

CAPTLDA: 基于胶囊网络和Transformer预测 LncRNA-疾病关联

张嘉辉, 谭建军*

北京工业大学化学与生命科学学院, 北京

收稿日期: 2025年11月18日; 录用日期: 2025年12月31日; 发布日期: 2026年1月9日

摘要

长链非编码RNA (lncRNA) 是一类长度超过200个核苷酸的转录物, 在多种疾病的发病机制中发挥关键作用。因此, 阐明lncRNA与疾病之间的关联对于理解潜在的发病机制和开发新的疾病预防、诊断和治疗策略至关重要。虽然传统的生物学实验对于预测长链非编码RNA-疾病关联(LDA)是有价值的, 但往往费用高昂且耗时。开发有效的LDA预测计算模型是有必要的。当前的计算方法在有效整合多源数据和捕获异质生物网络中的复杂高阶关系模式方面经常遇到限制。这项研究提出了一种新的计算框架命名为CAPTLDA, 将lncRNA、疾病和miRNA的相似性和关联整合到一个加权的异构网络邻接矩阵中, 引入了胶囊网络, 以增强特征学习。此外, 还采用Transformer编码器, 它结合了全局多头代理注意力机制和并行的多头局部注意力机制, 以全面捕获全局依赖关系和局部上下文信息, 最终实现准确的LDA预测。在两个基准数据集上进行的综合计算实验表明, 模型在性能上优于先进的现有方法。案例研究进一步验证了它在识别潜在疾病相关lncRNA方面的有效性。

关键词

LncRNA-疾病关联, 胶囊网络, Transformer, 代理注意力机制, 深度学习

CAPTLDA: Predicting LncRNA-Disease Associations Based on Capsule Networks and Transformers

Jiahui Zhang, Jianjun Tan*

College of Chemistry and Life Science, Beijing University of Technology, Beijing

Received: November 18, 2025; accepted: December 31, 2025; published: January 9, 2026

*通讯作者。

文章引用: 张嘉辉, 谭建军. CAPTLDA: 基于胶囊网络和 Transformer 预测 LncRNA-疾病关联[J]. 生物医学, 2026, 16(1): 11-24. DOI: 10.12677/hjbm.2026.161002

Abstract

Long non-coding RNAs (lncRNAs) are transcripts exceeding 200 nucleotides in length and play pivotal roles in the pathogenesis of various diseases. Therefore, elucidating the associations between lncRNAs and diseases is crucial for understanding underlying mechanisms and developing novel strategies for disease prevention, diagnosis, and treatment. While traditional biological experiments are valuable for predicting lncRNA-disease associations (LDA), they are often costly and time-consuming. Developing effective computational models for LDA prediction is therefore necessary. Current computational methods frequently encounter limitations in effectively integrating multi-source data and capturing complex higher-order relationship patterns in heterogeneous biological networks. This study proposes a novel computational framework named CAPTLDA, which integrates the similarities and associations of lncRNAs, diseases, and miRNAs into a weighted heterogeneous network adjacency matrix. A capsule network is introduced to enhance feature learning. Additionally, a Transformer encoder is employed, combining a global multi-head agent attention mechanism with parallel multi-head local attention mechanisms to comprehensively capture global dependencies and local contextual information, ultimately achieving accurate LDA prediction. Comprehensive computational experiments on two benchmark datasets demonstrate that our model outperforms existing advanced methods in performance. Case studies further validate its effectiveness in identifying potential disease-related lncRNAs.

Keywords

lncRNA-Disease Association, Capsule Network, Transformer, Agent Attention, Deep Learning

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

根据分子生物学的中心法则, 遗传信息主要编码在蛋白质编码基因中。然而这些基因仅占人类基因组的约 1.5%, 表明超过 98% 由非蛋白质编码序列组成[1]。长链非编码 RNA (long non-coding RNA, lncRNA) 被定义为长度超过 200 个核苷酸的非编码 RNA [2], 以前被视为转录噪声。然而, 有证据表明 lncRNA 在多种细胞过程中发挥重要的调控功能, 包括细胞周期控制、胚胎发育、基因表达的时空调控和细胞命运决定[3]。此外, lncRNA 与各种疾病的发病和进展越来越相关[4]。lncRNA 的失调或突变会导致复杂且通常难以治愈的疾病, 如心血管疾病[5]、阿尔茨海默病[6]和糖尿病[7]。lncRNA 表达异常与多种癌症的发生和发展也密切相关[8]。阐明 lncRNA 与人类疾病之间的关联已成为揭示潜在疾病机制[9]和确定新的治疗靶点[10]的关键研究焦点。尽管一些长链非编码 RNA-疾病关联(lncRNA-disease association, LDA)已经通过实验验证, 但传统的生物学方法通常费力、耗时和资源密集, 从而限制了 LDA 的全面识别。这一瓶颈突显了迫切需要有效的计算模型。利用现有的实验数据, 已经开发了许多计算方法来预测潜在的 LDA。这些方法可以大致分为三类: 基于生物网络的模型、矩阵补全和基于机器学习的方法。

基于生物网络的模型通过构建整合了实验验证的 LDA、疾病相似性和 lncRNA 相似性的异构网络来预测潜在的 LDA。这些方法通常采用随机游走和标签传播算法来捕获全局网络结构并推断潜在的关联。Sun 等人[11]提出了一种名为 RWRLncD 的方法, 该方法在 lncRNA 功能相似性网络上应用带重启的随机游走, 从而改善了对特征较少的 lncRNA 的预测。Chen 等人[12]开发了 KATZLDA, 整合了 LDA、lncRNA

表达、疾病语义相似性和高斯相互作用谱核相似性, 以提高预测准确性。KRWRH [13]通过 lncRNA 的表型信息和组织特异性表达细节来推断疾病-lncRNA 关联。

基于矩阵补全的方法构成了预测 LDA 的第二大类。这些方法利用归纳矩阵补全来恢复一个低秩矩阵, 该矩阵整合了 lncRNA 和疾病的先验知识, 从而补全 LDA 矩阵[14]。Lu 等人[15]提出了 SIMCLDA, 它计算 lncRNA 高斯相互作用谱核相似性和疾病功能相似性, 然后应用主成分分析提取特征进行归纳矩阵补全。GCRFLDA [16]从已知的 LDA 构建一个图, 并使用带有条件随机场和注意力的编码器-解码器框架来预测关联。MFLDA [17]可以通过为数据源分配不同的权重来选择和整合数据源, 进一步引入了一种迭代解决方案, 以同时优化权重和低秩矩阵。

基于机器学习的方法是 LDA 预测的第三个大类, 包括传统机器学习和深度学习方法。Zhu 等人[18]将增量主成分分析与随机森林相结合, 开发了 IPCARF, 这是一种集成多种相似性度量的 LDA 预测方法。CNNLDA [19]采用了一种具有注意力机制的双路径卷积神经网络来学习全局和关注特征表示, 然后将这些表示结合起来估计关联概率。GCNLDA [20]集成了图卷积网络和卷积神经网络来提取网络级别和局部特征。gGATLDA [21]通过为每个 lncRNA-疾病对提取子图、分配基于相似性的节点属性并采用图神经网络进行预。GTAN [22]利用多个图注意力网络以及卷积和多层感知器层对拓扑结构和节点属性进行编码。

现有方法在预测潜在 LDA 方面取得了显著进展, 但仍有很大改进空间。一方面, 直接利用原始的相似性和关联信息在计算上是低效的, 一些模型的计算可能会模糊节点嵌入的语义表示从而削弱分类性能。另一方面, 越来越复杂的模型架构所带来的性能提升相对有限, 这表明这些方法可能并没有完全符合 LDA 预测的基本假设。Geoffrey Hinton 等人[23]开发的胶囊网络克服了传统卷积神经网络的一些限制, 通过将特征信息表示为向量, 更好地捕获了节点的相对位置和方向。胶囊网络还引入了一种动态路由机制, 可以动态调整胶囊之间的连接权重, 从而在处理复杂的输入变化方面表现出更高的性能。Transformer 架构 [24]已被广泛应用于各个领域, 并越来越多地应用于 LDA 预测[25]。虽然 Transformer 中的自注意力机制有效地捕获了全局依赖关系, 但也存在固有的挑战, 它的二次计算复杂性限制了可扩展性, 并且在稀疏数据条件下性能通常会下降, 可能会导致关键信息的丢失。Han 等人[26]提出的一种新颖的全局注意力机制名为代理注意力机制, 可以在计算效率和表示能力之间实现最佳平衡。Luong 等人[27]提出了一种局部注意力机制, 它专注于输入数据的特定子集, 而不是整个输入数据。在 LDA 领域, 包含 lncRNA-疾病关联的特征矩阵具有高度稀疏性, 这可能会导致全局注意力机制忽略特定的局部相关模式。局部注意力机制的引入可以有效解决这个问题, 减少潜在的信息丢失从而增强模型捕获全局和关系信息的能力。

基于上述, 我们提出了一个全新的模型名为 CAPTLDA, 首先使用了 lncRNA、miRNA 的功能相似性还有疾病的语义相似性的信息数据, 构造了一个加权的邻接矩阵, 然后将邻接矩阵输入到胶囊网络中进行特征的学习与提取, 最后将胶囊网络的输出送到一个集成了全局多头代理注意力机制与多头局部注意力机制并行的 Transformer 编码器中进行信息整合以完成最终的预测输出。在两个数据集上采用多个评价指标对 CAPTLDA 的预测结果进行评估, 相较于其他几个先进的模型, CAPTLDA 具有更好的预测性能, 案例研究的分析进一步证明了模型潜在的预测价值。

2. 材料与方法

2.1. 数据集

为了严格评估所提出模型的有效性, 我们在两个受广泛认可的基准数据集上进行了评估:

数据集 1 源自 Fu 的研究[17]。该数据集广泛应用于 lncRNA 疾病关联预测的研究中。它包括 240 个 lncRNA、412 种疾病和 495 个 miRNA, 以及从 Lnc2Cancer [28]、LncRNADisease [29]和 GeneRIF [30]中收集的 2697 个实验验证的 lncRNA 疾病关联。此外, 它还包括来自 starBasev2.0 [31]的 1002 个 lncRNA-

miRNA 关联, 以及来自 HMDD v2.0 [32] 的 13,562 个 miRNA 疾病关联。

数据集 2 来源于 Zhou 等人的研究[33]。该数据集包括 665 个 lncRNA、316 种疾病和 295 个 miRNA。它整合了 3833 个实验支持的 LDA, 这些关联来自 Lnc2Cancer v3.0 [34] 和 LncRNADisease v2.0 [35] 的更新版本, 2108 个 lncRNA-miRNA 关联来自 starBase v2.0, 以及 8540 个 miRNA-疾病关联来自 HMDD v3.0 [36]。节点注释和分类分别来源于 MeSH² [37] 和 miRbase [38], 便于一致的生物学解释。

2.2. 模型架构

如图 1 所示, 本文构建了一个全新的模型 CAPTLDA 用于 lncRNA-疾病对的预测, 首先通过 lncRNA、miRNA、疾病三者之间的关联性与相似性构建出六个子矩阵, 然后将六个子矩阵融合为一个加权的邻接矩阵, 将该矩阵输入到胶囊网络中进行特征学习, 最后将胶囊网络的输出输入进混合注意力机制的 Transformer 编码器中进行全局信息整合输出分类预测结果。

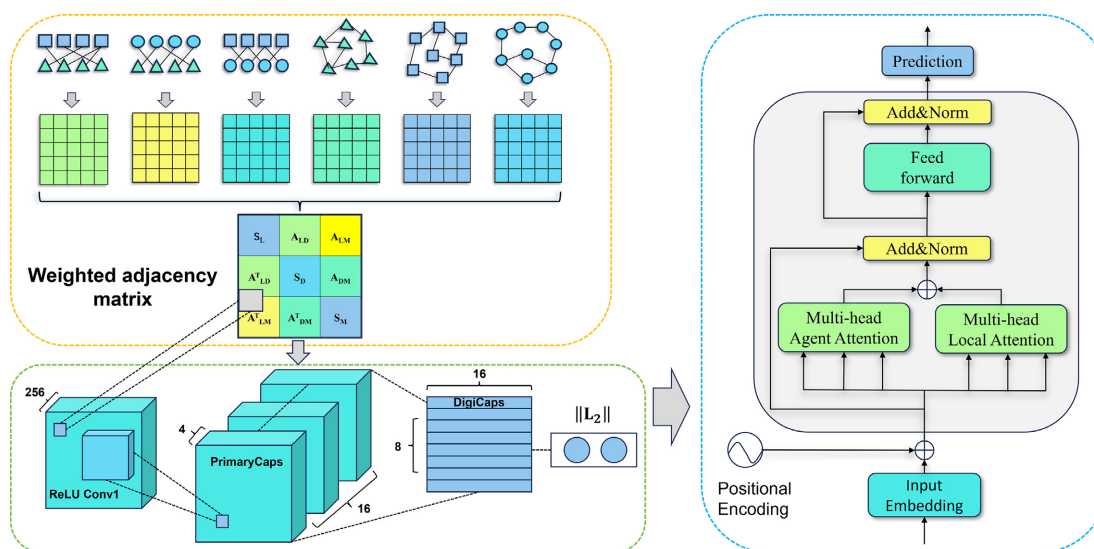


Figure 1. Flowchart of CAPTLDA

图 1. CAPTLDA 的流程图

2.3. 疾病语义相似性

疾病语义相似性(Disease semantic similarity, DSS)因其在不同疾病之间捕获表型相关性的有效性, 在 LDA 预测研究中得到了广泛应用。疾病语义信息来源于 Disease ontology, 使用有向无环图(DAG)结构对疾病之间的层次关系进行建模[39]。按照 Wang 等人提出的方法[40], 通过 Disease ontology 中疾病先祖的语义贡献来量化 DSS。具体来说, 对于给定的疾病 d_1 , 令 D_1 表示包括 d_1 本身以及 DAG 中所有祖先疾病的集合。然后, 正式定义 D_1 中每种疾病对 d_1 的语义贡献如下:

$$S_{d1}(d) = \begin{cases} 1, & d = d_1 \\ \max_{d' \in \text{children of } d} \left(\frac{S_{d1}(d')}{2} \right), & d \in D_1 \text{ and } d \neq d_1 \end{cases} \quad (1)$$

对于任何其他疾病 d_2 , d_1 和 d_2 之间的语义相似性定义为:

$$\text{Sim}(d_1, d_2) = \frac{\sum_{d \in D_1 \cap D_2} (S_{d1}(d) + S_{d2}(d))}{SV(d_1) + SV(d_2)} \quad (2)$$

其中, $SV(d_1) = \sum_{d \in D_1} S_{d1}(d)$ 表示集合 D_1 内所有疾病对疾病 d_1 的累积语义贡献。 $SV(d_2)$ 表示集合内所有疾病对疾病 d_2 的累积语义贡献。两种疾病之间的语义相似性是基于它们各自的先祖疾病集合之间的重叠程度来定量定义的。当共享的先祖疾病更接近目标疾病时, 相似性分数增加, 这反映了更精确的表型相关性。

2.4. lncRNA/miRNA 功能相似性

遵循 Wang 等人提出的方法[38], 通过整合疾病语义相似性与实验验证的 LDA 和 miRNA-疾病关联 (miRNA-disease association, MDA), lncRNA 功能相似性(lncRNA functional similarity, LFS)和 miRNA 功能相似性(miRNA functional similarity, MFS)进行定量评估。对于两个 lncRNA 或 miRNA, 分别表示为 r_1 和 r_2 , 令 n_1 和 n_2 分别表示与 r_1 和 r_2 相关的疾病数量。这些相关疾病定义为 $t_{1i}, (1 \leq i \leq n_1)$ 和 $t_{2j}, (1 \leq j \leq n_2)$ 。然后, r_1 和 r_2 之间的功能相似性表述如下:

$$Sim(r_1, r_2) = \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \max_{1 \leq j \leq n_2} (Sim(t_{1i}, t_{2j})) + \sum_{j=1}^{n_2} \max_{1 \leq i \leq n_1} (Sim(t_{2j}, t_{1i})) \right] \quad (3)$$

2.5. 加权邻接矩阵

考虑一个包含 l 个 lncRNA、 m 个 miRNA 和 d 种疾病的数据集。这些实体之间的类间关联关系由邻接矩阵 $A_{LD} \in R^{l \times d}$, $A_{LM} \in R^{l \times m}$ 和 $A_{DM} \in R^{d \times m}$ 表示。这些矩阵中的每个条目 (i, j) 被分配一个二进制值, 其中 1 表示相应实体之间存在已知关联, 否则为 0。类内相似性矩阵 $S_L \in R^{l \times l}$, $S_D \in R^{d \times d}$ 和 $S_M \in R^{m \times m}$ 对同一类实体之间的功能或语义相似性分数进行编码, 对角线上的值为自相似性值被设为 0。

$$A = \begin{bmatrix} S_L & A_{LD} & A_{LM} \\ A_{LD}^T & S_D & A_{DM} \\ A_{LM}^T & A_{DM}^T & S_M \end{bmatrix} \quad (4)$$

2.6. 胶囊网络

胶囊网络由六层神经网络组成, 包括胶囊卷积层、初级胶囊层、数字胶囊层和三个全连接层。当前一胶囊层的输出向量进入更高胶囊层时, 会首先与权重矩阵相乘, 并映射到与更高胶囊层神经元数量相同的空间中。计算过程如下:

$$\tilde{u}_{ji} = W_{ij} u_i \quad (5)$$

其中, W_{ij} 表示对应 u_i 的权重矩阵, $\tilde{u}_{ji}$ 表示预测向量。

动态路由利用权重来确定较低层的向量如何进入较高层的向量, 计算如下:

$$s_j = \sum_i c_{ij} \tilde{u}_{ji} \quad (6)$$

其中, c_{ij} 为耦合系数。

对于最大池化, 只有一个值可以进入该层, c_{ij} 就是相应的唯一的热点向量。计算方式如下:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad (7)$$

这里 b_{ij} 是胶囊 i 和胶囊 j 的对数概率。在迭代过程中, b_{ij} 不断更新, 其公式为:

$$b_{ij} = b_{ij} + v_j \tilde{u}_{ji} \quad (8)$$

Squash 是胶囊网络引入的一个重要创新, 是一种归一化操作, 它针对每个向量使得变化后长度介于 0 到 1 之间, 只改变大小而不影响方向, 计算过程如下:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \quad (9)$$

其中, v_j 表示胶囊 j 的输出向量, 而 s_j 则是胶囊 j 的总输入向量。

2.7. 混合注意力机制

2.7.1. 代理注意力机制

代理注意力具体形式化为一个四元向量 (Q, A, K, V) , 在传统的注意力模块中引入了一组额外的代理令牌 A 。代理令牌首先充当查询令牌 Q 的代理, 从键 K 和值 V 中聚合信息。随后, 聚合后的信息会被广播回原始的查询令牌 Q 。由于代理令牌的数量可以设计得远小于查询令牌的数量, 与广泛采用的 **softmax** 注意力相比, 代理注意力展示了更高的效率, 同时保持了全局上下文建模能力因此, 它无缝地结合了 **softmax** 注意力的强大表示能力和线性注意力的计算效率, 通过引入代理偏差, 可以进一步利用位置信息来增强机制的性能, 计算公式如下:

$$O = \sigma(QA^T + B_2) \sigma(AK^T + B_1) V + \text{DWC}(V) \quad (10)$$

其中 $B_1 \in \mathbb{R}^{n \times N}$ 和 $B_2 \in \mathbb{R}^{N \times n}$ 是代理偏置, $Q, K, V \in \mathbb{R}^{N \times C}$, $B_1 \in \mathbb{R}^{n \times N}$, $B_2 \in \mathbb{R}^{N \times n}$, 以及 $A = \text{Pooling}(Q) \in \mathbb{R}^{n \times C}$ 。

2.7.2. 局部注意力机制

局部注意力机制专注于输入数据的特定区域, 而不是整个输入数据。从输入特征开始, 一个 1×1 的可变形卷积层生成查询 Q , 而两个单独的可变形卷积层生成键(K)和值(V)。局部注意力模块的输出定义如下:

$$w_c = \sum \text{softmax}(qk) v \quad (11)$$

可变形卷积通过可训练的偏移量来适应采样网格, 使网络能够访问传统局部邻域之外的信息。这些偏移量是通过额外的卷积层从输入特征中学习得到的。因此, 可变形卷积增强了感受野, 并为特征建模引入了更大的灵活性(见图 2)。

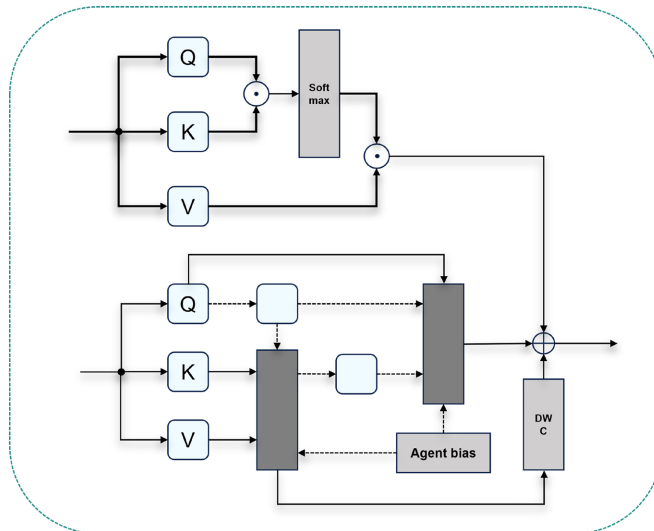


Figure 2. Schematic diagram of the hybrid attention mechanism
图 2. 混合注意力机制示意图

Transformer

Transformer 编码器由 N 个相同的层组成, 每个层包含两个子层: 一个多头注意力机制和一个前馈网络。对于每个子层, 都会加入残差连接, 然后进行层归一化。这个过程可以描述为:

$$H(X) = \text{layerNorm}(X + \text{subLayer}(X)) \quad (12)$$

其中 $\text{sublayer}(X)$ 表示由多头注意力机制或前馈网络实现的计算, $H(X)$ 表示每个子层的输出。前馈网络包括两个由 ReLU 激活函数分隔的全连接层:

$$\text{FFN}(X) = \max(0, XW_1 + b_1)W_2 + b_2 \quad (13)$$

其中, $W_1 \in R^{d_{\text{model}} \times (d_{\text{ff}} * d_{\text{model}})}$, $W_2 \in R^{(d_{\text{ff}} * d_{\text{model}}) \times d_{\text{model}}}$, d_{ff} 的值设置为 2。

预测层对来自注意力层的向量化输出应用线性变换, 然后采用 sigmoid 激活函数, 以获得预测的 LDA 概率 p :

$$p = \text{sigmoid}(\text{flatten}(X_{\text{encoded}})W_{\text{out}} + b_{\text{out}}) \quad (14)$$

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (15)$$

其中, $W_{\text{out}} \in R^{(l+d+m)*d_{\text{model}} \times 1}$, 最后, 二元交叉熵被用作优化模型的损失函数, 定义为:

$$\text{Loss} = -\sum [y \log(p) + (1-y) \log(1-p)] \quad (16)$$

其中 y 表示真实标签, 若数据集中的 lncRNA-疾病对有实验验证的关联则 y 等于 1, 否则 y 等于 0。

3. 实验结果

3.1. 交叉验证与评价指标

在这项研究中, 采用五折交叉验证来评估模型的性能。对于每个数据集, 已知的 LDA 被视为阳性样本, 而未知的 LDA 则构成阴性样本。所有阳性样本被划分为五个子集。四个子集用于训练, 结合等量的随机选择的阴性样本。其余的子集与等大小的随机选择的阴性样本集配对, 用于测试。采用了两个公认的指标来评估模型, 分别是 AUC, 定义为 ROC 曲线下的面积, 它量化了二进制分类器在所有分类阈值上的整体性能; AUPR, 该指标测量精确召回曲线下的面积, 表示分类器在处理不平衡数据集时的性能。此外, 还使用了四个评估指标: 准确率(Accuracy), 召回率(Recall), 精度(Precision), 特异性(Specificity), 以及精确率和召回率的调和平均值 F1 值。这些评价指标公式如下:

$$\text{Accuracy} = \frac{n_{TP} + n_{TN}}{n_{TP} + n_{TN} + n_{FP} + n_{FN}} \quad (17)$$

$$\text{Recall} = \frac{n_{TP}}{n_{TP} + n_{FN}} \quad (18)$$

$$\text{Precision} = \frac{n_{TP}}{n_{TP} + n_{FP}} \quad (19)$$

$$\text{Specificity} = \frac{n_{TN}}{n_{TN} + n_{FP}} \quad (20)$$

$$\text{F1} = 2 \frac{P_r R_c}{P_r + R_c} \quad (21)$$

其中, n_{TN} 是真实负样本对数, n_{TP} 是真正样本的对数, n_{FP} 是假正样本的对数以及 n_{FN} 是假负样本的对数。

3.2. 参数选择

这里展示了 CAPTLDA 的重要超参数在数据集一上的调整结果, 分别是胶囊个数(num_capsules)、胶囊网络中的卷积核大小(kernel size)、代理注意力和局部注意力的头数 agent_n_heads, local_n_heads)、Transformer 编码器的特征维度数(d_model)、前馈网络层数(d_ff)。胶囊个数从{4, 8, 16, 24}中选择, 卷积核大小从{3*3, 6*6, 9*9}中选择, 代理注意力和局部注意力的头数从{1, 2, 4, 8}中选择, 特征维度数从{4, 8, 16, 24}中选择, 前馈网络隐藏单元数从{0.5, 1, 1.5, 2}中选择。如表 1 所示, 当胶囊个数为 8, 卷积核大小为 9*9, 代理注意力头数为 8, 局部注意力头数为 4, Transformer 编码器特征维度数为 16, 前馈网络隐藏单元数为 1 时, 模型在数据集一的 AUC 和 AUPR 值达到最佳。此外, CAPTLDA 是参数不敏感的, 对于不同的超参数组合, 其所达到的 AUC 和 AUPR 值波动较小, 证明了模型良好的性能。

Table 1. The performance of the model when different hyperparameters are used on dataset 1 by CAPTLDA
表 1. CAPTLDA 在数据集 1 上超参数取不同值时模型的性能

参数名称	参数值	AUC	AUPR
num_capsules	4	0.9762	0.9823
	8	0.9898	0.9879
	16	0.9798	0.9810
	24	0.9725	0.9797
kernel_size	3*3	0.9716	0.9804
	6*6	0.9743	0.9738
	9*9	0.9898	0.9879
agent_n_heads	1	0.9798	0.9711
	2	0.9834	0.9692
	4	0.9884	0.9825
	8	0.9898	0.9879
local_n_heads	1	0.9771	0.9804
	2	0.9789	0.9797
	4	0.9898	0.9879
	8	0.9793	0.9805
d_model	4	0.9698	0.9685
	8	0.9775	0.9764
	16	0.9898	0.9879
	24	0.9802	0.9777
d_ff	0.5	0.9757	0.9844
	1	0.9898	0.9879
	1.5	0.9811	0.9817
	2	0.9766	0.9698

3.3. 与其他模型比较

为了验证 CAPTLDA 的性能优越性, 我们选择了五种先进的模型在两个数据集上与其进行对比, 分别是: SIMCLDA [15], 该方法采用主成分分析构建邻接矩阵, 并利用归纳矩阵补全预测潜在的 lncRNA-疾病关联。LDAformer [33], 一种结合了拓扑特征提取路径和 Transformer 编码器的模型, VGAELDA [41] 是一种结合了变分推断和图自动编码器的端到端模型, 具有两个图编码器, 分别是变分图自动编码器和标准图自动编码器。ACLNDA [42] 构建了一个三层异构图, 使用了 Top-K 层内相似性边缘构建方法, 并采用了一个非对称的图对比学习框架, 以最大化一跳邻域上下文和两跳相似性。SSCLMD [43] 是一种多任务预测模型, 该模型利用属性图和拓扑图的自监督对比学习来识别潜在的 LDA。

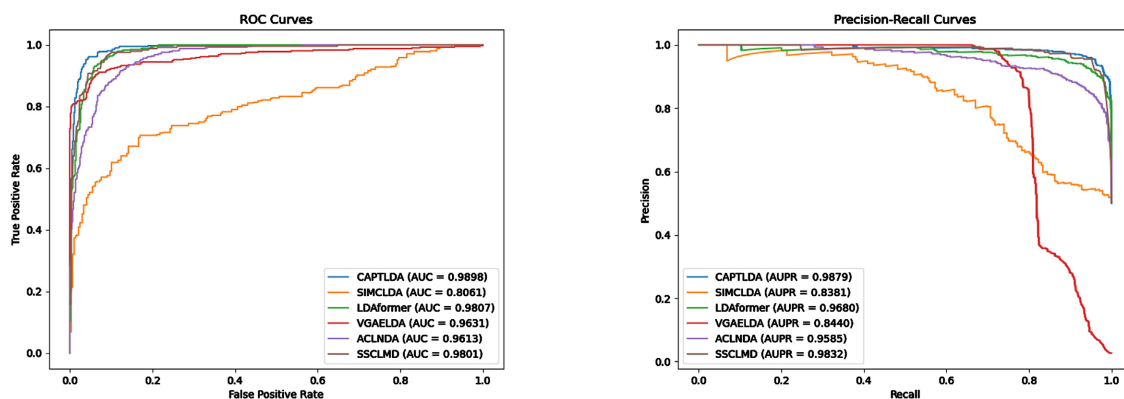


Figure 3. Comparison of ROC and PR curves between CAPTLDA and other baseline models on dataset 1

图 3. 在数据集一上 CAPTLDA 与其他基线模型的 ROC 和 PR 曲线对比

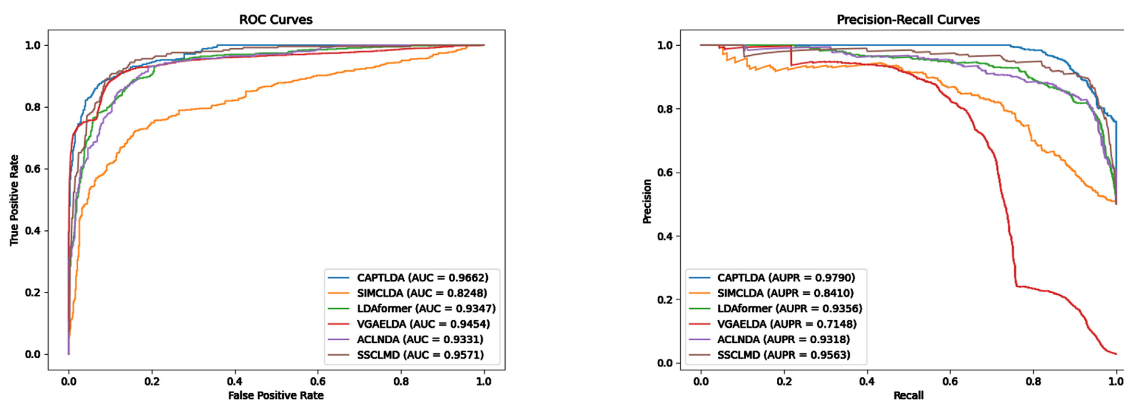


Figure 4. Comparison of ROC and PR curves between CAPTLDA and other baseline models on dataset 2

图 4. 在数据集二上 CAPTLDA 与其他基线模型的 ROC 和 PR 曲线对比

如图 3、图 4 所示, 相比于其他五个模型, CAPTLDA 在 AUC 和 AUPR 均达到了最优值, 在数据集一中, CAPTLDA 的 AUC 值为 0.9898, 比 SIMCLDA、LDAformer、VGAELDA、ACLNDA、SSCLMD 分别高出 18.37%、0.91%、2.67%、2.85%、0.97%; AUPR 值为 0.9879, 比 SIMCLDA、LDAformer、VGAELDA、ACLNDA、SSCLMD 分别高出 14.98%、1.99%、14.39%、2.94%、0.47%。在数据集二中, CAPTLDA 的 AUC 值为 0.9662, 比 SIMCLDA、LDAformer、VGAELDA、ACLNDA、SSCLMD 分别高出 14.14%、3.15%、2.08%、3.31%、0.91%; AUPR 的值为 0.9790, 比 SIMCLDA、LDAformer、VGAELDA、ACLNDA、SSCLMD 分别高出 13.8%、4.34%、26.42%、4.72%、2.27%。此外, 如图 5 所示, 在数据集一上, CAPTLDA 的 Accuracy、Recall 等五个评价指标的数值均高于其余五种模型, 在数据集二上,

CAPTLDA 的五种评价指标有四项均为六种模型中的最高值, 仅在 Precision 上略低于 SSCLMD, 这证明了 CAPTLDA 模型具有很强的预测 LDAs 的性能。

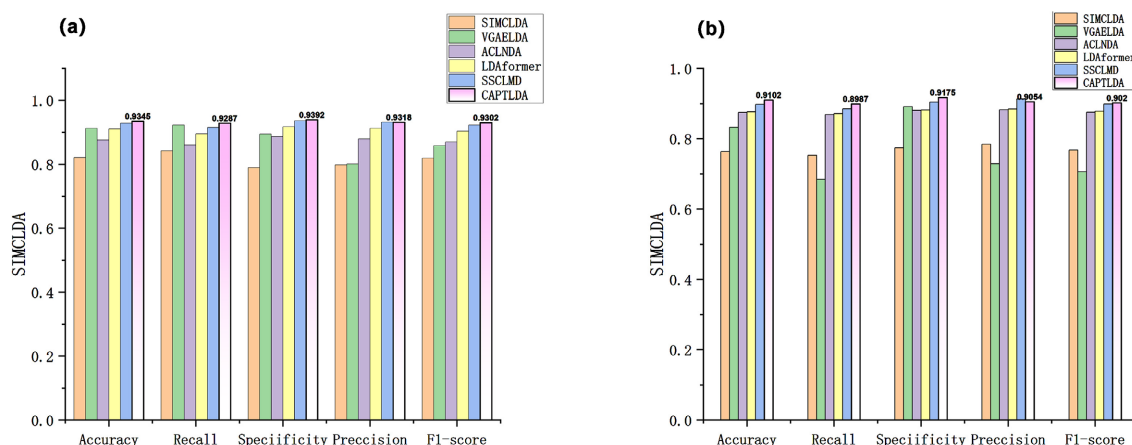


Figure 5. (a) Comparison of the accuracy, recall rate, precision, specificity and F1 value of the model in dataset 1; (b) Comparison of the accuracy, recall rate, precision, specificity and F1 value of the model in dataset 2

图 5. (a) 数据集一上模型的准确率, 召回率, 精度, 特异性和 F1 值的比较; (b) 数据集二上模型的准确率, 召回率, 精度, 特异性和 F1 值的比较

3.4. 消融实验

消融实验经常被用于在 LDAs 预测中验证模型各部分的存在有效性。为了验证胶囊网络的有效性, 以及将混合注意力机制集成到 Transformer 模型中的效果, 我们进行了消融研究。因此, CAPTLDA 与以下三个变体进行了比较, 分别是去掉了胶囊网络部分的变体 CAPTLDA_noCaps、只保留代理注意力机制的变体 CAPTLDA_AAo 以及采用原始 Transformer 自注意力机制的变体 CAPTLDA_SF。实验结果表明, CAPTLDA 在性能上优于其消融变体。如图 6、图 7 所示, 在两个数据集上, 所有三个变体的 AUC 和 AUPR 值均观察到明显下降, 这证实了每个模块都有效增强了模型的预测性能。此外, 变体 CAPTLDA_AAo 在两个数据集上的 AUC 和 AUPR 值均略高于 CAPTLDA_SF, 这表明在本研究中, 代理注意力机制优于自注意力机制。而与完整的 CAPTLDA 模型相比, 去掉胶囊网络的 CAPTLDA_noCaps 变体表现出较低的 AUC 和 AUPR 值。该观察结果证明了胶囊网络捕获特征的有效性以及突出了所提出架构在捕获多尺度特征方面的有效性。

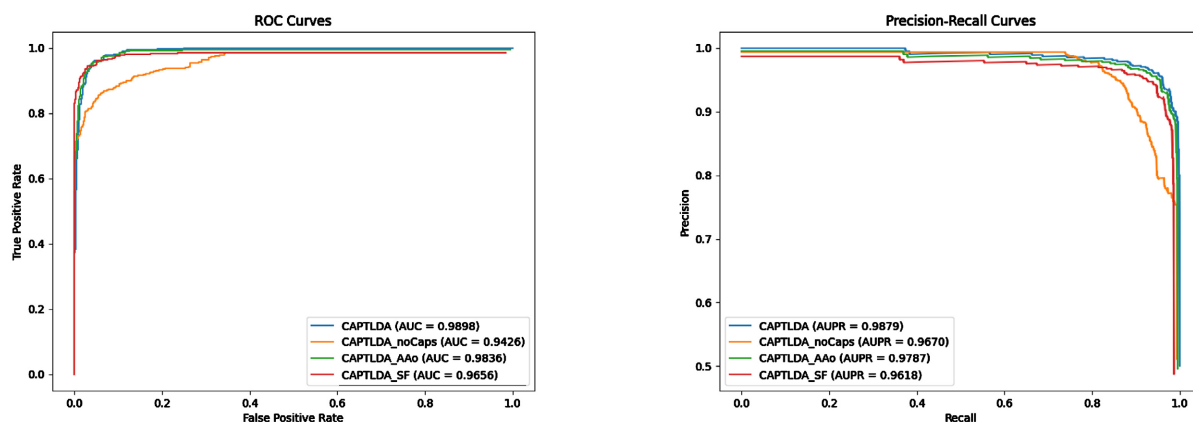


Figure 6. The comparison of ROC and PR curves between CAPTLDA and its three variants on dataset 1

图 6. CAPTLDA 与其三个变体在数据集一上的 ROC 和 PR 曲线对比

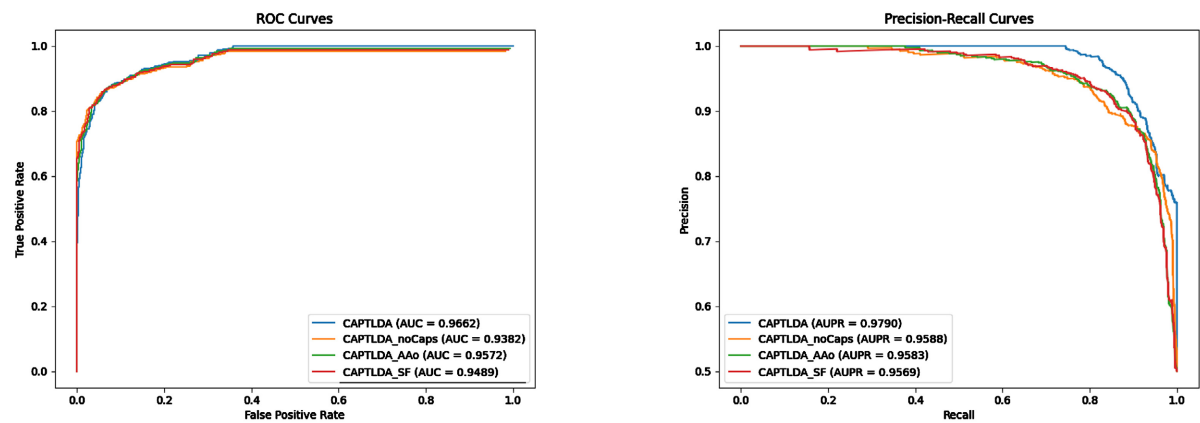


Figure 7. The comparison of ROC and PR curves between CAPTLDA and its three variants on dataset 2

图 7. CAPTLDA 与其三个变体在数据集二上的 ROC 和 PR 曲线对比

3.5. 案例分析

为了进一步验证 CAPTLDA 真实的预测能力,我们对数据集 1 和数据集 2 进行了案例分析研究。已知的 LDA 作为阳性样本,而相同数量的未知 LDA 被随机选择为阴性样本。测试样本根据模型生成的预测分数从高到低进行排序。对于每种疾病,记录排名前十的预测 lncRNA。最后,通过 PubMed 文献搜索来确定验证模型预测的 LDA 的真实性。对于数据集 1,选择了肝细胞癌(Hepatocellular carcinoma, HCC, DOID: 684),对于数据集 2,选择了乳腺癌(Breast Cancer, BC, DOID: 1612)。验证结果如表 2 所示。

Table 2. The ranking of the top ten lncRNAs related to HCC and BC as predicted by CAPTLDA in dataset 1 and dataset 2

表 2. 数据集一与数据集二中 CAPTLDA 预测与 HCC 和 BC 相关的前十个 lncRNA 排序

疾病	排名	LncRNA 名称	PubMedID
肝细胞癌	1	MIR155H	38850791
	2	DLX6-AS1	43039401
	3	CCDC26	Unknown
	4	UCA1	37854341
	5	ZEB1-AS1	33865414
	6	FEZF1-AS1	34023817
	7	LINC00176	28869604
	8	BANCR	34512168
	9	XIST	30091314
	10	MALAT1	32965597
乳腺癌	1	TP73-AS1	28857253
	2	CASC11	36204307
	3	MIR17HG	36943627
	4	DLEU1	31841195
	5	LINC-ROR	32335998

续表

乳腺癌	6	PVT1	36941464
	7	DGCR5	32521856
	8	LINC00261	33274565
	9	BANCR	29565494
	10	FOXD2-AS1	30628646

肝细胞癌已成为全球范围内的重大健康负担。作为一种异质性恶性肿瘤，其在全球范围内的患病率持续上升，发病率在所有癌症中排名第六[44]，是癌症相关死亡的第三大主要原因[45]。通过模型筛选出的所有十个肝细胞癌候选 lncRNA，有九个都得到了文献中的实验验证。例如，肝细胞癌病理标本中 lncRNA XIST 的表达明显下调。它通过吸附致癌 miRNA miR-155-5p 来抑制 HCC 的进展[46]。

乳腺癌是女性中常见的恶性肿瘤，是全球女性癌症相关死亡的主要原因之一[47]。模型预测的与乳腺癌相关的前十个 lncRNA 均已被现有文献证实。例如 PVT1 通过与乳腺癌细胞中的内源性 miR-145-5p 竞争性结合，重新编程代谢表型，从而调节糖酵解相关基因的表达[48]。

案例研究的结果表明，CCAPTLDA 能够有效地预测尚未被发掘的 LDA，为实验提供方向，减少不必要的损耗。

4. 总结与讨论

长链非编码 RNA 已被确定为疾病发病机制背后多种生物过程的关键调节因子。准确预测 LDA 可成为指导后续生物学实验的重要见解，并对推进疾病诊断和治疗的创新策略具有重要意义。本研究提出了一种新的 LDA 预测模型 CAPTLDA，整合了 lncRNA、miRNA、疾病相似性和类间关联。构建了一个加权邻接矩阵。随后，采用胶囊网络来捕获矩阵内的特征。最后，具有混合注意力机制的 Transformer 编码器整合特征路径之间的依赖关系，以预测潜在的 LDA。实验结果表明，CAPTLDA 优于其他先进的基线方法，能够准确有效地识别实际应用中的潜在 LDA。

研究同时具有部分局限性。所有实验都是使用平衡的正负样本进行的。丢弃一些负样本可能会遗漏潜在特征。因此，未来的工作可以探索模型在不平衡数据上的直接性能。其次，与现实相比，数据集只是所有 RNA 和疾病的一个子集。因此，未来的努力应该集中在提高数据集的可扩展性上。在特征构建方面，可以探索整合额外的生物关联，如 RNA 序列信息，以开发未来的模型，在预测准确性和计算效率之间取得最佳平衡。

参考文献

[1] Chen, X., Yan, C., Luo, C., Ji, W., Zhang, Y. and Dai, Q. (2015) Constructing lncRNA Functional Similarity Network Based on lncRNA-Disease Associations and Disease Semantic Similarity. *Scientific Reports*, 5, Article No. 12. <https://doi.org/10.1038/srep11338>

[2] Wapinski, O. and Chang, H.Y. (2011) Long Noncoding RNAs and Human Disease. *Trends in Cell Biology*, 21, 354-361. <https://doi.org/10.1016/j.tcb.2011.04.001>

[3] Chen, X., Yan, C.C., Zhang, X. and You, Z. (2016) Long Non-Coding RNAs and Complex Diseases: From Experimental Results to Computational Models. *Briefings in Bioinformatics*, 18, 558-576. <https://doi.org/10.1093/bib/bbw060>

[4] Wang, S.Y., Chen, Z.Y., Gu, J.Y., Chen, X. and Wang, Z.X. (2021) The Role of lncRNA PCAT6 in Cancers. *Frontiers in Oncology*, 11, Article ID: 701495. <https://doi.org/10.3389/fonc.2021.701495>

[5] Congrains, A., Kamide, K., Oguro, R., Yasuda, O., Miyata, K., Yamamoto, E., et al. (2012) Genetic Variants at the 9p21 Locus Contribute to Atherosclerosis through Modulation of ANRIL and CDKN2A/B. *Atherosclerosis*, 220, 449-455. <https://doi.org/10.1016/j.atherosclerosis.2011.11.017>

- [6] Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., *et al.* (2008) Expression of a Noncoding RNA Is Elevated in Alzheimer's Disease and Drives Rapid Feed-Forward Regulation of β -Secretase. *Nature Medicine*, **14**, 723-730. <https://doi.org/10.1038/nm1784>
- [7] Alvarez, M.L. and DiStefano, J.K. (2011) Functional Characterization of the Plasmacytoma Variant Translocation 1 Gene (PVT1) in Diabetic Nephropathy. *PLOS ONE*, **6**, e18671. <https://doi.org/10.1371/journal.pone.0018671>
- [8] Wu, D., Li, R.F., Liu, J.Y., Zhou, C.C. and Jia, R.P. (2022) Long Noncoding RNA LINC00467: Role in Various Human Cancers. *Frontiers in Genetics*, **13**, Article ID: 892009. <https://doi.org/10.3389/fgene.2022.892009>
- [9] Spagnolo, P., Kropski, J.A., Jones, M.G., Lee, J.S., Rossi, G., Karampitsakos, T., *et al.* (2021) Idiopathic Pulmonary Fibrosis: Disease Mechanisms and Drug Development. *Pharmacology & Therapeutics*, **222**, Article ID: 107798. <https://doi.org/10.1016/j.pharmthera.2020.107798>
- [10] Gavrillov, K. and Saltzman, W.M. (2012) Therapeutic siRNA: Principles, Challenges, and Strategies. *The Yale Journal of Biology and Medicine*, **85**, 187-200.
- [11] Sun, J., Shi, H.B., Wang, Z.Z., Zhang, C.J., Liu, L., Wang, L.T., *et al.* (2014) Inferring Novel LncRNA-Disease Associations Based on a Random Walk Model of a LncRNA Functional Similarity Network. *Molecular Biosystems*, **10**, 2074-2081. <https://doi.org/10.1039/c3mb70608g>
- [12] Chen, X. (2015) KATZLDA: KATZ Measure for the LncRNA-Disease Association Prediction. *Scientific Reports*, **5**, Article No. 16840. <https://doi.org/10.1038/srep16840>
- [13] Ganegoda, G.U., Li, M., Wang, W. and Feng, Q. (2015) Heterogeneous Network Model to Infer Human Disease-Long Intergenic Non-Coding RNA Associations. *IEEE Transactions on NanoBioscience*, **14**, 175-183. <https://doi.org/10.1109/tnb.2015.2391133>
- [14] Xi, J., Wang, M. and Li, A. (2017) Discovering Potential Driver Genes through an Integrated Model of Somatic Mutation Profiles and Gene Functional Information. *Molecular BioSystems*, **13**, 2135-2144. <https://doi.org/10.1039/c7mb00303j>
- [15] Lu, C., Yang, M., Luo, F., Wu, F., Li, M., Pan, Y., *et al.* (2018) Prediction of LncRNA-Disease Associations Based on Inductive Matrix Completion. *Bioinformatics*, **34**, 3357-3364. <https://doi.org/10.1093/bioinformatics/bty327>
- [16] Fan, Y.X., Chen, M.J. and Pan, X.Y. (2021) GCRFLDA: Scoring LncRNA-Disease Associations Using Graph Convolution Matrix Completion with Conditional Random Field. *Briefings in Bioinformatics*, **23**, bbab361. <https://doi.org/10.1093/bib/bbab361>
- [17] Fu, G., Wang, J., Domeniconi, C. and Yu, G. (2017) Matrix Factorization-Based Data Fusion for the Prediction of LncRNA-Disease Associations. *Bioinformatics*, **34**, 1529-1537. <https://doi.org/10.1093/bioinformatics/btx794>
- [18] Zhu, R., Wang, Y., Liu, J. and Dai, L. (2021) IPCARE: Improving LncRNA-Disease Association Prediction Using Incremental Principal Component Analysis Feature Selection and a Random Forest Classifier. *BMC Bioinformatics*, **22**, Article No. 175. <https://doi.org/10.1186/s12859-021-04104-9>
- [19] Xuan, P., Cao, Y., Zhang, T., Kong, R. and Zhang, Z. (2019) Dual Convolutional Neural Networks with Attention Mechanisms Based Method for Predicting Disease-Related LncRNA Genes. *Frontiers in Genetics*, **10**, Article No. 416. <https://doi.org/10.3389/fgene.2019.00416>
- [20] Xuan, P., Pan, S., Zhang, T., Liu, Y. and Sun, H. (2019) Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting LncRNA-Disease Associations. *Cells*, **8**, Article No. 1012. <https://doi.org/10.3390/cells8091012>
- [21] Wang, L. and Zhong, C. (2022) gGATLDA: LncRNA-Disease Association Prediction Based on Graph-Level Graph Attention Network. *BMC Bioinformatics*, **23**, Article No. 11. <https://doi.org/10.1186/s12859-021-04548-z>
- [22] Xuan, P., Zhan, L., Cui, H., Zhang, T., Nakaguchi, T. and Zhang, W. (2022) Graph Triple-Attention Network for Disease-Related LncRNA Prediction. *IEEE Journal of Biomedical and Health Informatics*, **26**, 2839-2849. <https://doi.org/10.1109/jbhi.2021.3130110>
- [23] Sabour, S., Frosst, N. and Hinton, G.E. (2017) Dynamic Routing between Capsules.
- [24] Vaswani, A., *et al.* (2017) Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.
- [25] Li, G.H., Bai, P.H., Liang, C. and Luo, J.W. (2024) Node-Adaptive Graph Transformer with Structural Encoding for Accurate and Robust LncRNA-Disease Association Prediction. *BMC Genomics*, **25**, Article No. 73. <https://doi.org/10.1186/s12864-024-09998-2>
- [26] Han, D., Ye, T., Han, Y., Xia, Z., Pan, S., Wan, P., *et al.* (2024) Agent Attention: On the Integration of Softmax and Linear Attention. In: Leonardis, A., *et al.*, Eds., *Computer Vision—ECCV 2024*, Springer, 124-140. https://doi.org/10.1007/978-3-031-72973-7_8
- [27] Luong, T., Pham, H. and Manning, C.D. (2015) Effective Approaches to Attention-Based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, September 2015, 1412-1421. <https://doi.org/10.18653/v1/d15-1166>

- [28] Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., *et al.* (2015) Lnc2Cancer: A Manually Curated Database of Experimentally Supported lncRNAs Associated with Various Human Cancers. *Nucleic Acids Research*, **44**, D980-D985. <https://doi.org/10.1093/nar/gkv1094>
- [29] Chen, G., Wang, Z.Y., Wang, D.Q., Qiu, C.X., Liu, M.X., *et al.* (2012) LncRNADisease: A Database for Long-Non-Coding RNA-Associated Diseases. *Nucleic Acids Research*, **41**, D983-D986. <https://doi.org/10.1093/nar/gks1099>
- [30] Lu, Z.Y., Cohen, K.B. and Hunter, L. (2007) GeneRIF Quality Assurance as Summary Revision. In: *13th Pacific Symposium on Biocomputing (PSB)*, World Scientific, 269.
- [31] Li, J., Liu, S., Zhou, H., Qu, L. and Yang, J. (2013) starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale Clip-Seq Data. *Nucleic Acids Research*, **42**, D92-D97. <https://doi.org/10.1093/nar/gkt1248>
- [32] Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., *et al.* (2013) HMDD V2.0: A Database for Experimentally Supported Human MicroRNA and Disease Associations. *Nucleic Acids Research*, **42**, D1070-D1074. <https://doi.org/10.1093/nar/gkt1023>
- [33] Zhou, Y., Wang, X., Yao, L. and Zhu, M. (2022) Ldaformer: Predicting LncRNA-Disease Associations Based on Topological Feature Extraction and Transformer Encoder. *Briefings in Bioinformatics*, **23**, bbac370. <https://doi.org/10.1093/bib/bbac370>
- [34] Gao, Y., Shang, S., Guo, S., Li, X., Zhou, H., Liu, H., *et al.* (2020) Lnc2cancer 3.0: An Updated Resource for Experimentally Supported lncRNA/circRNA Cancer Associations and Web Tools Based on RNA-Seq and ScRNA-Seq Data. *Nucleic Acids Research*, **49**, D1251-D1258. <https://doi.org/10.1093/nar/gkaa1006>
- [35] Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q. and Dong, D. (2018) LncRNADisease 2.0: An Updated Database of Long Non-Coding RNA-Associated Diseases. *Nucleic Acids Research*, **47**, D1034-D1037. <https://doi.org/10.1093/nar/gky905>
- [36] Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., *et al.* (2018) HMDD V3.0: A Database for Experimentally Supported Human MicroRNA-Disease Associations. *Nucleic Acids Research*, **47**, D1013-D1017. <https://doi.org/10.1093/nar/gky1010>
- [37] Wang, D., Wang, J., Lu, M., Song, F. and Cui, Q. (2010) Inferring the Human MicroRNA Functional Similarity and Functional Network Based on MicroRNA-Associated Diseases. *Bioinformatics*, **26**, 1644-1650. <https://doi.org/10.1093/bioinformatics/btq241>
- [38] Kozomara, A., Birgaoanu, M. and Griffiths-Jones, S. (2018) miRBase: From MicroRNA Sequences to Function. *Nucleic Acids Research*, **47**, D155-D162. <https://doi.org/10.1093/nar/gky1141>
- [39] Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., *et al.* (2013) Correction: Prediction of MicroRNAs Associated with Human Diseases Based on Weighted K Most Similar Neighbors. *PLOS ONE*, **8**, No. 9. <https://doi.org/10.1371/annotation/a076115e-dd8c-4da7-989d-c1174a8cd31e>
- [40] Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C. (2007) A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics*, **23**, 1274-1281. <https://doi.org/10.1093/bioinformatics/btm087>
- [41] Shi, Z., Zhang, H., Jin, C., Quan, X. and Yin, Y. (2021) A Representation Learning Model Based on Variational Inference and Graph Autoencoder for Predicting LncRNA-Disease Associations. *BMC Bioinformatics*, **22**, Article No. 136. <https://doi.org/10.1186/s12859-021-04073-z>
- [42] Fu, L.Y., Yao, Z.Y., Zhou, Y.Y., Peng, Q.K. and Lyu, H.Q. (2024) ACLNDA: An Asymmetric Graph Contrastive Learning Framework for Predicting Noncoding RNA-Disease Associations in Heterogeneous Graphs. *Briefings in Bioinformatics*, **25**, bbae533. <https://doi.org/10.1093/bib/bbae533>
- [43] Huang, L., Sheng, N., Gao, L., Wang, L., Hou, W., Hong, J., *et al.* (2025) Self-Supervised Contrastive Learning on Attribute and Topology Graphs for Predicting Relationships among lncRNAs, miRNAs and Diseases. *IEEE Journal of Biomedical and Health Informatics*, **29**, 657-668. <https://doi.org/10.1109/jbhi.2024.3467101>
- [44] Chen, Y., Zou, S., Xu, L., Chen, J., Wang, L., Shen, Y., *et al.* (2025) NSMCE2 Promotes the Occurrence and Development of HCC by Regulating the SUMOylation of PPAR α . *International Immunopharmacology*, **157**, Article ID: 114762. <https://doi.org/10.1016/j.intimp.2025.114762>
- [45] El-Serag, H.B. and Rudolph, K.L. (2007) Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. *Gastroenterology*, **132**, 2557-2576. <https://doi.org/10.1053/j.gastro.2007.04.061>
- [46] Lin, X.Q., Huang, Z.M., Chen, X., Wu, F. and Wu, W. (2018) XIST Induced by JPX Suppresses Hepatocellular Carcinoma by Sponging miR-155-5p. *Yonsei Medical Journal*, **59**, 816-826. <https://doi.org/10.3349/ymj.2018.59.7.816>
- [47] Nolan, E., Lindeman, G.J. and Visvader, J.E. (2023) Deciphering Breast Cancer: From Biology to the Clinic. *Cell*, **186**, 1708-1728. <https://doi.org/10.1016/j.cell.2023.01.040>
- [48] Qu, H., Li, X., Chen, F., Zhang, M., Lu, X., Gu, Y., *et al.* (2023) LncRNA PVT1 Influences Breast Cancer Cells Glycolysis through Sponging miR-145-5p. *Genes & Genomics*, **45**, 581-592. <https://doi.org/10.1007/s13258-023-01368-8>