MGDTI: Graph Transformer with Meta-Learning for Drug-Target Interaction Prediction

MGDTI: Meta-Learning Based Graph Transformer for Drug-Target Interaction Prediction

Zhenjiang Zhao*, Chengxin He*, Yuening Qu†, Huiru Zheng*, Duan Lei*, Jie Zuo*‡§
*School of Computer Science, Sichuan University, Chengdu, China
†The West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
‡Med-X Center for Informatics, Sichuan University, Chengdu, China
§Corresponding authors. Email: hechengxin@stu.scu.edu.cn, leiduan@scu.edu.cn

***Abstract*—Drug-target interaction prediction (DTI) is of great importance for drug discover and development. With the rapid development of biological and chemical technologies, Network-based method for DTI prediction is becoming the promising strategy. However, there are few methods explore to solve cold-start problem. Most of existing methods modeling requires modeling under the existing interaction that can't effectively capture information from new drugs and new targets which have few interaction in existing network. In this paper, we propose a meta-learning graph transformer model named MGDTI to fill the gap. We employ drug-drug similarity and target-target similarity to supplemental additional information for network to mitigate the scarcity of interaction. Besides, we train model via meta-learning to fast adapt new tasks. Moreover, we introduce graph transformer to prevent over-smoothing by capturing long-range dependencies. Experimental results on the benchmark dataset demonstrate that MGDTI is effective in DTI prediction. Case study reveals the effective of MGDTI for predicting potential drug-target pairs.**

***Index Terms*—Drug-target interaction prediction, meta-learning , graph transformer**

I. Introduction

Drug-target interaction prediction(DTI) is of great significance for drug discover and development. However, traditional experiment are normally time-consuming and labor-intensive process [1], [2]. In order to speed up drug discovery, computer-based methods are proposed and rapidly developing [3]. Benefiting from the rapid development of biological and chemical technologies, heterogeneous biological data are increasingly giving a good foundation for the development of computer-based methods. Because network can make better use of heterogeneous biological data, so network-based method has becoming a commonly used strategy for DTI prediction [4]–[6]. Graph neural network(GNN) has been widely used in heterogeneous information networks models because of its mechanism of aggregating neighbours with message passing and also a few GNNs have been used in network-based methods for DTI prediction and make great success [7], [8].

However, in real-world prediction scenarios, for new drug development, there are few or no interactions between new drug and targets [9]. This problem is often referred to as cold-start problem in recommender system for it is hard train a reasonable embedding using limited interaction [10]. Although network-based methods have widely used in DTI prediction, they still have shortcoming in some aspects.

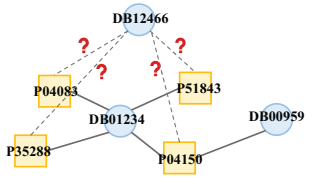

Fig. 1. cold-start example

Specifically, existing network-based methods mainly focus on the cases of existing drugs and existing targets(test drugs and targets are known in training set and have many edges in the network) ignoring the cold-start problem in DTI prediction, which makes model incapable of capturing information on cold-drugs and cold-targets. Precisely, cold-start problem in DTI prediction is classified as cold-drug task (predict new drugs interactions with targets) and cold-target task (predict new targets interactions with drugs). Otherwise, most network-based methods only focus on 1-hop neighbors information neglecting to extract long-range dependencies. Although GNN stacking can take advantage of information from distant nodes captured by the message passing mechanism, it can also suffer from problems such as over-smoothing [11].

To address the above problems, we use MGDTI, a meta-learning based graph transformer model to handle cold-start problem in DTI prediction. Our aim is to improve model generalization ability and capturing long-range dependencies ability on cold-start scenario. To enhance the ability of model's generalization, we introduce to train model via meta-learning to help it adapt fast to cold-drug task and cold-target task. It use drug-drug similarity matrix and target-target similarity matrix to supplement additional information to mitigate the scarcity of interaction. To prevent over-smoothing, we utilize node neighbour sampling method to derive the contextual sequence for each node and deliver them into graph transformer to capture the local structure information with context aggregation.

The main contributions are summarized as follows:

- We propose a meta-learning based graph transformer model to solve cold-start problem in DTI prediction named MGDTI.
- To address cold-start problem in DTI prediction, we

utilize similarity matrix to supplemental additional information for network to mitigate the scarcity of interaction and train model parameters through meta-learning to fast adapt cold-drug task and cold-target task to enhance the model's generalization capability. And we introduce graph transformer to capture for long-range dependencies for preventing model over-smoothing.
- We evaluate the performance of MGDTI on the benchmark dataset in cold-start scenario. Experimental results show that MGDTI is superior to state-of-the-art methods. And we design a case study on a real-world dataset to demonstrate that MGDTI is effective.

The rest of this paper is organized as follows. The related work is reviewed in Section I. Preliminaries and overall design are shown in Section III and Section IV, respectively. Then, the experimental setting and results are discussed detail in Section V. Finally, we conclude the paper in Section VI.

II. RELATED WORK

A. *Drug-target Interaction Prediction*

DTI is described as the binding of a drug molecule to a target (usually a protein), in which a drug interact a target to treat diseases. Improving the accuracy of DTI predictions could lead to faster drug development.

Most of the early work are computer-based which have many limitations and cost time. docking-based method [12] require three dimensional structure of the target while ligand-based method use known rules of interacting ligands to make predictions [13]. These methods are time-costing and not work well if some limitations can not be satisfied.

With the rapid advances in biological and chemical technologies, more sources of data are becoming available, such as drug-drug interactions and drug-target interactions, which could help the DTI task to predict more accurately. Because of the availability of data from more sources, network-based approaches have been created, which use graph-based technology to characterize the properties of drugs and targets to predict DTI tasks. DTINet [14] learns low-dimensional feature vectors of drugs and targets from heterogeneous networks, then finds the optimal projection from drug space to target space and predicts interactions. IMCHGAN [15] employs a two-level GAT strategy to learn drug and target latent feature representations from multiple networks and predicts DTI using inductive matrix completion. HGAN [5] , based on attentional mechanisms and diffusion techniques, captures complex structures and rich semantics in bioheterogeneous graphs for DTI prediction.

B. *Graph Transformer*

Recently, Graph transformer is growing popular for they could alleviate the limitations of Message-Passing-based GNN models such as over-smoothing, over-squashing and so on. Graph attention mechanism has shown its power in many different graph representation learning tasks. Transformers mainly integrate graph structural information into the transformer architecture to generalize graph-structured data. Some works, such as GraphTrans [16] , GraphiT [17] combine with GNNs to capture local structure information. And some works propose to add graph and structural encoding to complement topological information into Graph Transformer. HINormer [18], a Graph Transformer on heterogeneous information network utilize a local structure encoder and a heterogeneous encoder for node representation learning has achieved excellent results.

C. *Meta-learning*

Meta-learning [19] , an approach often regarded as "learn to learn", aims to enhance the model's ability to quickly adapt to new tasks. For example, MAML [20] is a popular deep neural network meta-learning framework, which learned cross-task generalities, serves as an initialisation of the neural network parameters. During the meta-testing phase, MAML can fine-tune the parameters with a small number of training examples so that the model can adapt efficiently to new learning tasks. Due to its quickly adapt ability, meta-learning has been widely used in the field of recommender systems to solve cold-start problems and have achieved great success [21], [22] .

With the rapid development of meta-learning, more and more bioscience-related tasks are also beginning to use it. META-DDIE [23] utilize meta-learning to solve few-shot drug-drug interaction. CML [9] , a meta-learning based approach for drug-target binding affinity prediction task.

III. PRELIMINARIES

Definition 1: **Drug-target information network:** Drug-target information network (DTN) is an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with a node type mapping function $\phi : V \rightarrow O$ and a relation type mapping function $\psi : E \rightarrow R$, where $O = \{\phi(v) | v \in V\}$ refers to the set of node types and R denotes the set of relations between nodes. Each node $v \in V$ belongs to a node type $\phi(v) \in O$, and each link $e \in E$ belongs to a relation $\psi(e) \in R$.

In the drug-target information network, node type set $O = \{drug, target\}$ and relation type set $R = \{drug - target - interaction, drug - drug - interaction, target - target - interaction\}$. Particularly, we denote $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ as drug set and $\mathcal{T} = \{t_1, t_2, ..., t_m\}$ as target set, $V = \mathcal{D} \cup \mathcal{T}$ obviously.

Definition 2: **Cold-drug set and Cold-target set:** We define $\mathcal{D}^c \subset \mathcal{D}$ as cold-drug set which have few interactions with node set V and $\mathcal{T}^c \subset \mathcal{T}$ as cold-target set which few interactions with node set V.

Definition 3: **Cold-start task Defination:**

- **Cold-drug task:** Given the drug-target information network $\mathcal{G}$, our goal is to learn a mapping function $\mathcal{F}$ which can predict the interaction probability between each pair of a drug $v \in \mathcal{D}^c$ and a target $v^{'} \in \mathcal{T}$.
- **Cold-target task:** Given the drug-target information network $\mathcal{G}$, our goal is to learn a mapping function $\mathcal{F}$ which can predict the interaction probability between each pair of a target $v^{'} \in \mathcal{T}^c$ and a drug $v \in \mathcal{D}$.

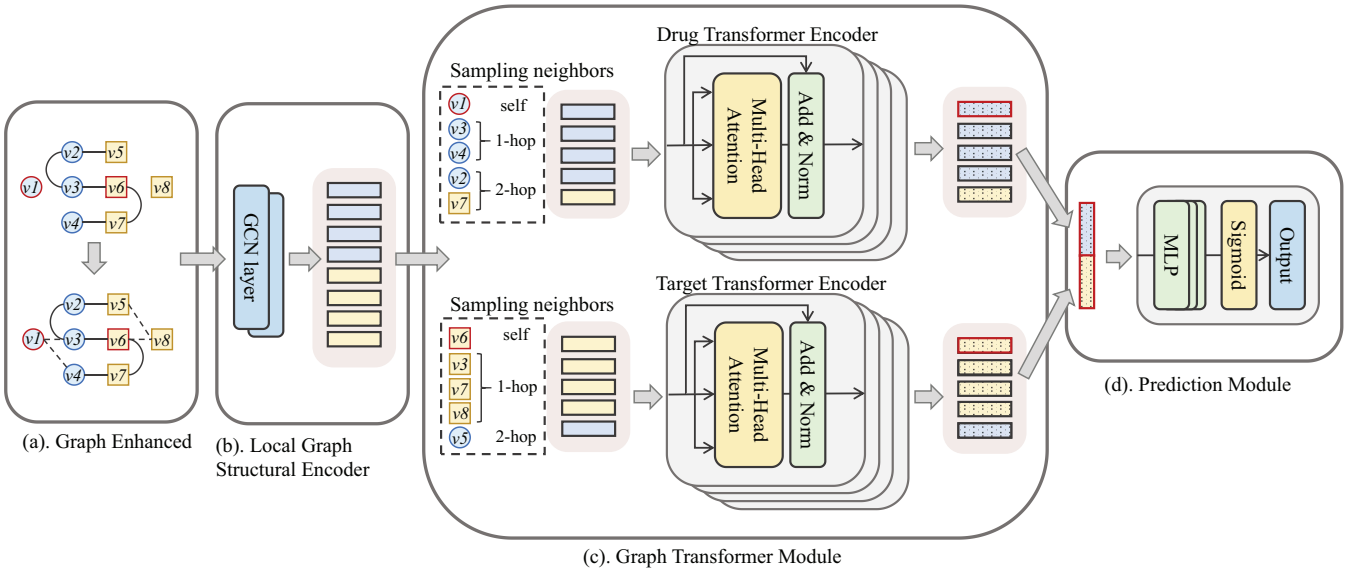

Fig. 2. **The overall architecture of MGDTI (a) Enhance graph by adding edges using similarity matrix (b) Using GCN to capture the local structural information and get all nodes' embedding (c) Do neighbor sampling for drugs and targets and feed them into different transformer to learn long dependicies respectivly (d) DTI prediction**

IV. Method

In this section, we presents the details of MGDTI. The framework of MGDTI is shown in 2.

A. Graph Enhanced Module

To help model solve the cold-start problem, MGDTI supplements the additional information with structural similarity between drugs and targets, respectively. For each drug $v \in \mathcal{D}$, we choose the top five drugs with highest structural similarity to drug v and add k edges to DTN $\mathcal{G}$. We do the same for each target $v^{'} \in \mathcal{T}$. After graph enhanced, we get the new DTN graph $\mathcal{G}^{'}$.

B. Local Graph Structural encoder

In this module, MGDTI utilize local graph structure encoder to learn nodes' embedding to fully capture local structural information from DTN $\mathcal{G}^{'}$. For each node $v \in V$, MGDTI randomly initialize its embedding into a d-dimensional latent space. In addition, we aggregate all nodes' embedding to form the embedding matrix $\mathbf{H}^{(0)} \in \mathbb{R}^{|\mathcal{V}| \times d}$ Graph Convolutional Network(GCN) [24] has been widely used in graph for its message-passing mechanism can effectively captures the local structural information. Formally, for the l-layer output

$$H^{(l+1)} = RELU(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

where $H^{(l)}$ is the feature representation at the $l-$th layer. Here, $\tilde{A} = A + I_N$ is the adjacency matrix of the undirected graph $\mathcal{G}^{'}$ with added self-connections where I_N is the identity matrix and $\tilde{D}$ is the degree matrix of $\tilde{A}$. $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ represents the normalized adjacency matrix which play a key role in solving gradient explosion problems during graph convolution process. And $W^{(l)}$ is $l-$layer trainable weight matrix. After Local Graph Structural encoder, we get nodes' new embedding H.

C. Graph Transformer Module

In this module, we aim to capture information from long-range dependencies through graph transformer because remote nodes own some useful information that neighbours don't have [25].

Inspired by HINormer [18], the model samples the fixed number (n) of node $v's$ neighbors as a neighborhood sequence S_v, whose embedding is used as the input to transformer encoder. For the node v we first sample itself, and then preferentially sample its 1-hop nodes into neighbourhood sequence S_v. If $|S_v| < n$, we sample from its 2-hop neighbours and so on, until $|S_v| = n$.

Formally, the neighborhood sequence S_v of node v is denoted as $S_v = [v, v_1, ..., v_{n-1}]$, so the embedding of neighborhood sequence is denoted as $\mathbf{H}_v^S = [H_v, H_{v_1}, ..., H_{v_{n-1}}] \in \mathbb{R}^{n \times d}$. Transformer [26] has been widely used in different fields due to its excellent learning ability for sequences. The standard transformer layer consists of two main components, multi-head self-attention module(MSA) and the feed forward network(FFN). We delete FFN in MGDTI, so we only briefly introduce MSA simplicity.

MSA allows model to learn multiple sets of attention weight in parallel to extract features from different subspaces and fuse them together to enhance model's representation power.

For the neighborhood sequence embedding of node v $\mathbf{H}_v^S$ to learn a set of attention weight $Attention$, the MSA firstly projects the input H_v^s to the query space, key space, and value space (denoted by Q, K, and V, respectively) via the three parameter matrices W_Q, W_K, W_V and

$$Q = H_v^s W_Q, K = H_v^s W_K, V = H_v^s W_V. \tag{2}$$

The attention weight is then calculated as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V. \quad (3)$$

We calculate two independent self-attention on $\mathbf{H}_v^S$ to obtain two sets of attention weight $Attention_1, Attention_2$ and we get MSA output by concatenating and linearly transforming them:

$$MSA(H_v^s) = Concat(Attention_1, Attention_2)W_M \quad (4)$$

where W_M is a learnable parameter. After that, the output of MSA will be connected to layer normalisation(LN) [27] and residual linkage. Formally, for each transformer encoder layer is denoted as:

$$\mathbf{H}^{j+1} = LN(MSA(\mathbf{H}^j) + \mathbf{H}^j). \quad (5)$$

where $\mathbf{H}^0 = \mathbf{H}_v^S$. After J layer transformer, the final output of transformer encoder is denoted as $\tilde{\mathbf{H}}_v^s \in \mathbb{R}^{n \times d}$.

For node v, we use the $\tilde{\mathbf{H}}_v^s[0]$ as its new embedding. After graph transformer module, MGDTI can learn the node feature based proximity between different positions of the neighborhood sequence. It is worth noting that when updating node's embedding of drug node and target node, we use different graph transformer module. Now, MGDTI gets drug node $v's$ embedding $Z_v = \tilde{\mathbf{H}}_v^s[0]$ and target node $v^{'}s$ embedding $Z_{v'} = \tilde{\mathbf{H}}_{v'}^s[0]$, respectively .

D. Prediction Module

MGDTI concatenates Z_v and $Z_{v'}$ as the input of the prediction module, which is a 3-layer MLP. The output of MLP is a prediction score indicating the probability of DTI, which is denoted as

$$\hat{y} = MLP(Z_v \oplus Z_v^{'}) \quad (6)$$

We convert the DTI prediction task into a binary classification task and user binary cross entropy loss:

$$\mathcal{L} = -y \cdot log(\hat{y}) - (1 - y) \cdot log(1 - \hat{y}) \quad (7)$$

where y is the ground truth, and $\hat{y}$ is the prediction value.

E. Meta-learning training method

To tackle the problem of data imbalance under cold-start scenarios, we introduce to use meta-learning to train model parameters.

Given a model f_θ with randomly initialized model parameter θ, the key idea of meta-learning is to learn the optimal parameters θ^* for different tasks so as to quickly adapt to new tasks. Firstly, for each epoch, we randomly divide training set into support set $\{X^p, Y^p\}$ and query set $\{X^q, Y^q\}$. And for each time, we copy the model parameter θ as $\hat{\theta}$, which will be updated by loss of support set $\mathcal{L}_p$:

$$\hat{\theta} \leftarrow \hat{\theta} - \alpha \nabla \hat{\theta} \mathcal{L}_p \quad (8)$$

TABLE I
STATISTICS OF THE DATASET

Node Type	# Nodes	Edge Type	# Edges	Source
		drug-target	1,923	DrugBank [28]
drug	708	drug-drug	10,036	DrugBank [28]
target	1512	target-target	7,363	HPRD [29]
side effect	4192	drug-disease	199,214	CTD [30]
disease	5603	target-disease	1,596,745	CTD [30]
		drug-side effect	80164	SIDER [25]

Then use the parameter $\hat{\theta}$ to train query set and get the loss of query set $\mathcal{L}_q^i$. We repeat the above steps k times, and get the average loss $\mathcal{L}_{mean}$, which is calculated as:

$$\mathcal{L}_{mean} = \frac{\sum_{i=1}^{k} \mathcal{L}_q^i}{k} \quad (9)$$

and use it to optimize the original parameters θ of model

$$\theta \leftarrow \theta - \beta \nabla \hat{\theta} \mathcal{L}_{mean} \quad (10)$$

where α and β are hyperparameter commonly called local update learning-rate and global update learning-rate, respectively. The detailed meta-learning process is shown with algorithm 1.

Algorithm 1 Meta-learning for MGDTI

Require: $\mathcal{T}$: training set
Require: α, β, k: step size hyperparameters
randomly initialize θ
while not done **do**
  $\hat{\theta} \leftarrow \theta$
  $\mathcal{L}_{tot} \leftarrow 0$
  Randomly divide $\mathcal{T}$ into P and Q
  for $i \leftarrow 1$ and $i <= k$ **do**
    Evaluate $\mathcal{L}_p$ using P and $\mathcal{L}$ in Equation 7
    update $\hat{\theta}$ with gradient descent in Equation 8
    get $\mathcal{L}_q^i$ using $\hat{\theta}$ and Q in equation 7
    add $\mathcal{L}_q^i$ to $\mathcal{L}_{tot}$: $\mathcal{L}_{tot} \leftarrow \mathcal{L}_{tot} + \mathcal{L}_q^i$
  end for
  $\mathcal{L}_{mean} \leftarrow \mathcal{L}_{tot}/k$
  using $\mathcal{L}_{mean}$ to update θ in Equation 10
end while

V. EXPERIMENTS

A. Experimental setting

1) Datasets: We utilize the heterogeneous network dataset which widely used in pervious studies [14]. The detail of dataset can be found in Table I. In MGDTI, we only use two types of nodes, drug and target with three types of edges drug-drug interaction, drug-target-interaction and target-target interaction in the original table. We also incorporate drug-drug structure similarity network by the dice similarities of the Morgan fingerprints [31] with radius 2 which computed by RDKit (http://www.rdkit.org) and target-target structure similarity network by protein sequence similarity network

based on pair-wise Smith-Waterman scores [32]. The data of structure similarity networks are obtained from DTINet [14]

Dataset prepare: To evaluate cold-start issue in DTI prediction, we perform distinct experimental setting split strategies for dataset. We divide the cold-start task into two categories, cold-drug task and cold-target task.

Taking cold-drug task as an example, firstly, we divided drugs into 10 parts to do 10-fold cross-validation. For each fold, we treat one part as the cold-drugs and the remaining nine parts as existing drugs. For each cold-drug, we mask a certain ratio of edges, including drug-drug-interaction and drug-target-interaction in the network to create a cold-drug scenario. Then, for each fold, we use masked edges of the drug-target and the same number of negative samples (i.e.,cold-drug and target have no interaction) as the test set while use unmasked edges in the network and same number of negative samples as training set. To evaluate the performance of model in different cold-drug scenario, we set different mask ratios 0.5, 0.7, 0.9 and 1.0 to simulate how much information that cold-drugs know. For the cold-target task, we do the same as above.

2) Baselines: We compared MGDTI with four advanced models with their default parameter setting:

- IMCHGAN [15]: Inductive Matrix Completion with Heterogeneous GAT for DTI prdiction. IMCHGAN employs a two-level GAT strategy to learn drug and target latent feature representations from multiple networks and predicts DTI using inductive matrix completion.
- SGCL-DTI [4]: Supervised graph co-contrastive learning for DTI prediction. SGCL-DTI contrasts topology structures and semantic features of drug-protein pair(DDP) network to predict DTI.
- MultiDTI [6]: Multi-modal representation learning for DTI prediction. MultiDTI uses joint learning in multimodal representation learning to combine similarity-based methods with network-based methods, mining not only structural information of drugs and targets, but also association information in heterogeneous networks to predict DTI. MultiDTI uses extra sequence data of drugs and targets, and can predict target for cold-drugs.
- HGAN [5]: Heterognous Graph Attention Network for DTI prediction. HGAN constructs the biological heterogeneous graph with both node information and edge type information and then utilize the graph attention diffusion module to learn deep representations for each bioentity node to predict DTI.

3) Experiment settings: To evaluate the effectiveness of MGDTI, we employ two commonly used metric : the area under the receiver operating characteristic curve (AUC). the area under the precision-recall curve (AUPR).

For MGDTI, the number of GCN layer parameter is 2, the number of transformer layer parameter is 3, the number of meta layer(k) is 5 and the dimension of representation is 256. We trained MGDTI by using Adam optimizer with local-update learning-rate(α) $1e^{-4}$, global-update learning-rate(β) $2e^{-4}$ and weight decay rate $1e^{-9}$. The epoch is set to 200. The sequence length n is set to 10.

TABLE II
COMPARISON OF RESULTS BETWEEN OUR MODEL AND BASELINES ON COLD-DRUGS

Mask-Rate	Metric	Multi DTI	SGCL-DTI	IMCH GAN	HGAN	MGF DTI
0.5	AUC	0.870	0.871	0.903	0.924	**0.927**
	AUPR	0.887	0.890	0.924	**0.936**	0.929
0.7	AUC	0.853	0.861	0.894	0.908	**0.923**
	AUPR	0.869	0.875	0.910	0.920	**0.932**
0.9	AUC	0.853	0.884	0.875	0.893	**0.913**
	AUPR	0.873	0.863	0.894	0.903	**0.921**
1.0	AUC	0.852	0.832	0.872	0.881	**0.907**
	AUPR	0.870	0.845	0.887	0.880	**0.915**

TABLE III
COMPARISON OF RESULTS BETWEEN OUR MODEL AND BASELINES ON COLD-TARGETS

Mask-Rate	Metric	Multi DTI	SGCL-DTI	IMCH GAN	HGAN	MG DTI
0.5	AUC	0.834	0.866	0.897	0.921	**0.936**
	AUPR	0.848	0.891	0.918	0.939	**0.940**
0.7	AUC	0.815	0.853	0.887	**0.919**	0.917
	AUPR	0.827	0.876	0.913	**0.936**	0.920
0.9	AUC	0.744	0.816	0.848	0.861	**0.897**
	AUPR	0.745	0.845	0.880	0.888	**0.904**
1.0	AUC	0.732	0.733	0.672	0.624	**0.854**
	AUPR	0.712	0.698	0.669	0.650	**0.859**

All experiments were conducted on a PC with four Intel Xeon E5-2698 GHz CPUs, four GeForce RTX 2080/3090 Ti GPUs and 512 GB memory, running Ubuntu 20.04. The algorithms were implemented in Python and compiled by Python 3.7. Our source code and the source codes of the compared methods, are publicly available on GitHub1.

B. Effectiveness

To evaluate the performance of models, we implemented 10-fold cross validation on the dataset, and show the average of ten fold results. We compare MGDTI with baselines on the the cold-start DTI prediction task. The results of cold-drug task is shown in II while the results of cold-target task is shown in III. It is shown that MLGTDTI achieves best results on most of experimental conditions on both cold-drug task and cold-target task. And for other experimental conditions, MLGTDTI achieves second best. According to the results, we can draw a conclusion that MLGTDTI is more suitable to tackle the cold-start scenarios than other baselines. As mask-rate increases, AUC and AUPR of all models have a certain degree of decline. We analyse that network-based methods require the aggregation of neighbors' information to get the representation of the drugs and targets, but in cold-start scenarios, cold-drugs and cold-targets has little of almost no

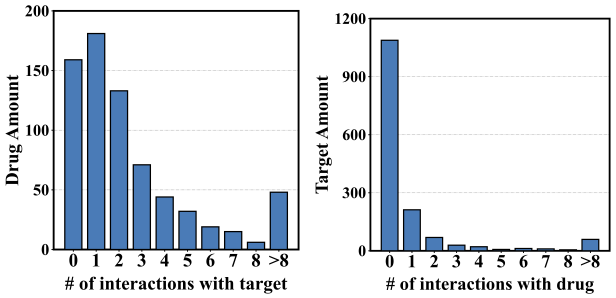

Fig. 3. drug-target interaction distributions

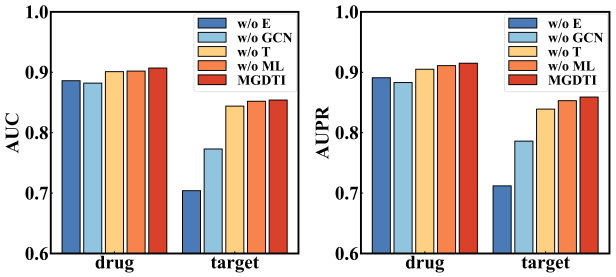

Fig. 4. Ablation study

interaction information in the network which will dramatically limit the performance of DTI prediction. Specifically, as the mask-rate increase, the experimental results of the cold-target task drop more significantly than the cold-target task. To explore the reasons for this case, we analyses the interaction between drugs and targets, which is shown in Figure 3. We also analysed the interaction between drugs and targets, the xx is shown in Fig. From the figure, we can see that most of drugs have at least one interaction with targets, while more than 70% of targets do not have any interaction with drugs in original DTN. It explains why the prediction performance of cold-target decreases much more than of cold-drugs while the mask rate increase.

C. Ablation Study

In order to understand the contribution of each component of MGDTI, we design four model variants as follows:(i) **w/o E:** MGDTI without graph enhanced module; (ii) **w/o GCN:** MGDTI without GCN layer; (iii) **w/o T:** MGDTI without graph transformer layer; (iv) **w/o ML:** MGDIT without meta-learning framework. In Figure 4, we can see that MGDTI outperforms the others and draw the following conclusions:

- Meta-learning framework can improve the model's performance of DTI prediction in cold-start scenarios, which because meta-learning can accumulate knowledge from experiences of learning similar task and seek rapid model adaption to unseen task.
- The graph enhanced module greatly affects the performance of the model, one potential reason is that graph enhanced module can complement the lack of information in cold-start scenarios. Specially, the graph enhanced module has a greater impact on cold-target task than on cold-drug task. It may because that cold-target originally has fewer edges in DTN, additional edge information is more necessary for cold-target task.
- Compared with w/o T and w/o GCN, MGDTI achieves better prediction results, proving that it is necessary to learn information in network through GCN layer and graph transformer layer.

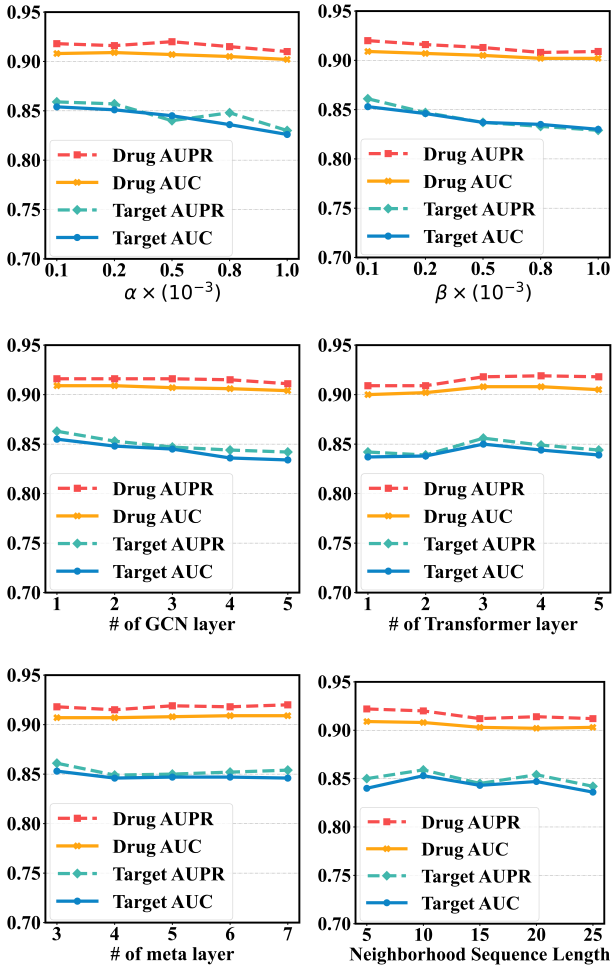

Fig. 5. parameter sensitivity

D. Parameters sensitivity study

MLDTDTI has six important hyper-parameters, including local-update learning rate α ,global-update leraning-rate β in meta-learning framework, num of meta-layer k, neighborhood length n, num of GCN layer L and num of Transformer layer J in MGDTI. In order to analyze the robustness of MGDTI, we conduct parameter sensitivity experiments to validate the impact of these hyper-parameters. When comparing a parameter, we keep the rest of hyper-parameters constant. performances are presented in Figure 3.

In order to analyze the influence of the num of GCN layer on MGDTI, we change the L in $\{1, 2, 3, 4, 5\}$. For cold-drug

TABLE IV
CASE-STUDY

UniProt ID	Target Name	DrugBank ID	Drug Name	MGLDTI Predict Score	Evidence
P37231	Peroxisome proliferator -activated receptor gamma	DB01050	Ibuprofen	0.993	NCT04334629/NCT04382768 [33]
O75469	Nuclear receptor subfamily 1 group I member 2	DB01234	Dexamethasone	0.992	NCT04509973/NCT04452565 [34]
P10415	Apoptosis regulator Bcl-2	DB01050	Ibuprofen	0.988	NCT04334629/NCT04382768 [33]
Q9BYF1	Angiotensin-converting enzyme 2	DB00608	Chloroquine	0.984	NCT04347798/NCT04303299 [35]
P04083	Annexin A1	DB00959	Methylprednisolone	0.976	NCT04244591/NCT04499313 [36]
Q07869	Peroxisome proliferator-activated receptor alpha	DB01050	Ibuprofen	0.968	NCT04334629/NCT04382768 [33]
P09488	Glutathione S-transferase Mu 1	DB00608	Chloroquine	0.820	NCT04347798/NCT04303299 [35]

task, the best performance is achieved when L=2, and drop slightly while L increases. And for cold-target task, the best performance is achieved when L=1, and drop obviously while L increases. The potential reason may be that the increase in the number of GCN layers may bring about the problem of over-smoothing, which is more abvious for cold-target task. neighborhood sequence length n, we change n in {5, 10, 15, 20, 25}. MGDTI gets best results when n=10 in both cold-drug and cold-target task. When n greater than 10, the performance of MGDTI become unstable, which may because too much information of neighbors may cause redundancy. And when n is set to 5, the information of neighbors is insufficient. To analyze the influence of the num of Transformer layer J, we set the J in the range {1, 2, 3, 4, 5}. The AUC and AUPR are the best for both cold-drug task and cold-target task when the Transformer layer J is 3.

We also assess the sensitivity of the num of meta-layer k when k changes in {3, 4, 5, 6, 7}. As shown in Figure 5, the overall performance is stable in both tasks, that is our results would not influence too much if k changes in an appropriate range. For cold-drug task, MGDTI achieves relatively better when k is 5, while achieves relatively better when k is 3 for cold-target task. We further analyze the changes in the local-update learning-rate α and global-update learning-rate β of meta-learning framework , and α and β varies from $1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 8 \times 10^{-4}, 1 \times 10^{-3}$. As we can see from Figure 6, for α, MLDTDTI has best AUC and AUPR in cold-target task, and relatively better AUC and AUPR in cold-drug task when $\alpha = 1 \times 10^{-4}$. And when $\beta = 1 \times 10^{-4}$, MGDTI performs best.

1) Case study: To evaluate the utility of model, we conduct a study of potential drug-target pairs which are real but not included in the benchmark dataset due to timeliness. We obtain 12 drug-target pairs from [37] which have no interaction in our benchmark. And then put them into our trained model for test. MGDTI successfully predict 7 of 12 pairs on 4 drugs and these drugs have appeared in some clinical studies and corresponding literatures. The result are shown in table IV.

VI. CONCLUSION

In this paper, we investigate the problem of cold-start scenarios in DTI prediction. To solve this problem, we proposed a novel model named MGDTI which train model parameters via meta-learning to fast adapt cold-drug task and cold-target task and utilize graph transformer to capture long-range dependencies for preventing over-smoothing. For feature work, we intent to capture the structural information of drugs and targets to find out which part of the structure works.

REFERENCES

[1] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of health economics*, vol. 22, no. 2, pp. 151–185, 2003.

[2] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve r&d productivity: the pharmaceutical industry's grand challenge," *Nature reviews Drug discovery*, vol. 9, no. 3, pp. 203–214, 2010.

[3] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: predicting drug–target binding affinity with graph neural networks," *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, 2020.

[4] Y. Li, G. Qiao, X. Gao, and G. Wang, "Supervised graph co-contrastive learning for drug-target interaction prediction," *Bioinformatics*, vol. 38, no. 10, pp. 2847–2854, 2022.

[5] M. Li, X. Cai, L. Li, S. Xu, and H. Ji, "Heterogeneous graph attention network for drug-target interaction prediction," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pp. 1166–1176, 2022.

[6] D. Zhou, Z. Xu, W. Li, X. Xie, and S. Peng, "MultiDTI: Drug-target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network," *Bioinformatics*, vol. 37, no. 23, pp. 4485–4492, 2021.

[7] F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions," *Bioinformatics*, vol. 35, no. 1, pp. 104–111, 2019.

[8] L. Jiang, J. Sun, Y. Wang, Q. Ning, N. Luo, and M. Yin, "Heterogeneous graph convolutional network integrates multi-modal similarities for drug-target interaction prediction," in *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Houston, TX, USA, December 9-12, 2021*, pp. 137–140, 2021.

[9] M. Li, S. Xu, X. Cai, Z. Zhang, and H. Ji, "Contrastive meta-learning for drug-target binding affinity prediction," in *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022, Las Vegas, NV, USA, December 6-8, 2022*, pp. 464–470, 2022.

[10] Y. Zhu, R. Xie, F. Zhuang, K. Ge, Y. Sun, X. Zhang, L. Lin, and J. Cao, "Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks," in *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pp. 1167–1176, 2021.

[11] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, "Measuring and relieving the over-smoothing problem for graph neural networks from the topological view," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 3438–3445, 2020.

[12] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, 2009.

[13] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature biotechnology*, vol. 25, no. 2, pp. 197–206, 2007.

[14] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Communications*, vol. 8, no. 1, 2017.

[15] J. Li, J. Wang, H. Lv, Z. Zhang, and Z. Wang, "IMCHGAN: Inductive matrix completion with heterogeneous graph attention networks for drug-target interactions prediction," *IEEE ACM Trans. Comput. Biol. Bioinform.*, vol. 19, no. 2, pp. 655–665, 2022.

[16] Z. Wu, P. Jain, M. A. Wright, A. Mirhoseini, J. E. Gonzalez, and I. Stoica, "Representing long-range context for graph neural networks with global attention," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 13266–13279, 2021.

[17] G. Mialon, D. Chen, M. Selosse, and J. Mairal, "GraphiT: Encoding graph structure in transformers," *CoRR*, vol. abs/2106.05667, 2021.

[18] Q. Mao, Z. Liu, C. Liu, and J. Sun, "HINormer: Representation learning on heterogeneous information networks with graph transformer," in *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pp. 599–610, 2023.

[19] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial Intelligence Review*, vol. 18, no. 2, pp. 77–95, 2002.

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, vol. 70, pp. 1126–1135, 2017.

[21] Z. Zhu, J. Kim, T. Nguyen, A. Fenton, and J. Caverlee, "Fairness among new items in cold start recommender systems," in *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pp. 767–776, 2021.

[22] W. Ouyang, X. Zhang, S. Ren, L. Li, K. Zhang, J. Luo, Z. Liu, and Y. Du, "Learning graph meta embeddings for cold-start ads in click-through rate prediction," in *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pp. 1157–1166, 2021.

[23] Y. Deng, Y. Qiu, X. Xu, S. Liu, Z. Zhang, S. Zhu, and W. Zhang, "META-DDIE: Predicting drug-drug interaction events with few-shot learning," *Briefings in Bioinformatics*, vol. 23, p. bbab514, 2022.

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, p. Poster, 2017.

[25] K. Michael, C. Monica, L. Ivica, J. L. Juhl, and B. Peer, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biology*, vol. 6, no. 1, p. 343, 2010.

[26] V. Ashish, S. Noam, P. Niki, U. Jakob, J. Llion, G. A. N, K. Łukasz, and P. Illia, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

[27] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016.

[28] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. Guo, and D. S. Wishart, "DrugBank 3.0: A comprehensive resource for Omics research on drugs," *Nucleic Acids Research*, vol. 39, no. Database-Issue, pp. D1035–D1041, 2011.

[29] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human protein reference database - 2009 update," *Nucleic Acids Research*, vol. 37, no. Database-Issue, pp. 767–772, 2009.

[30] A. P. Davis, C. G. Murphy, R. J. Johnson, J. M. Lay, K. Lennon-Hopkins, C. A. Saraceni-Richards, D. Sciaky, B. L. King, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: Update 2013," *Nucleic Acids Research*, vol. 41, no. Database-Issue, pp. 1104–1114, 2013.

[31] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[32] T. F. Smith, M. S. Waterman, *et al.*, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[33] E. Rinott, E. Kozer, Y. Shapira, A. Bar-Haim, and I. Youngster, "Ibuprofen use and clinical outcomes in COVID-19 patients," *Clinical Microbiology and Infection*, vol. 26, no. 9, pp. 1259.e5–1259.e7, 2020.

[34] B. M. Tomazini, I. S. Maia, A. B. Cavalcanti, O. Berwanger, R. G. Rosa, V. C. Veiga, A. Avezum, R. D. Lopes, F. R. Bueno, M. V. A. O. Silva, F. P. Baldassare, E. L. V. Costa, R. A. B. Moura, M. O. Honorato, A. N. Costa, L. P. Damiani, T. Lisboa, L. Kawano-Dourado, F. G. Zampieri, G. B. Olivato, C. Righy, C. P. Amendola, R. M. L. Roepke, D. H. M. Freitas, D. N. Forte, F. G. R. Freitas, C. C. F. Fernandes, L. M. G. Melro, G. F. S. Junior, D. C. Morais, S. Zung, F. R. Machado, L. C. P. Azevedo, and C. C.-. B. I. Investigators, "Effect of dexamethasone on days alive and ventilator-free in patients with moderate or severe acute respiratory distress syndrome and COVID-19: The CoDEX randomized clinical trial," *JAMA*, vol. 324, no. 13, pp. 1307–1316, 2020.

[35] S. Meo, D. Klonoff, and J. Akram, "Efficacy of chloroquine and hydroxychloroquine in the treatment of COVID-19," *European Review for Medical and Pharmacological Sciences*, vol. 24, no. 8, pp. 4539–4547, 2020.

[36] A. Papamanoli, J. Yoo, P. Grewal, W. Predun, J. Hotelling, R. Jacob, A. Mojahedi, H. A. Skopicki, M. Mansour, L. A. Marcos, and A. P. Kalogeropoulos, "High-dose methylprednisolone in nonintubated patients with severe COVID-19 pneumonia," *European Journal of Clinical Investigation*, vol. 51, no. 2, p. e13458, 2021.

[37] C. He, Y. Qu, J. Yin, Z. Zhao, R. Ma, and L. Duan, "Cross-view contrastive representation learning approach to predicting dtis via integrating multi-source information," *Methods*, 2023.