GraphCDA: a hybrid graph representation learning framework based on GCN and GAT for predicting disease-associated circRNAs

Qiguo Dai, Ziqiang Liu, Zhaowei Wang, Xiaodong Duan and Maozu Guo

摘要

动机: 环状 RNA (circRNA) 是一类具有高度保守性和稳定性的非编码 RNA，被认为是重要的疾病生物标志物和药物靶点。越来越多的证据表明，circRNA 在许多复杂疾病的发病和进展中起着至关重要的作用。由于生物实验费时费力，开发一种精确的计算预测方法对于识别与疾病相关的环状 rna 是非常分必要的。结果: 我们提出了一个名为 GraphCDA 的混合图表示学习框架，用于预测潜在的环状 rna-疾病关联。首先，构建环状 rna-环状 rna 相似网络和疾病-疾病相似网络，分别表征环状 rna 与疾病的关系。其次，引入一种结合图卷积网络和图注意网络的混合图嵌入模型，同时学习环状 rna 和疾病的特征表示。最后，将学习到的表征串联起来，用于构建预测模型，以识别环状 rna 与疾病的相关性。一系列实验结果表明，GraphCDA 在多个公共数据库上的性能优于其他当前最先进的方法。此外，当仅使用少量已知的环状 rna-疾病关联作为训练集时，GraphCDA 可以取得良好的性能。此外，对几种人类疾病的案例研究进一步证实了 GraphCDA 对潜在疾病相关环状 rna 的预测能力。总之，大量的实验结果表明，GraphCDA 可以作为探索 circRNA 在复杂疾病中的调控作用的可靠工具。

关键词：环状 rna -疾病关联；图卷积网络；图注意网络；特征学习

1 引言

环状 RNA (circRNA) 是一种特殊的单链环内源性非编码 RNA。长期以来，它被认为是异常剪切的产物，早在 1976 年，植物病毒[1]中就首次发现了异常剪切。近年来，随着高通量测序技术的发展，在许多生物体内发现了大量的环状 rna。circRNA 的主要特征是它具有闭环结构，没有 5’ 和 3’ 聚腺苷酸化尾[2]。现有的证据已经发现环状 rna 在细胞活性和基因调控中起着至关重要的作用。例如，circRNAs 可通过 microRNA 通路[3]调节食管鳞状细胞癌的辐射敏感性。此外，circrna 参与了许多复杂疾病的发生和发展，如癌症、心血管和神经系统疾病[4]。通过收集已证实的 circRNAs 与疾病之间的联系，提出了一些关键的数据库，如 circRNADisease[5]、circAtlas[6]、circ2Disease[7]、circFunbase[8]、circR2Disease[9]和 circR2Disease2.0[10]。尽管如此，仍有许多环状 rna 与疾病的关联尚未得到证实[11]。因此，有必要深入研究发现新的与疾病相关的环状 rna，有助于阐明疾病的发病机制和发展机制。

2 相关工作

最近，已经提出了许多计算方法来识别 circRNA 与疾病的潜在关联。之前的一些研究将机器学习模型应用于预测分类器，如 GBDT[12]、RWRKNN[13]等。这些方法基于从一组 circRNA 和疾病的相似网络中提取的特征来预测与疾病相关的 circRNA。尽管工作取得了巨大进展，但在学习 circRNA 与疾病关联的特征表示方面仍存在一些挑战。近十年来，深度学习在图像识别[14]、语音识别[15]和自然

语言处理[16]等领域取得了突出的成绩，受到了广泛的关注。基于深度学习的模型可以从输入数据中自动学习特征表示，而无需依赖传统的手工特征。目前，它在生物信息学中也得到了广泛的应用。例如，Wang 等[17]利用卷积神经网络 (CNN) 从多源信息中提取高级特征，然后预测 circrna 与疾病的关联。Deepthi 等人[18]开发了一种用于环状 rna -疾病相似性特征表示的深度自编码器，将其输入深度神经网络以预测疾病相关的环状 rna。虽然以上基于深度学习的工作取得了很好的效果，但是像 CNN 和 autoencoder 这样的深度学习模型并不适合于图数据，如 circRNA 相似度网络和疾病相似度网络。图神经网络 (GNNs) 可以通过消息传递[19]来模拟节点之间的邻接关系。图卷积网络 (Graph Convolutional Network, GCN) 是一种基于图的深度学习方法，它利用图内部的结构信息来聚合邻近节点的信息。它在文本分类和知识图[20]领域取得了优异的性能。在识别疾病相关非编码 RNA 方面，GCN 已被用于预测疾病相关 miRNAs 和长链非编码 RNA (lncRNAs)[21-22]。在 circrna-疾病关联预测方面，Wang 等[23]利用 FastGCN 融合疾病和 circrna 的相似信息，得到一个统一的描述符。提取的高水平特征被 ForestPA 分类器[24]用于预测 circrna 与疾病之间的关联。尽管 GCN 取得了良好的效果，但仍存在一定的局限性。

对于一个节点，它的不同邻居的特征重要性是不同的。注意机制可以对重要特征给予更多的关注，而对其他特征进行抑制。

在图学习中，图注意网络 (GAT)[25]是一种多头自注意机制，通过给邻域的不同节点分配不同的权重来学习图上节点的特征表示。它已被用于确定与良好表现相关的 mirna -疾病和 lncrna -疾病[26-27]。在之前的研究[28-29]中，GAT 也被用于学习用于表征 circrna -疾病关联的节点表示。

一般来说，GCN 和 GAT 分别用于预测 circRNA 与疾病的关联，它们可以从不同的角度表征节点的表征。因此，将这两种方法有效结合起来，可以学习到更合理的 circrna 和疾病的表示。

3 本文方法

3.1 本文方法概述

在这项研究中，提出了一个名为 GraphCDA 的混合图表示学习框架来预测 circRNA 与疾病的关联。通过使用 GCN 和 GAT 的组合，分别学习 circRNAs 和疾病的特征表示，它们被用于构建分类模型以产生最终预测结果。如图 1 所示，所提出的框架由以下模块组成：（i）从 CircR2Disease 数据库和疾病本体构建 circRNA 相似性网络和疾病相似性网络；（ii）通过融合 GCN 和 GAT 的混合模型，从相应的相似性网络中学习 circRNA 和疾病的特征表示，并将不同 GCN 层的输出组合起来，然后分别作为 circRNAs 和疾病的特性；（iii）将学习到的特征进一步串联，以产生 circRNA-疾病关联的描述符，这些描述符最终用于训练随机森林分类器，以预测 circRNA-疾病关联，如图 1所示：

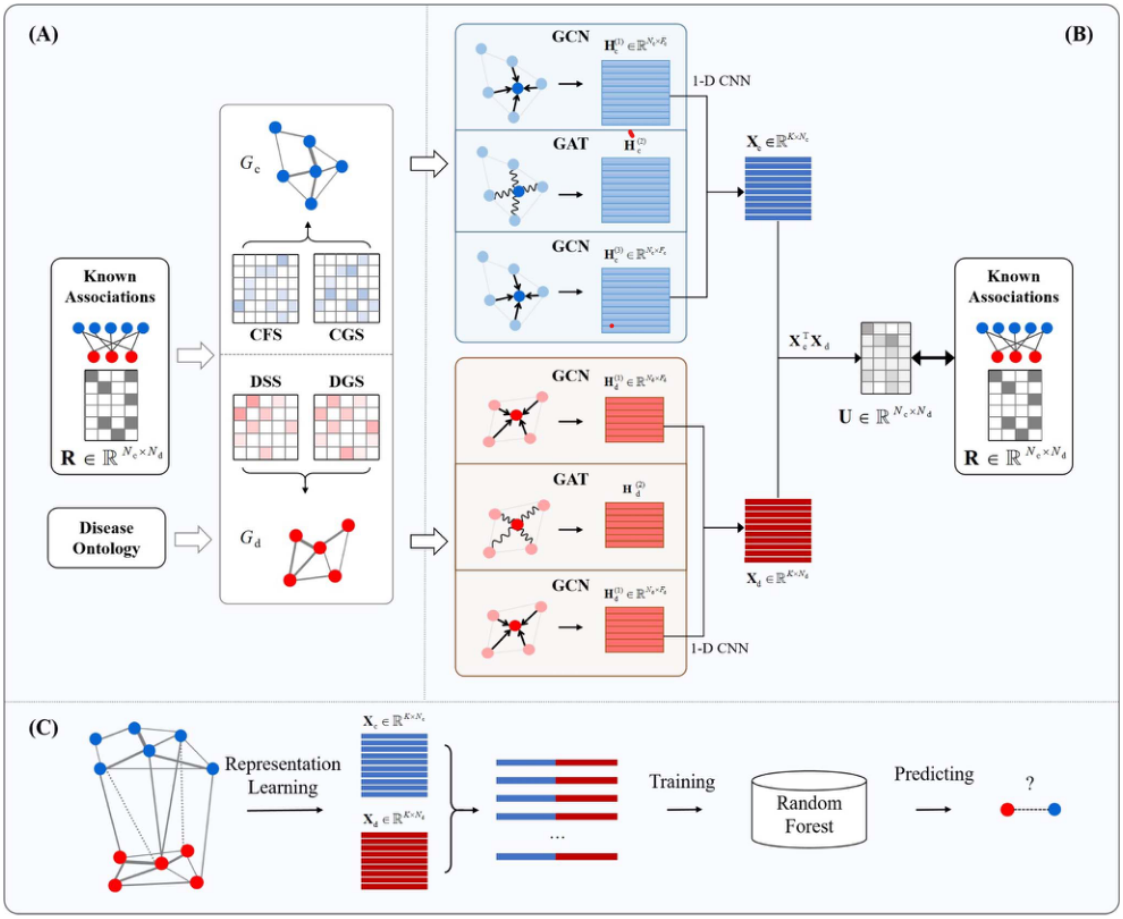

图 1: 方法示意图

3.2 数据来源

在这项研究中，广泛使用的 CircR2Disease 数据库[9]用于构建和测试 GraphCDA。基于该数据库，计算 circRNA 相似度和疾病相似度，并用于构建 circRNA-circRNA 集成相似度网络和疾病-疾病集成相似度网络。已验证的 circRNA−疾病关联下载了 circRNA-疾病[5]、circAtlas[6]、circ2Disease[7]、circFunbase[8]、circR2Disease[9]和 circR2Disease2.0[10]数据库中的 circRNA-疾病关联信息，用于构建 circRNA 与疾病关联矩阵。从这些数据库中删除了非人和重复数据，每个数据库的 circRNA、疾病和关联对的数量如表 1 所示。R 是 circRNA-疾病关联矩阵，其中 $R_{i,j}$=1，如果 circRNA ci 与数据库中的疾病 dj 相关；否则，$R_{i,j}$=0。

3.3 疾病语义相似性

为了构建疾病相似性网络，基于疾病本体计算不同疾病之间的语义相似性，其中使用有向无环图（DAG）组织疾病。数据库中每种疾病的疾病本体关系可从 https://disease-ontology.org 上获取。可以使用 DOSE 软件包中的 doSim 函数计算两种疾病的语义相似性，定义如下：

$$\mathrm{DSS}(d_i, d_j) = \frac{\sum_{x \in N_{d_i} \cap N_{d_j}} \left(S_{d_i}(x) + S_{d_j}(x)\right)}{\sum_{x \in N_{d_j}} S_{d_j}(x) + \sum_{x \in N_{d_j}} S_{d_j}(x)} \tag{1}$$

其中 N_{d_i} 由疾病 d_i 及其在 $DAG\ (d_i)$ 中的所有祖先疾病组成。$S_{d_j}(x))$ 表示疾病 x 对疾病 di 的语义贡献值，如下所示：

$$\begin{cases} S_{d_i}(x) = max\{\mu * S_{d_i}(x') \mid x' \in \text{ children of } d_i\} & \text{if } x \neq d_i \\ S_{d_i}(d_i) = 1 & \text{otherwise} \end{cases} \tag{2}$$

与相关工作[30]类似，语义贡献因子 μ 设置为 0.5。

3.4 CircRNA 功能相似性

在先前的相关研究[4,31]中，功能相似的 circRNA 被认为更可能与表型相似的疾病相关，反之亦然。因此，我们的研究还采用了 circRNA 功能相似性，这是从疾病语义相似性和 circRNA 与疾病关联数据中获得的。与之前的方法[30,32]类似，circRNA c_i 和 c_j 之间的功能相似性可以计算如下：

$$\mathrm{CFS}(c_i, c_j) = \frac{\sum_{1 \leqslant q \leqslant |D_i|} DS(d_q, D_j) + \sum_{1 \leqslant r \leqslant |D_j|} DS(d_r, D_i)}{|D_i| + |D_j|} \tag{3}$$

$$\mathrm{DS}(d_q, D_j) = \max_{1 \leqslant t \leqslant |D_j|} (\mathrm{DSS}(d_q, d_t)) \tag{4}$$

其中 D_i 表示与 circRNA c_i 相关的疾病集，$DS\ (d_q, D_j)$ 表示疾病 d_q 和 D_j 之间的语义相似性。

3.5 circRNA 与疾病的高斯相互作用谱核相似性 (GIP 相似性)

由于疾病本体中有一些疾病没有很好地注释，因此无法使用上述方法计算相应疾病的语义相似性及其相关 circRNA 的功能相似性。疾病语义相似性和 circRNA 功能相似性可以用高斯相互作用轮廓核相似性（GIP）代替。设 $IP\ (c_i)$ 表示 circRNAc_i 的二元相互作用轮廓向量，其对应于邻接矩阵 R 中的第 i 行。circRNA c_i 和 c_j 之间的 GIP 相似性可以如下计算：

$$\mathrm{CGS}\ (c_i, c_j) = \exp\left(-\rho \left\|\mathrm{IP}\ (c_i) - \mathrm{IP}\ (c_j)\right\|^2\right) \tag{5}$$

$$\rho = \frac{1}{\frac{1}{N_c} \sum_{i=1}^{N_c} \|IP(c_i)\|^2} \tag{6}$$

其中 ρ 是控制内核带宽的参数，N_c 是邻接矩阵 R 的行数。类似地，疾病 d_i 和 d_j 之间的 GIP 核相似性可以表示为:

$$\mathrm{DGS}(d_i, d_j) = \exp\left(-\rho \left\|\mathrm{IP}(d_i) - \mathrm{IP}(d_j)\right\|^2\right) \tag{7}$$

$$\rho = \frac{1}{\frac{1}{N_d} \sum_{i=1}^{N_d} \|\mathrm{IP}(d_i)\|^2} \tag{8}$$

其中 $IP\ (d_i)$ 表示疾病 d_i 的二元相互作用分布向量，其对应于邻接矩阵 R 中的第 i 列。因此，可以获得 circRNA（C）和 disease（d）的集成相似矩阵，其元素表示如下：

$$C_{i,j} = \begin{cases} \mathrm{CGS}(c_i, c_j) & \text{if } \mathrm{CFS}(c_i, c_j) = 0 \\ \mathrm{CFS}(c_i, c_j) & \text{otherwise;} \end{cases} \tag{9}$$

$$D_{i,j} = \begin{cases} \mathrm{DGS}(d_i, d_j) & \text{if } \mathrm{DSS}(d_i, d_j) = 0 \\ \mathrm{DSS}(d_i, d_j) & \text{otherwise} \end{cases} \tag{10}$$

其中矩阵 C 中 circRNA i 和 j 之间的相似性 $C \in \mathrm{R}^{N_c \times N_c}$ 作为 circRNA 相似网络 G_c 中节点 i 和 j 的边缘权重。矩阵 D 中疾病 i 和 j 之间的相似性 $D \in \mathrm{R}^{N_d \times N_d}$ 是相似网络 G_d 中疾病 i 和 j 的边缘权重。

3.6 circRNA 和疾病的图表征学习

两个由一组 GCN 和 GAT 层组成的图学习模块分别用于从 circRNA 和疾病的相似网络中学习特征表示。如图 1（B）所示，每个图学习模块包含两个图卷积层，它们之间有一个图注意层；并且使用融合层来集成不同图卷积层的输出。

使用 GCN 分别从 circRNA 相似网络和疾病相似网络中获得 circRNA 和疾病的特征表示。给定网络 G，其邻接矩阵和输入节点表示可以表示为 $C \in \mathrm{R}^{N \times N}$ 和 $H \in \mathrm{R}^{N \times F}$，N 是节点的数量，F 是节点特征的维度。输出节点表示 H^{news} 可以通过如下 GCN 层获得：

$$H^{new} = GCN(S, H) \tag{11}$$

$$\mathrm{GCN}(S, H) = \sigma\left(A^{-\frac{1}{2}} \tilde{S} A^{-\frac{1}{2}} H Q\right) \tag{12}$$

其中 $\tilde{S} = 1 + S$; $A = \sum_j \tilde{\mathcal{S}}_{i,j}$ 是度矩阵；$Q \in R^{F \times F}$ 是可训练权重矩阵；σ(.) 是 ReLU 激活函数。

GAT 是一种神经网络，它利用多头注意力根据相邻节点的重要性为其分配不同的权重。在 GraphCDA 中，GAT 层被引入到两个 GCN 层之间，旨在帮助后续 GCN 层提取 CircRNA 和疾病的高级特征。对于网络 G，GAT 层的输出节点表示 H^{new} 如下：

$$H^{new} = GAT(S, H) \tag{13}$$

$$\vec{H}_i^{new} = \sigma\left(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \neq i} \phi_{ij}^k W_k \vec{H}_i\right) \tag{14}$$

$\vec{H}_i^{new}$ 表示 H^{new} 中节点 i 的特征；K 是多头注意中注意机制的数量；H^{new} 是第 k 个注意机制的权重矩阵。H_i 是 circRNA 节点 i 的输入特征向量；ϕ_{ij}^k 表示为节点 i 和 t 之间的第 k 个注意力系数。

$$\phi_{ij}^k = \frac{\exp\left(\mathrm{LR}\left(a_k^T [W_k \vec{H}_i \,||\, W_k \vec{H}_t \,||\, B_k S_{\mathrm{it}}]\right)\right)}{\sum_{t \neq t} \exp\left(\mathrm{LR}\left(a_k^T [W_k \vec{H}_i \,||\, W_k \vec{H}_t \,||\, B_k S_{\mathrm{tt}}]\right)\right)} \tag{15}$$

其中 || 串联操作；LR 是 LeaklyReLU 激活函数；$a_k \in R^{2F+1}$ 是第 k 个注意力机制的权重向量；B_k 是要学习的边缘 S_{ij} 的权重。

基于上述 GCN 和 GAT 层，可以构建 circRNA 和疾病的表示学习模块，分别用于从对应的相似性网络构建节点的特征表示。对于输入 circRNA 相似性网络 G_c，其邻接矩阵表示为 C，其中 $C_{i,j} \in C$ 如等式 (9) 所示。和 $H_c^{(0)} \in R^{N_c \times F_c}$ 是网络的输入节点特征，其中 F 是 circRNA 特征的维度。上述交替引入 GCN 和 GAT 层，以获得不同水平 circRNA 节点的图形特征表示，如下所示：

$$\begin{cases} H_c^{(1)} = G\mathrm{CN}\left(C, H_c^{(0)}\right) \\ H_c^{(2)} = GAT\left(C, H_c^{(1)}\right) \\ H_c^{(3)} = GCN\left(C, H_c^{(2)}\right) \end{cases} \tag{16}$$

为了组合不同 GCN 层的输出特征 $H_c^{(1)}$ 和 $H_c^{(3)}$，使用 1D CNN 来产生 circRNA 表示 X_c。类似地，我们还利用 GCN 和 GAT 从疾病相似性网络 Gd 中学习多层次节点特征 $H_d^{(1)}$、$H_d^{(2)}$ 和 $H_d^{(3)}$。G_d 的邻接矩阵表示为 d，其中 $D_{i,j} \in D$ 如等式（10）所示。疾病网络的初始特征是 $H_d^{(0)} \in R^{N_d \times F_d}$。

$$\begin{cases} \mathrm{H}_d^{(1)} = \mathrm{GCN}\left(D, H_d^{(0)}\right) \\ \mathrm{H}_d^{(2)} = \mathrm{GAT}\left(D, H_d^{(1)}\right) \\ \mathrm{H}_d^{(3)} = \mathrm{GCN}\left(D, H_d^{(2)}\right) \end{cases} \tag{17}$$

如上所述，还使用 1D CNN 对 GCN 层 $\{H_d^{(1)}$，$H_d^{(3)}\}$ 的输出进行积分，以获得疾病表示 X_d。

3.7 表征学习的模型训练

基于 X_c 和 X_d 的表示，circRNA-疾病的预测偏好矩阵 U 可以获得为

$$\mathbf{U} = {X_c}^{\mathbb{T}} X_d \tag{18}$$

U 中的 U_{ij} 越高，circRNAi 与疾病 j 相关的可能性越大。使用二元交叉熵（BCE）来测量偏好矩阵 U 和已知邻接矩阵 R 之间的差异，将其作为训练图表示学习模型的损失函数。通过最小化 circRNA 与疾病关联的训练数据库上的损失函数，可以学习 X_c 和 X_d 的图形表征矩阵，这些矩阵用于预测新的 circRNA 和疾病关联。随机森林（RF）是一种强大的集成分类器，它利用多个决策树来缓解训练数据的过度拟合问题。它已被广泛用于解决生物信息学领域的预测问题。在之前的研究中，RF 也被用于预测 miRNA 与疾病的相关性。因此，在这项工作中，还基于上述融合描述符对 RF 进行了训练，以预测疾病相关的 circRNA。

4 复现细节

4.1 与已有开源代码对比

本次复现是参照论文所提出的模型以及作者源代码自己重新实现完成的，实验数据使用了论文提到的数据库所提供的数据。与开源代码相比，由于开源代码仅实现了作者所提出的模型算法，对于论文中提到的许多对比试验却没有实现。本人的工作除了实现该模型的代码外，还实现了许多论文提及的而开源代码没有实现的部分。并在此基础上，对作者提出的模型进行更改，尝试找到更好效果的模型。

4.2 实验环境搭建

- python (tested on version 3.8.13)
- pytorch (tested on version 1.11.0+cu115)
- torch-geometric (tested on version 2.1.0post1)
- numpy (tested on version 1.22.3)
- scikit-learn(tested on version 1.1.2)

4.3 界面分析与使用说明

重现结果：

- 运行 main.py 运行 GraphCDA。

文件夹:

- code：GraphCDA 的模型代码和训练代码。
- data：GraphCDA 所需的数据。

- datasets：几个公共数据库
- results：GraphCDA 运行的结果。

数据描述:

- d_d.csv: 疾病综合相似性
- c_c.csv: circRNA 综合相似性
- d_c.csv: 疾病-circRNA 关联表
- dss.csv: 疾病语义相似性
- cfs.csv: circRNA 功能相似性
- dgs.csv: 疾病高斯内核相似性
- cgs.csv: circRNA 高斯内核相似性
- disname.txt: 疾病名称列表
- circname.txt: circRNA 名称列表

4.4 创新点

作者在使用多层图神经网络计算得到环状 RNA 和疾病的特征后，使用随即森林分类器进行分类。由于作者的平均精准度已达到较高值，对模型的更改和参数的调整已很难再提高。故从最后分类方法寻求创新。随机森林和 lightGBM 都属于集成学习，集成学习本身不是一个单独的机器学习算法，而是通过构建并结合多个机器学习器来完成学习任务。根据基本学习器的生成方式，目前的集成学习方法大致分为两大类：即基本学习器之间存在强依赖关系、必须串行生成的序列化方法，以及基本学习器间不存在强依赖关系、可同时生成的并行化方法；前者的代表就是 Boosting，后者的代表是随机森林。因此，本人通过在最后分类的步骤中，新增论文中没有提到的另一个集成学习算法 lightGBM 分类器作为比较，对比 lightGBM 与随机森林的效果，结果对比将在下一章节显示。发现 lightGBM 与随机森林各有优劣，但都保持着较高值。

5 实验结果分析

为了验证 GraphCDA 中引入的混合图学习策略的有效性，我们将 GraphCDA 分解为两个模型，即 GraphCDA-nogat 和 GraphCDA-last。GraphCDA-nogat 指的是去除两个 GCN 层之间 GAT 的方法。GraphCDA 最后一个版本仅使用最后一个 GCN 层的输出进行特征提取，而不使用 GAT 层和 1D CNN 组合器。将训练后获得的 circRNAs 和疾病的节点特征矩阵输入到随机森林分类器中进行预测，并对 circRNADisease 进行了 100 次 5 倍交叉验证。结果如表 1所示。可以发现，GraphCDA-nogat 在部分指标上都显著优于 GraphCDAlast，这表明组合从 GCN 的不同层输出的节点特征可以有效地集成不同级别的表示。对于 GraphCDA 和 GraphCDA-nogat 的比较，可以看出 GraphCDA 在所有指标方面都优于 GraphCDA-nogat。这表明 GraphCDA 中的 GAT 层可以有效地提高 GraphCDA 中 circRNA 和疾病的表示学习能力。

在以下实验中，使用了 5 倍交叉验证来评估性能，相关工作中采用了几个评估指标，包括准确度（Acc.）、准确度（Pre.）、灵敏度（Sen.）、F1Score（F1）、马修斯相关系数（MCC）、接收机工作特性曲线下面积（AUROC）和准确度召回曲线（AUPR）[33-34]。除非另有说明，以下实验是在 5 倍交叉验证下

表 1: 在 circR2Disease 上使用不同网络的比较结果

Model	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
GraphCDA	**0.9485**	**0.9412**	**0.9564**	**0.9486**	**0.8969**	**0.9831**	**0.9845**
GraphCDA-nogat	0.9239	0.93	0.9176	0.9237	0.8477	0.9684	0.9761
GraphCDA-last	0.9146	0.9071	0.9249	0.9156	0.8299	0.9793	0.982

对 CircR2Disease 数据库中的人类 circRNA 疾病相关性进行测试的。表 2总结了 GraphCDA 的实验结果，其中显示了 5 倍交叉验证的性能。关于 5 个折叠的平均值，GraphCDA 在测试数据库上的准确度、精度、灵敏度、F1 评分和 MCC 方面分别达到了 0.9485、0.9412、0.9564、0.9486 和 0.8969。AUROC 和 AUPR 的平均值达到了 0.9831 和 0.9845。这些实验结果表明 GraphCDA 在测试数据库上表现良好，可以有效预测潜在的 circRNA 与疾病的关联。

表 2: GraphCDA 对 CircR2-疾病的 5 倍交叉验证测试结果

Validation set	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
1	0.95	0.9452	0.965	0.955	0.899	0.9779	0.9815
2	0.9308	0.935	0.92	0.9274	0.8614	0.9747	0.9783
3	0.95	0.9254	0.9764	0.9502	0.9014	0.9833	0.9866
4	0.9577	0.958	0.95	0.954	0.9149	0.9872	0.9814
5	0.9538	0.9424	0.9704	0.9562	0.9079	0.9923	0.9945
Avg	0.9485	0.9412	0.9564	0.9486	0.8969	0.9831	0.9845

如表 3所示，lightgbm、RF、SVM、DT、LR、AB 和 NB 的平均 AUROC 值分别为 0.9446、0.9485、0.8738、0.89、0.7492、0.7 和 0.63。可以发现，lightgbm 和 Random Forest 在各个方面取得了较好的结果。就整体表现而言，随机森林优于其他分类模型。如图 2 所示，但在 AUROC 值的分布方面，使用 lightgbm 作为最终分类器的 GraphCDA 也明显优于其他方法。

表 3: 在 circR2Disease 上使用不同分类器的 GraphCDA 的比较结果

Classifiers	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
lightgbm	0.9446	0.9316	**0.9601**	0.9455	0.8898	**0.9886**	**0.9908**
Random forest	**0.9485**	**0.9412**	0.9564	**0.9486**	**0.8969**	0.9831	0.9845
SVM	0.8738	0.8481	0.9118	0.8783	0.7503	0.9597	0.9603
Decision Tree	0.89	0.8638	0.925	0.8932	0.7814	0.8903	0.8371
Logistic Regression	0.7492	0.7273	0.7982	0.7602	0.5019	0.7828	0.6991
Adaptive Boosting	0.7	0.7126	0.6742	0.6922	0.3996	0.7636	0.7033
Naive Bayes	0.63	0.6061	0.7425	0.6662	0.2659	0.693	0.6712

为了研究不同 circRNA 相似性对 GraphCDA 性能的影响，我们将我们的方法与仅使用 circRNA 功能相似性的模型（GraphCDAfunc）和仅使用 cicircRNA GIP 高斯内核相似性的模式（GraphCDA-GIP）进行了比较。结果如表 4所示，结果表明，GraphCDA 的性能略为优于 GraphCDA-func 和 GraphCDA-gip，这表明整合两种类型的 circRNA 相似性有助于表示 circRNA 之间的关系并预测 circRNA 与疾病的关联。

表 4: 在 circR2Disease 上使用不同相似性的比较结果

Model	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
GraphCDA	**0.9485**	0.9412	**0.9564**	**0.9486**	**0.8969**	0.9831	0.9845
GraphCDAfunc	0.9475	**0.9663**	0.9276	0.9463	0.896	0.9818	0.984
GraphCDA-GIP	0.9346	0.9332	0.9371	0.9347	0.8697	**0.986**	**0.9878**

6 总结与展望

使用计算方法预测 circRNA 与疾病的相关性对于理解 circRNA 在病理机制、诊断和治疗人类复杂疾病中的作用至关重要。在这项研究中，提出了一种结合 GCN 和 GAT 的混合图表示学习框架，以识别潜在的疾病相关 circRNA，命名为 GraphCDA。首先，通过整合各种相似性，建立了 circRNA 相似性网络和疾病相似性网络。其次，将 GCN 和 GAT 相结合，以有效地学习 circRNA 和疾病的特征表示。由不同 GCN 层学习的节点特征与 1D CNN 集成。最后，基于 circRNA 和疾病的学习表示，基于随机森林的模型被训练用于预测 circRNA 与疾病的关联。为了验证 GraphCDA 的性能，在多个公共数据库上进行了一系列实验，并进行了 5 倍交叉验证。我们在几个公共数据库上将 GraphCDA 与其他最先进的方法进行了比较，结果表明，所提出的方法比其他最先进方法获得了更好的性能。

此外，对三种常见疾病的案例研究表明，当仅使用少数已知的相关 circRNA 作为训练集时，GraphCDA 表现出良好的泛化能力。总的来说，这项工作中提出的 GraphCDA 是预测潜在 circRNA 与疾病关联的有效且准确的方法。

参考文献

[1] SANGER H, KLOTZ G, RIESNER D, et al. VIROIDS ARE SINGLE-STRANDED COVALENTLY CLOSED CIRCULAR RNA MOLECULES EXISTING AS HIGHLY BASE-PAIRED ROD-LIKE STRUCTURES[J]. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 1976, 73(11): 3852-3856. DOI: 10.1073/pnas.73.11.3852.

[2] QU S, YANG X, LI X, et al. Circular RNA: A new star of noncoding RNAs[J]. CANCER LETTERS, 2015, 365(2): 141-148. DOI: 10.1016/j.canlet.2015.06.003.

[3] LIU J, XUE N, GUO Y, et al. CircRNA_100367 regulated the radiation sensitivity of esophageal squamous cell carcinomas through miR-217/Wnt3 pathway[J]. AGING-US, 2019, 11(24): 12412-12427. DOI: 10.18632/aging.102580.

[4] DEEPTHI K, JEREESH A S. Inferring Potential CircRNA-Disease Associations via Deep Autoencoder-Based Classification[J]. MOLECULAR DIAGNOSIS & THERAPY, 2021, 25(1): 87-97. DOI: 10.1007/s40291-020-00499-y.

[5] ZHAO Z, WANG K, WU F, et al. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations[J]. CELL DEATH & DISEASE, 2018, 9. DOI: 10.1038/s41419-018-0503-3.

[6] WU W, JI P, ZHAO F. CircAtlas: an integrated resource of one million highly accurate circular RNAs

from 1070 vertebrate transcriptomes[J]. GENOME BIOLOGY, 2020, 21(1). DOI: 10.1186/s13059-020-02018-y.

[7] YAO D, ZHANG L, ZHENG M, et al. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease[J]. SCIENTIFIC REPORTS, 2018, 8. DOI: 10.1038/s41598-018-29360-3.

[8] MENG X, HU D, ZHANG P, et al. CircFunBase: a database for functional circular RNAs[J]. DATABASE-THE JOURNAL OF BIOLOGICAL DATABASES AND CURATION, 2019. DOI: 10.1093/database/baz003.

[9] FAN C, LEI X, FANG Z, et al. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases[J]. DATABASE-THE JOURNAL OF BIOLOGICAL DATABASES AND CURATION, 2018. DOI: 10.1093/database/bay044.

[10] FAN C, LEI X, TIE J, et al. CircR2Disease v2.0: An Updated Web Server for Experimentally Validated circRNA-disease Associations and Its Application.[J]. Genomics, proteomics & bioinformatics, 2021. DOI: 10.1016/j.gpb.2021.10.002.

[11] WANG C C, HAN C D, ZHAO Q, et al. Circular RNAs and complex diseases: from experimental results to computational models[J]. BRIEFINGS IN BIOINFORMATICS, 2021, 22(6). DOI: 10.1093/bib/bbab286.

[12] LEI X, FANG Z. GBDTCDA: Predicting circRNA-disease Associations Based on Gradient Boosting Decision Tree with Multiple Biological Data Fusion[J]. INTERNATIONAL JOURNAL OF BIOLOGICAL SCIENCES, 2019, 15(13): 2911-2924. DOI: 10.7150/ijbs.33806.

[13] LEI X, BIAN C. Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association[J]. SCIENTIFIC REPORTS, 2020, 10(1). DOI: 10.1038/s41598-020-59040-0.

[14] STOIMCHEV M, IVANOVSKA M, STRUC V. Learning to Combine Local and Global Image Information for Contactless Palmprint Recognition[J]. SENSORS, 2022, 22(1). DOI: 10.3390/s22010073.

[15] LU X, SHI D, LIU Y, et al. Speech depression recognition based on attentional residual network[J]. FRONTIERS IN BIOSCIENCE-LANDMARK, 2021, 26(12): 1746-1759. DOI: 10.52586/5066.

[16] TSURUOKA Y. [Deep Learning and Natural Language Processing].[J]. Brain and nerve = Shinkei kenkyu no shinpo, 2019, 71(1): 45-55. DOI: 10.11477/mf.1416201215.

[17] WANG L, YOU Z H, HUANG Y A, et al. An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network[J]. BIOINFORMATICS, 2020, 36(13): 4038-4046. DOI: 10.1093/bioinformatics/btz825.

[18] DEEPTHI K, JEREESH A S. An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network[J]. GENE, 2020, 762. DOI: 10.1016/j.gene.2020.145040.

[19] NIU M, ZOU Q, WANG C. GMNN2CD: identification of circRNA-disease associations based on variational inference and graph Markov neural networks[J]. BIOINFORMATICS, 2022, 38(8): 2246-2253. DOI: 10.1093/bioinformatics/btac079.

[20] SPINELLI I, SCARDAPANE S, UNCINI A. Adaptive Propagation Graph Convolutional Network[J]. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 2021, 32(10): 4755-4760. DOI: 10.1109/TNNLS.2020.3025110.

[21] ZHU R, JI C, WANG Y, et al. Heterogeneous Graph Convolutional Networks and Matrix Completion for miRNA-Disease Association Prediction[J]. FRONTIERS IN BIOENGINEERING AND BIOTECHNOLOGY, 2020, 8. DOI: 10.3389/fbioe.2020.00901.

[22] XUAN P, PAN S, ZHANG T, et al. Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-Disease Associations[J]. CELLS, 2019, 8(9). DOI: 10.3390/cells8091012.

[23] WANG L, YOU Z H, LI Y M, et al. GCNCDA: A new method for predicting circRNA-disease associations based on Graph Convolutional Network Algorithm[J]. PLOS COMPUTATIONAL BIOLOGY, 2020, 16(5). DOI: 10.1371/journal.pcbi.1007568.

[24] ADNAN M N, ISLAM M Z. Forest PA: Constructing a decision forest by penalizing attributes used in previous trees[J]. EXPERT SYSTEMS WITH APPLICATIONS, 2017, 89: 389-403. DOI: 10.1016/j.eswa.2017.08.002.

[25] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph Attention Networks[EB/OL]. arXiv. (2018-02-04) [2022-10-25]. http://arxiv.org/abs/1710.10903. arXiv: 1710.10903[cs,stat].

[26] XUAN P, CAO Y, ZHANG T, et al. Dual Convolutional Neural Networks With Attention Mechanisms Based Method for Predicting Disease-Related lncRNA Genes[J]. FRONTIERS IN GENETICS, 2019, 10. DOI: 10.3389/fgene.2019.00416.

[27] TANG X, LUO J, SHEN C, et al. Multi-view Multichannel Attention Graph Convolutional Network for miRNA-disease association prediction[J]. BRIEFINGS IN BIOINFORMATICS, 2021, 22(6). DOI: 10.1093/bib/bbab174.

[28] BIAN C, LEI X J, WU F X. GATCDA: Predicting circRNA-Disease Associations Based on Graph Attention Network[J]. CANCERS, 2021, 13(11). DOI: 10.3390/cancers13112595.

[29] JI C, LIU Z, WANG Y, et al. GATNNCDA: A Method Based on Graph Attention Network and Multi-Layer Neural Network for Predicting circRNA-Disease Associations[J]. INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES, 2021, 22(16). DOI: 10.3390/ijms22168505.

[30] WANG D, WANG J, LU M, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. BIOINFORMATICS, 2010, 26(13): 1644-1650. DOI: 10.1093/bioinformatics/btq241.

[31] LI G, YUE Y, LIANG C, et al. NCPCDA: network consistency projection for circRNA-disease association prediction[J]. RSC ADVANCES, 2019, 9(57): 33222-33228. DOI: 10.1039/c9ra06133a.

[32] CHEN X, YAN C C, LUO C, et al. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity[J]. SCIENTIFIC REPORTS, 2015, 5. DOI: 10.1038/srep11338.

[33] SWETS J. MEASURING THE ACCURACY OF DIAGNOSTIC SYSTEMS[J]. SCIENCE, 1988, 240(4857): 1285-1293. DOI: 10.1126/science.3287615.

[34] BRADLEY A. The use of the area under the roc curve in the evaluation of machine learning algorithms [J]. PATTERN RECOGNITION, 1997, 30(7): 1145-1159. DOI: 10.1016/S0031-3203(96)00142-2.