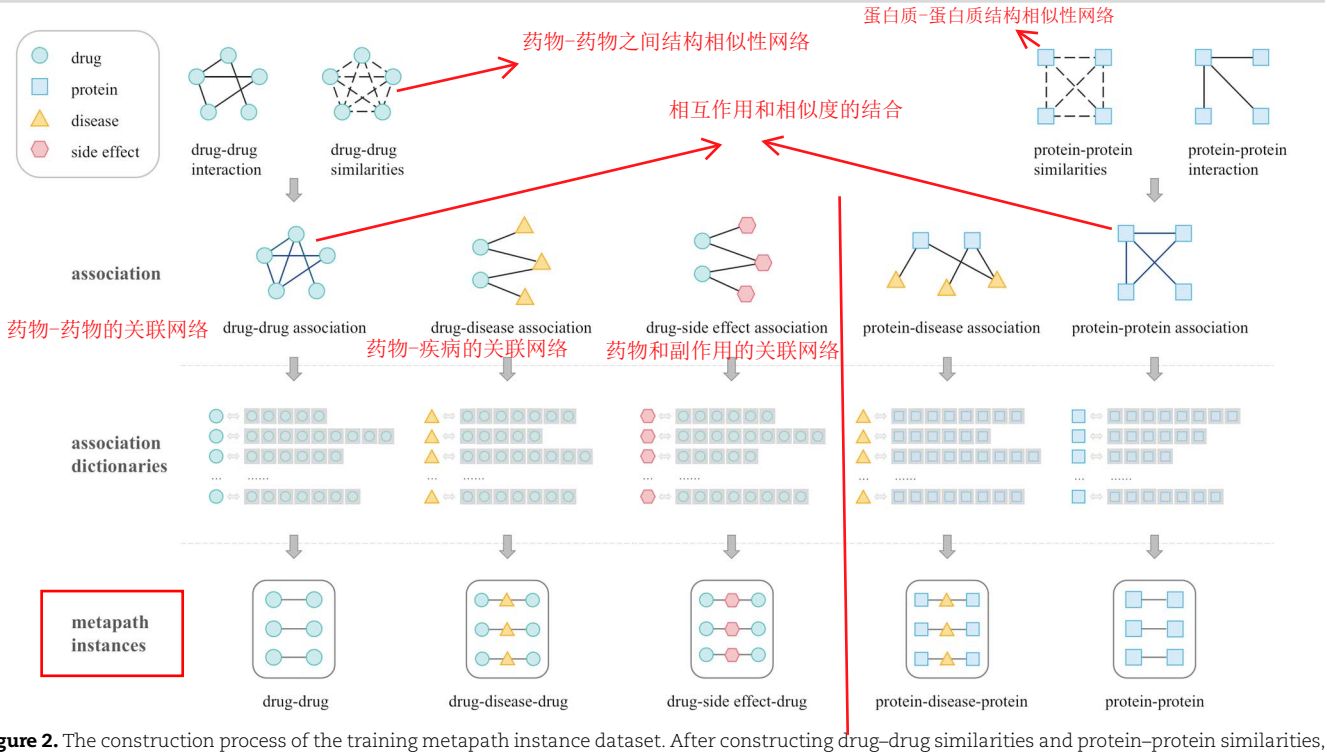


Figure 2. The construction process of the training metapath instance dataset. After constructing drug–drug similarities and protein–protein similarities, we combine the similarities and interaction to get drug–drug association and protein–protein association. Then, we construct corresponding association dictionaries based on five sets of association to obtain metapath instances in the training set.

训练元路径实例数据集的构建过程。在构建药物-药物相似度和蛋白质-蛋白质相似度之后，我们将相似度和相互作用结合起来，得到药物-药物关联和蛋白质-蛋白质关联。然后，基于5个关联集构建相应的关联字典，得到训练集中的元路径实例。

the drugs that produce the same side effects are also similar. Therefore, we introduce the concepts of metapath and metapath-based neighbor into the heterogeneous bioinformatics network, which greatly facilitates the learning of low-dimensional vector representations of drugs and targets, making more accurate DTI prediction.

adj. 合成的，复合的

Definition 2. (metapath). A metapath R is defined as a path in the form of $A_1 \xrightarrow{B_1} A_2 \xrightarrow{B_2} \cdots \xrightarrow{B_l} A_{l+1}$, where $B = B_1 \circ B_2 \circ \cdots \circ B_l$ describes a composite relation B with edge types B_1 to B_l between node types A_1 and A_{l+1}, and $\circ$ denotes the composition operator on relations.

Definition 3. (metapath instance). Given a metapath R of a heterogeneous graph G, a metapath instance $R_{(i,j)}$ of R is defined as a node sequence in the graph G following the schema defined by R with start node i and end node j.

Definition 4. (metapath-based neighbor). Given a metapath R of a heterogeneous graph G, the metapath-based neighbors N_v^R of a node v is defined as the set of nodes that include node v (itself) and connect with node v via metapath instances of R.

We construct seven types of metapath instances: *drug–drug*, *drug–protein–drug*, *drug–disease–drug*, *drug–side effect–drug*, *protein–protein*, *protein–drug–protein* and *protein–disease–protein*.

目前没有看到这两种元路径

Metapath instance-level transformer

Due to the heterogeneity of bioinformatics network, different types of nodes do not share their feature space. Drugs are in the chemical space, while proteins are in the genomic space. First, we perform feature transformation to project different types of node features into the same latent vector space, and obtain node i initial embedding $\mathbf{h}_i$.

To capture the hidden feature of a node, the most important thing is to model its context information, namely its metapath instances. Metapath instance-level transformer maps a metapath instance sequence into an abstract continuous representation, and then averages the representation to obtain the metapath instance embedding. Metapath instance-level transformer converts each metapath instance sequence $\mathbf{R}_k(i,j)$ of metapath $\mathbf{R}_k$ into the metapath instance embedding $\mathbf{h}_{\mathbf{R}_k(i,j)}$ as follows:

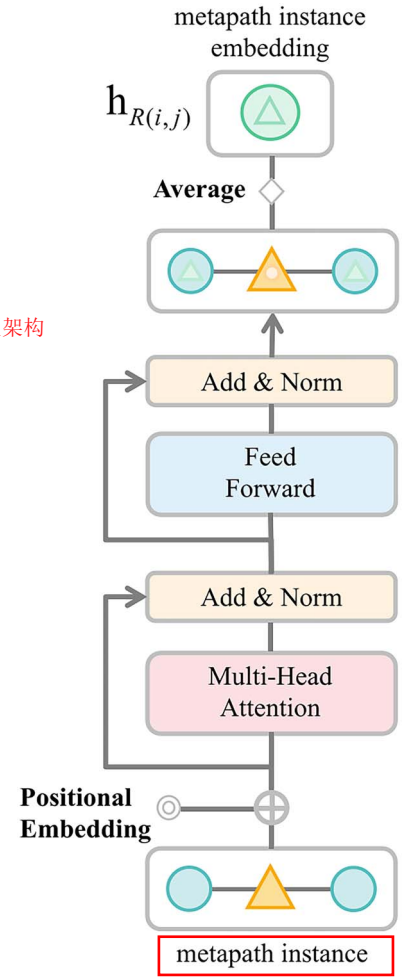
元路径实例序列Rk(i, j)是什么？

$$\mathbf{h}_{\mathbf{R}_{k(i,j)}} = transformer\left(\|(\mathbf{h}_t \mid \forall t \in \mathbf{R}_{k(i,j)})\right) \quad (1)$$

As Figure 3 shows, metapath instance-level transformer encoder consists of two major components. The first component implements multi-head attention mechanism, and receives different linearly projected versions of the queries, keys and values. Each produces outputs in parallel to generate a final result. The second component is a fully connected feed-forward neural network composed of two linear transformations with rectified linear unit (ReLU) activation in between. Each layer has a residual connection around it and is also succeeded by a normalization layer to normalize the sum computed between the input and the output generated by itself.

Note that the node positions in the metapath contain useful information for DTI prediction. For the central node, the metapath-based neighbors are far away, but they have the same node type as the central node, which have significant pharmacological significance. Consequently, metapath instance-level transformer encoder combines multi-head self-attention with positional embedding to capture the long-range dependencies.

In addition, attention mechanism can access all the nodes in the metapath and weight them according to a learned measure of



该结构有点像transformer模型架构中的编码器部分

Figure 3. The metapath instance-level transformer encoder architecture. The two major components in metapath instance-level transformer encoder are multi-head attention mechanism and feed-forward neural network.

relevance, providing relevant information about far-away nodes. Multi-head attention runs through an attention mechanism several times in parallel for attending to parts of the sequence differently to encode relevance relations that are meaningful to the task. Metapath instance-level transformer encoder applies four-head self-attention mechanism and performs well in DTI prediction.

Single-semantic attention

As shown in Figure 4, the central node i has a number of metapath instances of a certain metapath type, which contain some noise metapath instances, leading to significant differences in the importance of different metapath instances.

To obtain the single-semantic central node representation with critical and useful features, we not only aggregate the importance of different metapath instances, but also consider the importance of the central node itself. For drugs whose pharmacological properties have been fully explored, their targets prediction depends more on their own information. For some drugs that have not been fully understood, their targets prediction depends more on inference from the association with diseases and side effects. The central node weight plays a key role in reflecting the importance of central node information for DTI prediction.

It is remarkable that we introduce the central node weight $\alpha_{ii}^{R_k}$, which is an important indicator to measure the importance of the central node i. The metapath instance $R_{k(i,j)}$ weight $e_{ij}^{R_k}$ and the central node i weight $e_{ii}^{R_k}$ can be described as follows:

$$e_{ij}^{R_k} = \text{attention}_{single-semantic}(h_i, h_{R_{k(i,j)}}; R_k) \tag{2}$$

$$e_{ii}^{R_k} = \text{attention}_{single-semantic}(h_i; R_k) \tag{3}$$

We perform the softmax function to normalize the weight coefficient. According to Definition 4, $i \in N_i^{R_k}$, the weight coefficient $\alpha_{ij}^{R_k}$ and $\alpha_{ii}^{R_k}$ after normalization can be uniformly calculated as follows:

$$\alpha_{im}^{R_k} = \frac{\exp(e_{im}^{R_k})}{\sum_{n \in N_i^{R_k}} \exp(e_{in}^{R_k})} \tag{4}$$

To tackle the high variance of network data, we extend single-semantic attention to multi-head single-semantic attention for stable training. Given a certain metapath type R_k and central node i, we can repeat the attention for H times and concatenate the embeddings to learn the semantic-specific R_k node embedding $z_i^{R_k}$ as follows:

H：指的是注意力的头数

$$z_i^{R_k} = \Big\Vert_{h=1}^{H} \sigma\left(\sum_{j \in N_i^{R_k} - \{i\}} [\alpha_{ij}^{R_k}]_h \cdot h_{R_{k(i,j)}} + [\alpha_{ii}^{R_k}]_h \cdot h_i \right), \tag{5}$$

where $\sigma(\cdot)$ represents an activation function, $[\alpha_{ij}^{R_k}]_h$ denotes the normalized importance of metapath instance $R_{k(i,j)}$ to node i at the attention head h and $[\alpha_{ii}^{R_k}]_h$ denotes the normalized importance of central node i at the attention head h.

Multi-semantic attention

Different metapath types have different pharmacological semantics. To learn representations with multiple semantics, we need to aggregate multiple single-semantic node embeddings. Similarly, the importance of each metapath type to the central node is different, such as the importance of *drug–disease–drug* and *drug–side effect–drug* to the central drug embedding have obvious differences. The differences of metapath types need to be distinguished and reflected in the multi-semantic representation of the central node.

Therefore, as shown in Figure 4, we adopt a multi-semantic attention layer to combine the semantic information across all metapath types after aggregating node and edge information for each metapath type. We introduce the metapath type R_k weight β_{R_k} to the central node i to describe the importance difference as follows:

$$w_{R_k} = \text{attention}_{multi-semantic}(z_i^{R_k}) \tag{6}$$

Then we also get the normalized weight coefficient β_{R_k} through the softmax function as follows:

$$\beta_{R_k} = \frac{\exp(W_{R_k})}{\sum_{R_t \in \mathbb{R}} \exp(W_{R_t})} \tag{7}$$

Given central node i, we can fuse the embeddings $z_i^{R_k}$ under all the semantics to obtain the final central node embedding z_i as

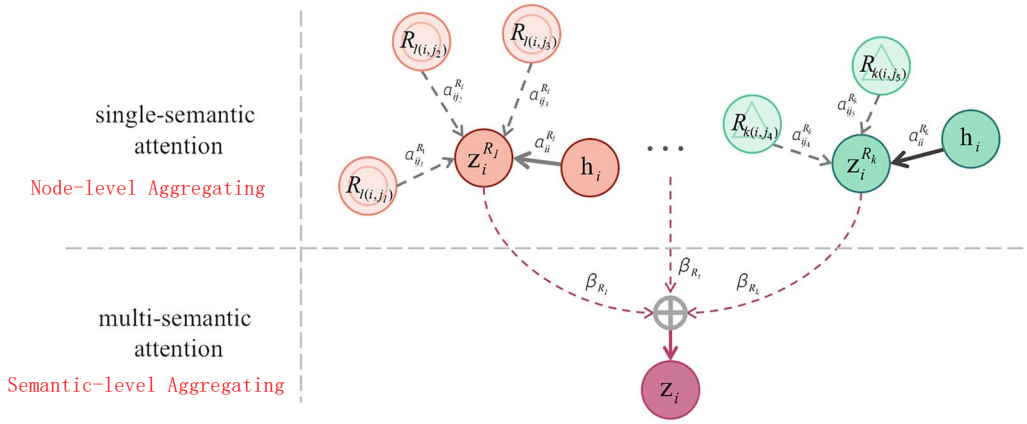


Figure 4. Explanation of single-semantic attention and multi-semantic attention. Single-semantic attention aggregates the central node and the metapath instances of the same type. Multi-semantic attention aggregates the node embeddings under different semantics.

follows:

$$z_i = \sum_{R_k \in \mathbb{R}} \beta_{R_k} \cdot z_i^{R_k}, \tag{8}$$

where $\mathbb{R}$ is the set of all metapaths.

DTI prediction

After obtaining the low-dimensional multi-semantic representations of drugs and proteins, we project the node embeddings into a space with node similarity measures for the downstream DTI prediction task. Given the drug d embedding $\mathbf{z}_d$ and the protein p embedding $\mathbf{z}_p$, we optimize the model weights by minimizing the following loss function through negative sampling:

$$\mathfrak{L} = -\sum_{(d,p)\in\Omega} \log\sigma(\mathbf{z}_d^{\mathrm{T}} \cdot \mathbf{z}_p) - \sum_{(d',p')\in\Omega^-} \log\sigma(-\mathbf{z}_{d'}^{\mathrm{T}} \cdot \mathbf{z}_{p'}), \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function, $\sigma(\mathbf{z}_d^{\mathrm{T}} \cdot \mathbf{z}_p)$ is the probability that drug d interacts with protein p, Ω is the set of the observed node pairs and Ω^- (the complement of Ω) is the set of non-existing drug–protein pairs sampled from all unobserved drug–protein pairs.

Results

Experiment preparation

To evaluate the performance of MHTAN-DTI, we test our model on the constructed metapath dataset. We treat the observed DTIs as positive samples, and select the non-existing DTIs as negative samples to generate the experimental dataset. In addition, the training set, validation set and test set are divided according to the ratio of 7:2:1, and the validation set and test set have the same number of randomly sampled positive and negative samples. MHTAN-DTI uniformly samples negative drug–protein pairs on the fly during the training stage.

DTI prediction can be regarded as a link prediction task, and the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) are used as evaluation indicators. Specifically, AUROC is an evaluation indicator to measure the performance of binary classification models, indicating the probability that the predicted positive example is ahead of the negative example. Furthermore, AUROC is widely used for link prediction and can be interpreted as the probability that a randomly selected missing link will get a higher score than a randomly selected non-existing link. AUPR is more suitable for measuring highly unbalanced datasets, ignoring the effects of skewed data.

In comparative experiments, we compare seven existing methods: BLMNII [25], NetLapRLS [63], CMF [28], NRLMF [64], DTINet [34], NeoDTI [42] and EEG-DTI [45] on the same dataset. Since MHTAN-DTI mainly utilizes the topology of the bioinformatics network, MolTrans [54] and Multi-TransDTI [52] are not selected for comparison, which additionally introduce pharmacological features of drugs and proteins, such as SMILES strings of drugs and local residues of proteins.

Parameter settings

We employ all experiments on CentOS release 7.9 system and use NVIDIA A100-PCIE GPU card with 40 GB memory for graphics acceleration. In the experiment, we set batch size to 32, dropout rate to 0.2 and epoch number to 120. The vector dimension and attention vector dimension are both set to 128, and the number of attention heads is set to 8. Adam optimizer is simple to implement with higher efficiency and less memory footprint. Its parameter update is not affected by gradient scaling transformation, and its hyperparameters have good interpretability. Therefore, we choose adam optimizer to adjust and calculate model parameters.

NeoDTI:https://academic.oup.com/bioinformatics/article/35/1/104/5047760?login=true

Performance evaluation

We validate the performance of MHTAN-DTI on the constructed metapath dataset and compare MHTAN-DTI with seven state-of-the-art DTI prediction methods: BLMNII, CMF, NetLapRLS, NRLMF, DTINet, NeoDTI and EEG-DTI. From Table 1 and Figure 5, the experimental results demonstrate that MHTAN-DTI achieves the best DTI prediction performance. Compared with the best results among the seven comparative experimental methods, AUROC and AUPR are improved by 3.01% (from 95.70 to 98.71%) and 2.71% (from 96.24 to 98.95%) by a large margin, respectively. Due to the homologous proteins or similar drugs in the testing set, the good performance of prediction methods might result from easy predictions. To ease the inflation of prediction performance, we conduct an additional experiment. Those easy predictions are excluded by removing the predicted DTIs similar to the known DTIs (drug chemical structure similarities>0.6 and protein sequence similarities>0.4). In this case, AUROC and AUPR

EEG-DTI:https://academic.oup.com/bib/article/22/5/bbaa430/6124914?login=true

DTINet:https://www.nature.com/articles/s41467-017-00680-8

Table 1. Performance evaluation on DTI prediction task in terms of the AUROC and AUPR scores

	BLMNII	CMF	NetLapRLS	NRLMF	DTINet	NeoDTI	EEG-DTI	MHTAN-DTI
AUROC	0.7156	0.8561	0.9441	0.9545	0.9308	0.9570	0.9509	**0.9871**
AUPR	0.6908	0.8966	0.9488	0.9564	0.9504	0.9544	0.9624	**0.9895**

Notes: The eight methods are BLMNII, CMF, NetLapRLS, NRLMF, DTINet, NeoDTI, EEG-DTI and MHTAN-DTI. The numbers in bold indicate the best performance.

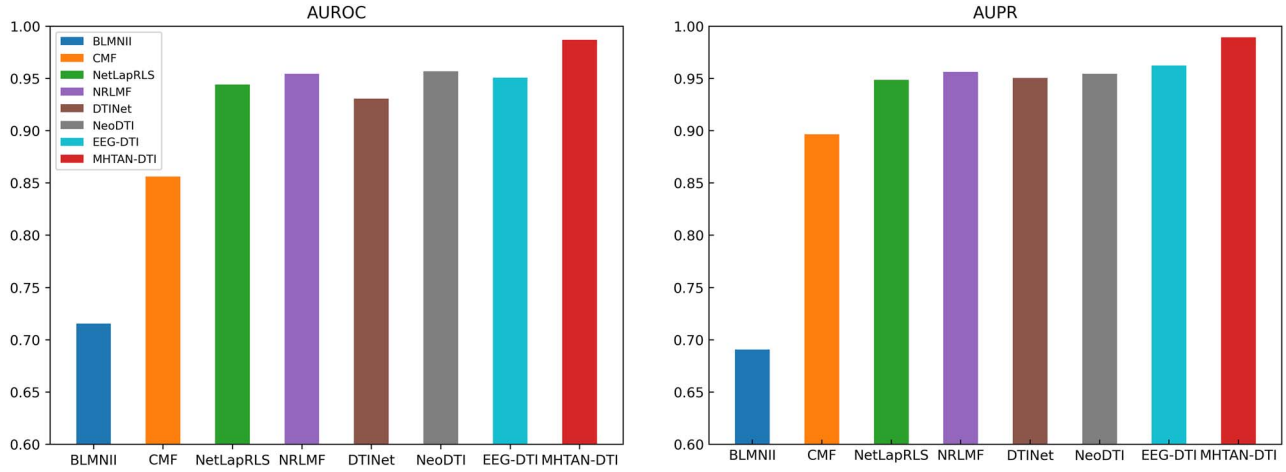


Figure 5. Comparison of DTI prediction experimental results. We compare MHTAN-DTI with other state-of-the-art methods, BLMNII, CMF, NetLapRLS, NRLMF, DTINet, NeoDTI and EEG-DTI on the constructed metapath dataset. The left one and the right one represent the AUROC and AUPR results of different methods, respectively.

of MHTAN-DTI achieve 98.14% and 98.43%, respectively, demonstrating the performance improvements are more dependent on MHTAN-DTI itself. BLMNII, CMF, NetLapRLS and NRLMF only utilize the similarities and interactions of drugs and proteins, which capture relatively limited associated features. DTINet contains two steps, feature learning and DTI prediction, which cannot be jointly trained. The embeddings learned from the unsupervised learning process are likely not the most suitable representations for DTI prediction, making its performance limited. NeoDTI and EEG-DTI achieve suboptimal performance after MHTAN-DTI. The end-to-end training framework improves the defects of DTINet, but they capture the features of drugs and targets without considering the impact of different neighbors on the central node. In contrast, MHTAN-DTI adopts a transformer encoder and two-layer attention network to capture the structural and semantic features in the bioinformatics network by assigning different weights. MHTAN-DTI ingeniously applies the attention mechanism and effectively captures the relations between multiple node types in the heterogeneous bioinformatics network, obtaining better low-dimensional vector representations and excellent DTI prediction results, which is convenient for rapid screening of new drug candidates to find 'a needle in a haystack'. In addition, MHTAN-DTI achieves 98.71% and 98.95% in AUROC and AUPR, respectively. The closeness of the two evaluation indicator values truly demonstrates the robustness and effectiveness of MHTAN-DTI.

Ablation studies

Our work mainly focuses on two stages: the construction of heterogeneous bioinformatics network and the DTI prediction with metapath-based hierarchical transformer and attention network. To further explore the significance of each component of the dataset and model for the experimental performance, we conduct ablation studies on metapath types, metapath instance encoders and central node embedding.

转移路径类型消融研究的定量结果

Table 2. The Quantitative results for ablation study of metapath types

Settings	AUROC	AUPR
all metapaths	**0.9871**	**0.9895**
without *drug-drug*	0.9585	0.9560
without *drug-disease-drug*	0.9707	0.9703
without *drug-side effect-drug*	0.9845	0.9890
without *protein-protein*	0.9692	0.9679
without *protein-disease-protein*	0.9761	0.9850

Notes: The numbers in bold indicate the best performance.

Ablation study on metapath types

To evaluate the impact of different metapath types introduced on DTI prediction performance, we perform ablation experiments on the constructed metapath dataset. Since *drug–protein–drug* and *protein–drug-protein* metapaths contain DTIs, we remove these two metapaths in training to avoid label leakage. There are only five effective metapaths in the training stage: *drug–drug*, *drug–disease–drug*, *drug–side effect–drug*, *protein–protein* and *protein–disease-protein*. Table 2 lists the experimental results of removing different metapaths.

As illustrated in Table 2, MHTAN-DTI shows significant performance advantages when all metapaths are introduced. Compared with the removal of *drug–side effect–drug*, the reduction of AUROC and AUPR are more obvious when *drug–disease–drug* or *protein–disease–protein* is removed, which seems reasonable and explainable. Side effects are only associated with drugs, while diseases are associated with both drugs and proteins, containing more potentially useful information. After removing *drug–drug* or *protein–protein*, AUROC and AUPR decrease significantly. *drug–drug* contains not only the interaction relationship between drugs, but also chemical-based, disease-related and side-effect-related similarity information. Similarly, *protein–protein* not only contains the interaction relationship between proteins, but also contains

Table 3. The Quantitative results for ablation study of node types

Settings	AUROC	AUPR
all metapaths	**0.9871**	**0.9895**
without disease-related metapaths	0.9346	0.9395
without side effect-related metapaths	0.9845	0.9890

Notes: The numbers in bold indicate the best performance.

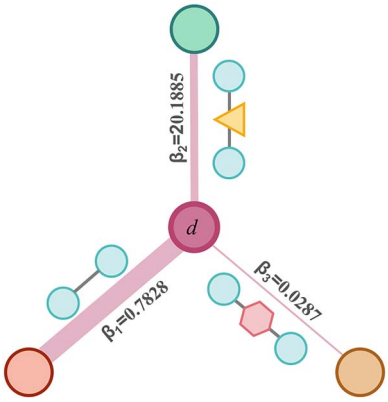


Figure 6. An example of attention mechanism for different metapath types. The width of the lines indicates the contribution of different metapath types to drug *d* representation

protein-sequence-based and disease-related similarity information. Since part of the information can also be obtained from other metapaths, some information loss can be compensated to some degree, which makes the AUROC and AUPR scores only decline within a certain range.

To evaluate the impact of introducing diseases and side effects into DTI prediction, we further conduct ablation experiments on removing node types in the bioinformatics network. In MHTAN-DTI, the essence of removing nodes is to remove related metapaths in the heterogeneous network. As listed in Table 3, AUROC and AUPR decrease significantly after removing disease-related metapaths, while AUROC and AUPR only drop slightly after removing side-effect-related metapaths. Disease is not only the cause of drug development, but also the result of protein changes, which is an important bridge between drug and protein. Side effect is often related to drug-off-target interactions, more focused on reflecting the pharmacological features of drug. It can be inferred that adding disease-related information can significantly improve DTI prediction performance, and adding side-effect-related information can also contribute to accurate DTI prediction in general.

We randomly selected a drug *d* and visualized the normalized weights of each metapath type, as shown in Figure 6. Since each drug is specific and contains only a subset of all metapath types, the weights of metapath types are not certainly positively correlated with the results in the ablation experiment. According to Figure 6, the weight of *drug–side effect–drug* is close to 0, while the weight of *drug–drug* accounts for the majority. The weights reflect that side effects are noise data without introducing key information for drug *d* target prediction, and *drug–drug* plays a dominant role. Attention mechanism is beneficial to analyzing and explaining DTI prediction results, which prompts researchers to discover which nodes or metapaths make the higher contributions for DTI prediction, and further improves the robustness of the model.

Table 4. The quantitative results for ablation study of metapath instance encoders

Settings	AUROC	AUPR
transformer encoder	**0.9871**	**0.9895**
BiLSTM encoder	0.9789	0.9865
average encoder	0.9721	0.9794
max pooling encoder	0.9754	0.9583
metapath-based neighbors*	0.8768	0.9062

Notes: The numbers in bold indicate the best performance. * Baseline: without any encoder.

Table 5. The quantitative results for ablation study of central node embedding

Settings	AUROC	AUPR
with central node embedding	**0.9871**	**0.9895**
without central node embedding	0.9753	0.9849

Notes: The numbers in bold indicate the best performance.

Ablation study on metapath instance encoders

In MHTAN-DTI, we introduce transformer encoder as a metapath instance aggregator. To evaluate the effects of different metapath instance encoders to MHTAN-DTI, we also run the model with bi-directional long short-term memory (BiLSTM) encoder, average encoder and max pooling encoder. In addition, we perform an extra experiment only utilizing the metapath-based neighbors without any metapath instance encoder.

As described in Table 4, AUROC and AUPR are dramatically reduced without the aggregation of intermediate nodes, indicating the final embeddings of drugs and proteins have lost important information. Furthermore, BiLSTM encoder achieves the second best performance after transformer encoder. Average encoder and max pooling encoder also get good experimental results. Transformer encoder achieves significant performance improvement over only considering metapath-based neighbors, which demonstrates the effectiveness of transformer encoder on metapath instances.

Ablation study on central node embedding

In single-semantic attention of MHTAN-DTI, we not only consider the metapath instance embeddings, but also aggregate the central node embedding. To explore the importance of the central node embedding for single-semantic attention, We perform ablation experiments on the central node embedding.

As listed in Table 5, AUROC and AUPR are improved by 1.18% (from 97.53 to 98.71%) and 0.46% (from 98.49 to 98.95%). Experimental results demonstrate that the central node plays a certain role in generating the expressive embeddings. Note that the performance improvement is closely related to the dataset itself. When more structure and property features of drugs and proteins are introduced, the performance improvement will be more significant.

Visualization 可视化DTI预测的结果

To show the model performance more intuitively, we conduct qualitative assessments by visualizing the DTI prediction results. We select positive and negative drug–protein pairs from the test set, and then use t-distributed stochastic neighbor embedding(t-SNE) to project drug and proteins into a 2D space. As shown in Figure 7, we integrate both the chemical space of drugs and the genomic space of targets into a unified space for DTI prediction.

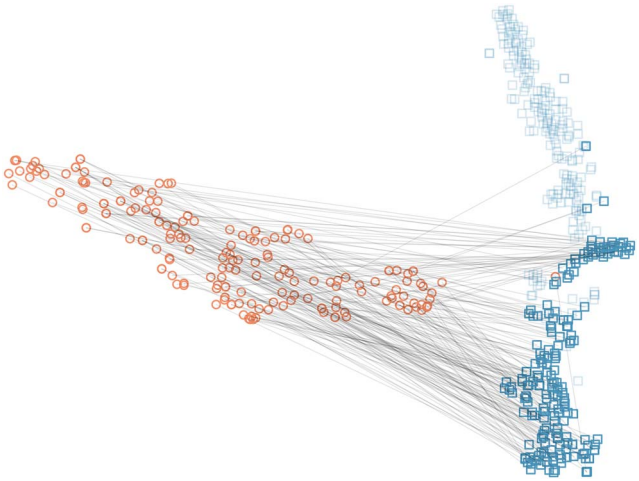


Figure 7. Visualization result of MHTAN-DTI. The orange circles and blue squares represent drugs and proteins, respectively. The gray lines represent the DTIs. The light markers indicate negative samples in the test set, while the dark markers indicate positive samples in the test set.

According to the visualization result, MHTAN-DTI obtains superior embedding results, with two well-separated drug and protein groups, and an aligned correlation of drug–protein pairs.

Conclusion

In this paper, we propose a DTI prediction method with metapath-based hierarchical transformer and attention network. We construct a heterogeneous bioinformatics network based on metapath. Metapath types are defined based on pre-exploration of bioinformatics network structures and domain expertise, making our model interpretable. Then we capture the rich structural and semantic information through transformer encoder and two-level attention mechanism to obtain low-dimensional vector representations of drugs and proteins for DTI prediction. Transformer encoder captures long-range dependencies of metapath instances, and the two-level attention network weakens the influence of noise data on DTI prediction performance to some extent. To verify the effectiveness and robustness of MHTAN-DTI, we compare MHTAN-DTI with seven state-of-the-art DTI prediction methods on the constructed metapath dataset. The results show that MHTAN-DTI exhibits significant performance improvements. In addition, we also perform ablation studies, proving the contribution of different metapath types and different components of MHTAN-DTI for improving DTI prediction performance. The visualization results of MHTAN-DTI intuitively show the performance of our model. All the results demonstrate that MHTAN-DTI can provide a powerful and robust tool to facilitate drug discovery and drug repositioning. In the future, we will expand MHTAN-DTI by integrating more heterogeneous attribute information, and further verify part of the prediction results through wet-lab experiments.

Key Points

- We construct a metapath instance dataset based on multiple types of nodes and association to better describe the relations between drugs and targets.
- We propose MHTAN-DTI, an end-to-end DTI prediction framework, to effectively capture the structural and semantic information.
- MHTAN-DTI introduces transformer to message passing between the nodes of different types, and adopts two-level attention mechanism to learn the underlying pharmacological knowledge, which outperforms various state-of-the-art DTI prediction methods.

Acknowledgments

This study is supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA16021400, XDB31000000, XDB38030300) and Informatization Plan of Chinese Academy of Sciences (CAS-WX2021SF-0101).

Data availability statement

The data and source code used in MHTAN-DTI are available at: https://cstr.cn/31253.11.sciencedb.01726 and https://github.com/ranzhran/MHTAN-DTI.

References

1. Drews J. Drug discovery: a historical perspective. *Science* 2000; **287**(5460): 1960–4.
2. Masoudi-Nejad A, Mousavian Z, Bozorgmehr JH. Drug-target and disease networks: polypharmacology in the post-genomic era. *In silico pharmacol* 2013; **1**(1): 1–4.
3. Masoudi-Sobhanzadeh Y, Omidi Y, Amanlou M, *et al*. Drug databases and their contributions to drug repurposing. *Genomics* 2020; **112**(2): 1087–95.
4. Cheng L, Han X, Zhu Z, *et al*. Functional alterations caused by mutations reflect evolutionary trends of sars-cov-2. *Brief Bioinform* 2021; **22**(2): 1442–50.
5. Chong CR, Sullivan DJ. New uses for old drugs. *Nature* 2007; **448**(7154): 645–6.
6. Zeng X, Liu L, Lü L, *et al*. Prediction of potential disease-associated micrornas using structural perturbation method. *Bioinformatics* 2018; **34**(14): 2425–32.
7. Zhang X, Zou Q, Rodriguez-Paton A, *et al*. Meta-path methods for prioritizing candidate disease mirnas. *IEEE/ACM Trans Comput Biol Bioinform* 2017; **16**(1): 283–91.
8. Shi Hua W, Yun ZZ, Zou Q. A discussion of micrornas in cancers. *Curr Bioinform* 2014; **9**(5): 453–62.
9. Zeng X, Liao Y, Liu Y, *et al*. Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Trans Comput Biol Bioinform* 2016; **14**(3): 687–95.
10. Shukla PK, Shukla PK, Sharma P, *et al*. Efficient prediction of drug–drug interaction using deep learning models. *IET Syst Biol* 2020; **14**(4): 211–6.
11. Lun H, Wang X, Huang Y-A, *et al*. A survey on computational models for predicting protein–protein interactions. *Brief Bioinform* 2021; **22**(5): bbab036.
12. Wang Z, Zou Q, Jiang Y, *et al*. Review of protein subcellular localization prediction. *Curr Bioinform* 2014; **9**(3): 331–42.
13. Zhao B-W, Lun H, You Z-H, *et al*. Hingrl: predicting drug–disease associations with graph representation learning on heterogeneous information networks. *Brief Bioinform* 2022; **23**(1): bbab515.

14. Zhao B-W, Xiao-Rui S, Peng-Wei H, *et al*. A geometric deep learning framework for drug repositioning over heterogeneous information networks. *Brief Bioinform* 2022; **23**(6): bbac384.
15. Ezzat A, Min W, Li X-L, *et al*. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform* 2019; **20**(4): 1337–57.
16. Wang T, Mian-Bin W, Zhang R-H, *et al*. Advances in computational structure-based drug design and application in drug discovery. *Curr Top Med Chem* 2016; **16**(9): 901–16.
17. Acharya C, Coop A, Polli JE, *et al*. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr Comput Aided Drug Des* 2011; **7**(1): 10–22.
18. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. *Methods* 2015; **74**:97–106.
19. Yamanishi Y. Chemogenomic approaches to infer drug–target interaction networks. *Data Min Syst Biol* 2013;97–113.
20. Zhang W, Zou H, Luo L, *et al*. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* 2016; **173**:979–87.
21. Zhang W, Chen Y, Liu F, *et al*. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinform* 2017; **18**(1): 1–12.
22. Van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PloS One* 2013; **8**(6):e66952.
23. Jian-Yu Shi and Siu-Ming Yiu. Srp: A concise non-parametric similarity-rank-based model for predicting drug-target interactions. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, DC, USA: IEEE, 2015. pp. 1636–41, https://doi.org/10.1109/BIBM.2015.7359921.
24. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009; **25**(18): 2397–403.
25. 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.
26. Buza K, Peška L. Drug–target interaction prediction with bipartite local models and hubness-aware regression. *Neurocomputing* 2017; **260**:284–93.
27. Gönen M. Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics* 2012; **28**(18): 2304–10.
28. Xiaodong Zheng, Hao Ding, Hiroshi Mamitsuka, and Shanfeng Zhu. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: Association for Computing Machinery, 2013; 1025–33, https://doi.org/10.1145/2487575.2487670.
29. Ezzat A, Zhao P, Min W, *et al*. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans Comput Biol Bioinform* 2016; **14**(3): 646–56.
30. Hao M, Bryant SH, Wang Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Sci Rep* 2017; **7**(1): 1–11.
31. Tabei Y, Pauwels E, Stoven V, *et al*. Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics* 2012; **28**(18): i487–94.
32. Xiao X, Min J-L, Wang P, *et al*. Igpcr-drug: a web server for predicting interaction between gpcrs and drugs in cellular networking. *PloS One* 2013; **8**(8):e72234.
33. 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.
34. 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.
35. Hirohara M, Saito Y, Koda Y, *et al*. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC Bioinform* 2018; **19**(19): 83–94.
36. Wang L, You Z-H, Chen X, *et al*. A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *J Comput Biol* 2018; **25**(3): 361–73.
37. 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.
38. Xiaoqing R, Wang L, Li L, *et al*. Exploration of the correlation between gpcrs and drugs based on a learning to rank algorithm. *Comput Biol Med* 2020; **119**:103660.
39. Cheng F, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012; **8**(5):e1002503.
40. 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.
41. Ba-Alawi W, Soufan O, Essack M, *et al*. Daspfind: new efficient method to predict drug–target interactions. *J Chem* 2016; **8**(1): 1–9.
42. 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.
43. Zeng X, Zhu S, Weiqiang L, *et al*. Target identification among known drugs by deep learning from heterogeneous networks. *Chem Sci* 2020; **11**(7): 1775–97.
44. Zhao T, Yang H, Valsdottir LR, *et al*. Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief Bioinform* 2021; **22**(2): 2141–50.
45. 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.
46. Wang H, Zhou G, Liu S, *et al*. Drug-target interaction prediction with graph attention networks arXiv preprint arXiv:2107.06099. 2021. https://doi.org/10.48550/arXiv.2107.06099.
47. 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.
48. Li Y, Qiao G, Wang K, *et al*. Drug–target interaction predication via multi-channel graph neural networks. *Brief Bioinform* 2022; **23**(1).
49. Liyi Y, Qiu W, Lin W, *et al*. Hgdti: predicting drug–target interaction by using information aggregation based on heterogeneous graph neural network. *BMC Bioinform* 2022; **23**(1): 1–18.
50. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *Advances in neural information processing systems 31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA 2017; **30**.
51. Zhang P, Wei Z, Che C, *et al*. Deepmgt-dti: transformer network incorporating multilayer graph information for

drug–target interaction prediction. *Comput Biol Med* 2022; **142**: 105214.
52. Wang G, Zhang X, Pan Z, *et al*. Multi-transdti: transformer for drug–target interaction prediction based on simple universal dictionaries with multi-view strategy. *Biomolecules* 2022; **12**(5): 644.
53. Liu J, Jiang T, Lu Y, Wu H. (2022). Drug-Target Interaction Prediction Based on Transformer. In: Huang DS, Jo KH, Jing J, Premaratne P, Bevilacqua V, Hussain A. (eds). *Intelligent Computing Theories and Application. ICIC 2022*. Springer, Cham: Lecture Notes in Computer Science, vol **13394**. https://doi.org/10.1007/978-3-031-13829-4_25.
54. Huang K, Xiao C, Glass LM, *et al*. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics* 2021; **37**(6): 830–6.
55. Zhang R, Wang Z, Meng Z, et al. Metapath-based heterogeneous bioinformatics network dataset[DS/OL]. *Science Data Bank, 2022[2023-02-28]*. https://cstr.cn/31253.11.sciencedb.01726. CSTR:31253.11.sciencedb.01726.
56. Knox C, Law V, Jewison T, *et al*. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 2010; **39**(suppl_1): D1035–41.
57. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database—2009 update. *Nucleic Acids Res* 2009; **37**(Database): D767–72.
58. Davis AP, Murphy CG, Johnson R, *et al*. The comparative toxicogenomics database: update 2013. *Nucleic Acids Res* 2013; **41**(D1): D1104–14.
59. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010; **6**(1): 343.
60. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010; **50**(5): 742–54.
61. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981; **147**(1): 195–7.
62. Jaccard P. The distribution of the flora in the alpine zone. 1. *New Phytol* 1912; **11**(2): 37–50.
63. Xia Z, Ling-Yun W, Zhou X, *et al*. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In: *BMC systems biology*, Vol. **4**. BioMed Central, 2010, 1–16. https://doi.org/10.1186/1752-0509-4-S2-S6
64. Liu Y, Min W, Miao C, *et al*. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput Biol* 2016; **12**(2):e1004760.