基于全局图注意力元路径异构网络的药物–疾病关联预测

郁 湧[1,2]，杨雨洁[1]，李虓晗[1]，高 悦[1]，于 倩[1*]

（1. 云南大学 软件学院，昆明 650504；2. 云南省软件工程重点实验室，昆明 650504）

摘要 提出了一个基于全局图注意力元路径异构网络模型（MHNGA）来进行药物–疾病关联预测。首先，收集整理药物和疾病数据，将已知的药物–疾病关联、药物相似性、疾病相似性构建为一个异构网络；其次，引入多个基于元路径的子图，使用图注意力神经网络提取这些子图的邻居节点的特征，并且通过通道注意力和空间注意力机制来增强特征；最后，通过十折交叉验证的评估，MHNGA 取得了 93.5% 的精确召回曲线下的面积和 99.4% 的准确率。

关键词 异构图；药物–疾病关联；预测；图注意力神经网络；元路径

中图分类号 TP181 **文献标志码** A DOI 10.12178/1001-0548.2023235

Drug-Disease Association Prediction Based on Meta-Path Heterogeneous Network with Global Graph Attention

YU Yong[1,2], YANG Yujie[1], LI Xiaohan[1], GAO Yue[1], and YU Qian[1*]

(1. School of Software, Yunnan University, Kunming 650504, China; 2. Key Laboratory in Software Engineering of Yunnan Province, Kunming 650504, China)

Abstract In this paper, a heterogeneous network model based on global graph attention meta-path, named MHNGA, is proposed for drug-disease association prediction. Firstly, the data of drugs and diseases are collected, and the known drug-disease association, drug similarity and disease similarity are constructed as a heterogeneous network. Secondly, multiple meta-path-based subgraphs are introduced, and the graph attention neural network is used to extract the features of the neighbor nodes of these subgraphs, and the features are enhanced by channel attention and spatial attention mechanisms. Finally, through the evaluation of ten-fold cross-validation, MHNGA achieves 93.5% of the area under the accurate recall curve and 99.4% of the accuracy.

Key words heterogeneous graph; drug-disease association; prediction; graph attention neural network; meta-path

在药物开发领域，耗时费力是常态。传统的药物开发通常包括 3 个阶段：发现、临床前和临床开发，每个阶段都需要巨额资金支持[1-3]。为了应对这些挑战，研究人员对现有药物进行重新应用，即药物重定位[4]，以降低开发成本，提高效率和降低安全隐患。最近，随着生物数据库的积累和计算方法的进步，利用大数据技术来发现药物和疾病之间的联系已经成为生物医学领域的研究热点[5]。

药物与疾病的相互关联可被视为网络。目前，已有大量计算方法用于识别药物与疾病之间的潜在联系[5]。基于推荐系统的方法通过协同过滤主要用于识别已知药物–疾病关联任务[6-9]，不过对于新药或新型疾病效果较差。基于机器学习的方法应用广泛[10-11]，但需要高质量地输入数据才能发挥作用，这在实际应用中较为困难。基于深度学习[12-13]的方法可以将原始数据转换为抽象的特征表示，从而解决人工标注数据的不完全性问题[14]，然而需要大量训练数据才能取得准确的结果，当数据较为稀疏时，容易出现过拟合。相较之下，基于网络的方法具有较好的预测性能[15-17]，可以捕获不同生物网络中有关药物与疾病的有效信息[18-20]，从而提高预测准确性。因此，引入异质网络以表示不同类型的生物网络，并在这些网络中保留相关特征，能有效发现药物与疾病之间的潜在关联。

注意力机制已经被广泛地应用于各种任务中，如自然语言处理、计算机视觉等，它的优点是能够

收稿日期：2023−09−12；修回日期：2024−01−20

基金项目：国家自然科学基金（62366058）；云南省科技厅面上项目（202001BB050063）；云南省软件工程重点实验室开放项目（2020SE315）

作者简介：郁湧，博士，副教授，主要从事社交网络分析、机器学习和推荐系统等方面的研究。

*通信作者 E-mail: yuqian@ynu.edu.cn

增强数据中的有效特征。在生物信息学方面，注意力机制也呈现出令人惊叹的表现，如使用层注意力机制[13]来预测药物疾病关联，使用多头注意力机制预测环状 RNA（circRNA）与疾病的关联[21]。

本文提出一种基于全局图注意力元路径异构网络模型（Drug-Disease Association Prediction based on Meta-path Heterogeneous Network with Global Graph Attention, MHNGA）来预测药物和疾病的关联。本文通过生物数据库收集药物、疾病的相关信息，构建了一个基准数据集。通过已知的药物疾病关联，并计算相关的药物相似性与疾病相似性，将之构建为一个异构网络，用于预测药物疾病关联。基于这个异构网络，引入多个基于元路径的子图，然后使用图形卷积注意力从这些同构和异构子图的邻居节点中学习药物和疾病的节点特征。然后使用通道注意力和空间注意力来增强来自图形卷积注意力运算的所有有用信息。最后通过一个基于整合嵌入的预测模块来预测未观察到的药物疾病关联。

1　药物−疾病异构网络构建

1.1　数据集

为了有效地评估本文所提出的模型，本文通过 CTD[22]、DrugBank[23] 等生物数据库收集相关数据，构建了一个基准数据集，即 A 数据集，其中包含 894 种药物、454 种疾病、7 199 条疾病−疾病的边、14 291 条药物−药物边以及 2 704 条药物−疾病已知关联。为了更好地评估 MHNGA，本文还构建了另外的一个基准数据集，即 B 数据集。在 B 数据集中包含 598 种疾病、269 种药物、4 043 条药物−药物的边、8 970 条疾病−疾病的边和 18 416 条已知药物−疾病关联，数据集统计信息如表 1 所示。其中，药物与药物之间相连的边表示两种药物具有相似性，它们会对同种疾病产生一定的疗效；同理，疾病与疾病之间相连的边表示两种疾病具有一定的相似性。

表 1　数据集统计信息

数据集	药物−药物	疾病−疾病	药物−疾病	总连接量
A_Dataset	14 291	7 199	2 704	26 898
B_Dataset	4 043	8 970	18 416	31 429

1.2　异构网络构建

1.2.1　药物−药物相似性

具有相似化学结构的药物分子通常具有类似的生物活性，从 DrugBank[23] 下载药物 SMILES 序列来获得其化学描述符，之后将 SMILES 序列转换为药物的拓扑指纹，根据指纹位点和谷本相似性来衡量药物之间的相似性。给定两种药物i和j，它们的指纹位点分别表示为dr_i和dr_j，则它们之间的相似性为：

$$\mathrm{Tm}=\frac{\mathrm{dr}_i\mathrm{dr}_j}{\mathrm{dr}_i^2+\mathrm{dr}_j^2-\mathrm{dr}_i\mathrm{dr}_j} \tag{1}$$

$$\mathrm{sim}^{\mathrm{dr}}(\mathrm{dr}_i,\mathrm{dr}_j)=1-\frac{\min(\mathrm{Tm})}{\max(\mathrm{Tm})-\min(\mathrm{Tm})} \tag{2}$$

最终，将$\mathbf{sim}^{\mathrm{dr}}$构建为一个药物相似性矩阵。

1.2.2　疾病−疾病相似性

医学主题词标识符（Medical Subject Headings, MESH）可用于计算疾病之间的语义相似度并作为疾病相互作用的概率，具体来说，疾病的 MESH 标识符可以构建为一个有向无环图（Directed Acyclic Graph, DAG）。在图 1 中，展示了牙源性囊肿的 DAG 图结构。

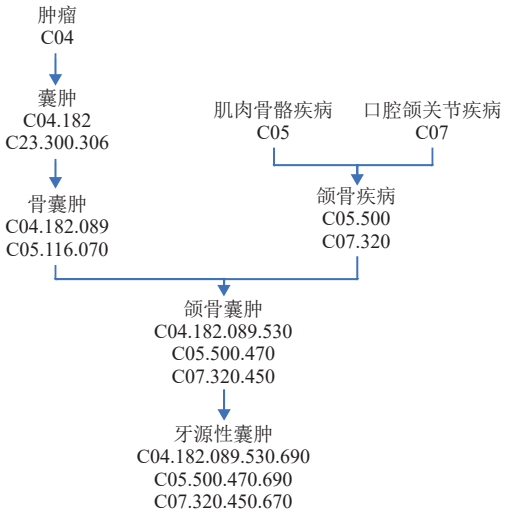

图 1　牙源性囊肿 DAG 图结构

DAG 图结构中的节点分为父节点和子节点两种类型，上下级间的节点都通过直线相互连接。每个节点都描述了关于一种疾病的 MESH 词条信息。给定疾病di_i，$\mathrm{DAG}_{\mathrm{di}_i}=(F(\mathrm{di}_i),E(\mathrm{di}_i))$，其中$F(\mathrm{di}_i)$表示疾病$\mathrm{di}_i$及其所有祖先节点的节点集，$E(\mathrm{di}_i)$表示$F(\mathrm{di}_i)$中疾病之间的所有关系的连接，一个疾病节点$\mathrm{di}_i\in F(\mathrm{di}_i)$对$\mathrm{di}_t$的语义贡献可以表示为：

$$D_{\mathrm{di}_i}(\mathrm{di}_t)=\begin{cases}1 & D_{\mathrm{di}_i}=\mathrm{di}_t\\ \max\{\Delta D_{\mathrm{di}_i}(\mathrm{di}_t')|\mathrm{di}_t'\in \mathrm{di}_i\} & D_{\mathrm{di}_i}\neq \mathrm{di}_t\end{cases} \tag{3}$$

式中，Δ为语义贡献因子，在本研究中，$\mathrm{DAG}_{\mathrm{di}_i}$中所有祖先节点对疾病di的语义贡献因子取 0.5，疾

病di对其本身的语义贡献值为 1。通过式 (3) 得知，di_t的贡献值主要是由di_t和di_i之间的距离决定。根据式 (3) 计算出所有F_{di_i}中所有祖先的贡献值，结合式 (4)，可以得到疾病di_i的语义值：

$$\mathrm{SV}(\mathrm{di}_i)=\sum_{\mathrm{di}_t\in\mathrm{di}_i} D_{\mathrm{di}_i}(\mathrm{di}_t) \tag{4}$$

结合式 (3) 和式 (4)，疾病di_i和di_j之间的语义相似度可以表述为:

$$\mathrm{di}_{\mathrm{sim}}(\mathrm{di}_i,\mathrm{di}_j)=\frac{\sum_{\mathrm{di}_t\in F(\mathrm{di}_i)\cap F(\mathrm{di}_j)} D_{\mathrm{di}_i}(\mathrm{di}_t)+D_{\mathrm{di}_j}(\mathrm{di}_t)}{\mathrm{SV}(\mathrm{di}_i)+\mathrm{SV}(\mathrm{di}_j)} \tag{5}$$

式中，di_t对di_i和di_j的贡献分别表示为$D_{\mathrm{di}_i}(\mathrm{di}_t)$和$D_{\mathrm{di}_j}(\mathrm{di}_t)$。

1.2.3 异构网络构建

本文的异构网络是基于药物–疾病关联表示、药物–药物相似性以及疾病–疾病相似性网络构建而形成的。为了更好地表示药物–疾病异构网络，通过一种参数化的形式来表示异构网络。将药物–疾病关联表示为一个二元网络$A\in\{0,1\}^{N\times M}$，其中，N表示疾病的数量，M表示药物的数量。如果药物dr_i和疾病di_j相关联，那么$A_{\mathrm{dr}_i,\mathrm{di}_j}=1$，否则$A_{\mathrm{dr}_i,\mathrm{di}_j}=0$。药物与药物之间的相似性表示为相似性矩阵$\mathbf{sim}^{\mathrm{dr}}$，$\mathrm{sim}^{\mathrm{dr}}_{i,j}$表示药物$i$与药物$j$之间的相似性。同理，疾病与疾病之间的相似性表示为相似性矩阵$\mathbf{sim}^{\mathrm{di}}$，$\mathrm{sim}^{\mathrm{di}}_{i,j}$表示疾病$i$与疾病$j$之间的相似性。最后，构建异构网络模型为：

$$\boldsymbol{H}=\begin{bmatrix}\mathbf{sim}^{\mathrm{dr}} & \boldsymbol{A}\\ \boldsymbol{A}^{\mathrm{T}} & \mathbf{sim}^{\mathrm{di}}\end{bmatrix}\in\mathbb{R}^{(N+M)\times(N+M)} \tag{6}$$

2 MHNGA 模型和方法

MHNGA 方法包括 3 个部分：1）元路径特征节点提取；2）节点特征聚集及增强；3）药物–疾病关联预测。MHNGA 的工作流程如图 2 所示。

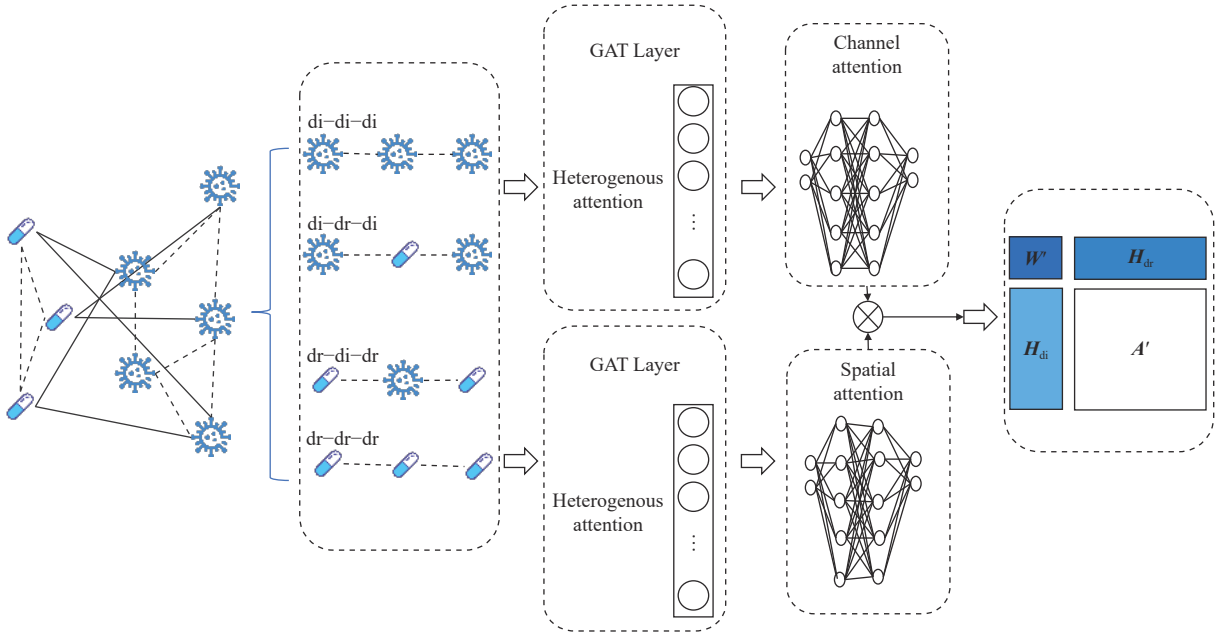

图 2 模型总体架构图

2.1 元路径特征节点提取

异构网络中两个不同的实体类型可以通过不同的路径相连，称为元路径，其中节点之间的连接由相应的关系类型决定。本文基于元路径理论对异构图上潜在的药物–疾病关联进行研究。元路径可以定义为以$c_1\xrightarrow{R_1}c_2\xrightarrow{R_2}\cdots\xrightarrow{R_l}c_{l+1}$的形式出现的路径，其中$c_1,c_2,\cdots,c_{l+1}$表示为节点类型，不同类型的节点通过复合关系$R=R_1\circ R_2\circ\cdots\circ R_l$连接，$\circ$表示关系上的复合运算符。在异构图中，一个元路径可以捕获异构图上的特定语义信息，不同的元路径表示不同的语义信息。如药物–疾病–药物的元路径表示两个药物对该疾病都起作用，因此两个药物可能具有一定的相似功能。对于一个异构图G，将其拆分为多个子图，每个子图G'都是由通过元路径连接的所有相邻对构成，这些子图均称为基于元路径的子图。也就是说，对于某种类型的节点，每个

子图与其邻居的连接可以看作是节点之间的某种交互模式，它可以揭示节点之间的语义关系和潜在的模式，从而更好地理解节点的功能和特性，如图 3 所示，当一条元路径的第一个节点和最后一个节点相同时，将之称为齐次子图；反之，当第一个节点和最后一个节点不相同时，将之称为异质子图。

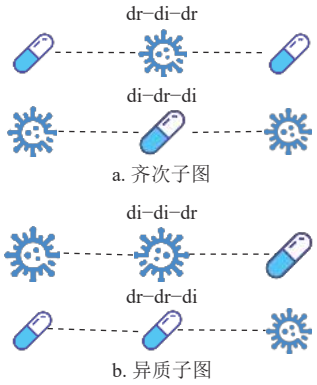

图 3　齐次子图和异质子图示例图

GAT[24] 是一种针对图结构的创新型图神经网络，它通过引入自注意力机制来学习节点之间的关联程度，提高了图数据的处理和表达能力。本文利用 GAT[24] 来学习节点的表示。首先，GAT 会对每一个节点赋予一个初始表示向量，然后对每个节点，计算节点间的相似度得到注意力权重，使用注意力权重对邻居节点的表示进行加权聚合，使用 soft max 激活函数更新节点的表示，最后重复上一步骤直到节点表示达到稳定状态，将最终的节点表示作为模型输出。特别地，本文对于齐次子图和异质子图使用了不同的注意力评分方法。基于元路径的齐次子图 $G^{\mathrm{hom}o}$，节点 (i,j) 的重要性可表示为：

$$e_{i,j}^{G^{\mathrm{hom}o}} = \mathrm{LeakyReLU}(\boldsymbol{\mu}_{G^{\mathrm{hom}o}}^{\mathrm{T}} \left[\boldsymbol{h}_i \,\|\, \boldsymbol{h}_j\right]) \tag{7}$$

在异质子图 G^{hetro} 中，节点 (i,j) 的重要性可表示为：

$$e_{i,j}^{G^{\mathrm{hetro}}} = \mathrm{LeakyReLU}(\boldsymbol{\mu}_{G^{\mathrm{hetro}}}^{\mathrm{T}} \left[\boldsymbol{h}_i \,\|\, \boldsymbol{h}_j\right]) \tag{8}$$

式中，$e_{i,j}^{G^{\mathrm{hom}o}}$ 和 $e_{i,j}^{G^{\mathrm{hetro}}}$ 分别表示为齐次子图和异质子图中节点 j 对节点 i 的重要性；$\boldsymbol{h}$ 表示为初始的一组节点输入特征；T 表示矩阵转置；$\boldsymbol{\mu}_{G^{\mathrm{hetro}}}^{\mathrm{T}}$ 和 $\boldsymbol{\mu}_{G^{\mathrm{hom}o}}^{\mathrm{T}}$ 分别表示异质子图和齐次子图的节点级注意参数，$\|$ 表示连接操作。为了便于比较不同节点之间的注意力系数，采用 softmax 激活函数来对注意力系数进行标准化，这样有助于提升模型的健壮性和性能表现：

$$a_{i,j}^{G'} = \begin{cases} \mathrm{soft\,max}(e_{i,j}^{G^{\mathrm{hom}o}}) & e_{i,j} = e_{i,j}^{G^{\mathrm{hom}o}} \\ \mathrm{soft\,max}(e_{i,j}^{G^{\mathrm{hetro}}}) & e_{i,j} = e_{i,j}^{G^{\mathrm{hetro}}} \end{cases} \tag{9}$$

式中，G' 表示子图，$N_i^{G'}$ 表示子图 G' 中节点 i 的一阶邻居。在获得调整后的注意力系数后，将会进行标准化处理，形成一个注意力的得分矩阵。根据注意力得分矩阵，使用一个非线性激活函数 $\sigma(\cdot)$ 将邻居节点 $N_i^{G'}$ 的投影特征聚合起来，形成子图 G' 的节点 i 的嵌入特征：

$$\boldsymbol{h}_i^{G'} = \sigma(\sum_{j \in N^i} a_{i,j}^{G'} \boldsymbol{W} \boldsymbol{h}_j) \tag{10}$$

式中，$\boldsymbol{W} \in \mathbb{R}^{F'} \times \mathbb{R}^{F'}$ 为权重矩阵，被应用于每个节点上，F 表示输入特征 $\boldsymbol{h}$ 的初始特征，F' 表示新生成的特征 (F 与 F' 可能有不同的基数)。为了减少单一的自注意力机制的影响，稳定注意力系数，增加了多头注意力机制。具体而言，将自注意力机制重复 K 次，然后把它们的输出特征连接起来，就得到以下的输出特征表示：

$$\boldsymbol{h}_i^{G'} = \|_{k=1}^{K} \sigma(a_{i,j}^{k} \boldsymbol{W} \boldsymbol{h}_j) \tag{11}$$

每个节点的输入特征 F 通过一组权重矩阵 $\boldsymbol{W}^k$ 进行线性变换，然后将结果输入到注意力机制中，得到归一化的注意力系数 $a_{i,j}^k$。最终 $\boldsymbol{h}_i'$ 由每个节点的 KF' 特征组成，其中 KF' 通过将每个节点的 F' 特征与注意力系数相乘并加权求和得到。输出特征 $\boldsymbol{h}'$ 中的每个维度对应一个节点的 KF' 特征。

2.2　节点特征聚集及增强

这个模块的目标是设计一个能够增强模型对重要特征的关注程度的方法。通过聚合通道注意力和空间注意力[25] 两个部分，对于给定的 H 个元路径的药物或疾病节点 i，在基于这些元路径的子图 $G' = \{G'_1, G'_2, \cdots, G'_H\}$ 的情况下获得 H 组嵌入 $H = \{\boldsymbol{h}_l^{G'_1}, \boldsymbol{h}_l^{G'_2}, \cdots, \boldsymbol{h}_l^{G'_H}\}$。不同的元路径包含了不同的语义信息，学习不同元路径的重要性，从基于元路径的子图中获取节点嵌入作为输入，通过对节点嵌入进行线性变换，用于学习通道注意力权重，再通过 ReLU 激活函数，增强非线性表达能力：

$$a_{\mathrm{channels}} = \mathrm{ReLU}(M_c \boldsymbol{h}_i^{G'_H} + b_c) \tag{12}$$

式中，$M_c \in \mathbb{R}^{F'}$ 和 $b_c \in \mathbb{R}^{F'}$ 是学习的非线性变换参数，在通过通道注意力获得元路径语义节点嵌入的

重要性之后，使用两个卷积层进行空间信息融合，用于突出重要的空间位置，抑制不重要的空间位置：

$$a_{\text{spatial}} = M_s(a_{\text{channels}}) \otimes a_{\text{channels}} \tag{13}$$

式中，channels表示通道注意力；spatial表示空间注意力；⊗表示元素相乘。

2.3 药物–疾病关联预测

在经过双层注意力之后，最终得到药物–疾病的嵌入向量$\begin{vmatrix} \boldsymbol{H}_{\text{di}} \\ \boldsymbol{H}_{\text{dr}} \end{vmatrix}$，为了重新构建药物–疾病之间的关联矩阵，采用了双线性内积解码器：

$$H' = \text{soft}\max(\boldsymbol{H}_{\text{di}} \boldsymbol{W}' \boldsymbol{H}_{\text{dr}}) \tag{14}$$

式中，$\boldsymbol{W}' \in \mathbb{R}^{M \times N}$是一个可训练的权重矩阵，药物和疾病之间的关联预测得分由相应的(i, j)决定，每个元素$H'_{i,j}$表示药物i和疾病j之间的关联得分。

3 实验分析与案例研究

3.1 实验设置

3.1.1 实验参数

为了评估 MHNGA 的准确性，本文使用了 10 次交叉验证来减少因数据分裂而引起的随机误差，为了尽可能全面地评估模型的性能，采用了多个指标包括 AUC、AUPR、F1-评分、准确率、召回率、特异性和精确度。其中，AUC（曲线下面积）是评估分类器质量的一种经典指标，它衡量了分类器在不同阈值下的真阳性率和假阳性率之间的权衡；AUPR（召回率–精确率的曲线下面积）评估分类器在不同阈值下准确性和召回率之间的权衡；F1-评分则是准确性和召回率的调和平均值，它将二者的指标综合考虑；准确率、召回率、特异性和精确度分别衡量了分类器在不同方面的性能。

在 MHNGA 中有几个超参数， 如嵌入维度、初始学习率lr、注意力头数、MHNGA 的总训练周期以及随机失活 (dropout) 率。参考文献 [26] 的研究并且在多组参数值中进行了实验，通过多次实验比对，发现当嵌入维度为 64、注意力头数为 5、学习率为 0.005 时效果最好，由此来设置 MHNGA 模型的参数。

3.1.2 实验环境

本论文实验使用的框架为 Python 3.8 版本，搭配 PyTorch 2.0.1 和 DGL 1.1.0 等开源工具进行实现。

3.2 MHNGA 实验结果分析

3.2.1 和其他算法进行比较

本文将 MHNGA 与其他 6 种药物疾病关联算法进行比较， 以证明 MHNGA 模型的有效性。这些方法包括 REDDA[26]、DEEPDR[9]、NIMGCN[27]、HINGRL[28]、LAGCN[13] 和 DRWBNCF[29]，实验结果如表 2 和表 3 所示。

表 2 7 种方法在 A 数据集上的表现

Methods	AUC	AUPR	F1	Accuracy	Recall	Specificity	Precision
REDDA	0.922	**0.444**	**0.495**	**0.994**	0.451	**0.998**	**0.552**
DEEPDR	0.821	0.135 3	0.199	0.940	0.295	0.957	0.291
NIMGCN	0.668	0.181	0.260	0.993	0.197	**0.998**	0.388
HINGRL	0.870	0.241	0.283	0.992	0.248	0.997	0.341
LAGCN	0.850	0.314 9	0.310 0	0.960 2	0.353 9	0.975 9	0.275 8
DRWBNCF	0.790	0.352	0.416	**0.994**	0.347	**0.998**	0.524
MHNGA	**0.935**	0.404	0.486	**0.994**	**0.459**	**0.998**	0.516

表 3 7 种方法在 B 数据集上的表现

Methods	AUC	AUPR	F1	Accuracy	Recall	Specificity	Precision
REDDA	0.829	**0.491**	**0.482**	0.989	0.531	0.990	**0.469**
DEEPDR	0.809	0.121	0.170	0.935	0.246	0.930	0.289
NIMGCN	0.652	0.132	0.221	0.987	0.165	**0.991**	0.298
HINGRL	0.835	0.198	0.256	0.982	0.224	0.989	0.320
LAGCN	0.829	0.297	0.291	0.921	0.309	0.972	0.264
DRWBNCF	0.730	0.328	0.393	**0.990**	0.299	**0.991**	0.484
MHNGA	**0.837**	0.469	0.471	0.860	**0.545**	0.901	0.416

3.2.2　MHNGA 的泛化能力

为了研究 MHNGA 模型的泛化能力，进行了额外的实验。本文并没有通过传统的使用 A 数据集训练 MHNGA，然后用 MHNGA 去预测 B 的方法，而是通过去除一定比例的药物疾病关联，来分析具有不同稀疏性的异构网络的泛化能力。

实验中，从 A 数据集中逐步去除药物疾病的关联比例，每次去除数据集的十分之一，直到十分之九。使用 MHNGA 模型进行训练，并得到了表 4 和表 5 的结果。根据实验结果得知，当训练中涉及更多的药物疾病关联时，MHNGA 的性能得到了提高，这是因为当在训练数据中包含更多的药物–疾病关联信息时，MHNGA 能够学习到更多的知识。MHNGA 的 AUC、AUPR、F1 评分比原模型下降了 10%～15%，这验证了 MHNGA 的泛化能力。总的来说，训练和测试数据集中的药物和疾病的数量对 MHNGA 的泛化能力有一定的影响，但考虑到 GAT 对网络结构的约束性，仍然能表明 MHNGA 具有较好的泛化能力。

表 4　本文模型的泛化能力在 A 数据集上的比较

Fold/%	AUC	AUPR	F1-Score
10	0.750	0.298	0.512
20	0.751	0.299	0.519
30	0.759	0.301	0.523
40	0.760	0.304	0.529
50	0.773	0.309	0.533
60	0.787	0.322	0.535
70	0.790	0.327	0.540
80	0.799	0.337	0.543
90	0.802	0.352	0.551

表 5　本文模型的泛化能力在 B 数据集上的比较

Fold/%	AUC	AUPR	F1-Score
10	0.732	0.245	0.501
20	0.736	0.244	0.503
30	0.740	0.252	0.511
40	0.743	0.255	0.518
50	0.751	0.260	0.523
60	0.754	0.267	0.527
70	0.762	0.270	0.534
80	0.766	0.273	0.539
90	0.773	0.287	0.546

3.2.3　消融实验

为了证明 MHNGA 模型结构的重要性和必要性，本文进行了一系列的消融实验，提出并测试了一些可以看作是 MHNGA 简化版的设计，以下是这些变化的详细情况。

MHNGA-GAT：没有图注意力神经网络的 MHNGA。

MHNGA-GCN：将图注意力网络替换成图卷积网络的 MHNGA。

MHNGA-Layer：没有全局注意力模块的 MHNGA。

MHNGA-SE：将全局注意力网络改为压缩激活注意力的 MHNGA。

如表 6 所示，MHNGA 的性能在 AUC、Accuracy、Recall 和 Specificity 上都得到了最高分，表明 MHNGA 的模型结构是最佳的。相比之下，MHNGA-Layer 和 MHNGA-SE 的表现不如 MHNGA，这说明层注意力能够很好地在数据中提取有用特征并提高药物–疾病关联任务的性能。同时，与 MHNGA-GAT 相比，MHNGA-Layer 和 MHNGA-SE 的表现都要更优秀。这表对选特征的筛选有助于更好地利用生物信息。总的来说，这些结果证明了 MHNGA 模型结构的合理性。

表 6　A 数据集上本文模型与其 3 种简化版本的比较

Methods	AUC	Accuracy	Recall	Specificity
MHNGA-GAT	0.802	**0.994**	0.446	0.997
MHNGA-Layer	0.901	0.990	0.324	0.994
MHNGA-SE	0.902	0.985	0.265	0.990
MHNGA	**0.935**	**0.994**	**0.459**	**0.998**

3.3　案例研究

为了验证 MHNGA 发现新的药物–疾病关联的能力，本文利用已知的药物–疾病关联来训练 MHNGA 模型，利用这个模型预测新的关联。为了验证结果的有效性，本文使用公共文献或其他可靠来源进行验证。MHNGA 模型预测的前 10 个药物疾病关联如表 7 所示，其中前 4 个已经被证实[30]。

表 7　本文模型预测的前 10 名药物–疾病关联

Drug	Disease	Evidence
clobazam	Hypophosphatemia	CTD
rofecoxib	Thrombosis	CTD
Valsartan	Fibrosis	CTD
Dasatinib	Asthenia	CTD
Acebutolol	Hemangioma	NA
Doxorubicin	Ataxia	文献[31]
Phenformin	Vomiting	文献[32]
Rifampin	Diarrhea	PMID: 3 571 889
Sulfasalazine	Coma	PMC10145436
Mefloquine	Malaria	文献[33]

此外，检查了氯米帕明（Clomipramine）的前

5 个候选疾病和乳腺肿瘤的前 5 个候选药物。表 8 和表 9 显示了实验结果，其中一些预测可以得到证实。如氯米帕明是一种用于治疗强迫症、恐慌症、重度抑郁以及慢性疼痛等疾病的药物。文献 [34] 通过实验证明氯米帕明的基本作用与丙咪嗪相同，证实氯米帕明可用于治疗部分性癫痫。文献 [35] 通过对 32 例乳腺癌患者的他莫昔芬治疗，发现有 12 例完全缓解或部分缓解，同时还发现含有雌激素受体的肿瘤对他莫昔芬有反应，证明了他莫昔芬可治疗乳腺癌。

表 8　本文模型预测的氯米帕明前 5 个候选疾病

Disease	Evidence
Cataplexy	文献[36]
Nausea	CTD
Seizures	文献[37]
Hepatitis	文献[38]
Arthralgia	PMID: 7 009 255

表 9　本文模型预测的乳腺癌前 5 个候选药物

Drug	Evidence
ochratoxin A	CTD
Tamoxifen	PMID: 9 045 313
Dinoprostone	CTD
bisphenol A	CTD
Cacodylic Acid	CTD

案例研究表明，MHNGA 可以帮助确定新的关联以及相关疾病（相关药物）的给定药物（疾病）。

4　结束语

本文建立了一个 MHNGA 的模型来发现药物–疾病关联。采用了药物–疾病关联、药物–药物相似性、疾病–疾病相似性组成的异构网络。MHNGA 通过将图注意力机制和全局注意力机制相结合，成功预测了药物–疾病关联。从实验结果可以看出，本文方法优于其他药物–疾病关联预测方法。

生物网络是一个互联互通的大型网络，未来将考虑加入更多的生物网络，如基因、蛋白质、靶标，共同建立一个包含更多节点类型和边类型的异构网络，以便进行药物–疾病关联预测。虽然 GAT 是一个强大的图神经网络，能够有效提取网络中的节点信息，但它丢失了网络中的结构信息，未来我们将考虑加入图嵌入等方法解决这一问题。

参 考 文 献

[1] TAMIMI N A M, ELLIS P. Drug development: From concept to marketing![J]. Nephron Clinical Practice, 2009, 113(3): c125-c131.

[2] DIMASI J A, HANSEN R W, GRABOWSKI H G. The price of innovation: New estimates of drug development costs[J]. Journal of Health Economics, 2003, 22(2): 151-185.

[3] DICKSON M, GAGNON J P. Key factors in the rising cost of new drug discovery and development[J]. Nature Reviews Drug Discovery, 2004, 3(5): 417-429.

[4] HENRY G. Are the economics of pharmaceutical research and development changing?[J]. PharmacoEconomics, 2004, 22(2): 15-24.

[5] PUSHPAKOM S, IORIO F, EYERS P A, et al. Drug repurposing: Progress, challenges and recommendations[J]. Nature Reviews Drug Discovery, 2019, 18(1): 41-58.

[6] DAI W, LIU X, GAO Y, et al. Matrix factorization-based prediction of novel drug indications by integrating genomic space[J]. Computational and Mathematical Methods in Medicine, 2015, DOI: 10.1155/2015/275045.

[7] HUANG F, QIU Y, LI Q, et al. Predicting drug-disease associations via multi-task learning based on collective matrix factorization[J]. Frontiers in Bioengineering and Biotechnology, 2020, 8: 218.

[8] LUO H, LI M, WANG S, et al. Computational drug repositioning using low-rank matrix approximation and randomized algorithms[J]. Bioinformatics, 2018, 34(11): 1904-1912.

[9] ZENG X, ZHU S, LIU X, et al. deepDR: A network-based deep learning approach to in silico drug repositioning[J]. Bioinformatics, 2019, 35(24): 5191-5198.

[10] GOTTLIEB A, STEIN G Y, RUPPIN E, et al. Predict: A method for inferring novel drug indications with application to personalized medicine[J]. Molecular Systems Biology, 2011, 7(1): 496.

[11] WANG Y, CHEN S, DENG N, et al. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data[J]. PloS One, 2013, 8(11): e78518.

[12] LI Z, HUANG Q, CHEN X, et al. Identification of drug-disease associations using information of molecular structures and clinical symptoms via deep convolutional neural network[J]. Frontiers in Chemistry, 2020, 7: 1-14.

[13] YU Z, HUANG F, ZHAO X, et al. Predicting drug-disease associations through layer attention graph convolutional network[J]. Brief Bioinform, 2021, 22(4): bbaa243.

[14] ZENG X, ZHU S, LU W, et al. Target identification among known drugs by deep learning from heterogeneous networks[J]. Chemical Science, 2020, 11(7): 1775-1797.

[15] LUO H, WANG J, LI M, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm[J]. Bioinformatics, 2016, 32(17): 2664-71.

[16] LUO Y, ZHAO X, ZHOU J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous

information[J]. Nature Communications, 2017, 8(1): 573.
[17] CHU Y, WANG X, DAI Q, et al. MDA-GCNFTG: Identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph[J]. Briefings in Bioinformatics, 2021, 22(6): bbab165.
[18] HU L, CHAN K C C, YUAN X, et al. A variational Bayesian framework for cluster analysis in a complex network[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(11): 2115-2128.
[19] HU L, CHAN K C C. Fuzzy clustering in a complex network based on content relevance and link structures[J]. IEEE Transactions on Fuzzy Systems, 2015, 24(2): 456-470.
[20] HU L, ZHANG J, PAN X, et al. HiSCF: Leveraging higher-order structures for clustering analysis in biological networks[J]. Bioinformatics, 2021, 37(4): 542-550.
[21] PENG L, YANG C, CHEN Y, et al. Predicting CircRNA-disease associations via feature convolution learning with heterogeneous graph attention network[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(6): 3072-3082.
[22] DAVIS A P, GRONDIN C J, JOHNSON R J, et al. Comparative toxicogenomics database (CTD): Update 2021[J]. Nucleic Acids Res, 2021, 49(D1): D1138-D1143.
[23] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018[J]. Nucleic Acids Research, 2018, 46(D1): D1074-D1082.
[24] VELIKOVI P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. [2023-05-10]. https://arxiv.org/pdf/1710.10903.
[25] LIU Y C, SHAO Z R, HOFFMANN N. Global attention mechanism: Retain information to enhance channel-spatial interactions[EB/OL]. [2023-05-15]. https://arxiv.org/abs/2112.05561v1.
[26] GU Y, ZHENG S, YIN Q, et al. REDDA: Integrating multiple biological relations to heterogeneous graph neural network for drug-disease association prediction[J]. Computers in Biology and Medicine, 2022, 150.
[27] LI J, ZHANG S, LIU T, et al. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction[J]. Bioinformatics, 2020, 36(8): 2538-2546.
[28] ZHAO BW, HU L, YOU ZH, et al. HINGRL: Predicting drug–disease associations with graph representation learning on heterogeneous information networks[J]. Briefings in Bioinformatics, 2022, 23(1): bbab515.
[29] MENG Y, LU C, JIN M, et al. A weighted bilinear neural collaborative filtering approach for drug repositioning[J]. Brief Bioinform, 2022, 23(2): bbab581.
[30] DAVIS A P, GRONDIN C J, JOHNSON R J, et al. The Comparative Toxicogenomics Database: Update 2019[J]. Nucleic Acids Res, 2019, 47(D1): D948-D954.
[31] WENNINGMANN N, KNAPP M, ANDE A, et al. Insights into doxorubicin-induced cardiotoxicity: molecular mechanisms, preventive strategies, and early monitoring[J]. Molecular Pharmacology, 2019, 96(2): 219-232.
[32] ROMANKIEWICZ J A. Phenformin-associated lactic acidosis: A review[J]. American Journal of Health-System Pharmacy, 1975, 32(5): 502-507.
[33] SCHLAGENHAUF P, ADAMCOVA M, REGEP L, et al. The position of mefloquine as a 21st century malaria chemoprophylaxis[J]. Malaria Journal, 2010, 9(357): 1-15.
[34] SAKAKIHARA Y, OKA A, KUBOTA M, et al. Reduction of seizure frequency with clomipramine in patients with complex partial seizures[J]. Brain and Development, 1995, 17(4): 291-293.
[35] KIANG D T, KENNEDY B J. Tamoxifen (antiestrogen) therapy in advanced breast cancer[J]. Annals of Internal Medicine, 1977, 87(6): 687-690.
[36] SCHACHTER M, PARKES J D. Fluvoxamine and clomipramine in the treatment of cataplexy[J]. Journal of Neurology, Neurosurgery, and Psychiatry, 1980, 43: 171-174.
[37] SAKAKIHARA Y, OKA A, KUBOTA M, et al. Reduction of seizure frequency with clomipramine in patients with complex partial seizures[J]. Brain & Development 1995, 17: 291-293.
[38] ALDERMAN C P, ATCHISON M M, MCNEECE J I, et al. Concurrent agranulocytosis and hepatitis secondary to clomipramine therapy[J]. British Journal of Psychiatry, 1993, 162(5): 688-689.

编　辑　叶　芳