Hans Journal of Computational Biology 计算生物学, 2026, 16(2), 64-78
Published Online June 2026 in Hans. https://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2026.162006

Hans汉斯

SSAL：一种基于序列与结构特征及注意力模型的circRNA-RBP相互作用位点预测方法

刘　晨

大连交通大学理学院，辽宁　大连

收稿日期：2026年5月11日；录用日期：2026年6月4日；发布日期：2026年6月10日

摘　要

预测环状RNA (circRNA)与RNA结合蛋白(RBP)之间的相互作用位点，对于揭示疾病调控机制及开发新型治疗靶点具有重要意义。随着基因组范围内circRNA结合事件数据的日益丰富，计算模型已成为高效预测circRNA-RBP相互作用位点的主流工具。然而，如何通过有效提取circRNA的多尺度特征来提升预测准确率，仍是该领域面临的核心挑战。为解决这一问题，本研究提出了一种名为SSAL的深度学习模型，旨在实现大规模数据集上circRNA-RBP相互作用位点的精准预测。该模型的核心模块包括：首先，系统性地提取circRNA的序列特征与二级结构特征；随后，利用注意力机制对多尺度序列特征进行融合。为增强模型的稳定性和泛化能力，本文构建了一个集成学习框架，通过整合多个子模型的预测结果，有效缓解了单一模型固有的误差与随机性。为验证SSAL的性能，本文在14个大规模circRNA数据集上进行了全面评估，并将其与当前主流方法进行了对比。实验结果表明，SSAL的平均曲线下面积(AUC)达到97.66%，不仅充分证实了其在效率与鲁棒性方面的优势，且在预测准确率上均优于所有对比方法。

关键词

相互作用位点预测，注意力机制，深度学习，多尺度特征

SSAL: A Prediction Method for circRNA-RBP Interaction Sites Using an Attention Model Based on Sequence and Structural Features

Chen Liu

School of Science, Dalian Jiaotong University, Dalian Liaoning

Received: May 11, 2026; accepted: June 4, 2026; published: June 10, 2026

文章引用: 刘晨. SSAL: 一种基于序列与结构特征及注意力模型的circRNA-RBP相互作用位点预测方法[J]. 计算生物学, 2026, 16(2): 64-78. DOI: 10.12677/hjcb.2026.162006

Abstract

Predicting the interaction sites between circular RNA (circRNA) and RNA-binding proteins (RBPs) is of significant importance for deciphering disease regulatory mechanisms and developing novel therapeutic targets. With the increasing accumulation and availability of genome-wide circRNA binding event data for computational analysis, computational models have become mainstream tools for efficiently predicting circRNA-RBP interaction sites. However, enhancing prediction accuracy by effectively extracting multi-scale features of circRNA remains a key challenge in this field. To address this issue, this study proposes a deep learning model named SSAL, which can accurately predict circRNA-RBP interaction sites on large-scale datasets. The core modules of this model include: first, systematic extraction of circRNA sequence features and secondary structure features; Subsequently, an attention mechanism was employed to fuse multi-scale sequence features. To enhance model stability and generalization capability, we constructed an ensemble learning framework that integrates predictions from multiple sub-models, effectively mitigating the errors and randomness inherent in individual models. To validate the performance of SSAL, we conducted comprehensive evaluations on 14 large-scale circRNA datasets and compared it with current mainstream methods. The results demonstrate that SSAL achieved an average AUC of 97.66%, not only fully confirming its advantages in efficiency and robustness but also surpassing all comparison methods in prediction accuracy.

Keywords

Interaction Site Prediction, Attention Mechanism, Deep Learning, Multi-Scale Features

1. 引言

环状 RNA (circRNA)是一类通过反向剪接形成的共价闭合环状非编码 RNA。由于其结构缺乏 5′端帽子和 3′端 poly (A)尾，这种特性赋予了 circRNA 极高的稳定性，并能抵抗外切核酸酶的降解[1]。circRNA 在真核细胞中广泛表达，具有结构稳定、序列保守和来源丰富等特点，预示着其具有重要的生物学功能[2]。circRNA 通过多种机制发挥作用：作为 miRNA 海绵，它们可以参与肿瘤细胞的增殖、远端转移及化疗耐药等多种活动[3]；此外，部分 circRNA 具备招募核糖体的能力，可作为翻译模板编码功能性蛋白[4]。同时，RNA 的功能与 RNA 结合蛋白(RBPs)密切相关，后者是一类广泛参与基因转录与翻译、控制胞内 RNA 加工、转运及降解等过程的蛋白质。circRNA 是 RBP 发挥调控功能的主要 RNA 靶标之一[5]-[7]。因此，深入研究 circRNA-RBP 相互作用位点，对于阐明癌症等疾病的发病机制、开发新型治疗靶点以及鉴定稳定的液态活检生物标志物具有重要意义[8]-[10]。

随着高通量测序技术的进步，RNA-RBP 结合位点数据库相继建立，为计算分析奠定了基础。本文数据来源于 CIRCpedia、CircR2Disease 及 CircInteractome [11]-[14]。在此背景下，Zhang 等人于 2019 年提出了 CRIP 模型[15]，该模型率先将堆叠密码子表示法引入环状 RNA 序列编码中。通过整合卷积神经网络(CNN)与循环神经网络(RNN)的混合架构，CRIP 有效捕捉了序列内的局部依赖性与长程上下文信息，显著提升了预测性能。受深度学习在生物信息学领域成功的启发，次年 Jia 等人开发了 PASSION 方法[16]。其核心创新在于提出了增量特征选择策略并结合 XGBoost 算法，从六种编码方案中筛选出最具辨

识度的最优特征子集，从而规避了冗余信息的干扰。在此基础上，PASSION 利用混合神经网络对所选特征进行深层建模，实现了高精度的位点预测。2021 年，Yang 团队推出了 iCircRBP-DHN 模型[17]，该框架结合了双向门控循环单元(BiGRU)与注意力机制，构建了深度多尺度残差网络，有效捕获了不同层级的核苷酸依赖性。2022 年，Niu 等人提出了 CRBPDL 模型[18]，利用深度多尺度残差网络(MSRN)和 BiGRU 表征序列，并采用 Adaboost 集成策略优化预测结果。同年，Yang 等提出的 HCRNet [19]采用了深度时空卷积网络架构，通过同时捕捉序列的空间结构特征与时间依赖性，显著增强了 circRNA 结合事件的识别精度。2023 年，曹研究团队在前期工作基础上，通过引入自注意力机制对 iCircRBP-DHN 模型进行了重大改进[20]。该创新使模型能够自适应学习序列中不同位置特征的重要性权重，从而更精准地捕获关键结合模式，为大规模识别位点提供了技术支持。

尽管现有方法已取得良好效果，但仍存在若干待解决的问题。首先，多数方法仅将 circRNA 的一维核苷酸序列作为输入，未能有效利用结构特征，而结构信息理论上对于精准识别结合位点至关重要。近年来虽有研究尝试预测 RNA 的三维结构[21]-[23]，但受限于结构的复杂性，此类方法通常面临技术挑战且误差较大。此外，这些方法在临床转化(如生物标志物发现与靶向药物开发)方面的实际贡献仍有限。其次，现有模型未能充分挖掘不同特征间的深层关联。多维特征往往被直接用于预测，缺乏充分的交互与融合，导致特征间的互补性与表达力不足，进而削弱了分类器的性能[24]-[26]。

为更准确地识别 circRNA-RBP 相互作用位点，本研究设计并实现了 SSAL 计算模型。在特征构建阶段，模型多维度提取信息：一方面通过核苷酸编码、密码子堆叠编码及间隔密码子编码提取序列信息；另一方面利用 CDPfold 生成 RNA 碱基配对矩阵以捕捉结构特征。为有效整合这些异构特征，SSAL 引入注意力机制进行多尺度特征的交互融合，使模型能够自适应学习深层依赖关系。融合后的特征被输入精心构建的分类模块，该模块集成了多层感知机(MLP)与 Softmax 函数，利用强大的非线性映射能力进行深度分析，最终输出每个核苷酸位点与 RBP 结合的预测概率。

实验结果表明，SSAL 充分利用了序列与结构信息。为提升泛化能力，本研究构建了基于多个子模型的集成模型。在 14 个大规模数据集上的五折交叉验证结果显示，SSAL 表现卓越，平均 AUC 值高达 99.7%，显著优于当前主流方法。这种优势在多个独立测试集中得到了验证，充分证明了模型的稳定性。最后，本文对预测结果进行了可视化处理。综上所述，SSAL 是一个高效、鲁棒的 circRNA-RBP 结合位点预测模型。

2. 方法

SSAL 模型的总体架构如图 1 所示。其核心设计由四个关键构建模块组成，构成了系统化的预测框架。这四个模块分别为：序列特征提取模块、结构特征提取模块、注意力融合模块以及输出预测模块。接下来的章节将详细介绍每个模块的设计原理及其功能实现。

2.1. 数据集

本研究的实验数据源自 CircInteractome 数据库(URL: https://circinteractome.nia.nih.gov/) [27] [28]。该数据库提供了 37 个公认的环状 RNA-RBP 相互作用数据集，为模型的训练与评估奠定了可靠的数据基础。为了确保数据质量并防止冗余序列引入训练偏差，本文首先利用 CD-HIT 工具对所有数据集进行了预处理。通过设定合理的相似性阈值，剔除高度同源的序列。经过严格的去重步骤，共获得 32,216 条高质量 circRNA-RBP 相互作用数据。这些样本涵盖了多种 circRNA 类型及其对应的 RBP 结合信息，构成了后续特征提取与模型训练的核心数据集。随后，读取每个 CLIP-seq 的峰值(peaks)并定位至相应的结合位点，提取峰值上游和下游各 50 个核苷酸，最终得到长度为 101 个核苷酸的 circRNA 片段[29] [30]。正

样本来源于经实验验证的结合位点，负样本则从未经验证的 circRNA 片段中随机抽取。为确保数据集平衡，正负样本比例维持在 1:1。在处理后的 37 个数据集中，本文筛选出 14 个规模较大的数据集(样本量均超过 20,000 条)。这些数据最终用于评估本模型的性能，并与其他主流模型进行对比测试，这 14 个数据集的详细信息如表 1 所示。

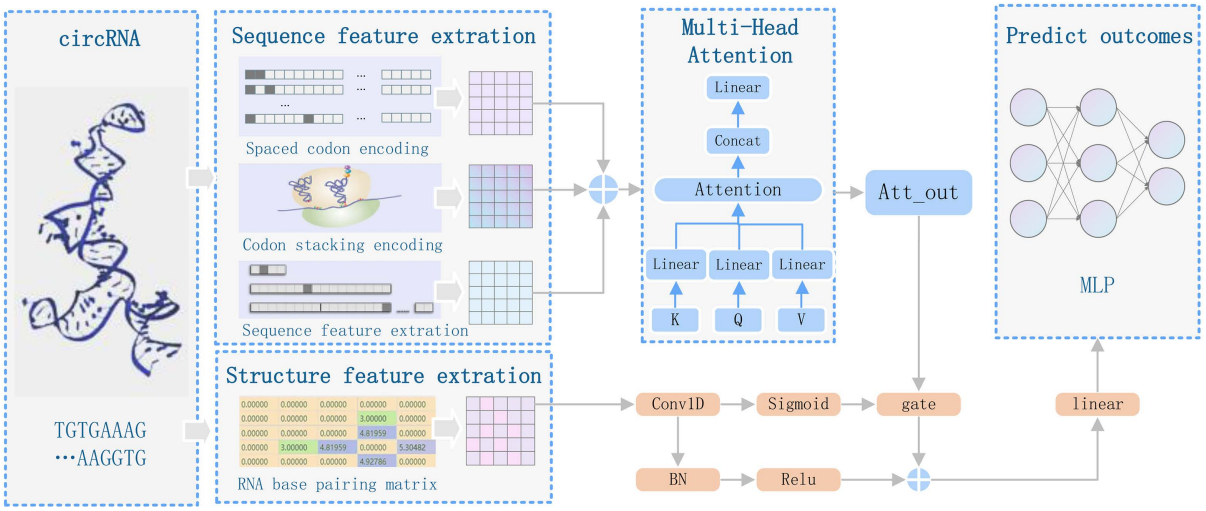

Figure 1. The overall framework of SSAL
图 1. SSAL 的整体架构

Table 1. Datasets statistic
表 1. 数据集统计

CircRNA	SIZE	CircRNA	SIZE
AGO1	34,636	HUR	40,000
AGO2	40,000	IGF2BPI	40,000
DGCR8	40,000	IGF2BP2	20,000
EIF4A3	40,000	IGF2BP3	40,000
FMRP	40,000	LIN28A	36,554
FUS	40,000	PTB	40,000
HNRNPC	28,448	ZC3H7B	26,238

2.2. circRNA 序列特征提取

首先，采用了密码子堆叠编码模式。该模式能够提取多种短程序列依赖信息和局部语义特征，在保留原始序列模式的同时，显著弥补了独热编码(One-hot encoding)的局限性。

以长度为 L 的特定 circRNA 序列为例，通过滑动窗口沿序列选取 k 个连续核苷酸，从而产生具有不同组合的核苷酸基团。利用独热编码来表示每组由 k ($k = 1, 2, 3$)个相邻核苷酸构成的组合。为了捕获完整信息，通过对这些密码子进行堆叠处理，最终为每个 circRNA 生成一个 84 维的独热编码向量。为了更清晰地阐述这一编码过程，给出如下形式化定义：

$$X_k = \left[x_1^k, x_2^k, \cdots, x_{4^k}^k \right] \tag{1}$$

$$X_n = \text{concat}\left(X_1, X_2, X_3 \right) \tag{2}$$

这里， $X_k \in R^{l \times 4^k}$ 表示 circRNA 序列的单核苷酸、双核苷酸和三核苷酸编码特征。其中，符号 concat()表示拼接操作，而 $X_n \in R^{l \times 84}$ 则表示特征编码后的序列表示。

基于 JLCRB 描述的方法，引入了密码子堆叠机制，以解析三核苷酸组并进一步提升编码能力。具体而言，该方法通过密码子将每三个核苷酸视为一组，并将其翻译为相应的氨基酸，从而生成原始 circRNA 的伪氨基酸序列表示。随后，利用独热编码(One-hot encoding)对每个氨基酸进行编码，最终从长度为 L 的 circRNA 序列中生成一个 L × 21 维的独热矩阵 X_4。最后，通过拼接编码后的特征 X_n 和 X_4，circRNA 序列特征可以表示为：

$$X_m = \text{concat}\left(\left(W_m X_n + b_m \right), X_4 \right) \tag{3}$$

其中， $X_m \in R^{l \times 105}$ 表示 circRNA 的序列特征， W_m 和 b_m 分别表示模型参数的权重矩阵和偏置项。

在序列上下文特征(Sequence Context Features)的初始化过程中，本文不仅采用了 k-mer 编码，还引入了组合 k 间隔核苷酸对(CKSNAP)编码来提取原始序列特征[31]。这些编码方法在多种生物信息学应用中均展现出卓越的序列信息提取能力。在本研究中，CKSNAP 用于表示 k 个核苷酸对的频率分布。本文采用了 k 值分别为 0、1、2、3、4 和 5 的编码方案，最终生成了一个 96 维的特征向量($6 \times 16 = 96$)。上述特征提取过程均采用 iLearn 工具包[32]完成，该工具包包含了这些编码方法的详细说明。在特征处理最后阶段，对编码后的特征序列 X_m 和 circRNA 序列特征 X_5 进行拼接操作从而得到后续建模的 circRNA 序列特征：

$$X_s = \text{concat}\left(\left(W_s X_m + b_s \right), X_5 \right) \tag{4}$$

其中， $X_s \in R^{l \times 121}$ 是 circRNA 序列的特征表示， W_s 和 b_s 分别表示模型参数的权重矩阵和偏置项。

2.3. circRNA 结构特征提取

本文通过计算内部碱基配对概率来确定环状 RNA 的二级结构信息，并构建 RNA 碱基配对矩阵，以进一步细化精准的二级结构细节。具体而言，给定长度为 L 的环状 RNA 序列，利用开源工具 CDPfold 预测并生成一个 $L \times L$ 维的碱基配对矩阵 M_t [33]。该矩阵由 L 行和 L 列组成，其中第 i 行、第 j 列的元素代表序列中第 i 个和第 j 个碱基之间的配对概率。具体而言，CDPfold 方法根据氢键数量为不同的碱基对分配权重(例如，A-U 权重为 2，G-C 权重为 3)。为了进一步评估每个碱基在茎区(Stemregions)形成稳定配对的潜力，CDPfold 引入了局部加权线性回归(Locally weighted linear regression)的概念。该方法采用高斯函数作为权重函数，综合考虑邻近碱基配对状态的影响。由此最终生成碱基配对矩阵 M_t，该矩阵反映了 circRNA 的结构特征。

基于前述构建的 RNA 碱基配对矩阵，本文设计了一个专门的结构特征提取模块。该模块由多个堆叠的卷积层(Convolutional layers)和批归一化层(Batch normalization layers)组成：卷积层从碱基配对概率图中提取局部结构模式——例如茎(Stems)和环(Loops)等二级结构单元；批归一化层对特征图(Feature maps)进行标准化处理，以防止梯度消失或梯度爆炸。该模块不仅能有效捕获 circRNA 的局部结构特征，还能将其聚合为全局结构特征表示，从而全面表征 RNA 分子内部的碱基相互作用模式。该结构的特征提取过程如下：

$$X_a = \sigma\left(\text{CNN}_{1\text{D}}\left(M_t \right) \right) \tag{5}$$

$$X_b = \mathrm{RL}\left(\mathrm{BN}\left(\mathrm{CNN}_{1\mathrm{D}}\left(M_t\right)\right)\right) \tag{6}$$

其中，$X_a \in R^l$ 表示用于交互的结构特征分数，X_m 表示 circRNA 结构特征。CNN 表示采用卷积神经网络，其中“1D”特指使用了一维卷积模块。在模型架构图中，BN 专门指代批归一化(Batch Normalization)层，其主要功能是对卷积层的输出进行标准化处理，以确保数据分布的稳定性。RL 代表 ReLU 激活函数，通过保留正值并抑制负值来增强特征的判别能力。σ 则表示 Sigmoid 激活函数，位于网络末端，用于输出最终的综合概率预测结果。

2.4. 输出连接和注意力模块

如前所述，通过多种特征提取方法，最终获得了 circRNA 的序列特征 X_s、结构特征 X_b 以及结构特征评分 X_a。在常规方法中，多源特征通常在输入全连接层进行分类前直接进行拼接。然而，本文的分析表明，circRNA 的序列特征与结构特征之间存在显著的相关性[34] [35]。这意味着简单的拼接操作可能会导致特征间互补增强信息的丢失。为解决这一问题，引入了注意力机制，通过增强关键序列特征的贡献并降低次要特征的影响，从而实现更高效的特征融合：

$$\begin{aligned} Q_i &= X_s \times W_i^Q; \\ K_i &= X_s \times W_i^K; \\ V_i &= X_s \times W_i^V \end{aligned} \tag{7}$$

$$\begin{aligned} head_i &= \mathrm{attention}\left(Q_i, K_i, V_i\right) \\ &= \mathrm{softmax}\left(\frac{Q_i \times K_i^{\mathrm{T}}}{\sqrt{d_{k_i}}}\right)V_i \end{aligned} \tag{8}$$

$$\mathrm{Multi\text{-}attention}\left(Q, K, V\right) = \mathrm{concat}\left(head_1, \cdots, head_n\right)W_0 \tag{9}$$

其中，X_s 表示注意力模块的输入，W_i^Q, W_i^K, W_i^V 表示注意力模块的参数，d_{k_i} 表示矩阵 k_i 的维度。

随后，将 circRNA 的序列特征 X_s，输入多头自注意力(Multi-head self-attention)机制，以获取全新的交互式序列特征 X_f。最后，通过将多头自注意力交互后的序列特征与结构特征进行简单拼接，生成融合了两种信息的联合表示(Joint representation)

$$X_C = \mathrm{concat}\left(X_a \times X_f, X_b\right) \tag{10}$$

通过多层级特征提取与融合，本文获取了 circRNA 特征中最核心的信息，并将其作为最终的向量表示(Vector representation)。

2.5. 模型训练和评估

本模型采用轻量化架构，其核心由卷积层和少量线性层组成。为了在不增加训练资源消耗的前提下提升预测性能，构建了一个集成模型(Ensemble model)。在构建预测模型的过程中，为增强模型对多样化数据集的适应性并提高泛化性能，采用 K 折交叉验证法对数据集进行处理。具体而言，数据集被均匀划分为 K 个互斥子集。在每一轮训练–测试迭代中，选取 $K-1$ 个子集用于模型训练以学习数据特征，剩余的一个子集则作为测试集用于评估模型性能。完成 K 轮训练后，获得了 K 个具有不同参数的子模型[36]。随后，利用集成学习策略将这些子模型整合为一个大型集成模型。该集成模型的最终输出由所有子模型的输出取平均值得到，并将其作为最终预测结果。为确定 K 的最优值，在合理范围内利用网格搜索(Grid search)进行了优化[37]。最终，经过全面评估，选定 5 折交叉验证方案进行实际模型训练。实践证明，该集成策略有效结合了各子模型的优势，显著提升了模型的预测性能。

为了有效指导模型参数的优化与更新，本文在训练阶段选择二元交叉熵(Binary Cross-Entropy, BCE)作为损失函数[38]。BCE 专门针对二分类场景设计，其核心原理在于衡量模型预测概率与实际二进制标签之间的差异：当模型对正样本的预测概率接近 1 且对负样本的预测概率接近 0 时，损失值趋于零；反之，若预测结果与实际结果相左，损失值将显著增加。这一机制使 BCE 能够为模型提供明确的优化方向。该损失函数的定义如下：

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right) \tag{11}$$

其中，N 表示训练样本的总数，y_i 代表第 i 个样本的真实标签，取值为 0 或 1 (0 表示负样本，1 表示正样本)，$p_i \in (0,1)$ 表示模型预测该样本属于正样本的概率。在此基础上，损失值(Loss value)衡量了模型预测结果与真实标签之间的差距；损失值越小，表明模型的性能越优。

2.6. 实验设置

SSAL 模型基于 Python 3.8 和 PyTorch 1.11.0 框架实现，并在以下硬件配置上进行了训练：

CPU：16 vCPU AMD EPYC 9654 96 核处理器

GPU：NVIDIA GeForce RTX 4090 (24 GB) × 1

3. 实验结果与分析

3.1. 评估指标

本文采用五项常用于评估环状 RNA-RBP 结合预测问题的指标：受试者工作特征曲线下面积(AUC)、准确率(ACC)、精确率(Precision)、召回率(Recall)以及 F1 分数[39]-[41]。这些指标被用于全面评估 SSAL 模型的性能。其具体定义如下

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{12}$$

$$\mathrm{FRP} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{13}$$

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{14}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{15}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{16}$$

$$\mathrm{F1} = 2\times\frac{\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} \tag{17}$$

具体而言，TP (真阳性)表示真阳性样本的数量，即实际属于正类且被模型正确预测为正类的样本；TN (真阴性)对应于真阴性样本的数量，即实际属于负类且被模型准确预测为负类的样本。FP (假阳性)代表假阳性样本的数量，即实际为负类但被模型错误预测为正类的样本；FN (假阴性)则表示假阴性样本的数量，即实际为正类但被模型错误分类为负类的样本。

为了评估 SSAL 模型在不同 circRNA-RBP 相互作用场景下的适用性与鲁棒性，本文在 14 个公开基准数据集上进行了 5 折交叉验证实验。在交叉验证的每一折过程中，优化算法均会对参数进行精细调整。一旦训练集参数完成优化，模型便会对相应的验证集进行严谨的性能评估，为后续调整提供关键参考。通过计算五折实验的平均值与标准差，对模型性能进行了全面衡量。实验数据汇总于表 2。

Table 2. The experimental results of SSAL on 14 different datasets
表 2. SSAL 在 14 个不同数据集上的实验结果

数量	AUC	ACC	Precision	Recall	F1
AGO1	0.9973	0.9825	0.9859	0.9794	0.9826
AGO2	0.9967	0.9844	0.9758	0.9935	0.9846
DGCR8	0.9981	0.9857	0.9826	0.9898	0.9862
EIF4A3	0.9986	0.95	0.9986	0.901	0.9473
FMRP	0.9979	0.9771	0.9572	0.9995	0.9779
FUS	0.9964	0.9811	0.9806	0.9814	0.981
HNRNPC	0.9988	0.9938	0.9913	0.9965	0.9939
HUR	0.9974	0.9846	0.9802	0.9893	0.9847
IGF2BP1	0.997	0.9858	0.9785	0.9942	0.9863
IGF2BP2	0.9939	0.9738	0.9675	0.9813	0.9743
IGF2BP3	0.998	0.9855	0.9767	0.9951	0.9858
LIN28A	0.9949	0.9822	0.9784	0.9859	0.9821
PTB	0.998	0.9851	0.9769	0.9937	0.9852
ZC3H7B	0.998	0.9888	0.9871	0.9905	0.9888

结果表明，SSAL 在所有 14 个数据集上的 AUC 值均超过了 0.99，展现出卓越的排序能力且不依赖于分类阈值。此外，模型的准确率(ACC)、精确率(Precision)、召回率(Recall)及 F1 分数均超过了 0.9。高召回率与高精确率的双重表现表明，该模型在有效捕捉真实结合位点测的同时，能较好地控制假阳性率，充分证明了特征融合与注意力机制的设计优势。这些实验结果从多个维度证实了 SSAL 在 circRNA-RBP 结合位点预测任务中的有效性与先进性。

3.2. 与其他方法的比较

为了全面评估 SSAL 模型的预测性能，利用 5 折交叉验证，在 14 个大规模 circRNA 数据集上将 SSAL 与五种最先进(State-of-the-art)的环状 RNA 结合位点预测模型进行了系统对比。所选的基准模型涵盖了多种特征提取方法与深度学习架构，具体说明如下：

HCRNet：是一种基于深度时间卷积网络(Deep Temporal Convolutional Network)识别 circRNA-RBP 结合位点的框架[19]。

JLCRB：是一种基于多视图融合(Multi-view fusion)的 circRNA-RBP 结合位点预测方法，集成了 HNF、CircRNA2vec、PSTNP 和 DNA-BERT 等多种编码方式对 circRNA 序列进行表征[42]。

CRBPDL：利用双向门控循环单元(BiGRU)算法和自注意力机制来识别 circRNA-RBP 相互作用位点[18]。

PASSION：是一种采用混合神经网络(Hybrid neural networks)预测 circRNA-RBP 结合位点的方法[16]。

CRIP：通过基于密码子(Codon-based)的表示法对 circRNA 进行编码，并使用混合深度神经网络预测 circRNA-RBP 的结合位点[15]。

如图2和表3所示，SSAL模型在所有数据集上的表现均优于其他五种基准模型。在针对14个circRNA数据集的评估中，SSAL模型的AUC值始终高于HCRNet、JLCRB、CRBPDL、PASSION和CRIP，其平均AUC达到了0.997。与JLCRB方法相比，SSAL在14个大规模circRNA数据集上均实现了AUC值的提升，展现了卓越的性能。

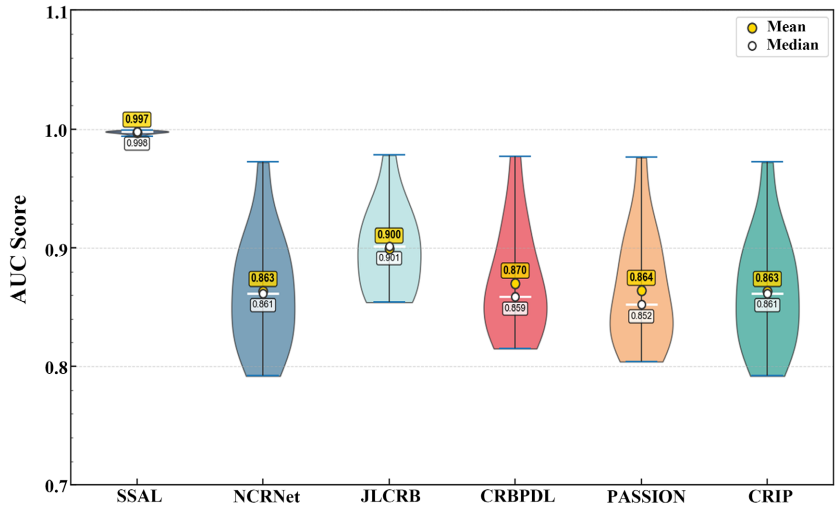

Figure 2. Average AUC values distribution of SSAL and baseline models depicted by the violin plot
图 2. SSAL与基准模型在小提琴图中的平均AUC值分布情况

Table 3. System resulting data of standard experiment
表 3. 标准试验系统结果数据

数量	SSAL	HCRNet	JLCRB	CRBPDL	PASSION	CRIP
AGO1	0.997	0.905	0.941	0.923	0.909	0.905
AGO2	0.997	0.881	0.854	0.823	0.822	0.881
DGCR8	0.998	0.914	0.930	0.924	0.917	0.914
EIF4A3	0.999	0.812	0.857	0.853	0.823	0.812
FMRP	0.998	0.898	0.930	0.897	0.900	0.898
FUS	0.996	0.858	0.900	0.862	0.859	0.858
HNRNPC	0.999	0.972	0.978	0.977	0.976	0.972
HUR	0.997	0.874	0.911	0.876	0.879	0.874
IGF2BP1	0.997	0.843	0.902	0.855	0.845	0.843
IGF2BP2	0.994	0.821	0.886	0.843	0.827	0.821
IGF2BP3	0.998	0.822	0.878	0.823	0.831	0.822
LIN28A	0.995	0.865	0.903	0.875	0.875	0.865
PTB	0.998	0.826	0.861	0.835	0.829	0.826
ZC3H7B	0.998	0.792	0.863	0.815	0.804	0.792

这种优异的性能主要归功于 SSAL 模型在特征提取阶段的创新设计。具体而言，引入了 CDPfold 工具深度挖掘 circRNA 的二级结构信息，生成了反映碱基配对概率的结构特征矩阵。与仅依赖序列信息的方法相比，CDPfold 提供的结构特征揭示了 RNA 分子内部的空间构象(Spatial conformation)与碱基相互作用模式(Base interaction patterns)，为模型补充了序列数据之外的关键维度。图 2 直观地展示了 SSAL 相对于其他模型无与伦比的预测准确性与稳定性。

3.3. 消融实验

在本节中，围绕 SSAL 模型的构建与优化进行了系统的实验分析。首先，为探究不同序列编码策略对 circRNA-RBP 结合位点预测任务的影响，对比了多种序列编码方法，以确定最优的序列表征方案。其次，在选定编码方案的基础上，进一步评估了特征提取模块与特征融合策略的有效性，并通过消融实验验证了各核心组件的必要性。最后，对所提集成模型进行了深入分析，从基学习器构成、集成策略优越性以及模型鲁棒性等方面揭示了其性能提升的内在机制。下文将从这三个研究维度展开详细阐述。

3.3.1. 不同编码方法的消融

为了评估各种环状 RNA 序列编码方法对模型性能的影响，利用规模为 20,000 的数据集(IGF2BP2)进行实验，并开展了基于 5 折交叉验证的消融研究。从编码方法库中选取了四种方法：k-核苷酸频率(KNF) [43]、异质核苷酸频率(HNF) [44]、CKSNAP [31]以及 CDPfold [33]。这四种编码方法的组合方式详见表 4。

Table 4. Ablation results on different encoding methods
表 4. 不同编码方法的消融结果

	KNF	HNF	CKSNAP	CDPfold	AUC
Encoding-1		✓	✓	✓	0.941
Encoding-2	✓		✓	✓	0.931
Encoding-3	✓	✓		✓	0.933
Encoding-4	✓	✓	✓		0.983
SSAL	✓	✓	✓	✓	0.994

为了评估不同序列编码策略的影响，对比了 SSAL 在多种组合编码方法下的表现，结果如表 4 所示。在 IGF2BP2 数据集上，SSAL 实现了 0.994 的最高 AUC 值，分别比 Encoding-1 和 Encoding-3 显著高出 5.63%和 6.54%。这一显著差异证明了编码方法的选择直接影响模型捕获序列信息的能力，同时也验证了环状 RNA 序列中确实包含与 RBP 结合相关的价值信息。然而，仅依赖 KNF 或 CKSNAP 进行特征编码虽能显著提高编码效率，但仍显不足。只有当两种序列编码方法融合时，才能全面捕获序列中的有效信息。仅利用序列信息进行特征编码是不充分的，SSAL 整合了涵盖序列与结构特征的更全面信息。Encoding-4 的 AUC 值为 0.983，在各种编码方法中名列前茅，但仍落后于 SSAL。该结果进一步突显了结构信息对提升预测性能的关键贡献，表明仅依赖序列层面的编码策略难以完全揭示 circRNA-RBP 相互作用的复杂模式。同时，SSAL 较 Encoding-2 实现了 6.66%的 AUC 提升，充分验证了将环状 RNA 序列编码为伪氨基酸序列(Pseudo-amino acid sequences)的有效性。这种方法证明其能够有效传递与蛋白质识别相关的生物学特征(如密码子偏好性)。综上所述，本研究采用的多维组合编码方法显著提升了模型的优选预测性能。

3.3.2. 不同模块的消融

进一步以 IGF2BP2 数据集为基准开展消融研究，旨在深入探究不同特征提取模块、特征融合策略以及集成模型对 SSAL 性能的影响。通过对比各变体模型在该数据集上的性能变化，可以清晰地识别出对提高预测准确性起关键作用的设计元素。

模型 1：移除序列特征提取模块，仅将结构特征输入预测模块，以评估序列信息对结合位点预测的贡献。

模型 2：移除结构特征提取模块，仅使用序列特征进行预测，以测试结构信息的必要性。

模型 3：移除注意力融合机制，在输入预测模块前将序列特征与结构特征直接拼接，以验证特征融合策略的有效性。

模型 4：不采用集成模型架构，而是直接将单个子模型的输出作为最终预测结果，旨在验证集成学习策略对提升模型稳定性和准确性的贡献。

图 3 展示了 SSAL 模型与其各变体模型在 AUC 值上的对比。在 IGF2BP2 数据集上，SSAL 实现了最高 AUC 值(0.997)，分别比模型 1 和模型 2 高出 1.5%和 57.7%，这有力地证明了整合 circRNA 序列特征与结构特征的重要性。模型 3 的 AUC 值为 0.965，比 SSAL 低 3.3%；而模型 4 达到的 AUC 为 0.991，比 SSAL 低 0.6%。

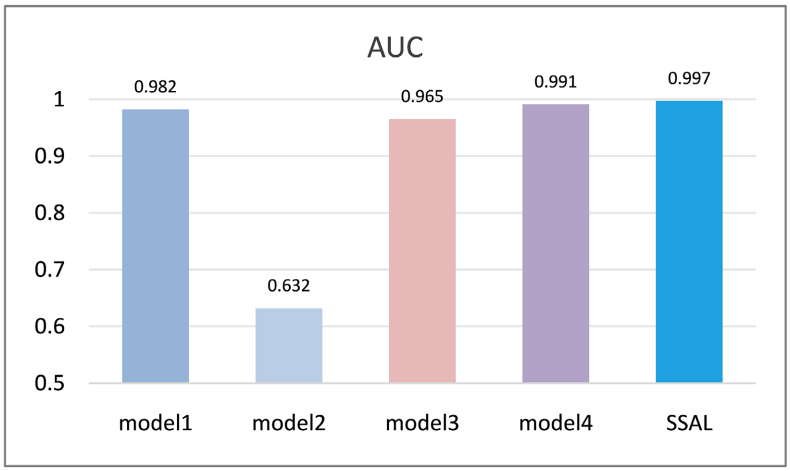

Figure 3. Ablation results on different modules
图 3. 在不同模块上的消融结果

上述结果表明，基于注意力的特征交互融合模块能够学习特征间更深层次的依赖关系，从而有效提升模型性能。同时，与模型 4 相比，SSAL 的 AUC 提升了 1.76%，这表明通过多个子模块构建集成模型可以获得更好的泛化能力和更卓越的预测表现。综上所述，SSAL 是通过整合特征提取模块、特征融合方法以及集成建模策略构建而成的。

3.4. 结果可视化

为了更直观地评估 SSAL 模型区分结合位点与非结合位点的能力，利用 t-分布邻域嵌入(t-SNE)算法，对 IGF2BP2 数据集的预测结果进行了可视化降维分析。如图 4 所示，将模型从 circRNA 序列空间和结构空间中学习到的联合特征表示映射到二维平面上，从而构建了一个统一的结合位点预测空间。图中红色点代表正样本(真实结合位点)，蓝色点代表负样本(非结合位点)。

可视化结果清晰地表明，SSAL 模型能够有效地将两类样本划分为截然不同的簇(Clusters)。正负样本

之间的边界清晰，仅存在极少数重叠区域。这种分布模式直观地证实了 SSAL 模型卓越的分类性能，表明其具备从高维特征空间中学习具有判别力(Discriminative)的特征表示的能力。

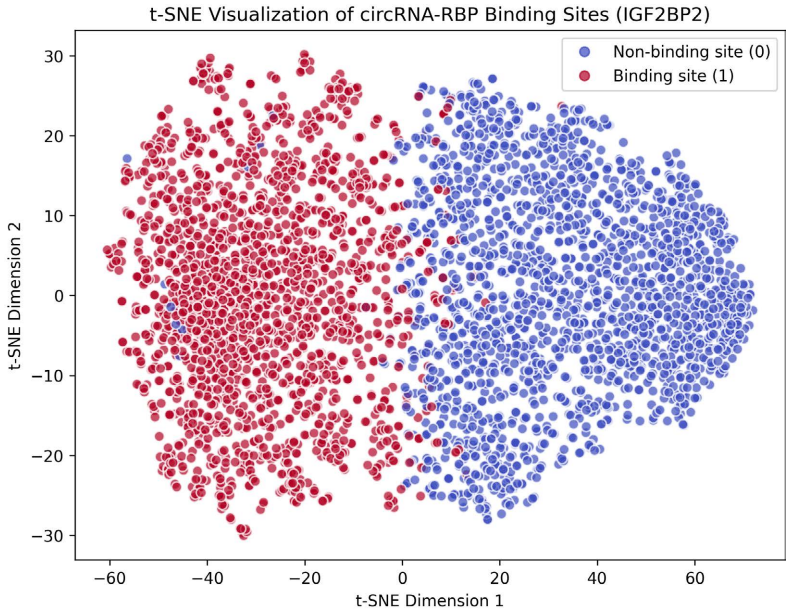

Figure 4. Visualization result of SSAL on the IGF2BP2 dataset
图 4. SSAL 在 IGF2BP2 数据集上的可视化结果

4. 结论和讨论

近年来，随着测序技术的飞速发展，不仅积累了大量与 circRNA 结合事件相关的数据集，深度学习模型在预测 circRNA 与 RBP 相互作用位点方面的应用也日益广泛[45] [46]。然而，如何灵活且高效地提取 circRNA 的多尺度特征以进一步提升预测准确性，仍然是当前面临的核心挑战。为解决这一问题，本研究提出了一种名为 SSAL 的新型 circRNA-RBP 相互作用位点预测模型，其核心功能在于高效整合 circRNA 的序列特征与结构特征。在 14 个大规模 circRNA 数据集上进行了实验，并将 SSAL 与当前最先进的预测模型进行了对比。结果表明，SSAL 在预测准确性及其他评估指标上均优于现有的先进方法。实验结果充分证实了 SSAL 预测模型的高效性与鲁棒性。

SSAL 的卓越性能可归功于三个核心因素。首先，SSAL 模型不仅提取了 circRNA 的序列特征，还引入了结构特征，这种多尺度特征显著增强了特征表征能力，使得模型能够更全面、深入地从数据中挖掘有效信息，为准确预测奠定了坚实基础。其次，SSAL 采用基于注意力机制的交互式特征融合方法处理不同序列特征，该方法能够准确识别并利用不同特征间的内在相关性，从而发掘更具生物学意义的关键特征，进一步提升了特征质量。最后，SSAL 通过集成多个独立的子模型构建了大规模集成模型。这种集成方法有效降低了仅使用单一模型预测所产生的偏差，显著增强了模型的泛化能力与预测精度。

尽管如此，SSAL 的性能仍存在进一步优化的空间。一方面，SSAL 目前使用 CDPfold 工具生成的碱基配对矩阵作为结构特征。近年来，RNA 结构信息的表征方法不断增加与优化，这意味着理论上可能存在更优的模型或方法能够更精确地捕捉 circRNA 的结构特征[47]-[49]。另一方面，目前已知的 RBP 结合位点数据有限，可能导致数据集中正负样本分布不均，从而影响预测结果。因此，未来的研究可以通过收集更多关于 circRNA-RBP 相互作用位点的生物实验数据来扩充现有数据集，同时持续探索更合适的 circRNA 结构表征方法，以进一步提升 SSAL 的预测性能。

基金项目

本研究得到了辽宁省省属高校基本科研业务费专项资金(项目编号：LJ212410150016)的资助。

参考文献

[1] Yao, Z., *et al*. (2020) Advances in the Effects of Related Specific Molecules in Circular RNA on Bladder Cancer Cells. *Journal of Clinical Urology*, **35**, 417-420.

[2] Liu, C. and Chen, L. (2022) Circular RNAs: Characterization, Cellular Roles, and Applications. *Cell*, **185**, 2016-2034. https://doi.org/10.1016/j.cell.2022.04.021

[3] Zhou, H., Zhang, H., Yan, R., *et al*. (2024) Mechanism and Role of CircRNA in Occurrence and Development of Hepatocellular Carcinoma. *Cancer Research on Prevention and Treatment*, **49**, 496-502.

[4] Sinha, T., Panigrahi, C., Das, D. and Chandra Panda, A. (2021) Circular RNA Translation, a Path to Hidden Proteome. *WIREs RNA*, **13**, e1685. https://doi.org/10.1002/wrna.1685

[5] Kelaini, S., Chan, C., Cornelius, V.A. and Margariti, A. (2021) RNA-binding Proteins Hold Key Roles in Function, Dysfunction, and Disease. *Biology*, **10**, Article 366. https://doi.org/10.3390/biology10050366

[6] Zang, J., Lu, D. and Xu, A. (2018) The Interaction of CircRNAs and RNA Binding Proteins: An Important Part of CircRNA Maintenance and Function. *Journal of Neuroscience Research*, **98**, 87-97. https://doi.org/10.1002/jnr.24356

[7] Su, J., Chen, S., Yang, S. and Deng, Z. (2024) RNA-Binding Proteins Regulate Osteoarthritis via RNA Metabolism Regulation. *Journal of Central South University* (*Medical Sciences*), **49**, 1973-1982.

[8] Peng, T. and Xu L. (2023) Crosstalk between Epigenetic Modification and CircRNA in Colorectal Cancer: Recent Advances. *Journal of Shanghai Jiao Tong University* (*Medical Science*), **43**, 237-243.

[9] Li. H., Wang, Wei. and Hao, M. (2021) Progress of CircRNAs as A Promising Biomarker and Therapeutic Target for Cervical Cancer. *Journal of International Obstetrics and Gynecology*, **48**, 322-327.

[10] Zhu, H., Jia, J., and Yu, L. (2021) Research Progress on CircRNA in Liquid Biopsy of Gastric Cancer. *Cancer Research on Prevention and Treatment*, **48**, 1023-1029.

[11] Lee, J. (2023) The Principles and Applications of High-Throughput Sequencing Technologies. *Development & Reproduction*, **27**, 9-24. https://doi.org/10.12717/dr.2023.27.1.9

[12] Kuwamoto-Imanishi, S. and Fujii, H. (2025) Online Databases in Circular RNAs. In: Xiao, J., Ed., *Advances in Experimental Medicine and Biology*, Springer, 43-57. https://doi.org/10.1007/978-981-96-9428-0_4

[13] Fan, C., Lei, X., Fang, Z., Jiang, Q. and Wu, F. (2018) CircR2Disease: A Manually Curated Database for Experimentally Supported Circular RNAs Associated with Various Diseases. *Database*, **2018**, bay044. https://doi.org/10.1093/database/bay044

[14] Meng, X., Hu, D., Zhang, P., Chen, Q. and Chen, M. (2019) CircFunBase: A Database for Functional Circular RNAs. *Database*, **2019**, baz003. https://doi.org/10.1093/database/baz003

[15] Zhang, K., Pan, X., Yang, Y. and Shen, H. (2019) CRIP: Predicting CircRNA-RBP-Binding Sites Using a Codon-Based Encoding and Hybrid Deep Neural Networks. *RNA*, **25**, 1604-1615. https://doi.org/10.1261/rna.070565.119

[16] Jia, C., Bi, Y., Chen, J., Leier, A., Li, F. and Song, J. (2020) PASSION: An Ensemble Neural Network Approach for Identifying the Binding Sites of RBPs on CircRNAs. *Bioinformatics*, **36**, 4276-4282. https://doi.org/10.1093/bioinformatics/btaa522

[17] Yang, Y., Hou, Z., Ma, Z., Li, X. and Wong, K. (2020) iCircRBP-DHN: Identification of CircRNA-RBP Interaction Sites Using Deep Hierarchical Network. *Briefings in Bioinformatics*, **22**, bbaa274. https://doi.org/10.1093/bib/bbaa274

[18] Niu, M., Zou, Q. and Lin, C. (2022) CRBPDL: Identification of CircRNA-RBP Interaction Sites Using an Ensemble Neural Network Approach. *PLOS Computational Biology*, **18**, e1009798. https://doi.org/10.1371/journal.pcbi.1009798

[19] Yang, Y., Hou, Z., Wang, Y., Ma, H., Sun, P., Ma, Z., *et al*. (2022) HCRNet: High-Throughput CircRNA-Binding Event Identification from CLIP-Seq Data Using Deep Temporal Convolutional Network. *Briefings in Bioinformatics*, **23**, bbac027. https://doi.org/10.1093/bib/bbac027

[20] Cao, C., Yang, S., Li, M. and Li, C. (2023) CircSSNN: CircRNA-Binding Site Prediction via Sequence Self-Attention Neural Networks with Pre-Normalization. *BMC Bioinformatics*, **24**, Article No. 220. https://doi.org/10.1186/s12859-023-05352-7

[21] Wang, X., Yu, S., Lou, E., Tan, Y. and Tan, Z. (2023) RNA 3D Structure Prediction: Progress and Perspective. *Molecules*, **28**, 5532. https://doi.org/10.3390/molecules28145532

[22] Shen, T., Hu, Z., Sun, S., Liu, D., Wong, F., Wang, J., *et al*. (2024) Accurate RNA 3D Structure Prediction Using a

Language Model-Based Deep Learning Approach. *Nature Methods*, **21**, 2287-2298. https://doi.org/10.1038/s41592-024-02487-0

[23] Mukherjee, S., Moafinejad, S.N., Badepally, N.G., Merdas, K. and Bujnicki, J.M. (2024) Advances in the Field of RNA 3D Structure Prediction and Modeling, with Purely Theoretical Approaches, and with the Use of Experimental Data. *Structure*, **32**, 1860-1876. https://doi.org/10.1016/j.str.2024.08.015

[24] Zheng, J., Liu, H., Feng, Y., Xu, J. and Zhao, L. (2023) CASF-Net: Cross-Attention and Cross-Scale Fusion Network for Medical Image Segmentation. *Computer Methods and Programs in Biomedicine*, **229**, Article 107307. https://doi.org/10.1016/j.cmpb.2022.107307

[25] Xia, S., Zhang, X., Meng, H. and Jiao, L. (2024) Ternary Modality Contrastive Learning for Hyperspectral and Lidar Data Classification. *IEEE Transactions on Geoscience and Remote Sensing*, **62**, 1-17. https://doi.org/10.1109/tgrs.2024.3417011

[26] Hamed, S.K., Ab Aziz, M.J. and Yaakub, M.R. (2023) A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion. *Heliyon*, **9**, e20382. https://doi.org/10.1016/j.heliyon.2023.e20382

[27] Dudekula, D.B., Panda, A.C., Grammatikakis, I., *et al.* (2016) CircInteractome: A Web Tool for Exploring Circular RNAs and Their Interacting Proteins and microRNAs. *RNA Biology*, **13**, 34-42. https://doi.org/10.1080/15476286.2015.1128065

[28] Wei, Y., Zhang, Q. and Liu, L. (2025) The Improved De Bruijn Graph for Multitask Learning: Predicting Functions, Subcellular Localization, and Interactions of Noncoding RNAs. *Briefings in Bioinformatics*, **26**, bbae627. https://doi.org/10.1093/bib/bbae627

[29] Uhl, M., Houwaart, T., Corrado, G., *et al.* (2017) Computational Analysis of CLIP-seq Data. *Methods*, **118**, 60-72. https://doi.org/10.1016/j.ymeth.2017.02.006

[30] Zhang, M., Wang, T., Xiao, G. and Xie, Y. (2020) Large-Scale Profiling of Rbp-CircRNA Interactions from Public Clip-Seq Datasets. *Genes*, **11**, Article 54. https://doi.org/10.3390/genes11010054

[31] Li, M., Fan, Y., Zhang, Y. and Lv, Z. (2022) Using Sequence Similarity Based on CKSNP Features and a Graph Neural Network Model to Identify MiRNA-Disease Associations. *Genes*, **13**, Article 1759. https://doi.org/10.3390/genes13101759

[32] Amerifar, S., Norouzi, M. and Ghandi, M. (2022) A Tool for Feature Extraction from Biological Sequences. *Briefings in Bioinformatics*, **23**, bbac108. https://doi.org/10.1093/bib/bbac108

[33] Mishra, R. (2024) Deep Learning Based Convolute Neural Approach in the Prediction of RNA Structure. 2024 *IEEE International Conference on Big Data & Machine Learning* (*ICBDML*), Bhopal, 24-25 February 2024, 86-90. https://doi.org/10.1109/icbdml60909.2024.10577366

[34] Wei, Y., Tan, Z. and Liu, L. (2025) Cr-Deal: Explainable Neural Network for CircRNA-RBP Binding Site Recognition and Interpretation. *Interdisciplinary Sciences*: *Computational Life Sciences*, **17**, 463-476. https://doi.org/10.1007/s12539-025-00694-7

[35] Liu, L., Wei, Y., Tan, Z., Zhang, Q., Sun, J. and Zhao, Q. (2024) Predicting CircRNA-RBP Binding Sites Using a Hybrid Deep Neural Network. *Interdisciplinary Sciences*: *Computational Life Sciences*, **16**, 635-648. https://doi.org/10.1007/s12539-024-00616-z

[36] Vu, H.L., Ng, K.T.W., Richter, A. and An, C. (2022) Analysis of Input Set Characteristics and Variances on K-Fold Cross Validation for a Recurrent Neural Network Model on Waste Disposal Rate Estimation. *Journal of Environmental Management*, **311**, Article 114869. https://doi.org/10.1016/j.jenvman.2022.114869

[37] Sun, Y., Ding, S., Zhang, Z. and Jia, W. (2021) An Improved Grid Search Algorithm to Optimize SVR for Prediction. *Soft Computing*, **25**, 5633-5644. https://doi.org/10.1007/s00500-020-05560-w

[38] Li, Q., Jia, X., Zhou, J., Shen, L. and Duan, J. (2024) Rediscovering BCE Loss for Uniform Classification. arXiv:2403.07289.

[39] Obi, J.C. (2023) A Comparative Study of Several Classification Metrics and Their Performances on Data. *World Journal of Advanced Engineering Technology and Sciences*, **8**, 308-314. https://doi.org/10.30574/wjaets.2023.8.1.0054

[40] Naidu, G., Zuva, T. and Sibanda, E.M. (2023) A Review of Evaluation Metrics in Machine Learning Algorithms. In: Silhavy, R. and Silhavy, P., Eds., *Lecture Notes in Networks and Systems*, Springer International Publishing, 15-25. https://doi.org/10.1007/978-3-031-35314-7_2

[41] Chicco, D. and Jurman, G. (2023) The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification. *BioData Mining*, **16**, Article No. 4. https://doi.org/10.1186/s13040-023-00322-4

[42] Du, X. and Xue, Z. (2022) JLCRB: A Unified Multi-View-Based Joint Representation Learning for CircRNA Binding Sites Prediction. *Journal of Biomedical Informatics*, **136**, Article 104231. https://doi.org/10.1016/j.jbi.2022.104231

[43] Gibertini, E. and Magagnin, L. (2022) PEDOTS:PSS@KNF Wire-shaped Electrodes for Textile Symmetrical Capacitor. *Advanced Materials Interfaces*, **9**, Article 2200513. https://doi.org/10.1002/admi.202200513

[44] Orenstein, Y., Wang, Y. and Berger, B. (2016) RCK: Accurate and Efficient Inference of Sequence- and Structure-Based Protein-RNA Binding Models from Rnacompete Data. *Bioinformatics*, **32**, i351-i359. https://doi.org/10.1093/bioinformatics/btw259

[45] Wang, Z., Lei, X., Zhang, Y., Wu, F. and Pan, Y. (2025) Recent Progress of Deep Learning Methods for RBP Binding Sites Prediction on CircRNA. *Current Bioinformatics*, **20**, 487-505. https://doi.org/10.2174/0115748936308564240712053215

[46] Wang, Z. and Lei, X. (2021) Prediction of RBP Binding Sites on CircRNAs Using an LSTM-Based Deep Sequence Learning Architecture. *Briefings in Bioinformatics*, **22**, bbab342. https://doi.org/10.1093/bib/bbab342

[47] Liu, X., Wang, S., Sun, Y., Liao, Y., Jiang, G., Sun, B., *et al*. (2025) Unlocking the Potential of Circular RNA Vaccines: A Bioinformatics and Computational Biology Perspective. *eBioMedicine*, **114**, Article 105638. https://doi.org/10.1016/j.ebiom.2025.105638

[48] Chen, X. and Huang, L. (2022) Computational Model for NcRNA Research. *Briefings in Bioinformatics*, **23**, bbac472. https://doi.org/10.1093/bib/bbac472

[49] Wang, X., Yu, C., You, Z., Qiao, Y., Li, Z. and Huang, W. (2023) An Efficient CircRNA-MiRNA Interaction Prediction Model by Combining Biological Text Mining and Wavelet Diffusion-Based Sparse Network Structure Embedding. *Computers in Biology and Medicine*, **165**, Article 107421. https://doi.org/10.1016/j.compbiomed.2023.107421