Hierarchical and Dynamic Graph Attention Network for Drug-Disease Association Prediction

Shuhan Huang, Minhui Wang, Xiao Zheng, Jiajia Chen, and Chang Tang, *Senior Member, IEEE*

***Abstract*—In the realm of biomedicine, the prediction of associations between drugs and diseases holds significant importance. Yet, conventional wet lab experiments often fall short of meeting the stringent demands for prediction accuracy and efficiency. Many prior studies have predominantly focused on drug and disease similarities to predict drug-disease associations, but overlooking the crucial interactions between drugs and diseases that are essential for enhancing prediction accuracy. Hence, in this paper, a resilient and effective model named Hierarchical and Dynamic Graph Attention Network (HDGAT) has been proposed to predict drug-disease associations. Firstly, it establishes a heterogeneous graph by leveraging the interplay of drug and disease similarities and associations. Subsequently, it harnesses the capabilities of graph convolutional networks and bidirectional long short-term memory networks (Bi-LSTM) to aggregate node-level information within the heterogeneous graph comprehensively. Furthermore, it incorporates a hierarchical attention mechanism between convolutional layers and a dynamic attention mechanism between nodes to learn embeddings for drugs and diseases. The hierarchical attention mechanism assigns varying weights to embeddings learned from different convolutional layers, and the dynamic attention mechanism efficiently prioritizes inter-node information by allocating each node with varying rankings of attention coefficients for neighbour nodes. Moreover, it employs residual connections to alleviate the over-smoothing issue in graph convolution operations. The latent drug-disease associations are quantified through the fusion of these embeddings ultimately. By conducting 5-fold cross-validation, HDGAT's performance surpasses the performance of existing state-of-the-art models across various evaluation metrics, which substantiates the exceptional efficacy of HDGAT in predicting drug-disease associations.**

***Index Terms*—Drug-disease association prediction, graph attention network, hierarchical attention mechanism, bidirectional long short-term memory networks, residual connections.**

I. Introduction

THE analysis of drug-disease association prediction holds paramount importance in the domain of biomedicine [1], [2]. Its outcomes frequently wield substantial influence over the decisions made by medical researchers during clinical trials, ultimately shaping the timely and effective alleviation and resolution of human diseases through pertinent drugs.

In the realm of drug-disease association prediction, conventional wet lab experiments are marred by inefficiency, time consumption, and often yield suboptimal accuracy. In response to the stringent accuracy requirements in medical research, a surge of researchers are now delving into the formulation of efficient computational techniques to surmount the challenge of low accuracy in experimental findings.

The prevailing algorithms for forecasting drug-disease associations can be broadly classified into four categories: 1. Literature extraction-based methods; 2. Similarity-based methods; 3. Network fusion-based methods; 4. Deep learning-based methods.

Literature extraction-based methods commonly leverage natural language processing techniques to extract drug-disease associations from extensive biomedical literature. The gleaned information is subsequently harnessed to construct pertinent feature representations and train models using suitable algorithms for forecasting drug-disease associations. For instance, Karaa et al. [3] introduced a method employing natural language processing and UMLS(Unified Medical Language System) ontology, coupled with a support vector machine classifier, to extract semantic relationships between drugs and diseases, leading to substantial enhancements in performance. Besides, Wang et al. [4] employed pattern matching and network embedding algorithms to autonomously extract extensive and precise drug-disease pairs from medical literature, providing vital support for drug repurposing endeavors. Although literature extraction-based methods can leverage the wealth of medical literature to extract drug-disease associations with commendable interpretability, their predictive efficacy is considerably influenced by

Manuscript received 5 October 2023; revised 19 December 2023, 24 January 2024, and 28 January 2024; accepted 2 February 2024. Date of publication 6 February 2024; date of current version 5 April 2024. This work was supported by the National Natural Science Foundation of China under Grant 62076228. *(Corresponding authors: Jiajia Chen and Chang Tang.)*

Shuhan Huang and Chang Tang are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: huangshuhan@cug.edu.cn; tangchang@cug.edu.cn).

Minhui Wang is with the Department of Pharmacy, Lianshui People's Hospital of Kangda College Affiliated to Medical University, Huai'an 223300, China (e-mail: minhwang@163.com).

Xiao Zheng is with the School of Computer, National University of Defense Technology, Changsha 410073, China (e-mail: zhengxiao@nudt.edu.cn).

Jiajia Chen is with the Department of Pharmacy, The Affiliated Huai'an Hospital of Xuzhou Medical University, Xuzhou 221006, China, and also with The Second People's Hospital of Huai'an, Huai'an 223002, China (e-mail: jjiachen@outlook.com).

Digital Object Identifier 10.1109/JBHI.2024.3363080

the literature's quality and encounters various challenges during the information extraction process.

Similarity-based methods commonly establish a bipartite graph encompassing drugs and diseases. These methods subsequently employ similarity measurement techniques to evaluate the likeness between drugs and diseases, in conjunction with the structural attributes of the bipartite graph. Afterwards, they prognosticate latent drug-disease associations via graph algorithms operating on the similarity-based graph. For instance, Zhang et al. [5] introduced a novel computational approach that exclusively utilizes established drug-disease associations to foresee unobserved ones. This method employs linear neighbour similarities to compute the similarities between drugs and diseases, and subsequently employs a label propagation process to predict latent drug-disease associations within the similarity-based graph. Similarly, Di et al. [6] devised a novel approach for amalgamating drug-disease associations, drug and chemical data, drug target domain information, and target annotation data to facilitate drug repositioning. They introduced interaction profiles of drugs and diseases within a network, treating them as label information for training models to predict new candidates. Similarity-based methods adeptly account for the network's structural characteristics connecting drugs and diseases, thereby enhancing the comprehension of their associations. Nonetheless, in cases where drug-disease associations are sparse, these methods may encounter prediction challenges and struggle to unearth potential associations.

Network fusion-based methods commence by amalgamating data regarding drug and disease associations and generating a heterogeneous graph. Subsequently, network diffusion algorithms are employed to extract features from diffused nodes, and suitable machine learning models are utilized for training, culminating in performance evaluation. In recent studies, many researchers have employed neural networks based on heterogeneous graphs[7], [8], [9], [10], [11] for tasks such as Drug-Drug Interaction (DDI) and Drug-Target Interaction (DTI) predictions. For instance, Tanvir et al. [8] adopted HAN-DDI, a heterogeneous graph attention network to predict drug-drug interactions in an end-to-end mode. Peng et al. [11] developed the (EEG)-DTI model, which constructs a heterogeneous graph with different nodes representing drug-protein interactions, drug-disease interactions, and more. The model adopts GCNs to aggregate node information and make DTI predictions in an end-to-end mode. Despite the capability of network fusion-based methods to comprehensively consider the network structural attributes between drugs and diseases and capture their associations, these methods might overlook latent associations due to their heavy reliance on network propagation, potentially yielding incomplete prediction outcomes.

Deep learning methods are frequently employed across diverse practical tasks owing to their efficiency and high accuracy, including pattern recognition and object detection [12]. Similarly, these methods find application in drug-disease association prediction. Within this domain, deep learning methods generally encompass data preprocessing and feature extraction, followed by the selection of suitable deep learning models for training-such as CNNs, RNNs, GCNs, and more. Ultimately, the model undergoes evaluation, and hyperparameters are optimized. For example, Liu et al. [13] proposed a technique leveraging a deep neural network based on a heterogeneous drug-disease network to predict novel drug-disease associations, which involves constructing drug-drug and disease-disease similarity networks, and then integrating established drug-disease associations to extract topological features from the heterogeneous network for training the DNN model. Meanwhile, Xuan et al. [14] proposed a creative method for predicting drug and disease association by utilizing graph convolutional and fully-connected autoencoders with attention mechanisms to integrate drug-disease associations, disease similarities, multiple drug similarities, and drug node attributes. Although deep learning methods excel in numerous tasks, their interpretability often remains limited.

In this study, we present a novel hierarchical and dynamic graph attention network for the prediction of drug-disease associations. Firstly, we construct a heterogeneous graph utilizing drug-drug similarities, disease-disease similarities, and drug-disease associations. Secondly, we pioneer the incorporation of graph convolutional neural networks and bidirectional long short-term memory networks to enhance the aggregation of both node-specific and structural information. Thirdly, a hierarchical attention mechanism [15] and residual connections are innovatively integrated across diverse network levels, while the introduction of dynamic attention mechanisms within each layer. The feature embeddings of drug and disease are finally obtained through these processes. Ultimately, the model's performance is assessed through the evaluation of undetected drug-disease associations, facilitated by the fusion of embeddings. Significantly, HDGAT surpasses other baseline models, achieving an impressive area under the precision-recall curve of 0.2665 and a remarkable prediction accuracy of 0.9614 on main dataset.

Fig. 1 shows the whole workflow and details of HDGAT model. In the future, drug-disease association prediction will be widely applied to optimize clinical trials, and its results can scientifically guide doctors in devising more suitable treatment plans for patients. Moreover, in future research, employing large language models to extract drug-disease associations from literature data may also be a promising approach, which utilizes increased computational power to capture potential drug-disease associations in literature, offering a straightforward and efficient approach. Similarly, integrating further exploration of heterogeneous associations between drugs and diseases will be a future research direction, aiming to delve more profoundly into the fundamental and intricate connections between drugs and diseases. In the context of HDGAT, its application extends to drug repositioning, disease mechanism analysis, and related domains, thereby enhancing the efficiency in identifying suitable drugs for the treatment of diseases.

II. The Proposed Model

A. Graph Preprocessing

To capture the intricate interplay between drugs and diseases more effectively, we opted to create a heterogeneous graph encompassing drug-drug similarities, disease-disease similarities, and drug-disease associations. This heterogeneous graph encapsulates nodes and edges that exhibit a multitude of connections, thus enabling the propagation of node information and the acquisition of embeddings via graph convolution operations. As a result, it aids in amplifying the efficacy of the model. Thus, the construction of the heterogeneous graph emerges as a pivotal facet in the holistic advancement of the model [16], [17].

1) Calculations of Drug-Drug Similarities: To begin, we establish and elucidate the drug-drug similarity data. Due to the diverse features present in drugs, we employ distinct sets of

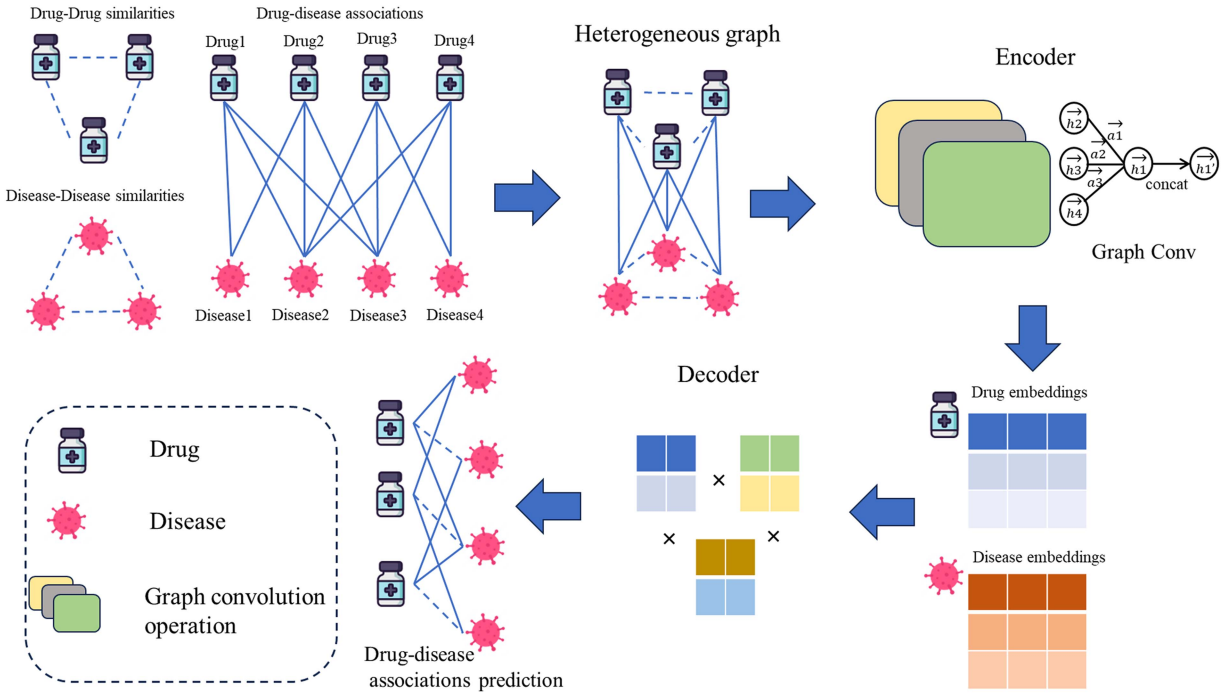


Fig. 1. Whole workflow and details of HDGAT: It contains three components: 1. Heterogeneous Network Construction: It incorporates drug-drug similarities, disease-disease similarities and drug-disease associations to construct a heterogeneous network; 2. Encoder for Embedding Learning: The heterogeneous network, as input graph, is processed by Encoder to learn drug and disease embeddings; 3. Adjacency Matrix Reconstruction: The Decoder reconstruct the adjacency matrix containing drug and disease associations based on drug embeddings and disease embeddings.

drug features to represent different categories of drugs. Utilizing existing data on diverse drug features, we analyze all possible features that a specific drug may have. If the drug possesses a particular feature, we set the corresponding value in the vector to 1; otherwise, it is set to 0. This approach enables us to obtain the binary feature vector of a drug. Besides, distinct binary feature vectors serve to distinguish various drugs, aiding in the representation of diverse drug entities. The collection of different drug features forms the feature matrix for that drug. Subsequently, various methods of similarity calculation can be employed to quantify the drug-drug similarity based on this feature matrix. Prominent techniques encompass the Jaccard index [18], cosine similarity [19], Euclidean distance [20], Manhattan distance [21], and other comparable approaches. Then we use a two-dimensional matrix, denoted as M, to represent the matrix of drug-drug similarities, where each row and column correspond to different types of drugs. The value M_{ij}^{r} represents the similarity between drug r_i and drug r_j.

The computation approach for evaluating drug-drug similarities using the Jaccard index is outlined as follows:

$$\mathbf{M_{ij}^{r}} = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} \tag{1}$$

where $|x_i \cap x_j|$ signifies the count of situations where elements in drug feature vector x_i and corresponding elements in feature vector x_j are both equal to 1, while $|x_i \cup x_j|$ denotes the count of situations where elements in drug feature vector x_i or corresponding elements in feature vector x_j are equal to 1.

The calculation method for cosine similarity in assessing drug-drug similarities is as follows:

$$\mathbf{M_{ij}^{r}} = \frac{x_i \cdot x_j}{\|x_i\| \cdot \|x_j\|} \tag{2}$$

where x_i and x_j represent drug feature vectors, $\|x_i\|$ and $\|x_j\|$ denote the L2-norm of x_i and x_j, respectively.

The calculation method for Euclidean distance in assessing drug-drug similarities is as follows:

$$\mathbf{M_{ij}^{r}} = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \tag{3}$$

where $x_{i,k}$ and $x_{j,k}$ represent the values of the k^{th} dimension in the feature vectors x_i and x_j, respectively.

The calculation method for Manhattan Distance in assessing drug-drug similarities is as follows:

$$\mathbf{M_{ij}^{r}} = \sum_{k=1}^{n}|x_{i,k} - x_{j,k}| \tag{4}$$

where $x_{i,k}$ and $x_{j,k}$ represent the values of the k^{th} dimension in the feature vectors x_i and x_j respectively.

Comparing the advantages and disadvantages of the four methods mentioned above, we have found that the evaluation method of Jaccard index is not influenced by feature scales. Unlike several other methods that employ distance measures, Jaccard index disregards the numerical values of elements. This insensitivity to numerical differences is advantageous when dealing with potential variations between different drugs and diseases since the Jaccard index focuses exclusively emphasizes

the presence or absence of elements. As the Jaccard index method is more suitable in this task, we adopt it to calculate drug-drug similarities.

2) Calculations of Disease-Disease Similarities: From the description in [22], we can infer that the MeSH (Medical Subject Headings) database can be utilized for naming diseases. Additionally, disease-disease similarities can be represented using directed acyclic graphs (DAGs). Within these DAGs, a disease A can be depicted in the form: $DAG_A = (S, D_A, E_A)$, where D_A represents the set of nodes including node A and all its ancestor nodes, and E_A is the collection of direct links between parent and child nodes. Based on the aforementioned DAG structure, we define $W_A(d)$ as the semantic contribution of a specific disease d to disease A within DAG_A, and its calculation method is as follows:

$$W_A(d) = \begin{cases} 1, & \text{if } d = A \\ \max\{\Delta \cdot W_A(d') \mid d' \in \text{children of } d\}, & \text{if } d \neq A \end{cases} \quad (5)$$

where Δ is a parameter representing the semantic contribution value of the edge link between disease d and its child disease d', ranging from 0 to 1. Based on previous experimental results, we set it to 0.5 here. Utilizing (5), we define the semantic value of disease A as $DV(A)$:

$$DV(A) = \sum_{d \in D_A} W_A(d) \quad (6)$$

When measuring the semantic similarities between two diseases, A_i and A_j, we take their positional relationship within the DAG into consideration. We hypothesize that having a greater number of common ancestors in the DAG tends to imply higher semantic similarities. We define the semantic similarities value between A_i and A_j as S_{ij}^A. Therefore, we derive the following calculation formula:

$$S_{ij}^A = \frac{\sum_{d \in D_{A_i} \cap D_{A_j}} (W_{A_i}(d) + W_{A_j}(d))}{DV(A_i) + DV(A_j)} \quad (7)$$

Equation (7) illustrates the semantic relationship among diseases A_i, A_j and their ancestor diseases.

3) Drug-Disease Associations: Drug-disease association data is a two-dimensional matrix denoted as $\mathbf{R}$, with dimensions $N * M$, where N represents the number of drug types and M represents the number of disease types. The element R_{ij} in the matrix corresponds to the value at the intersection of the i^{th} row and j^{th} column [13]. This value signifies the association between the i^{th} drug type and the j^{th} disease type. The values are as follows:

$$\mathbf{R_{ij}} = \begin{cases} 1, & \text{if drugs } i \text{ is associated with disease } j \\ 0, & \text{if drugs } i \text{ is not associated with disease } j \end{cases} \quad (8)$$

4) Construction of Heterogeneous Graph: The drug-disease heterogeneous graph is constructed based on theses three parts mentioned above.

The pairwise similarities among the N types of drug are represented using the similarity matrix $\mathbf{S^M} \in \mathbb{R}^{N*N}$, while the pairwise similarities among the M types of disease are represented using the similarity matrix $\mathbf{S^A} \in \mathbb{R}^{M*M}$. Here, S_{ij}^M and S_{ij}^A denote the elements at the indices (i, j) of these two matrices respectively. Furthermore, we normalize the two similarity matrices, $\tilde{\mathbf{S}}^{\mathbf{M}}$ and $\tilde{\mathbf{S}}^{\mathbf{A}}$, as follows:

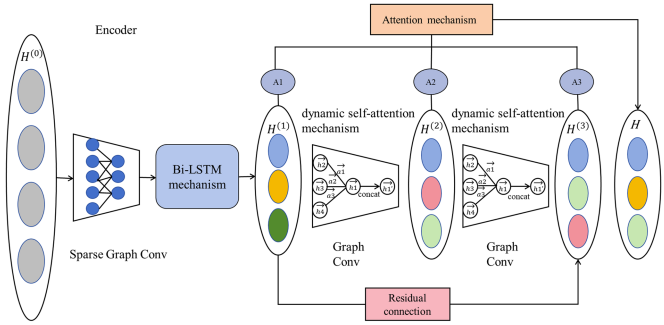


Fig. 2. Module of Encoder: It employs a hierarchical and dynamic graph attention network, coupled with Bi-LSTM mechanisms and residual connections, to process input data, capturing latent interactions between nodes and the structural information of a heterogeneous network.

$$\tilde{\mathbf{S}}^{\mathbf{M}} = (\mathbf{D_M})^{-\frac{1}{2}} S_{ij}^M (\mathbf{D_M})^{-\frac{1}{2}} \quad (9)$$

$$\tilde{\mathbf{S}}^{\mathbf{A}} = (\mathbf{D_A})^{-\frac{1}{2}} S_{ij}^A (\mathbf{D_A})^{-\frac{1}{2}} \quad (10)$$

where $\mathbf{D_M} = \text{diag}(\sum_j S_{ij}^M)$, $\mathbf{D_A} = \text{diag}(\sum_j S_{ij}^A)$. Based on these works and previous learning experiences [23], we construct the drug-disease heterogeneous graph:

$$\mathbf{H} = \begin{bmatrix} \tilde{\mathbf{S}}^{\mathbf{M}} & \mathbf{R} \\ \mathbf{R^T} & \tilde{\mathbf{S}}^{\mathbf{A}} \end{bmatrix} \quad (11)$$

Finally, in order to control the contribution of drug and disease similarities during the subsequent graph convolution propagation, we introduce a penalty factor λ. As a result, we obtain the ultimate input graph:

$$\mathbf{G} = \begin{bmatrix} \lambda \cdot \tilde{\mathbf{S}}^{\mathbf{M}} & \mathbf{R} \\ \mathbf{R^T} & \lambda \cdot \tilde{\mathbf{S}}^{\mathbf{A}} \end{bmatrix} \quad (12)$$

B. Encoder

Fig. 2 illustrates the encoder's processing procedure. It introduces a hierarchical mechanism among embeddings acquired from each layer and a dynamic attention mechanism among nodes. Furthermore, a Bi-LSTM module and a residual connection module are utilized to integrate node and structural information within the heterogeneous network. Detailed descriptions of these modules are provided below.

1) Graph Convolutional Network: In the prediction of associations between drugs and diseases, graph convolutional networks excel in learning low-dimensional node information. They efficiently aggregate information from neighbouring nodes around the central node and propagate it, thereby capturing graph structural features effectively, which confers an advantage in link prediction tasks.

Our input graph is a sparse two-dimensional matrix with a shape of $(N+M)*(N+M)$, where N and M represent the numbers of drug and disease, respectively. Due to its sparsity, we employ graph convolutional networks, which is suitable for sparse graphs, to aggregate neighbour node information and the topological structure of heterogeneous graphs. The propagation mechanism for each layer in the graph convolutional

network [24] is computed as follows:

$$\mathbf{H}^{(\mathbf{l+1})} = \sigma\left(\mathbf{D}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{G}\mathbf{D}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{H}^{(\mathbf{l})}\mathbf{W}^{(\mathbf{l})}\right) \tag{13}$$

where $\mathbf{H}^{(\mathbf{l+1})}$ and $\mathbf{H}^{(\mathbf{l})}$ represent the node embeddings of the $(l+1)^{th}$ and l^{th} layers respectively, D is the degree matrix of graph $\mathbf{G}$, defined as $\mathbf{D} = \text{diag}(\sum_j G_{ij})$, $\mathbf{W}^{(\mathbf{l})}$is the learnable weight matrix for layer l, and we initialize $\mathbf{W}^{(\mathbf{l})}$ using Xavier initialization. $\sigma(\cdot)$ denotes a nonlinear activation function.

During the application of graph convolutional network operations on a heterogeneous graph, information dissemination occurs among nodes throughout the graph. This process empowers central nodes to effectively collect insights from a diverse array of neighbouring nodes, leading to the refinement of their individual information. Furthermore, the utilization of the exponential linear unit as the activation function within the graph convolutional layer contributes to the augmentation of the model's capacity for generalization.

2) Dynamic Self-Attention Mechanism in Convolutional Layers: In traditional convolutional networks, the process of gathering information from neighbouring nodes usually involves assigning equal weights to all these neighbours. Nevertheless, in actual heterogeneous graphs characterizing drug-disease associations, the associations of different neighbour nodes to a central node evolves across layers. Thus, the assignment of distinct weights to diverse neighbour nodes becomes indispensable. To address this, we integrate a self-attention mechanism [25] into the graph convolutional layer, where the fundamental procedure for determining attention weights is outlined as follows:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^{\mathbf{T}} \cdot [\mathbf{W}h_i || \mathbf{W}h_j]\right) \tag{14}$$

$$\alpha_{ij} = \text{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in N_i}\exp\left(e_{ik}\right)} \tag{15}$$

In (14), e_{ij} represents the significance score of neighbour node j to node i's features. Meanwhile, h_i and h_j represent sets of node representations serving as inputs to a given layer. The $||$ denotes a concatenation operation. Both $\mathbf{a}$ and $\mathbf{W}$ are learnable parameters, initialized using Xavier initialization. $LeakyReLU(\cdot)$ represents a non-linear activation function. In (15), α_{ij} denotes the attention coefficient, N_i denotes the set of neighbouring nodes for node i, and $softmax_j(\cdot)$ is the normalization function applied.

In the attention calculation approach discussed above, attention coefficients exhibit a uniform ordering across all nodes within the graph and remain uninfluenced by the specific node. However, in the context of our heterogeneous graph depicting drug-disease associations, in order to more effectively capture the attributes of nodes and edges belonging to various types within this heterogeneous structure, we adopt a dynamic attention strategy [26]. To be precise, we enhance the mechanism by refining the formulation outlined in (16) as follows:

$$e_{ij} = \mathbf{a}^{\mathbf{T}}\text{LeakyReLU}\left(\mathbf{W}[h_i || h_j]\right) \tag{16}$$

Within this approach, unique key nodes are selected for different query nodes, thus yielding distinct neighbour nodes with different scores. This technique excels in effectively prioritizing relative information, surpassing the capabilities of static attention mechanisms, where the ranking of attention scores remains unconditioned on the query node, and showcasing enhanced resilience. Drawing upon the dynamic attention mechanism elucidated earlier, the weights attributed to diverse neighbour nodes are computed. Following this, the central node assimilates information from its neighbours contingent on these varied weights. Subsequent to a comprehensive normalization operation, the node's output feature vector is generated by amalgamating the node's associated features with attention coefficients and subjecting them to an activation function, as explicated in (17):

$$h_i' = \sigma\left(\sum_{j\in N_i}\alpha_{ij}\,\mathbf{W}h_j\right) \tag{17}$$

By integrating a dynamic attention mechanism into the graph convolutional layer, the model becomes adept at dynamically refining the central node's information in accordance with the varying significance of its neighbouring nodes. This augmentation not only fortifies the model's ability to generalize, but also effectively enhances the robustness when dealing with noisy data.

3) Bi-LSTM Module: The Bi-LSTM module [27], [28] is composed of two LSTM modules designed for processing input data. In the context of input data stemming from a heterogeneous graph of drug-disease associations, the first LSTM within the Bi-LSTM module is harnessed to manage the time-ordered data of the initial round. Concurrently, the second LSTM is responsible for processing the time-ordered data of the second round in a reversed sequence. At last,the outcome of the module is obtained through a multilayer perceptron. This dual-round LSTM processing strategy effectively captures long-range dependencies between drug and disease nodes embedded within the heterogeneous graph, thus enhancing prediction accuracy. Moreover, recognizing the potential existence of unobserved values within the heterogeneous graph, the Bi-LSTM module exhibits an elevated capacity to handle missing values during the processing of time-ordered data. This adaptive feature also serves to mitigate the impact of data incompleteness. Compared to the regular LSTM module, Bi-LSTM incorporates a process of acquiring node information from the reverse sequence in its structure. Through bidirectional processing of sequential data, it comprehensively captures contextual semantic information, thereby unveiling potential associations between drug and disease nodes. The distinct operational sequence of the Bi-LSTM is illustrated in Fig. 3.

4) Residual Connection Module: The utilization of a residual connection module [29] has proven effective in addressing challenges related to gradient vanishing and network degradation. In the context of this task, as the depth of graph convolutional networks increases, there is a susceptibility to encountering gradient vanishing issues during the propagation of node information through convolutional layers. To counteract this, we incorporate a residual connection module, which involves adding the initial input data to the output obtained from deeper convolutional layers. This integration serves to alleviate the predicaments associated with gradient vanishing. The computational process can be succinctly described as follows: $x_{l+1} = F(x) + x$, where x_{l+1} denotes the output derived from the deeper convolutional layer, x signifies the original input, and $F(x)$ represents the output subsequent to undergoing regular convolutional layer transformations.

The incorporation of residual connections can effectively mitigate the concern of over-smoothing [30] that often arises

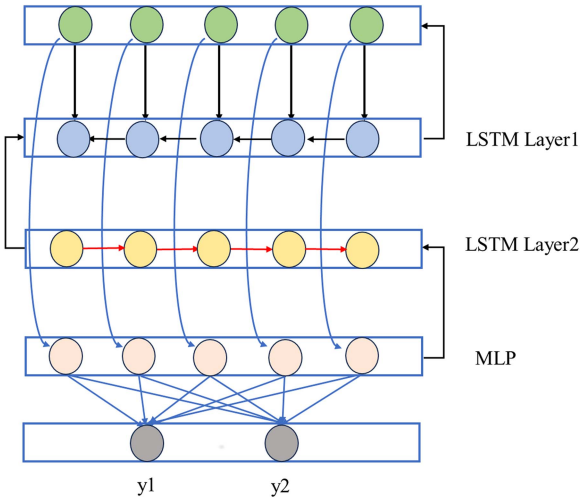


Fig. 3. Workflow of Bi-LSTM: The first LSTM Layer processes input sequences in a forward sequence and then the second LSTM layer processes the time-ordered data in a reversed sequence. Through a MLP, a new sequence aggregating more information of the context is obtained.

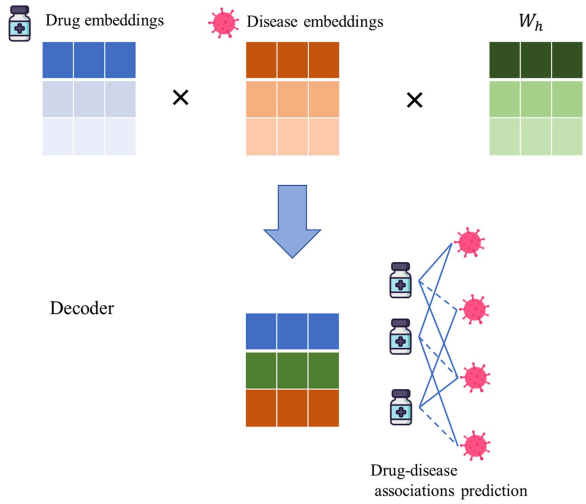


Fig. 4. Module of Decoder: It is responsible for reconstructing the adjacency matrix containing drug-disease associations. This is achieved using a bilinear decoder that relies on drug embeddings and disease embeddings.

during the execution of standard graph convolution operations. Moreover, it can also accelerate our model's convergence speed.

5) Hierarchical Attention Mechanism: In graph convolutional layers, each layer learns distinct node embeddings that encapsulate the structural characteristics of the drug-disease heterogeneous graph across multiple dimensions. Nevertheless, the embeddings across different layers might lack continuity, and the contributions of output embeddings from various layers to the final output embedding can vary. Building upon earlier studies [23], [31], we introduce a hierarchical attention mechanism to address this challenge. This mechanism assigns diverse weights to node embeddings originating from different layers, ultimately computing the resultant output node embedding. The ultimate representations of drug and disease node embeddings are formulated as follows:

$$\begin{bmatrix} \mathbf{E_M} \\ \mathbf{E_A} \end{bmatrix} = \sum \alpha^l \mathbf{H}^{(l)} \tag{18}$$

where $\mathbf{E_M}$ represents the final drug node embedding, $\mathbf{E_A}$ represents the final disease node embedding, α^l stands for the learnable weight of the l^{th} layer, and $\mathbf{H}^{(l)}$ denotes the node embeddings of the l^{th} layer.

Hierarchical attention mechanism comprehensively consolidates graph structural information embeddings acquired from diverse layers by attributing distinct weights to embeddings of each layer. This facilitates the acquisition of an augmented hierarchy of contextual node information, thereby amplifying the model's capacity for generalization and predictive precision.

C. Decoder

Once the encoder has generated the drug and disease node embeddings, the decoder is instrumental in reconstructing the adjacency matrix that captures the drug-disease associations. This begins with the extraction of separate node embeddings for drugs and diseases from the obtained node embeddings. Following this, a bilinear decoder is employed for the reconstruction task. The mathematical formulation for the bilinear decoder is as follows:

$$\mathbf{S}' = sigmoid(E_M \mathbf{W}' {E_A}^T) \tag{19}$$

where $\mathbf{S}'$ is the matrix of predicted association scores between drugs and diseases and its element s'_{ij} represents the predicted association score between drug i and disease j. $\mathbf{W}'$ is a trainable weight parameter, initialized using Xavier initialization [32], and $sigmoid(\cdot)$ represents a nonlinear activation function. The process of decoder is illustrated in Fig. 4.

D. Model Optimization

Based on previous learning experiences [23], [33], for a dataset consisting of N drugs and M diseases, we select drug-disease association pairs as positive instances and all other combinations as negative instances. We denote drug positive instances and negative instances as x^+ and x^- respectively. As the observed number of associations is significantly smaller than the number of unobserved associations, we opt for weighted cross-entropy [33] as the loss function, computed as follows:

$$\text{Loss} = -\frac{1}{N \times M} \left(\lambda \times \sum_{(i,j) \in x^+} \log s'_{ij} + \sum_{(i,j) \in x^-} \log \left(1 - s'_{ij}\right) \right) \tag{20}$$

where $\lambda = \frac{|x^-|}{|x^+|}$, $|x^+|$ and $|x^-|$ denote the quantities of x^+ and x^- respectively, (i, j) represents a drug-disease pair with drug M_i and disease A_j. The parameter λ acts as a weight factor to mitigate the impact of data imbalance.

Following this, we apply the Adam optimizer [34] to minimize the loss function. The Adam optimizer iteratively updates the neural network parameters to minimize the loss. Furthermore, to address overfitting concerns, we incorporate node dropout and regular dropout within the graph convolutional layers. Moreover, we adopt a cyclic learning rate [35] approach, which cyclically adjusts the learning rate magnitude to enhance the convergence speed of the model.

TABLE I
TWO DATASET

Dataset	Drugs	Disease	Known associations	Drug features				
				Target	Enzyme	Drug-drug interactions	Pathway	Substructure
Main dataset	269	598	18416	623	247	2086	465	881
Therapeutic dataset	269	598	6244	623	247	2086	465	881

E. Model Interpretability

In HDGAT, the information update for a specific node within a heterogeneous graph is influenced by the information emanating from neighboring nodes within the same graph. Concurrently, neighboring nodes have the capacity to convey higher-dimensional node information to the central node, facilitating improved aggregation of information across the entire heterogeneous graph. Following the application of GCNs in HDGAT, the input node feature vectors undergo a process that maps nodes to low-dimensional vectors. The dynamic attention mechanism embedded in GCN dynamically assigns varying weights to the neighboring nodes of the central node. Additionally, the hierarchical attention mechanism allocates distinct weights to the embeddings of low-dimensional feature vectors learned by different layers. These processes enhance the precision of feature fusion from other relevant nodes, consequently augmenting the accuracy of the model's predictive decisions.

F. Model Scalability

To validate the scalability of the model, two datasets of distinct scales, which contain 269 types of drugs and 598 types of disease, were deliberately chosen. The smaller dataset encompasses 6244 known drug-disease associations, while the larger dataset incorporates 18416 known drug-disease associations. After handling the heterogeneous graph, the dimensions of the input graph are 867*867. This selection facilitated the observation of the model's predictive performance in scenarios characterized by a higher number of nodes and edges within the heterogeneous graph. The experimental results illustrate that HDGAT consistently delivers outstanding performance across datasets of diverse scales. This observation attests to the commendable scalability of HDGAT, affirming its capacity to undergo stable training and prediction on larger-scale datasets.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Dataset

The dataset employed in this study is sourced from prior learning endeavours [23], [36], originating from the Comparative Toxicogenomics Database (CTD) [37]. This dataset encompasses 18,416 drug-disease associations, involving 269 distinct drugs and 598 different diseases. Pertinent details pertaining to the drugs, including targets, enzymes, pathways, drug-drug interactions, and substructures, are meticulously extracted from the DrugBank database [38]. Furthermore, disease similarities are assessed based on their MeSH (Medical Subject Headings). The data in the table represents the number of drug and disease categories, and the number of different features of drugs. Different features of drugs can effectively distinguish the functions and relationships among different drugs. Specifically, different categories of drug substructures signify diverse structural units or functional groups within drug molecules, exerting notable influence on drug-drug interactions. Hence, we calculate drug-drug similarities based on their different features. In order to validate the generality and robustness of our model, we utilized a distinct therapeutic dataset comprising 6244 annotated therapeutic drug-disease associations from CTD for performance comparison. For a comprehensive overview of the main dataset and therapeutic dataset, refer to Table I.

TABLE II
EXPERIMENTAL SETTING

Hyper parameter						
k	L	lr	n	dp	dp_n	α
64	3	0.01	400	0.4	0.6	6

B. Experimental Parameter Configuration

Before conducting our experiments, it was necessary to set the parameters of the model, including the embedding dimension k, the number of convolutional layers L, the learning rate lr, the training epochs n, node dropout dp_n, regular dropout dp, penalty factor α, and so on.

In the process of configuring experimental parameters, we conducted experiments to fine-tune parameters within specified ranges, such as $lr \in \{0.001, 0.005, 0.05, 0.1\}$, $dp, dp_n \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$, $L \in \{2, 3, 4, 5\}$, and so on. The learning rate governs the extent of updating network weights, where an excessively large learning rate may result in the model failing to converge, while a too small learning rate may lead to slow model convergence. The dropout rate serves as a strategy to mitigate model overfitting, with an appropriately chosen dropout rate enhancing the model's generalization performance. However, an excessively large dropout rate may cause significant information loss, diminishing the model's learning capabilities. The number of graph convolutional layers in the model represents a critical hyperparameter, where a too small value may yield insufficient learning capacity, while an excessively large value can induce the oversmoothing phenomenon, characterized by node representations learned by deep graph convolution becoming increasingly homogeneous. After multiple rounds of experimentation to fine-tune these parameters, we established the initial values for the hyperparameters, as shown in Table II.

Upon model training, a 5-fold cross-validation approach is employed to assess its performance. This technique, widely adopted in statistical evaluation, involves the random partitioning of the dataset into five exclusive subsets. Across five rounds, four subsets function as training data while the remaining subset is allocated as validation data. During each round, the model is trained on the training data and evaluated on the validation data. After these five rounds are completed, the performance metrics garnered from all iterations are averaged, culminating in a conclusive evaluation of the model's performance. The application of the 5-fold cross-validation technique effectively gauges the model's capacity to generalize to novel datasets, concurrently illuminating potential issues of overfitting or underfitting.

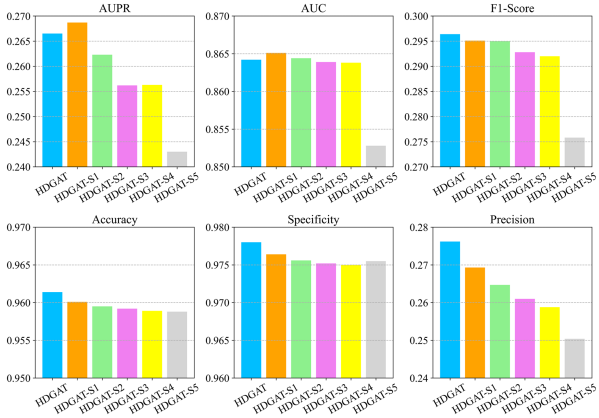


Fig. 5. Model performance with different modules: HDGAT-S1:Replace Bi-LSTM with LSTM on HDGAT; HDGAT-S2:Remove LSTM on HDGAT-S1; HDGAT-S3:Remove residual connection on HDGAT-S2; HDGAT-S4:Replace dynamic attention with static attention on HDGAT-S3; HDGAT-S5:Remove static attention on HDGAT-S4.

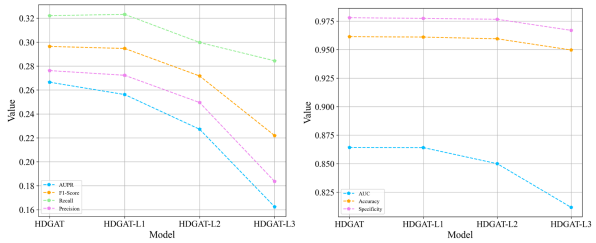
C. Ablation Experiment

To gain a comprehensive understanding of the distinct contributions made by various components of the model to the overall efficacy of drug-disease association prediction, we conducted ablation experiments to discern the impact of different combinations of modules within the HDGAT framework. Ablation experiments constitute a pivotal aspect of our experimental approach, as they serve to validate the effectiveness of specific modules in enhancing performance for the designated task, ultimately guiding refinements aimed at achieving heightened predictive accuracy.

In this pursuit, we formulated five distinct sets of module configurations in conjunction with our original HDGAT model for rigorous experimentation. In the initial set, we substituted Bi-LSTM with LSTM within the framework of HDGAT, denoted as HDGAT-S1. In the subsequent set, we excluded LSTM from HDGAT-S1, resulting in HDGAT-S2. In the third set, we eliminated residual connections from HDGAT-S2, identified as HDGAT-S3. In the fourth set, we converted the dynamic attention mechanism into a static attention mechanism based on HDGAT-S3, labeled as HDGAT-S4. Finally, in the fifth set, we omitted the static attention mechanism from HDGAT-S4, designated as HDGAT-S5. Through rigorous experimentation involving HDGAT and these three model variants, we garnered outcomes delineated in Fig. 5, offering crucial insights into the impact of diverse module combinations on the overall performance.

D. Impact of Different Layer Embeddings

From the results of our experimental analysis, it is evident that HDGAT-S1 exhibits a notable reduction in both accuracy and precision in comparison to the original HDGAT model. This finding underscores the beneficial influence of the dual-round LSTM processing strategy on the accuracy of drug-disease association prediction. HDGAT-S2 showcases a decrease in AUPR, Precision when contrasted with HDGAT-S1, which manifests that LSTM can capture the latent associations between drugs and diseases. Notably, HDGAT-S3 displays a decrease across all



(a) AUC, Accuracy, Specificity (b) AUPR, F1-Score, Recall, Precision

Fig. 6. Model performance of different layer's embedding.

metrics. This shift indicates that the inclusion of the residual connection module has a favorable effect on predicting positive instances. Furthermore, comparing HDGAT-S4 and HDGAT-S5, we observed a significant decline in all metrics for HDGAT-S5 except for Accuracy and Specificity. From this, it can be inferred that assigning different weights to nodes can effectively enhance node information aggregation capability.

In summation, the outcomes collectively underscore that HDGAT emerges as the superior performer among the four models under consideration. This observation corroborates that the amalgamation of modules integrated into HDGAT distinctly contributes to the model's overall performance enhancement. Besides, ablation experiments serve to elucidate the contributions of various modules to HDGAT.

To investigate the impact of different layer embeddings within the graph convolutional networks on the overall predictive performance of the model, we conducted a series of experiments specifically designed to assess the predictive capabilities of embeddings derived from various layers. As outlined in our previous experimental configurations, we have employed a neural network with a total of 3 layers. The rationale behind selecting this architecture is rooted in the notion that a 3-layer neural network delivers commendable predictive performance. Employing an excessive number of layers could potentially trigger the problem of gradient vanishing during training, whereas utilizing too few layers might fail to capture the intricate high-dimensional associations prevalent within the heterogeneous graph.

Hence, we proceeded to leverage the node embeddings acquired from each of the first, second, and third layers of the network as standalone entities for prediction purposes. These single-layer models are denoted as HDGAT-L1, HDGAT-L2, HDGAT-L3, respectively. Through rigorous evaluation encompassing 5-fold cross-validation, we subjected these models to meticulous scrutiny and juxtaposed their performances against the original model. The comprehensive findings stemming from these experiments are meticulously presented in Table III and visually depicted in Fig. 6.

The experimental outcomes distinctly reveal that the predictive prowess exhibited by the HDGAT-L1 and HDGAT-L2 models markedly surpasses that of the HDGAT-L3 model. This compelling evidence strongly suggests that, in contrast to the embeddings originating from the third layer, the embeddings stemming from the first and second layers encompass a wealth of pertinent information concerning graph structure and inter-node relationships. This discernible pattern strongly implies that the HDGAT framework could potentially encounter certain constraints when it comes to effectively capturing intricate high-dimensional associations prevalent among nodes.

TABLE III
MODEL PERFORMANCE OF DIFFERENT LAYER'S EMBEDDING

Module	AUPR	AUC	F1-Score	Accuracy	Recall	Specificity	Precision
HDGAT	0.2665	0.8642	0.2964	0.9614	0.3221	0.9780	0.2762
HDGAT-L1	0.2562	0.8641	0.2947	0.9610	0.3232	0.9774	0.2723
HDGAT-L2	0.2272	0.8500	0.2717	0.9595	0.2997	0.9766	0.2495
HDGAT-L3	0.1624	0.8117	0.2219	0.9497	0.2844	0.9669	0.1836

TABLE IV
MODEL PERFORMANCE WITH DIFFERENT AGGREGATE METHODS

Module	AUPR	AUC	F1-Score	Accuracy	Recall	Specificity	Precision
HDGAT	0.2665	0.8642	0.2964	0.9614	0.3221	0.9780	0.2762
HDGAT-CONCAT	0.2467	0.8620	0.2901	0.9583	0.3376	0.9744	0.2555
HDGAT-AVG	0.2413	0.8600	0.2869	0.9605	0.3157	0.9772	0.2678

TABLE V
PERFORMANCE OF DIFFERENT METHODS ON MAIN DATASET AND THERAPEUTIC DATASET

Dataset	Module	AUPR	AUC	F1-Score	Accuracy	Recall	Specificity	Precision
Main dataset	TL-HGBI	0.0665	0.7029	0.1266	0.9114	0.2545	0.9284	0.0843
	DRRS	0.1321	0.8429	0.2178	0.9324	0.3276	0.9468	0.1631
	DeepDR	0.1351	0.8211	0.1991	0.9400	0.2959	0.9567	0.1500
	NIMCGCN	0.2002	0.8533	0.2661	0.9572	0.3083	0.9739	0.2341
	LAGCN	0.2287	0.8486	0.2701	0.9588	0.3021	0.9758	0.2445
	HDGAT	0.2665	0.8642	0.2964	0.9614	0.3221	0.9780	0.2762
Therapeutic dataset	TL-HGBI	0.0388	0.7401	0.0720	0.9761	0.1151	0.9830	0.0524
	DRRS	0.1383	0.8754	0.2249	0.9849	0.2726	0.9907	0.1914
	DeepDR	0.1011	0.8572	0.1610	0.9806	0.2327	0.9866	0.1231
	NIMCGCN	0.0899	0.8075	0.1525	0.9798	0.2225	0.9859	0.1160
	LAGCN	0.2777	0.8810	0.1910	0.9713	0.4216	0.9757	0.1236
	HDGAT	0.2707	0.8877	0.2204	0.9783	0.3829	0.9830	0.1549

E. Impact of Hierarchical Attention Mechanism

The data presented in Table III undeniably highlights HDGAT's superior performance relative to the other three models, conclusively illustrating that embeddings derived from distinct layers exert disparate influences on the final predictive efficacy of the model. Hence, it becomes imperative to amalgamate embeddings from various dimensions to effectively amalgamate the acquired structural attributes and association insights. To address this objective, we adopted three distinct methods for embedding aggregation: 1. HDGAT; 2. HDGAT-AVG; 3. HDGAT-CONCAT. In the HDGAT approach, we harnessed the hierarchical attention mechanism for embedding amalgamation. In HDGAT-AVG, we aggregated embeddings through summation followed by averaging. Conversely, HDGAT-CONCAT involved a direct concatenation of node embeddings. The comprehensive results of these experiments are meticulously detailed in Table IV.

The outcomes presented in Table IV illuminate HDGAT's supremacy over the other two methods concerning predictive prowess. This observation underscores the inherent benefits of employing the hierarchical attention mechanism to manage the divergent contributions originating from embeddings of distinct layers. The experimental results illustrated in Fig. 6 further corroborate this assertion, revealing that node embeddings from lower layers encompass a more substantial repository of graph structural attributes and node association insights, while those from higher layers contain comparatively diminished information. Thus, the hierarchical attention mechanism judiciously assigns greater weights to lower-layer embeddings and correspondingly lighter weights to their higher-layer counterparts, manifestly striving to attain optimal predictive performance for the model.

F. Comparisons With Other Models

In this section, we compare the HDGAT model with two baseline models and three existing excellent models for drug-disease association prediction. We selected NIMCGCN [39] and LAGCN [23] as our baseline model due to their shared foundation as GCN-based models. Based on the experimental parameter configuration above and the reported data from prior studies [23], we conducted comparative experiments on main dataset. Besides, we also used therapeutic dataset to validate our model's generality. The results are shown in Table V.

TL-HGBI [5]: It introduces an algorithm that iteratively updates to calculate the length of drug-disease pairs in a heterogeneous graph network integrating drug and disease information, and predicts drug-disease associations.

DRRS [40]: It first constructs a heterogeneous graph interaction network by combining drug-drug, disease-disease, and drug-disease networks. Then, it utilizes a rapid Singular Value Thresholding (SVT) algorithm based on recommendation systems to predict unknown drug-disease associations.

DeepDR [39]: It learns higher-order features of drugs from the heterogeneous graph network using a multi-modal deep autoencoder. It encodes and decodes these learned features along with known drug-disease pairs using a variational autoencoder to infer drug-disease associations.

NIMCGCN [41]: It initially employs Graph Convolutional Networks (GCN) to learn latent feature representations of miRNAs and diseases from a similarity network. Then, it utilizes a novel neural inductive matrix completion approach with the learned features to generate the association matrix.

LAGCN [23]: It employs GCN to learn drug-disease node embeddings in a heterogeneous graph network, containing drug and disease similarities and associations, and applies a layer

attention mechanism to these embeddings for predicting drug-disease associations.

In the aforementioned experiments, the training data for drug-disease associations in the two datasets exhibit disparate distributions, encompassing variations in both the types of drug associations and the sizes of the datasets. Consequently, the experiments provide a robust observation of HDGAT's generalization capability across diverse data distributions. The experimental results in Table V demonstrate that HDGAT consistently achieves high accuracy across multiple datasets, outperforming both baseline models and existing state-of-the-art models in terms of overall performance. These results affirm the robust generalization capacity of HDGAT. Notably, in terms of AUPR and Precision, HDGAT exhibits significant improvement,which indicates that our model is capable of maintaining a relatively high recall while preserving a high precision. Compared to the baseline models, the introduction of residual connections helps overcome certain limitations inherent in NIMCGCN and LAGCN, particularly alleviating the over-smoothing issue encountered during the training of multi-layer GCNs. Besides, HDGAT employs a dynamic attention mechanism that allows for more precise assignment of weights to nodes, thereby accurately aggregating information from different nodes. This substantiates the enhanced robustness exhibited by HDGAT when contrasted with other models.

During the process of model training, the stacking of an excessive number of layers in GCNs results in the persistence of issues such as gradient vanishing and oversmoothing. This occurrence leads to a gradual reduction in the learning capacity of the model. Consequently, the challenges posed by the oversmoothing problem and the gradient vanishing issue introduced by GCN persist as significant obstacles for HDGAT in its development.

G. Case Studies

In this paper, we utilize HDGAT to predict new drug-disease associations based on existing known drug-disease associations. Previously, we evaluate the model's predictive performance solely based on the predicted association scores. However, assessing predictive accuracy solely based on prediction scores might not provide a comprehensive measure of accuracy. Therefore, we compare the predicted drug-disease associations from our model with the existing medical data and associations present in databases to assess its practical applicability and reliability.

In this experiment, we employ HDGAT to predict drug-disease associations. The top 10 drug-disease association pairs with the highest prediction scores are calculated by HDGAT. To explore the association scores of these 10 drugs with the 10 diseases, we generated a heatmap as depicted in Fig. 7.

Based on the predicted associations from the model, we try to seek relevant medical evidence to determine whether there is a connection between them. For instance, trandolapril is an angiotensin-converting enzyme (ACE) inhibitor that is widely used for the treatment of patients with Hypertension [42]. ACE inhibitors help control hypertension by reducing the formation of angiotensin II, and they can also reduce the cardiac load and improve cardiac function. Maprotiline is an antidepressant, however, the use of such antidepressants often comes with side effects, including Tremor [43].

Furthermore, we conducted specific experiments on certain drugs and diseases to explore the associations in more detail. We

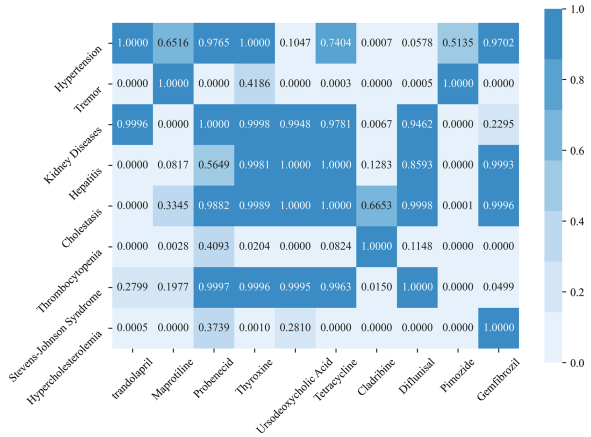


Fig. 7. Association scores between the top 10 pairs of drug-disease associations predicted by HDGAT.

TABLE VI
TOP 10 RELEVANT DISEASES ASSOCIATED WITH SPECIFIC DRUG PREDICTED BY HDGAT

Drug	Disease	Probability
clobazam	Seizures	0.99954560
	Hallucinations	0.97442794
	Bipolar Disorder	0.97149930
	Substance Withdrawal Syndrome	0.96609294
	Delirium	0.96327174
	Psychoses, Substance-Induced	0.94725070
	Myoclonus	0.93702210
	Anxiety Disorders	0.93016540
	Cognition Disorders	0.88776463
	Chemical and Drug Induced Liver Injury	0.86777973

TABLE VII
TOP 10 RELEVANT DRUGS ASSOCIATED WITH SPECIFIC DISEASE PREDICTED BY HDGAT

Disease	Drug	Probability
Adenocarcinoma of lung	Doxorubicin	0.99999980
	docetaxel	0.99999964
	vinorelbine	0.99999920
	gefitinib	0.99999905
	Epirubicin	0.99997944
	Methotrexate	0.99995940
	Fluorouracil	0.99992420
	Erlotinib Hydrochlorides	0.99958885
	Indomethacin	0.99904263
	Midazolam	0.99895513

investigated the top ten diseases most associated with the drug clobazam and the top ten drugs most associated with the disease Adenocarcinoma of lung. The results of these experiments are presented in Tables VI and VII.

For the drug-disease associations identified, we also validate them by consulting relevant literature. For example, clobazam [44] has been confirmed as an effective antiepileptic drug and is widely used for both pediatric and adult epilepsy patients in various countries. Epilepsy is a neurological disorder that can lead to various types of seizures, and clobazam can reduce the frequency of these seizures. Similarly, in the context of advanced treatments for Adenocarcinoma of lung, Doxorubicin and docetaxel are frequently utilized [45], [46] as chemotherapy drugs. They are commonly used in cancer treatment, including Adenocarcinoma of lung.

However, in the practical process of model prediction, HDGAT falls short of achieving flawless and precise predictions for all drug-disease associations. The intricacies of the relationships between drugs and diseases, with some associations being direct and others indirect, present formidable challenges to the model's predictive capabilities. Moreover, the quality of medical literature and the timeliness of medical research play pivotal roles in influencing the accurate discernment of drug-disease predictions. Consequently, encountering false positives and false negatives during the prediction process is at times inevitable. In response to such instances, we opt to amplify the scale of the dataset and broaden the scope of model predictions, with the goal of fortifying the model's robustness through more extensive training data. Furthermore, validation using more authoritative and official medical literature serves as an effective means to corroborate the accuracy of the prediction results.

These case studies provide compelling support for the conclusions drawn from our model's predictions, demonstrating that the HDGAT model is capable of effectively predicting drug-disease association information.

IV. Conclusion

In this paper, we introduce HDGAT, a novel model designed for the prediction of drug-disease associations. The methodology comprises three primary stages:

1. The initial phase involves the construction of a heterogeneous graph by amalgamating networks representing drug-drug similarities, disease-disease similarities, and drug-disease associations.
2. Subsequently, a graph convolutional network equipped with hierarchical and dynamic attention mechanisms, in conjunction with a Bi-LSTM module, is employed to acquire node embeddings and capture the structural information within the heterogeneous network.
3. In addition, residual connections are integrated into HDGAT. The model's performance is assessed using a rigorous five-fold cross-validation approach, supplemented by case studies to validate predictive outcomes.

HDGAT excels in capturing the intricate network structure of the heterogeneous graph and relevant associations between drug and disease over long distance. The incorporation of residual connections mitigates over-smoothing issues associated with graph convolutions while enhancing both convergence speed and generalization capabilities. Besides, dynamic attention mechanisms assist in dynamically refining the central node's information to more effectively capture the characteristics of nodes and edges within a heterogeneous graph. The experimental results robustly demonstrate HDGAT's superiority across numerous evaluation metrics in the context of drug-disease association prediction tasks when compared to existing state-of-the-art models.

However, compared with homogeneous graphs, HDGAT exhibits certain limitations in aggregating information from different nodes and edges in heterogeneous graphs, where connected nodes often exhibit similar properties. Besides, while residual connection modules prove effective in mitigating the oversmoothing phenomenon introduced by GCNs, the challenge of gradient vanishing persists, particularly when stacking an excessive number of layers in graph neural networks.

In future research, enhancing GCN-based models by integrating additional considerations for heterogeneous nodes and edges, based on the foundation of the HDGAT framework, emerges as a promising avenue for improvement. Furthermore, investigating node information aggregation through meta-paths [7] has the potential to mitigate oversmoothing phenomenon, making it another viable direction for extending future models.

V. Materials

1. The drug-disease associations database: http://ctdbase.org.
2. The drug features database: https://www.drugbank.ca/.
3. The disease MeSH database: https://meshb.nlm.nih.gov/.
4. All the materials and codes in this paper are available at https://github.com/37918273918/HDGAT.

References

[1] L. Liu et al., "Multi-view contrastive learning hypergraph neural network for drug-microbe-disease association prediction," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, vol. 2023, pp. 4829–4837.
[2] Z. Chu et al., "Hierarchical graph representation learning for the prediction of drug-target binding affinity," *Inf. Sci.*, vol. 613, pp. 507–523, 2022.
[3] W. Ben Abdessalem Karaa, E. H. Alkhammash, and A. Bchir, "Drug disease relation extraction from biomedical literature using nlp and machine learning," *Mobile Inf. Syst.*, vol. 2021, pp. 1–10, 2021.
[4] P. Wang, T. Hao, J. Yan, and L. Jin, "Large-scale extraction of drug–disease pairs from the medical literature," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 11, pp. 2649–2661, 2017.
[5] W. Zhang et al., "Predicting drug-disease associations based on the known association bipartite network," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2017, pp. 503–509.
[6] Y.-Z. Di, P. Chen, and C.-H. Zheng, "Similarity-based integrated method for predicting drug-disease interactions," in *Proc. Intell. Comput. Theories Appl.: 14th Int. Conf.*, 2018, pp. 395–400.
[7] F. Tanvir, M. I. K. Islam, and E. Akbas, "Predicting drug-drug interactions using meta-path based similarities," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol.*, 2021, pp. 1–8.
[8] F. Tanvir, K. M. Saifuddin, M. Ifte Khairul Islam, and E. Akbas, "Predicting drug-drug interactions using heterogeneous graph attention networks," in *Proc. 14th ACM Int. Conf. Bioinf., Comput. Biol., Health Inf.*, 2023, pp. 1–6.
[9] F. Tanvir, K. M. Saifuddin, T. Hossain, A. Bagavathi, and E. Akbas, "HeTriNet: Heterogeneous graph triplet attention network for drug-target-disease interaction," 2023, *arXiv:2312.00189*.
[10] K. M. Saifuddin, B. Bumgardner, F. Tanvir, and E. Akbas, "HyGNN: Drug-drug interaction prediction via hypergraph neural network," in *Proc. IEEE 39th Int. Conf. Data Eng.*, 2023, pp. 1503–1516.
[11] J. Peng et al., "An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction," *Brief. Bioinf.*, vol. 22, no. 5, 2021, Art. no. bbaa430.
[12] C. Tang et al., "DeFusionNET: Defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 955–968, Feb. 2022.
[13] H. Liu, W. Zhang, Y. Song, L. Deng, and S. Zhou, "HNet-DNN: Inferring new drug–disease associations with deep neural network based on heterogeneous network features," *J. Chem. Inf. Model.*, vol. 60, no. 4, pp. 2367–2376, 2020.
[14] P. Xuan, L. Gao, N. Sheng, T. Zhang, and T. Nakaguchi, "Graph convolutional autoencoder and fully-connected autoencoder with attention mechanism based method for predicting drug-disease associations," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1793–1804, May 2021.
[15] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.
[16] C. Tang et al., "Spatial and spectral structure preserved self-representation for unsupervised hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531413.

[17] C. Tang, X. Zheng, W. Zhang, X. Liu, X. Zhu, and E. Zhu, "Unsupervised feature selection via multiple graph fusion and feature weight learning," *Sci. China Inf. Sci.*, vol. 66, no. 5, pp. 1–17, 2023.
[18] X. Zeng et al., "Measure clinical drug–drug similarity using electronic medical records," *Int. J. Med. Inf.*, vol. 124, pp. 97–103, 2019.
[19] C. Yan, G. Duan, Y. Zhang, F.-X. Wu, Y. Pan, and J. Wang, "Predicting drug-drug interactions based on integrated similarity and semi-supervised learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 168–179, Jan./Feb., 2022.
[20] X. Yang, G. Yang, and J. Chu, "The neural metric factorization for computational drug repositioning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 1, pp. 731–741, Jan./Feb. 2023.
[21] J. Yu, J. Amores, N. Sebe, P. Radeva, and Q. Tian, "Distance learning for similarity estimation," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 451–462, Mar., 2008.
[22] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on micrornaassociated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.
[23] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug–disease associations through layer attention graph convolutional network," *Brief. Bioinf.*, vol. 22, no. 4, 2021, Art. no. bbaa243.
[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2016.
[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[26] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2021.
[27] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *Proc. IEEE Int. Conf. Big Data*, 2019, pp. 3285–3292.
[28] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 793–803.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.
[30] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3538–3545.
[31] S. Zhang, X. Xu, Y. Pang, and J. Han, "Multi-layer attention based CNN for target-dependent sentiment classification," *Neural Process. Lett.*, vol. 51, pp. 2089–2103, 2020.
[32] J. Sirignano and K. Spiliopoulos, "Scaling limit of neural networks with the xavier initialization and convergence to a global minimum," 2019, *arXiv:1907.04108*.
[33] Y. S. Aurelio, G. M. De Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," *Neural Process. Lett.*, vol. 50, pp. 1937–1949, 2019.
[34] D. P. Kingma, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2014.
[35] J. Li and X. Yang, "A cyclical learning rate method in deep learning training," in *Proc. IEEE Int. Conf. Comput., Inf. Telecommun. Syst*, 2020, pp. 1–5.
[36] W. Zhang et al., "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinf.*, vol. 19, pp. 1–12, 2018.
[37] A. P. Davis et al., "The comparative toxicogenomics database: Update2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D972–D978, 2017.
[38] V. Law et al., "DrugBank 4.0: Shedding new light on drug metabolism," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1091–D1097, 2014.
[39] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "deepDR: A network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, 2019.
[40] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, "Computational drug repositioning using low-rank matrix approximation and randomized algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1904–1912, 2018.
[41] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction," *Bioinformatics*, vol. 36, no. 8, pp. 2538–2546, 2020.
[42] L. N. C. Duc and H. R. Brunner, "Trandolapril in hypertension: Overview of a new angiotensin-converting enzyme inhibitor," *Amer. J. Cardiol.*, vol. 70, no. 12, pp. D27–D34, 1992.
[43] J. Bouchard et al., "Citalopram versus maprotiline: A controlled, clinical multicentre trial in depressed patients," *Acta Psychiatrica Scandinavica*, vol. 76, no. 5, pp. 583–592, 1987.
[44] Y.-t. Ng and S. D. Collins, "Clobazam," *Neurotherapeutics*, vol. 4, no. 1, pp. 138–144, 2007.
[45] W. M. Jordan et al., "Treatment of advanced adenocarcinoma of the lung with ftorafur, doxorubicin, cyclophosphamide, and cisplatin (FACP) and intensive IV hyperalimentation," *Cancer Treat. Rep.*, vol. 65, no. 3–4, pp. 197–205, 1981.
[46] H. Yasuda et al., "Nitroglycerin treatment may enhance chemosensitivity to docetaxel and carboplatin in patients with lung adenocarcinoma," *Clin. Cancer Res.*, vol. 12, no. 22, pp. 6748–6757, 2006.