# 通过circRNA的多源特性来预测circRNA亚细胞定位

# 摘要

环状核糖核酸（circRNA）是新型的非编码RNA，尽管在人类的基于中存在大量的circRNA并且参与多种人类生理过程，但是大多数的circRNA的功能仍旧尚未被发掘。迄今为止，wet-lab experiments方法被用于检测circRNA的亚细胞位置，然而通过wet-lab experiments方法检测circRNA的亚细胞位置很昂贵，考虑到近年来深度学习的广泛应用以及其在其它非编码RNA定位预测方面取得的进展，本研究开发了一种预测circRNA亚细胞定位的新型计算方法，称为CircLoc。由于许多的circRNAs具有多个亚细胞定位，因此CircLoc本质上是一种多标签分类器。本研究中采用了circRNA的几个特性，例如circRNA序列、circRNA生物特征、circRNA-disease、circRNA-drug、circRNA-miRNA关联网络生成的信息丰富的circRNA特征，并且采用了强大的node2vec算法和图注意力自动编码器（GATE），产生了8种特征类型。这些特征被作为输入送入到自注意力和全连接层中进行预测，并采用交叉验证来验证结果，结果表明CircLoc的准确率高于0.8，AU-ROC超过0.77，AUPR超过0.42。进一步的测试证明，所有使用的特征类型都可以提高CircLoc的性能，而 GATE 和自注意力层有助于提高性能。

**关键词：**circRNA；亚细胞定位；node2vec；图注意力；circRNA-disease关联、circRNA-drug关联；circRNA-miRNA关联；GATE

# Introduction

非编码核糖核酸（ncRNAs）现在被认为是十分重要功能，它具有调节神经的作用[[1](#_ENREF_1)]而并非垃圾序列。研究表明，ncRNAs在细胞分化、基因组印记[[2](#_ENREF_2)]、基因表达调控等[[3](#_ENREF_3)]多种生物学过程中发挥关键作用。根据序列特征、结构特征，ncRNAs可以被分类多个亚类，Figure 1表示了ncRNAs的广泛分类。其中环状核糖核酸（circRNA）作为长链非编码RNA（lncRNA）的一种，由于它在细胞过程中的作用，以及作为miRNA海绵的能力等功能，circRNA已经成为RNA研究领域的重要成员[[4](#_ENREF_4)]。

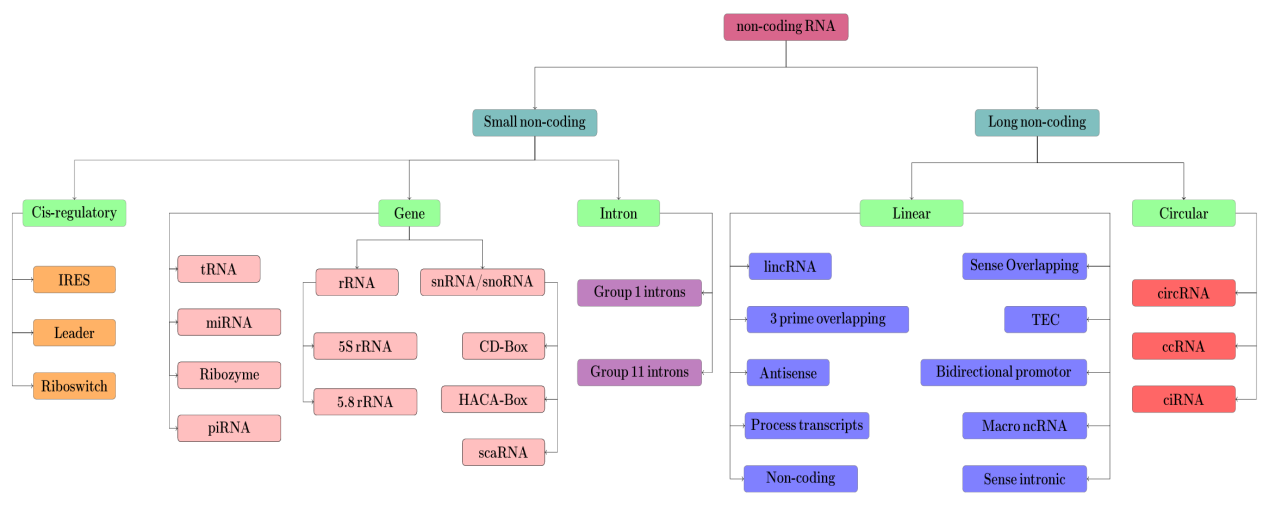


Figure 1．非编码RNA的分层分类

尽管早在60年前circRNAs存在哺乳动物、植物类病毒中就以及被报道了[[5](#_ENREF_5)]，但是它长期被看作RNA剪接过程中产生的低丰富度的副产物[[6](#_ENREF_6), [7](#_ENREF_7)]，且被普遍认为其在生物学关键过程中缺乏关键性的贡献。因此，截至2010年，仅有少量新型circRNA被发现[[8](#_ENREF_8)]。然而随着生物医学技术的革新和高通量RNA测序的发展，近年来的突破性研究彻底重塑了学界对于circRNA生物学价值的认知[[9](#_ENREF_9), [10](#_ENREF_10)]。

研究证明，circRNA可以通过多重分子机制发挥调控作用：1)作为RNA结合蛋白(RBP)与miRNA分子海绵通过竞争性结合抑制靶向分子的生物学效应[[11-13](#_ENREF_11)]。2)通过调节亲本基因表达、介导可变剪接及转录调控参与基因表达网络的重编程。部分circRNA甚至具备翻译潜能，可以生成功能性多肽或蛋白质[[11-13](#_ENREF_11)]，值得注意的是，circRNA的异常表达与多种病理过程密切相关，包括肿瘤发生(如结直肠癌特异性表达谱[[14](#_ENREF_14)])、动脉粥样硬化性血管病变及神经系统疾病[[15](#_ENREF_15)]，其作为胃癌诊断标志物[[16](#_ENREF_16)]、衰老相关分子标记及唾液疾病检测靶点的潜力也逐步显现[[16](#_ENREF_16)]。这些发现凸显circRNA在发育生物学、疾病起始与演进中的独特调控地位，并为其转化为临床诊断工具与治疗靶点提供了理论依据[[11](#_ENREF_11)]。

由于circRNA亚细胞定位预测方法十分少，传统的蛋白质定位和miRNA定位预测的方法可以为我们提供思路，现在对于蛋白质预测有四类方法：基于序列特征的传统机器学习方法、基于进化信息的预测方法、基于深度学习的端到端预测方法和多标签定位预测。例如Li等人使用GNN预测细菌效应蛋白定位[[17](#_ENREF_17)]。而miRNA亚细胞定位预测分为三类：基于序列特征的传统机器学习方法，基于深度学习的序列建模方法和基于网络和多源数据融合的方法，Xu等人使用基于mRNA定位推断miRNA定位[[18](#_ENREF_18)]，Bai等人结合序列和miRNA-疾病关联数据预测miRNA定位[[19](#_ENREF_19)]。

上述方面在各自的领域表现良好，鉴于circRNA与miRNA、疾病、药物的高度关联性，以及其自身学列所携带的信息，我们整合这些信息将有助于构建更强大的预测模型。本研究开发了新型多标签分类器CircLoc（基于多源特征的circRNA亚细胞定位预测器），融合circRNA序列、相似性、疾病关联、药物关联及miRNA关联（含miRNA亚细胞定位）等多维特性。通过node2vec、图注意力自编码器（GATE）等先进算法及新设计特征处理方案，生成8类信息丰富的circRNA特征，最终经自注意力层和全连接层进行预测。交叉验证显示，CircLoc在七个亚细胞定位上的准确率超0.8，AU-ROC超过0.77，AUPR超过0.42。进一步实验证实所有特征类型均对预测有积极贡献，GATE和自注意力层能有效提升性能。

# Materials and methods

## dataset

定义明确的数据集对于构建高效的分类器非常重要，在这项研究中，我们使用人类circRNA亚细胞定位数据集，该数据集是从RNALocate v3.0和circBase中获得的，原始数据集中包含134185条人类circRNA定位信息，它们属于16个不同的亚细胞位置，其中包含7类亚细胞定位的数据量仅仅15个样本，样本数量非常少，并且在预测circRNA的亚细胞位置时会混淆分类器，因此我们丢弃了这些序列，而剩余的9类亚细胞定位中的2类样本数量共131998，使得数据集极度不平衡，因此我们丢弃了这些序列，剩余的7类定位共有1486个样本数量，其亚细胞定位分别是：Chromatin，Cytoplasm，Cytosol，Membrane，Nucleolus，Nucleus，Nucleoplasm。其中定位Cytoplasm的样本数量是622条，Cytosol的样本数量是429条，Nucleus的样本数量是416，Membrane的样本数量是164，Nucleolus的样本数量是63，Nucleoplasm的样本数量是47，Chromatin的样本数量是46条。所以，在数学表达式中具有单一亚细胞定位的circRNA可以被定义为独立子集用，，，，，，表示。可以用以下公式表示基准数据集：

可以看出一些circRNA可以具有多个亚细胞定位，为了阐述这一情况，我们绘制了一张反转图Figure 2。因此，当亚细胞定位被视为标记而 circRNA 被视为样本时，将 circRNA分类为亚细胞定位是一个多标签分类问题。

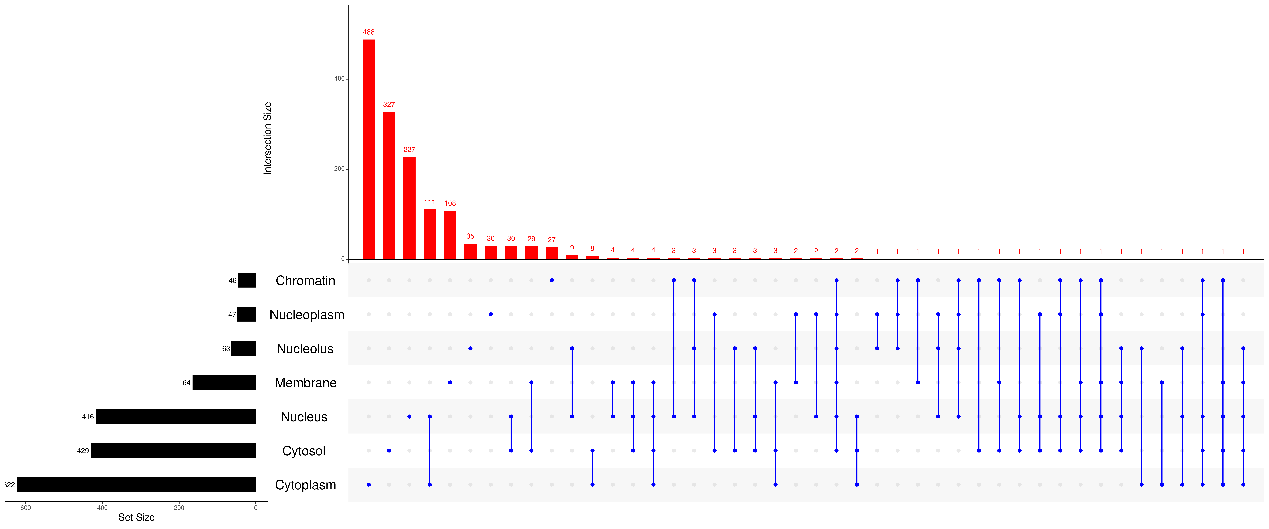


Figure 2. Upset graph to show the intersections of human circRNAs in seven subcellular localizations. Lots of circRNAs have multiple subcellular localizations.

## circRNA 的多源特性

最近，使用对象的多个属性来构建高效的分类器非常流行。迄今为止，已经在一些公共数据库中发现并收集了几种circRNA特性，从而可以利用这些特性预测 circRNA亚细胞定位。本研究采用了circRNA的多种特性，介绍如下。

### 序列编码方法

序列编码方法的核心目标是将基因组或蛋白质组序列转化为具有丰富数学表征的形式，从而有效揭示其与目标特性的关键关联[[20](#_ENREF_20)]。给定一个生物序列S，其原始线性表达可表示为：

(1)

其中 R 代表残基，下标表示该残基在长度为 L 的序列中的位置。由于所有算法都需固定长度的特征向量作为输入[[20](#_ENREF_20)]，因此我们必须将表(1)中给出的序列表达式转换为实值向量。考虑到生物序列的特异性和残基分布规律，研究者开发了多种序列描述符，用于将序列的生物学特征编码为统计特征向量[[20](#_ENREF_20)]。统计表征学习是序列分析流程中的关键环节。优秀的描述符应同时考虑残基频率和语义信息，以生成具有强判别性的特征向量。反之，过于简化的表征方式可能导致不同类别序列的特征相似度过高，严重影响分类器的泛化能力。由于序列描述符的性能取决于序列特性及残基分布模式[[20](#_ENREF_20)]，其在不同类型序列中的表现存在显著差异。例如适用miRNA[[21](#_ENREF_21)]或piRNA[[22](#_ENREF_22)]等短序列(17-25个残基)的描述符，可能无法有效捕捉IncRNA或circRNA等长序列的特征，对于长序列，需要能够捕获残基长程依赖关系的描述符。在本研究中，我们使用了两种不同的序列描述符，这些序列描述符被广泛用于编码特征上与 circRNA 序列相似的 ncRNA 序列[[23](#_ENREF_23)]。本研究中使用的不同序列描述符将在以下小结中简要描述。

#### K-mer

在生物序列编码方法中，K-mer算法因其简洁高效的特点，已成为DNA、RNA和蛋白质序列表征的基础技术手段[[24](#_ENREF_24)]。如图Figure 2所示，当窗口大小设为3且步长为1时，K-mer算法通过滑动窗口机制对序列进行特征提取。该技术已成功应用于增强子识别、人类基因调控序列预测[[25](#_ENREF_25), [26](#_ENREF_26)]以及调控序列特征分析[[27](#_ENREF_27)]等多个基因组学研究领域。针对circRNA序列分析，我们采用K-mer算法进行特征编码。考虑到circRNA由A、U、C、G四种碱基组成，当K取值为4时，理论上可产生4K种不同的特征组合。对于长度为m的circRNA序列，其K-mer子序列数量可通过公式m-k+1精确计算。该算法通过统计K邻接残基的出现频率，为每个circRNA序列构建对应的特征向量[[24](#_ENREF_24)]，具体实现过程包含以下关键步骤：(1)序列遍历：采用滑动窗口技术系统扫描整个circRNA序列。(2)特征提取：记录所有可能的K-mer组合及其出现频次。(3)向量构建：根据预设的K值生成固定维度的特征表示。



Figure 3．Process of Generating Sequence K-mers (e.g., 3-mers), Where each Particular Color Frame Denotes a Unique 3-mer.

#### Reverse Complimen K-mer

反向互补K-mer编码作为传统K-mer方法的改进形式，已被广泛应用于DNA调控区预测[[28](#_ENREF_28)]、DNA N4-甲基胞嘧啶位点识别[[29](#_ENREF_29)]以及RNA亚细胞定位分析[[30](#_ENREF_30)]等研究领域。该方法通过以下两个关键步骤实现序列特征优化：(1)序列转换处理：首先对RNA序列进行反向排列，随后按照碱基互补原则（A↔U，C↔G）进行替换，最终生成具有生物学意义的反向互补序列[[25](#_ENREF_25)]。(2)特征降维优化：以4种标准碱基（A、G、C、U）为例，当K=2时，传统K-mer方法产生16种组合，通过去除反向互补重复特征（如AC与GU），精简为10个独特特征(AA, AC, AG, AU, CA, CC, CG, GA, GC, UA)。本研究采用系统性实验策略，通过生成2至5-mer的多尺寸特征，全面考察不同K值对circRNA序列特征的捕获能力。

### RNAErnie model

近期，预训练语言模型在分析核苷酸序列方面展现出巨大潜力，特别是在RNA序列的结构与功能研究中日益受到关注。在本研究中，我们引入了RNAErnie[[31](#_ENREF_31)]模型作为特征提取模块的重要组成部分。RNAErnie是一种基于Transformer架构的RNA专用预训练语言模型，其设计融合了两项关键创新策略。首先，在预训练阶段，RNAErnie引入了“多尺度掩码策略”，即在碱基级别、子序列级别以及生物学上具有重要意义的motif级别上同时进行掩码操作，显著增强了模型对RNA局部结构与功能元素的捕捉能力。与传统的token掩码方式不同，该策略借助motif数据库对RNA结构信息进行有效建模，从而提高了模型在结构预测与序列分类任务中的表现。其次，在微调阶段，RNAErnie提出了“类型引导微调架构”（type-guided fine-tuning），通过预先预测RNA类型并将其作为附加输入，以引导模型在特征空间中更好地适配特定类型的RNA。这种方法尤其适用于目标任务与预训练分布存在差异的场景，具备良好的泛化能力与迁移性能。

综合以上优势，RNAErnie不仅在多项下游任务（如RNA序列分类、RNA-RNA相互作用预测、RNA二级结构预测等）中取得了优异的性能表现，而且其统一的预训练基础架构为多任务RNA分析提供了支撑。在本研究中，我们充分利用了RNAErnie模型对circRNA多源特征的建模能力，将其嵌入到整体分析框架中，以提升特征表达的质量与下游预测模型的准确性

### circRNA sequences and similarity network

上文使用circRNA的序列提取生物信息，对于构建图神经网络circRNA序列也起着十分重要的作用。对于从RNALocate v3.0和circBase获得的circRNA序列采用SmithWaterman算法[[32](#_ENREF_32)]来测量两个circRNA的序列相似性，对于circRNA 和，它们的序列相似性可以通过以下方式计算：

(2)

基于公式(2)，我们构建了一个circRNA序列相似性网络，用表示，1486个circRNA作为图节点，而且当且仅当它们的序列相似性分数大于0时两个circRNA节点才相连接。此外，相连节点的circRNA的序列相似性分数作为边的权重。为方便起见，这个网络用表示。

### circRNA-disease association network

近年来，越来越多的研究表明circRNA在多种疾病的发生与发展过程中具有重要作用，其与多种人类疾病之间存在显著的相关性[[33-36](#_ENREF_33)]。在疾病关联研究中，circRNA的功能性常被视为其在特定疾病背景下的调控作用，因此circRNA–disease的关联信息对于理解其生物学功能具有重要价值。本研究基于circRAN Disease[37]中收集了已知的circRNA-disease关联的数据集，该数据集中涵盖了337个circRNA与177种疾病之间的实验验证关系。结合前文中构建的circRNA相似性网络(网络)，我们进一步构建了一个circRNA–disease异构关联网络，用于综合建模circRNA的结构相似性与功能关联性。该网络包含1486个circRNA节点与177个疾病节点。circRNA节点之间的边按照网络中定义的相似性度量建立；同时，基于circR2Disease中提供的337条circRNA–disease配对信息，我们为相应的circRNA节点与疾病节点之间添加了关联边。为便于描述，该网络在后续分析中记为，其中“”表示circRNA节点，“”表示疾病节点。

### circRNA-drug association network

与疾病类似，已有研究证实circRNA在多种药物反应机制中也发挥着关键作用。例如，circ-PVT1 被发现可促进胃癌细胞对紫杉醇的耐药性[[37](#_ENREF_37)]，而circCELSR1 的高表达则显著降低卵巢癌细胞对紫杉醇的敏感性 [[38](#_ENREF_38)]。这些发现提示，circRNA与药物之间的关联关系不仅揭示了其潜在的生物学功能，也可能为预测其亚细胞定位提供有力支持。基于上述，本研究引入circRNA–药物敏感性信息以辅助构建亚细胞定位分类器。具体而言，我们从[[39](#_ENREF_39)]中获取了circRNA–药物敏感性关联数据，并进一步筛选出与本研究基准数据集（即具有定位标签的circRNA）相关的条目。经过筛选，我们最终获得了包含64个circRNA与217种药物之间的有效关联信息。

在此基础上，我们构建了circRNA–药物异构关联网络，其中circRNA节点之间的边仍然依据网络中的相似性原则建立；同时，64个circRNA与217个药物之间的已知关联关系则决定了是否在节点间添加边。为了便于后续描述，该网络在本文中记为，其中“”表示circRNA节点，“”表示药物节点。

### circRNA-miRNA association network and subcellular localizations of miRNAs

miRNA（microRNA）是一类研究较为深入的非编码RNA，其亚细胞定位信息已被系统性地揭示。近年来，circRNA与miRNA之间的特异性调控关系也逐渐被研究所揭示。已有研究表明，circRNA可通过吸附miRNA的方式调节其对靶基因的调控功能，即充当“miRNA海绵”，从而阻断miRNA对下游靶基因的抑制作用[[40](#_ENREF_40)]。这一机制进一步揭示了circRNA与miRNA在功能层面上的紧密关联[[41-43](#_ENREF_41)]。

基于此生物学背景，本研究引入circRNA–miRNA之间的相互作用信息以及miRNA的亚细胞定位数据，以此辅助构建更具生物学解释力的circRNA亚细胞定位分类器。我们从circBank数据库中提取了circRNA–miRNA关联信息，并筛选出与本研究数据集中的circRNA存在交集的记录，最终获得包含620个circRNA和1781个miRNA的有效关联对。

随后，我们构建了circRNA–miRNA异构关联网络。该网络中，circRNA节点之间的边依然依照前述网络的方法基于序列相似性建立；miRNA节点则作为新的节点类型加入网络，circRNA与miRNA之间的边由两者的关联关系决定。为便于后续叙述，我们将该网络记为，其中“”代表circRNA节点，“”代表miRNA节点。

除了网络拓扑结构外，我们还引入了miRNA的亚细胞定位信息，作为分类器输入特征的一部分，用以进一步提升预测性能。1781个miRNA的定位信息来自RNALocate数据库，涵盖七种亚细胞定位类别：Cytoplasm、Exomere、Extracellular exosome、Extracellular vesicle、Microvesicle、Mitochondrion 以及 Supermere。每类定位均包含不少于50个miRNA，部分miRNA具有多个亚细胞定位，形成了具有重叠关系的子集结构。Figure 4展示了这七种亚细胞定位之间的交集情况，揭示了miRNA在细胞不同亚结构中的多重分布特性。

综上所述，在本节中我们综合利用了多种类型的circRNA特征信息来构建分类模型，具体包括三个异构关联网络（Table 1）、circRNA序列相似性网络、预训练模型RNAErnie、基于K-mer和RCK-mer的序列特征表达，以及miRNA的亚细胞定位信息。该多模态信息融合策略有望进一步提升circRNA亚细胞定位的预测准确性和模型的泛化能力。

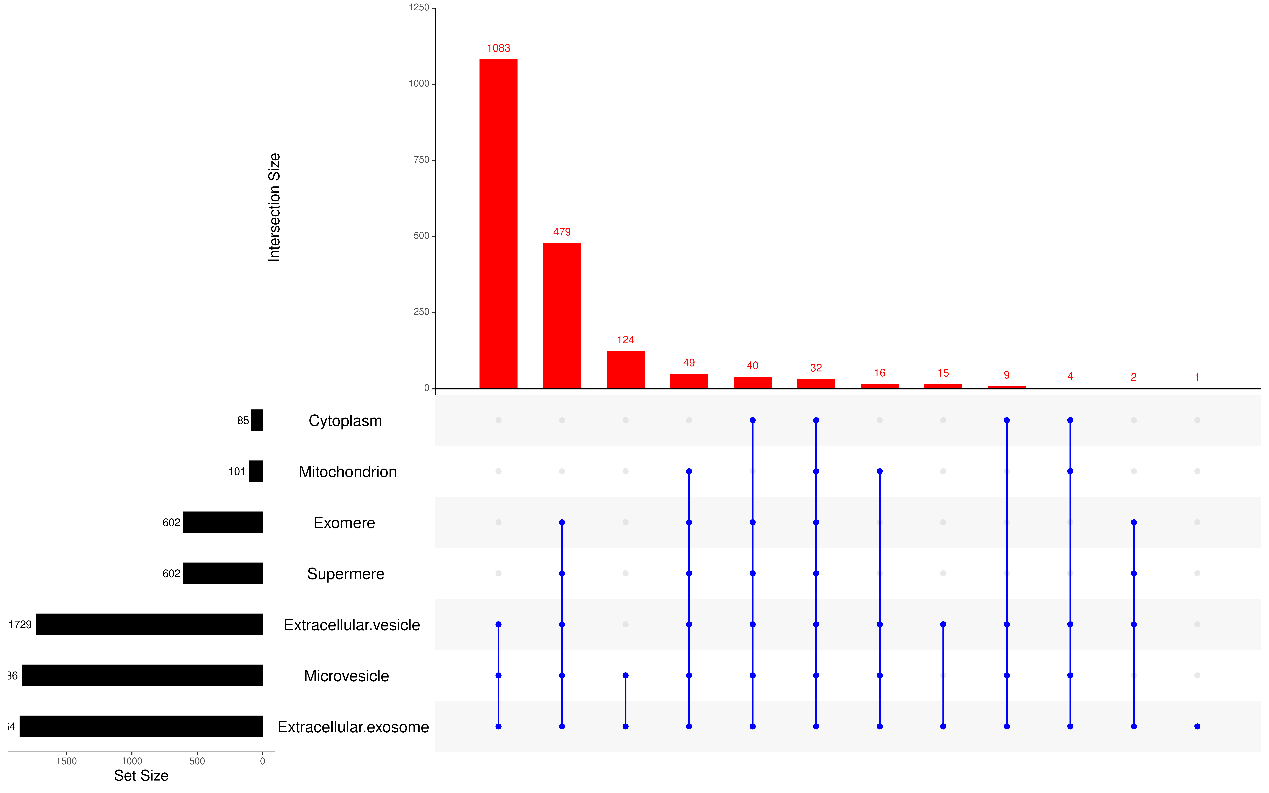


Figure 4. miRNA定位信息

|  |  |  |  |
| --- | --- | --- | --- |
| Association network | Number of circRNAs | Number of other objects | Number of associations |
| circRNA-disease | 337 | 166 | 1006 |
| circRNA-durg | 64 | 217 | 850 |
| circRNA-miRNA | 620 | 1781 | 6893 |

Table 1. Overview of three association networks for circRNAs in dataset S.

## circRNA多源特性特征

circRNA的几个特性在 circRNA的多源特性部分介绍。如何将它们完美地融合到分类器中是一个具有挑战性的问题。本研究中采用了下面这些现有或新设计的方法来解决这些特性。在此方法的帮助下获得了信息丰富的circRNA特征。

### Node2vec生成的特征

节点嵌入特征已成为近年来处理网络类数据问题的一种主流方法。本研究中使用的三类miRNA关联网络、miRNA序列相似性网络及功能相似性网络均属于此类非欧几里得结构数据。为有效处理此类数据，近年来提出了多种高效的图结构建模方法，其中，网络嵌入算法是一类核心技术，它能够将图中每一个节点映射为低维数值向量，从而便于后续的特征融合与模型训练。

在众多网络嵌入方法中，node2vec是目前应用最广泛的算法之一[[44](#_ENREF_44)]。该方法通过灵活的随机游走策略在保持网络结构信息的同时提升嵌入的表达能力.在本研究中，我们使用node2vec对多个circRNA网络进行编码，从而提取circRNA的结构化特征信息。具体而言，我们从circRNA序列相似性网络()中提取了每个circRNA的64维嵌入向量；同时，在circRNA–disease、miRNA–drug和circRNA–miRNA三类关联网络中，分别为每个circRNA生成了128维的嵌入表示。该多源嵌入特征不仅保留了不同网络结构中的局部与全局信息，也为后续模型提供了更具判别力的表示基础。

### Features yielded by GATE

除了传统的网络嵌入算法外，图卷积网络(Graph Convolutional Network, GCN)[47]也是当前处理网络数据的重要方法之一。GCN能够综合节点的原始表示与其邻居节点的信息，学习出更具表达力的节点嵌入表示。在本研究中，我们采用了一种特殊的GCN变体——GATE(Graph Auto-Encoder)[[45](#_ENREF_45)]，用于提取miRNA的高级语义特征。

GATE模型包含编码器（encoder）与解码器（decoder）两个核心模块。其中，编码器在聚合节点邻居信息的基础上更新节点的表示；而解码器则试图尽可能准确地重建每个节点的原始表示，从而实现对网络结构和节点特性的双重建模。

在本研究中，我们将通过node2vec从circRNA–disease、circRNA–drug以及circRNA–miRNA三类关联网络中获得的原始circRNA表示作为输入，进一步输入至GATE模型以获得更高层次的特征表达。同时，我们还将circRNA序列相似性网络())引入GATE中，以进一步增强模型的语义理解能力。

最终，GATE模型为每个circRNA节点生成了三组128维的高层特征向量，分别对应于circRNA–disease、circRNA–drug以及circRNA–miRNA三种关联网络。这些高级嵌入特征在保留原始网络结构信息的基础上，进一步提升了特征的表达能力与判别性能，为后续的分类任务提供了坚实的数据基础。

### Features derived from circRNA–miRNA association network and subcellular localizations of miRNAs

在Susovan Sadhukhan等人的研究中，他们通过“海绵”机制验证了circRNA与miRNA在亚细胞上存在共同定位。在这里，我们设计了一个方案从circRNA-miRNA关联和miRNA亚细胞定位中提取基本特征。

给定一个，它的相关miRNA可以从circRNA-miRNA关联网络中提取，即该网络中的邻居，他们构成miRNA集合，使用表示。如上述circRNA-miRNA关联网络和miRNA的亚细胞定位部分所述，这些miRNA所属七个亚细胞定位：Cytoplasm，Exomere，Extracellular exosome，Extracellular vesicle，Microvesicle，Mitochondrion，Supermere。分别标记为：，，，，，，。其中包含了中具有亚细胞定位在Cytoplasm的miRNA，其余的集合定义方式与它类似。但是如果一个集合的miRNA数量远远大于其他的集合，则具有相应的亚细胞定位概率也会对应上升，所以直接使用上述的集合并不是一种完美的方式，在本研究中我们将上述集合进一步处理，将结果限制在0-1之间，这会更加适合表示，我们按照下面方法处理集合：

(3)

通过上述处理，我们将ciRNA亚细胞定位信息转换为归一化向量表示，有效增强了circRNA亚细胞定位预测过程中的上下文生物学解释性。

最终，每个circRNA的特征表示融合了八种类型的多源信息，具体包括：第一种特征类型是通过node2vec从circRNA序列相似性网络中获得的，第二种特征是通过K-mer方法从circRNA序列中获得的，第三种特征是通过反向互补K-mer方法从circRNA序列中获得的，并且我们将第二种和第三种特征通过GATE和circRNA详细网络进一步处理。第四种特征是通过预训练的语言模型RNAErnie模型获得，第五种特征通过node2vec从circRNA-disease关联网络种获得，并且通过GATE和circRNA相似网络进一步加工。第六种和第七种特征类型与第五种类似，分别来自于circRNA-drug和circRAN-miRNA关联网络。为了获取更多信息特征，还采用了miRNA亚细胞定位的信息来生成这种特征。

### CircLoc概述

在本研究中，我们提出了一种基于多源特征融合的多标签分类模型，命名为 CircLoc，用于预测circRNA的亚细胞定位信息。模型的整体构建流程如Figure xxx所示。正如“circRNA多源特性”部分所述，我们综合利用了多种circRNA相关信息，包括circRNA的序列相似性网络、序列本身的特征，以及其与疾病、药物和miRNA的关联网络。

具体而言，我们从circRNA序列中提取了两类序列特征：一类是通过K-mer与反向K-mer方法获得的统计特征，另一类是借助预训练语言模型RNAErnie提取的深层语义特征；此外，我们还利用circRNA–disease、circRNA–drug、circRNA–miRNA三类异构网络中节点之间的拓扑关系提取了网络结构特征。综合上述不同来源的信息，我们共生成了8种不同类型的特征用于表征每一个circRNA。所有这些特征最终被拼接成一个统一的高维特征向量，作为模型的输入。

CircLoc模型由两部分组成：自注意力机制层 与 全连接神经网络层。在自注意力层中，模型通过自注意力机制学习各类特征之间的相对重要性，从而捕捉特征之间的深层依赖关系，增强表示能力。随后，更新后的特征向量被输入至包含三个ReLU激活的隐藏层与一个Sigmoid输出层的全连接网络中进行分类预测。损失函数采用二元交叉熵（binary cross-entropy），模型参数由Adam优化器进行训练与更新[[46](#_ENREF_46)]。整个多标签分类器基于TensorFlow 2.11.0 与 Scikit-learn[[47](#_ENREF_47)]实现。

需要指出的是，实验所用数据集中存在类别分布不均的情况。以亚细胞定位为例，Cytoplasm类别下共有622个circRNA样本，而Chromatin类别下仅有46个样本，类别间显著的不平衡可能导致模型倾向于预测样本数量较多的类别。为缓解该问题，我们引入了类别加权策略：对每一类样本赋予相应权重，样本数量越多，其对应的权重越小；反之，样本较少的类别权重则相对较大。该策略旨在促使模型关注难以预测的低频类别，从而提升整体分类性能与少数类的识别能力。

Figure xxx

### 评估指标

交叉验证（Cross-validation）是一种广泛应用于分类器性能评估的标准方法[55]。在本研究中，我们采用10折交叉验证（10-fold cross-validation）对所提出的分类器进行系统评估。交叉验证的结果通过多种评价指标进行量化，包括、accuracy、recall和Precision估指标。

除上述总体性能指标外，我们还针对每个亚细胞定位标签绘制了ROC曲线（Receiver Operating Characteristic Curve）和PR曲线（Precision–Recall Curve），并计算其对应的面积指标，即AUC（Area Under the ROC Curve）和AUPR（Area Under the PR Curve）。这些曲线和指标可以更细致地反映模型在不同标签上的分类性能，尤其是在数据不平衡的背景下更具解释力。

# Result and discussion Performance of CircLoc

在10折交叉验证对CircLoc进行评估，结果显示，其accuracy，F1，recall和precision分别是0.807，0.480，0.645，0.464，此外，我们统计了CircLoc在其中亚细胞定位中的性能，以AUC和AUPR衡量，其ROC和PR曲线如Figure 5和Figure 6所示，在Chromatin，Nucleoplasm，Nucleolus，Membrane，Nucleus，Cytosol，Cytoplasm上的AUC值分别是0.750，0.790，0.807，0.798，0.728，0.727，0.824。绝大部分的AUC值均高于0.75，这表明CircLoc在这些亚细胞定位中表现良好，而在Chromatin，Nucleoplasm，Nucleolus，Membrane，Nucleus，Cytosol，Cytoplasm上的AUPR分别是0.234，0.259，0.254，0.386，0.660，0.482，0.759。在Chromatin，Nucleoplasm，Nucleolus，Membrane上表现不佳，通过观察Figure 2可知，其样本数量较少，且负样本远多于正样本，由于AUPR对样本不平很问题比AUC更敏感，这些因素导致了它的AUPR值偏低。进一步计算平均AUC和AUPR，结果分别是0.775和0.433

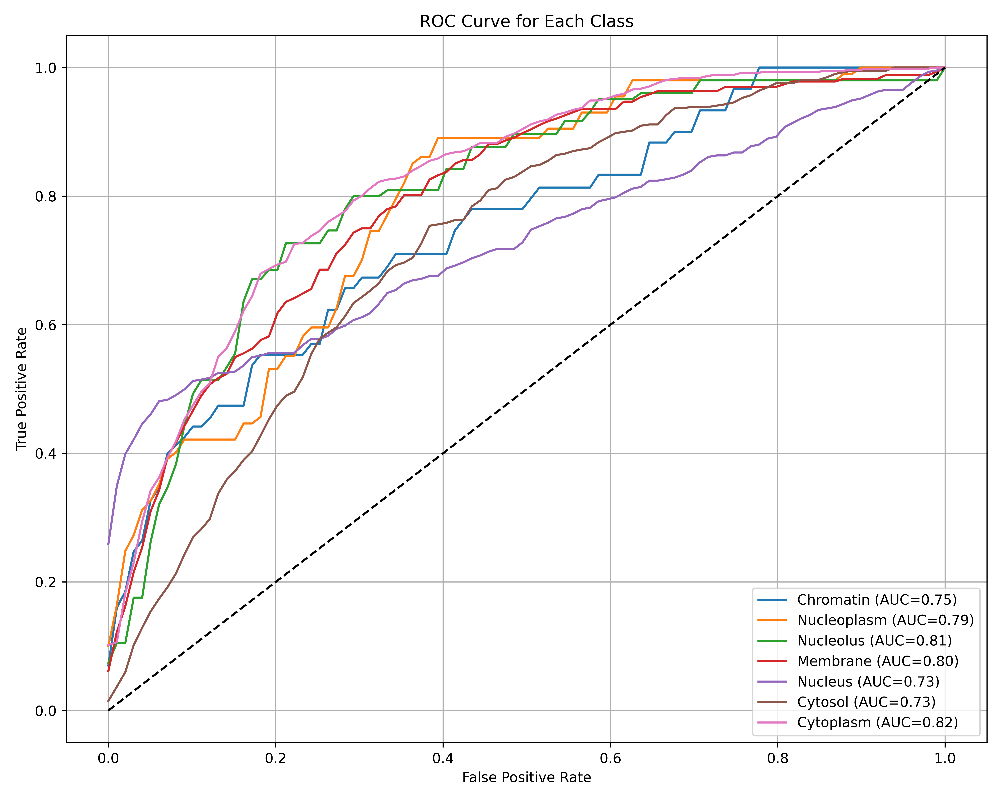


Figure 5. ROC curves

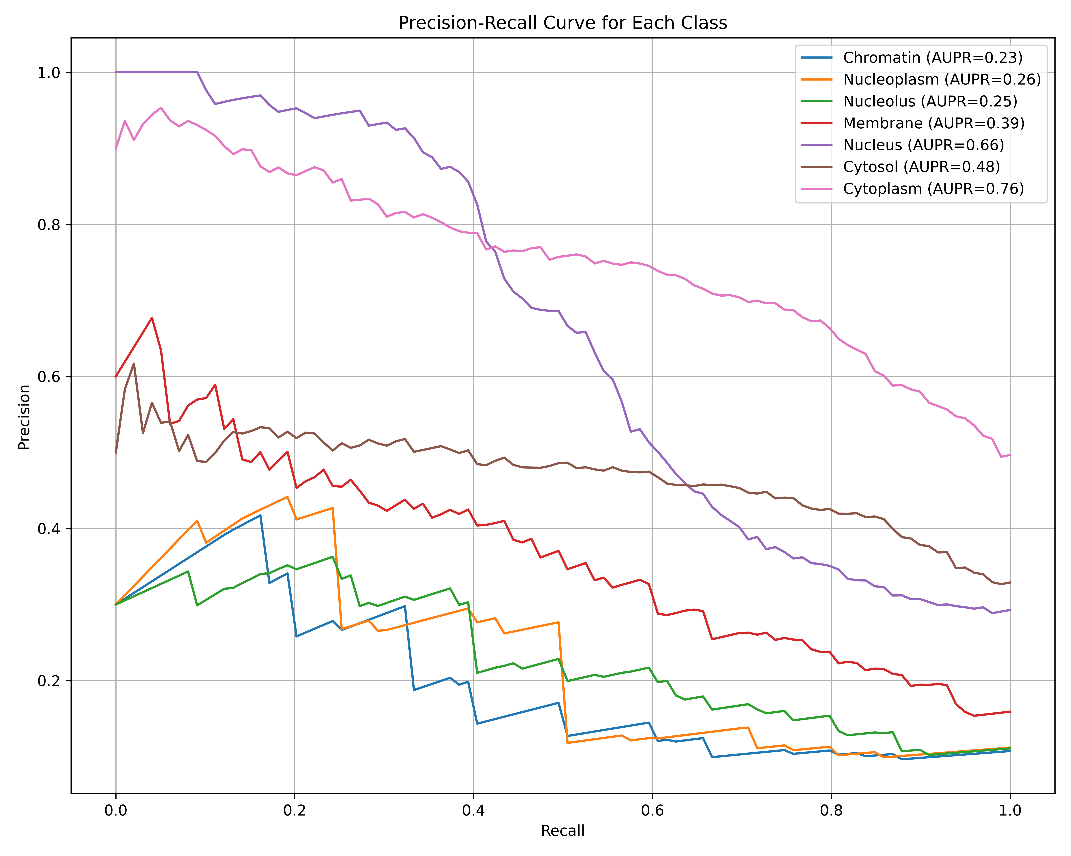


Figure 6. PR curves

## 八种特征类型的有效性分析

CircLoc分类器整合了八种特征类型：序列相似度特征、Kmer、RCKmer、RNAErine、疾病关联特征、药物关联特征、miRNA网络特征及miRNA共定位特征。本节通过特征消融实验验证各特征对模型的贡献度，并证明多种特征融合能提升预测性能。考虑到八种特征能够生成255中组合，由于组合数过多，我们将其分为四类，分别是通过传统的序列编码方法K-mer和RCKmer，预训练语言模型特征，序列相似度网络特征以及和疾病、miRNA、drug等相关联网络特征，这里分别记作α、β、γ、η。通过分类我们可能得到15种组合。针对每种组合，我们构建了分类器并进行10折交叉验证，其性能指标见Table 2

Table . 各分类器指标得分

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **α** | **β** | **γ** | **η** | **Average AUC** | **Average AUPR** | **Average Accuracy** | **Average F1** | **Average Recall** | **Average Precision** |
| √ |  |  |  | 0.544 | 0.234 | 0.577 | 0.332 | 0.681 | 0.263 |
|  | √ |  |  | 0.621 | 0.263 | 0.635 | 0.358 | 0.708 | 0.288 |
|  |  | √ |  | 0.716 | 0.351 | 0.752 | 0.406 | 0.629 | 0.407 |
|  |  |  | √ | 0.699 | 0.350 | 0.764 | 0.412 | 0.620 | 0.380 |
| √ | √ |  |  | 0.612 | 0.263 | 0.661 | 0.369 | 0.648 | 0.306 |
| √ |  | √ |  | 0.709 | 0.343 | 0.743 | 0.402 | 0.643 | 0.382 |
| √ |  |  | √ | 0.699 | 0.348 | 0.748 | 0.410 | 0.618 | 0.398 |
|  | √ | √ |  | 0.751 | 0.408 | 0.801 | 0.463 | 0.622 | 0.481 |
|  | √ |  | √ | 0.735 | 0.400 | 0.814 | 0.459 | 0.595 | 0.454 |
|  |  | √ | √ | 0.730 | 0.369 | 0.790 | 0.426 | 0.599 | 0.411 |
| √ | √ | √ |  | 0.711 | 0.368 | 0.761 | 0.438 | 0.622 | 0.427 |
| √ | √ |  | √ | 0.727 | 0.378 | 0.781 | 0.442 | 0.633 | 0.404 |
| √ |  | √ | √ | 0.736 | 0.378 | 0.792 | 0.437 | 0.606 | 0.431 |
|  | √ | √ | √ | 0.746 | 0.405 | 0.800 | 0.459 | 0.615 | 0.458 |
| √ | √ | √ | √ | **0.775** | 0.433 | **0.807** | **0.480** | **0.645** | **0.464** |

实验结果表明，采用全部的CircLoc在AUC、AUPR、F1、Recall、Precision分数指标上均取得最高值，Accuracy取得第二高，这证实多特征协同能显著增强分类器效能。

此外，我们统计了它在7种亚细胞定位上的表现，AUC和AUPR值分别见Table 3和Table 4。对比显示，CircLoc在5种定位(Chromatin、Nucleoplasm、Membrane、Nucleus、Cytoplasm)上获得最高AUC值，同时在5种定位(Nucleoplasm、Membrane、Nucleus、Cytosol 、Cytoplasm)上取得最高的AUPR值。这表明CircLoc显著优于其他缺失部分特征的分类器，每个特征类型都对模型具有独特贡献。

为了探究数量与性能的关系，我们把15个分类器按特征类型数量分为四组，CircLoc作为第4组(4类特征)，Figure 7和Figure 8显示，在7种亚细胞定位的AUC和AUPR指标上均呈现严格单调上升趋势(除Nuleolus在AUPR上几乎持平外)，证明多特征融合能持续提升分类器性能。

Table 3. 各亚细胞定位AUC分数

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **α** | **β** | **γ** | **η** | **Chromatin** | **Nucleoplasm** | **Nucleolus** | **Membrane** | **Nucleus** | **Cytosol** | **Cytoplasm** |
| √ |  |  |  | 0.478 | 0.556 | 0.542 | 0.555 | 0.528 | 0.526 | 0.572 |
|  | √ |  |  | 0.678 | 0.603 | 0.712 | 0.643 | 0.624 | 0.518 | 0.617 |
|  |  | √ |  | 0.595 | 0.684 | 0.690 | 0.644 | 0.714 | 0.675 | 0.793 |
|  |  |  | √ | 0.528 | 0.655 | 0.670 | 0.702 | 0.722 | 0.695 | 0.786 |
| √ | √ |  |  | 0.634 | 0.558 | 0.732 | 0.639 | 0.598 | 0.513 | 0.587 |
| √ |  | √ |  | 0.615 | 0.644 | 0.682 | 0.649 | 0.685 | 0.644 | 0.761 |
| √ |  |  | √ | 0.578 | 0.669 | 0.770 | 0.677 | 0.710 | 0.688 | 0.789 |
|  | √ | √ |  | 0.765 | 0.713 | 0.764 | 0.743 | 0.751 | 0.708 | 0.811 |
|  | √ |  | √ | 0.701 | 0.712 | 0.753 | 0.740 | 0.726 | 0.710 | 0.796 |
|  |  | √ | √ | 0.639 | 0.677 | 0.723 | 0.749 | 0.729 | **0.744** | 0.820 |
| √ | √ | √ |  | 0.664 | 0.728 | 0.764 | 0.714 | 0.674 | 0.667 | 0.765 |
| √ | √ |  | √ | 0.671 | 0.710 | **0.814** | 0.726 | 0.713 | 0.673 | 0.780 |
| √ |  | √ | √ | 0.616 | 0.762 | 0.779 | 0.736 | 0.720 | 0.722 | 0.814 |
|  | √ | √ | √ | 0.733 | 0.725 | 0.768 | 0.750 | 0.729 | 0.714 | 0.804 |
| √ | √ | √ | √ | **0.750** | **0.790** | 0.807 | **0.798** | **0.728** | 0.727 | **0.824** |

α: K-mer features+RCK-mer features; β:RNAErine features; γ:sequence features; η: disease features +drug features +miRNA features +mRNA co-localization features.

Table 4. 各亚细胞定位AUPR分数

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **α** | **β** | **γ** | **η** | **Chromatin** | **Nucleoplasm** | **Nucleolus** | **Membrane** | **Nucleus** | **Cytosol** | **Cytoplasm** |
| √ |  |  |  | 0.065 | 0.123 | 0.180 | 0.167 | 0.330 | 0.326 | 0.484 |
|  | √ |  |  | 0.122 | 0.103 | 0.134 | 0.214 | 0.401 | 0.315 | 0.528 |
|  |  | √ |  | 0.068 | 0.089 | 0.115 | 0.184 | 0.657 | 0.414 | 0.748 |
|  |  |  | √ | 0.085 | 0.115 | 0.116 | 0.239 | 0.656 | 0.460 | 0.697 |
| √ | √ |  |  | 0.131 | 0.105 | 0.200 | 0.221 | 0.386 | 0.315 | 0.497 |
| √ |  | √ |  | 0.091 | 0.203 | 0.145 | 0.216 | 0.598 | 0.406 | 0.698 |
| √ |  |  | √ | 0.098 | 0.113 | 0.269 | 0.243 | 0.642 | 0.456 | 0.694 |
|  | √ | √ |  | **0.247** | 0.205 | 0.253 | 0.258 | 0.656 | 0.452 | 0.757 |
|  | √ |  | √ | 0.130 | 0.192 | 0.268 | 0.287 | 0.662 | 0.460 | 0.717 |
|  |  | √ | √ | 0.082 | 0.088 | 0.182 | 0.290 | 0.677 | 0.504 | 0.776 |
| √ | √ | √ |  | 0.176 | 0.168 | 0.236 | 0.289 | 0.595 | 0.417 | 0.694 |
| √ | √ |  | √ | 0.191 | 0.138 | **0.261** | 0.274 | 0.645 | 0.431 | 0.705 |
| √ |  | √ | √ | 0.099 | 0.190 | 0.205 | 0.278 | 0.653 | 0.481 | 0.741 |
|  | √ | √ | √ | 0.198 | 0.188 | 0.312 | 0.276 | 0.654 | 0.471 | 0.723 |
| √ | √ | √ | √ | 0.234 | **0.259** | 0.254 | **0.386** | **0.660** | **0.582** | **0.759** |

α: K-mer features+RCK-mer features; β:RNAErine features; γ:sequence features; η: disease features +drug features +miRNA features +mRNA co-localization features.

Figure 7. performance on seven subcellula localizations measured by AUC

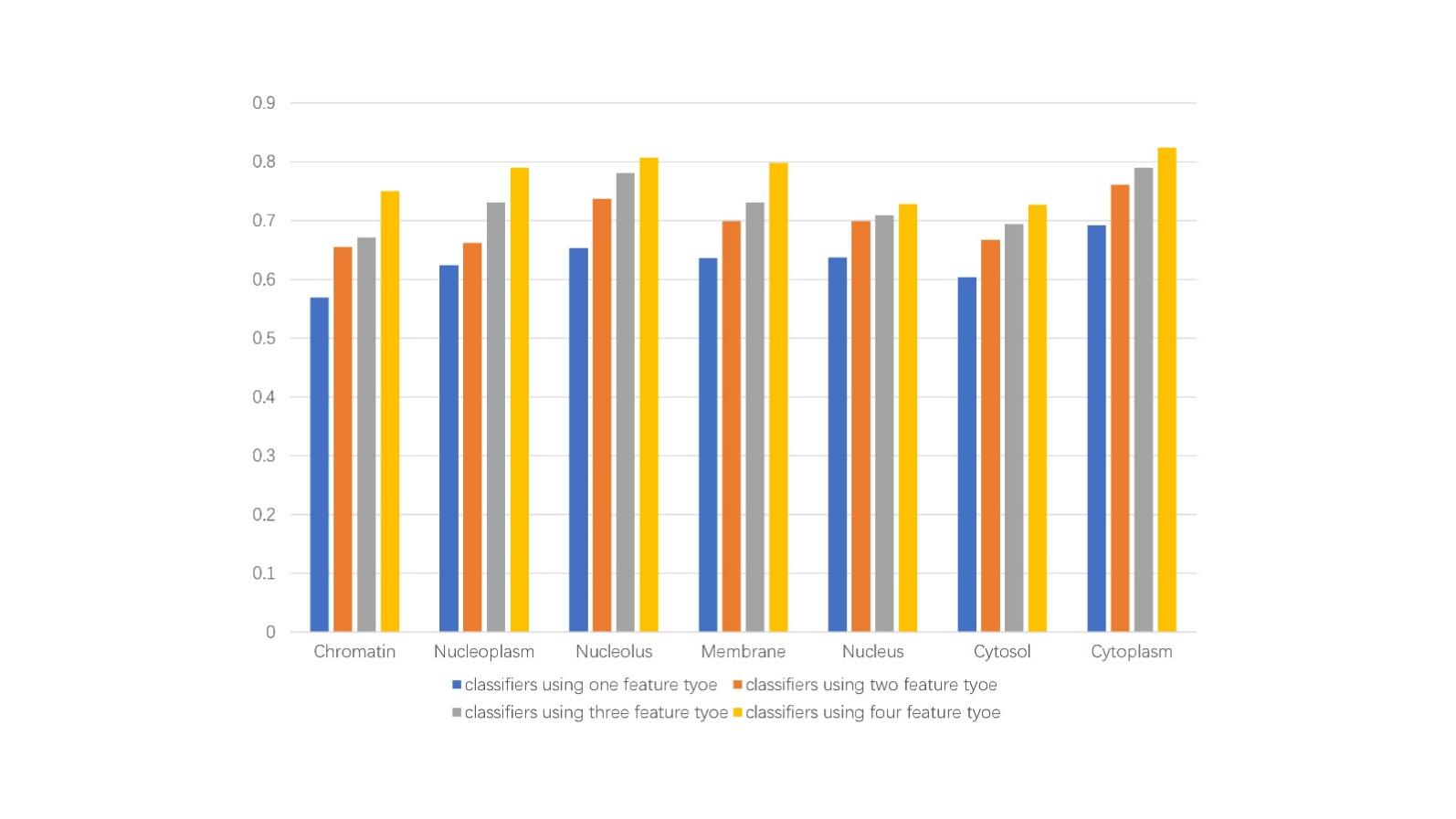
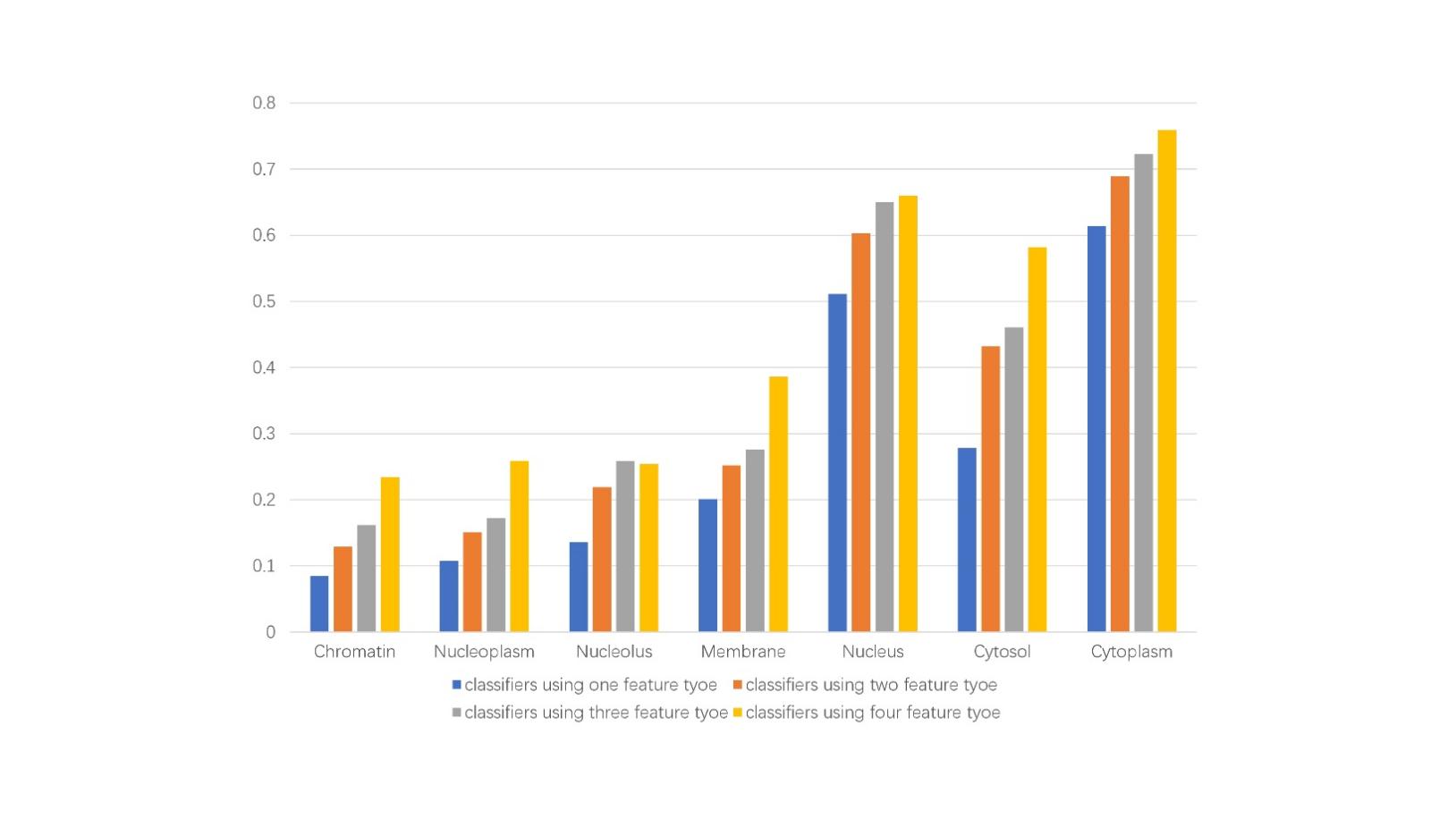


Figure 8. performance on seven subcellula localizations measured by AUPR



## Effectiveness of GATE

为提升源自circRNA-疾病关联、circRNA-药物作用及circRNA-miRNA互作网络的原始特征质量，本研究采用了图注意力增强（GATE）模块。需验证该模块对CircLoc性能的实际提升作用，我们移除了GATE模块并重建分类器，同样采用10折交叉验证进行评估（结果见Table 5）。表中同步列出CircLoc的测量值以进行对比分析。

实验数据显示，去除GATE模块后，分类器的Accuracy、F1、Recall、Precision、AUC以及AUPR均出现下降，在7种亚细胞定位的预测性能上，AUC值相比较完整的CircLoc全面下滑，这些结果证明，未配置GATE模块的分类器性能明显逊色于CircLoc，充分验证了GATE模块对于构建高性能分类器的必要性。

## Effectiveness of self-attention layer

自注意力层的设计旨在进一步提升最终预测前的特征质量。为验证该模块对CircLoc的贡献，我们移除了自注意力层并重建分类器，通过10折交叉验证进行评估(结果见Table 5)，与完整版CircLoc相比，去除该模块后，所有指标均下降，在7种亚细胞定位的预测中，所有AUC值均下降，下降幅度0.001-0.02，AUPR值也呈现相似程度的衰减。这些结果表明，缺失自注意力层的分类器性能显著弱于CircLoc，证实了该模块对模型性能的提升具有实质性作用。

## Effectiveness of focal\_loss

在CircLoc模型中，Focal Loss被用于解决类别不平衡问题，通过调整损失函数，使模型更关注难以分类的样本（尤其是少数类）。为了验证其有效性，我们进行了消融实验，通过10折交叉验证进行评估(结果见Table 5)，对比了使用Focal Loss和不使用Focal Loss时的模型性能差异。在各项指标上，绝大部分指标都有不同程度的衰减，在7种亚细胞定位的预测中，除Nucleus外的AUC分数均出现不同程度的衰减，AUPR也出现了类似的程度的衰减。这些结果表明，缺失focal\_loss的分类器性能显著弱于CircLoc，证实了该模块对模型性能的提升具有实质性作用。

Table 5. Ablation test results on graph attention auto-encoder,self-attention layer and focal\_loss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Measurements** |  | **Classifier without graph attention auto-encoder** | **Classifier without self-attention layer** | **Classifier without focal\_loss** | **CircLoc** |
| Average AUC |  | 0.755 | 0.737 | 0.746 | **0.775** |
| Average AUPR |  | 0.411 | 0.391 | 0.418 | **0.433** |
| Average Accuracy |  | 0.798 | 0.788 | **0.832** | 0.807 |
| Average F1 |  | 0.459 | 0.448 | 0.455 | **0.480** |
| Average Recall |  | 0.627 | 0.614 | 0.532 | **0.645** |
| Average Precision |  | 0.444 | 0.451 | 0.445 | **0.464** |
| AUC |  |  |  |  |  |
|  | Chromatin | 0.641 | 0.651 | 0.692 | **0.750** |
|  | Nucleoplasm | 0.773 | 0.766 | 0.706 | **0.790** |
|  | Nucleolus | 0.801 | 0.788 | 0.800 | **0.807** |
|  | Membrane | 0.753 | 0.729 | 0.760 | **0.798** |
|  | Nucleus | **0.742** | 0.727 | 0.735 | 0.728 |
|  | Cytosol | **0.751** | 0.692 | 0.712 | 0.727 |
|  | Cytoplasm | 0.825 | 0.804 | 0.817 | **0.824** |
| AUPR |  |  |  |  |  |
|  | Chromatin | 0.118 | 0.125 | 0.233 | **0.234** |
|  | Nucleoplasm | 0.187 | 0.179 | 0.196 | **0.259** |
|  | Nucleolus | **0.330** | 0.263 | 0.276 | 0.254 |
|  | Membrane | 0.281 | 0.332 | 0.332 | **0.386** |
|  | Nucleus | 0.686 | 0.657 | 0.654 | **0.660** |
|  | Cytosol | 0.503 | 0.435 | 0.465 | **0.582** |
|  | Cytoplasm | 0.770 | 0.746 | 0.758 | **0.795** |

## Performance comparison with previous prediction methods

由于目前没有使用深度学习模型进行circRNA亚细胞预测的检测方法，考虑到同为ncRNA的miRNA亚细胞定位预测，已有多种方法被提出，本研究中我们采用三种代表性的方法：MiRLoc[[18](#_ENREF_18)]，MirLocPredictor[[21](#_ENREF_21)]，PMLocMSCAM[[48](#_ENREF_48)]，使用本研究的数据集进行比较预测比较，结果如Table 6和Table 7显示：

Table . Performance comparison in terms of AUC

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subcellular localization | MiRLoc | MirLocPredictor | PMLocMSCAM | circLoc |
| Chromatin | 0.622 | 0.532 | 0.617 | **0.750** |
| Nucleoplasm | 0.720 | 0.512 | 0.669 | **0.790** |
| Nucleolus | 0.520 | 0.509 | 0.766 | **0.807** |
| Membrane | 0.704 | 0.579 | 0.679 | **0.798** |
| Nucleus | 0.677 | 0.553 | 0.723 | 0.728 |
| Cytosol | 0.550 | 0.532 | 0.595 | **0.727** |
| Cytoplasm | 0.588 | 0.645 | 0.812 | **0.824** |

Table . Performance comparison in terms of AUPR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subcellular localization | MiRLoc | MirLocPredictor | PMLocMSCAM | circLoc |
| Chromatin | 0.043 | 0.185 | 0.138 | **0.234** |
| Nucleoplasm | 0.051 | 0.233 | 0.128 | **0.259** |
| Nucleolus | 0.052 | 0.207 | 0.244 | **0.254** |
| Membrane | 0.189 | 0.325 | 0.270 | **0.386** |
| Nucleus | 0.423 | 0.447 | 0.666 | 0.660 |
| Cytosol | 0.301 | 0.442 | 0.379 | **0.582** |
| Cytoplasm | 0.469 | 0.597 | 0.722 | **0.759** |

## 本研究局限性分析

尽管CircLoc模型在circRNA亚细胞定位预测中表现出良好的性能，仍存在若干局限性：物种适用性受限：当前模型仅针对人类circRNA进行测试，尚未验证其跨物种预测能力。数据完整性依赖：虽然整合多源circRNA特征可增强表征能力，但部分circRNA存在特征缺失问题，导致模型对此类样本的预测效能下降。数据集不平衡问题：所用人类circRNA数据集存在显著类别不均衡现象，未采用过采样或欠采样方法进行校正，致使某些定位类别的预测性能不平衡。未来研究将致力于突破上述限制，开发适用性更广的高效预测模型。

## 研究结论

本研究开发了一种新型多标签分类器CircLoc，通过融合多维度circRNA特征和序列特征并设计专用特征提取方案，实现了circRNA亚细胞定位的预测。交叉验证表明，该模型拥有良好的表现，且具有以下优势：多特征融合策略显著提升预测效能、图注意力自编码器（GATE）与自注意力层发挥关键作用、对疾病/药物/mRNA关联性较弱的miRNA仍能保持准确预测、该模型有望成为circRNA亚细胞定位研究的重要工具，数据集与代码已开源至xxxxxxxxxx

# 引用

1. Oddo, J.C., et al., *Conservation of context-dependent splicing activity in distant Muscleblind homologs.* Nucleic Acids Res, 2016. **44**(17): p. 8352-62.

2. Cheng, L. and K.S. Leung, *Quantification of non-coding RNA target localization diversity and its application in cancers.* J Mol Cell Biol, 2018. **10**(2): p. 130-138.

3. Frias-Lasserre, D. and C.A. Villagra, *The Importance of ncRNAs as Epigenetic Mechanisms in Phenotypic Variation and Organic Evolution.* Front Microbiol, 2017. **8**: p. 2483.

4. Meng, X., et al., *Circular RNA: an emerging key player in RNA world.* Brief Bioinform, 2017. **18**(4): p. 547-557.

5. Cocquerelle, C., et al., *Splicing with inverted order of exons occurs proximal to large introns.* EMBO J, 1992. **11**(3): p. 1095-8.

6. Sanger, H.L., et al., *Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures.* Proc Natl Acad Sci U S A, 1976. **73**(11): p. 3852-6.

7. Cocquerelle, C., et al., *Mis-splicing yields circular RNA molecules.* FASEB J, 1993. **7**(1): p. 155-60.

8. Zaphiropoulos, P.G., *Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping.* Proc Natl Acad Sci U S A, 1996. **93**(13): p. 6536-41.

9. Li, Z., et al., *Exon-intron circular RNAs regulate transcription in the nucleus.* Nat Struct Mol Biol, 2015. **22**(3): p. 256-64.

10. Salzman, J., et al., *Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types.* PLoS One, 2012. **7**(2): p. e30733.

11. Geng, X., et al., *Circular RNA: biogenesis, degradation, functions and potential roles in mediating resistance to anticarcinogens.* Epigenomics, 2020. **12**(3): p. 267-283.

12. Liu, J., et al., *Emerging Roles and Potential Biological Value of CircRNA in Osteosarcoma.* Front Oncol, 2020. **10**: p. 552236.

13. Holdt, L.M., A. Kohlmaier, and D. Teupser, *Molecular roles and function of circular RNAs in eukaryotic cells.* Cell Mol Life Sci, 2018. **75**(6): p. 1071-1098.

14. Bachmayr-Heyda, A., et al., *Correlation of circular RNA abundance with proliferation--exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues.* Sci Rep, 2015. **5**: p. 8057.

15. Li, F., et al., *Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/beta-catenin pathway.* Oncotarget, 2015. **6**(8): p. 6001-13.

16. Li, P., et al., *Using circular RNA as a novel type of biomarker in the screening of gastric cancer.* Clin Chim Acta, 2015. **444**: p. 132-6.

17. Xue, L., et al., *DeepT3: deep convolutional neural networks accurately identify Gram-negative bacterial type III secreted effectors using the N-terminal sequence.* Bioinformatics, 2019. **35**(12): p. 2051-2057.

18. Xu, M., et al., *MiRLoc: predicting miRNA subcellular localization by incorporating miRNA-mRNA interactions and mRNA subcellular localization.* Brief Bioinform, 2022. **23**(2).

19. Bai, T., K. Yan, and B. Liu, *DAmiRLocGNet: miRNA subcellular localization prediction by combining miRNA-disease associations and graph convolutional networks.* Brief Bioinform, 2023. **24**(4).

20. Chen, Z., et al., *iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization.* Nucleic Acids Res, 2021. **49**(10): p. e60.

21. Asim, M.N., et al., *MirLocPredictor: A ConvNet-Based Multi-Label MicroRNA Subcellular Localization Predictor by Incorporating k-Mer Positional Information.* Genes (Basel), 2020. **11**(12).

22. Monga, I. and I. Banerjee, *Computational Identification of piRNAs Using Features Based on RNA Sequence, Structure, Thermodynamic and Physicochemical Properties.* Curr Genomics, 2019. **20**(7): p. 508-518.

23. Niu, M., et al., *CirRNAPL: A web server for the identification of circRNA based on extreme learning machine.* Comput Struct Biotechnol J, 2020. **18**: p. 834-842.

24. Lee, D., R. Karchin, and M.A. Beer, *Discriminative prediction of mammalian enhancers from DNA sequence.* Genome Res, 2011. **21**(12): p. 2167-80.

25. Gupta, S., et al., *Predicting human nucleosome occupancy from primary sequence.* PLoS Comput Biol, 2008. **4**(8): p. e1000134.

26. Noble, W.S., et al., *Predicting the in vivo signature of human gene regulatory sequences.* 2005. **21**(suppl\_1): p. i338-i343.

27. Fletez-Brant, C., et al., *kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W544-56.

28. Zuo, Y., H. Zhou, and Z. Yue. *Prorice: An ensemble learning approach for predicting promoters in rice*. in *Proceedings of the 4th International Conference on Computer Science and Application Engineering*. 2020.

29. Xu, H., P. Jia, and Z. Zhao, *Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning.* Brief Bioinform, 2021. **22**(3).

30. Wang, H., et al., *Multi-label learning for identi cation of RNA-associated subcellular localizations.* 2020.

31. Wang, N., et al., *Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning.* Nature Machine Intelligence, 2024. **6**(5): p. 548-557.

32. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

33. Meng, S., et al., *CircRNA: functions and properties of a novel potential biomarker for cancer.* 2017. **16**: p. 1-8.

34. Gao, J.-L., et al., *CircRNA as a new field in human disease research.* 2018. **43**(3): p. 457-462.

35. Guarnerio, J., et al., *Oncogenic role of fusion-circRNAs derived from cancer-associated chromosomal translocations.* 2016. **165**(2): p. 289-302.

36. Shang, Q., et al., *The novel roles of circRNAs in human cancer.* 2019. **18**: p. 1-10.

37. Wei, L., et al., *Noncoding RNAs in gastric cancer: implications for drug resistance.* Mol Cancer, 2020. **19**(1): p. 62.

38. Cui, C., et al., *Functions and mechanisms of circular RNAs in cancer radiotherapy and chemotherapy resistance.* Mol Cancer, 2020. **19**(1): p. 58.

39. Deng, L., et al., *Predicting circRNA-drug sensitivity associations via graph attention auto-encoder.* 2022. **23**(1): p. 160.

40. Chen, Y., et al., *RGCNCDA: relational graph convolutional network improves circRNA-disease association prediction by incorporating microRNAs.* 2022. **143**: p. 105322.

41. Zhang, M., et al., *A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis.* 2018. **37**(13): p. 1805-1814.

42. He, W., et al., *Circ\_0019693 promotes osteogenic differentiation of bone marrow mesenchymal stem cell and enhances osteogenesis-coupled angiogenesis via regulating microRNA-942-5p-targeted purkinje cell protein 4 in the development of osteoporosis.* 2022. **13**(2): p. 2181-2193.

43. Fan, X., et al., *Hsa\_circRNA\_0045861 promotes renal injury in ureteropelvic junction obstruction via the microRNA-181d-5p/sirtuin 1 signaling axis.* 2021. **9**(20): p. 1571.

44. Grover, A. and J. Leskovec. *node2vec: Scalable feature learning for networks*. in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016.

45. Salehi, A. and H.J.a.p.a. Davulcu, *Graph attention auto-encoders.* 2019.

46. Kingma, D.P. and J.J.a.p.a. Ba, *Adam: A method for stochastic optimization.* 2014.

47. Abraham, A., et al., *Machine learning for neuroimaging with scikit-learn.* 2014. **8**: p. 14.

48. Jiang, J. and C.J.I.S.B. Yan, *PMLocMSCAM: Predicting miRNA Subcellular Localisations by miRNA Similarities and Cross‐Attention Mechanism.* 2025. **19**(1): p. e70023.