计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(7): 1726−1742, 2026

人工智能在 RNA 研究中的应用

顾睿初[1] 梁雅琳[2] 温 翰[1,3]
[1]（北京科学智能研究院 北京 100084）
[2]（北京大学信息管理系 北京 100871）
[3]（北京大学核糖核酸北京研究中心 北京 100871）
（sjtugrc@163.com）

Applications of Artificial Intelligence in RNA Research

Gu Ruichu[1], Liang Yalin[2], and Wen Han[1,3]
[1]（*AI for Science Institute*, *Beijing* 100084）
[2]（*Department of Information Management*, *Peking University*, *Beijing* 100871）
[3]（*Beijing Advanced Center of RNA Biology* (*BEACON*), *Peking University*, *Beijing* 100871）

Abstract Ribonucleic acid (RNA) plays a central role in the regulation of biological processes, and its systematic investigation is essential for understanding disease mechanisms at molecular level and developing innovative therapeutic strategies. Nevertheless, progress in RNA research has been hindered by the structural and functional complexity of RNA molecules, the explosive growth of high-throughput sequencing data, and the intrinsic limitations of conventional experimental techniques. In recent years, artificial intelligence (AI) has emerged as a powerful enabling technology, offering novel solutions through its strengths in pattern recognition, high-dimensional data analysis, and predictive modeling. This review provides a comprehensive overview on the major applications of AI in RNA research, encompassing RNA sequence element optimization, RNA modification sites prediction, RNA secondary/tertiary structures prediction, and computational modeling of RNA-protein interactions. Furthermore, the key challenges facing AI-driven RNA studies are critically examined, including the limited availability of high-quality annotated datasets and the insufficient interpretability of complex learning models. Finally, future perspectives are outlined, emphasizing integrated data resource development, algorithmic and model-level innovations, and the construction of closed-loop translational frameworks, with the goal of fostering deeper integration between AI methodologies and RNA biology, accelerating both fundamental discoveries and the development of next-generation RNA-based therapeutics.

Key words ribonucleic acid (RNA); artificial intelligence (AI); sequence optimization; structure prediction; RNA modification

摘 要 核糖核酸（ribonucleic acid, RNA）作为生命活动的核心调控分子，其研究对于疾病机制解析与创新药物研发至关重要。然而，RNA 结构与功能的复杂性、高通量测序产生的海量数据，以及传统实验方法的局限，共同制约了 RNA 研究的深入发展。近年来，人工智能技术（artificial intelligence, AI）凭借其强大的模式识别、高维数据挖掘与预测建模能力，为 RNA 研究提供了创新解决方案。系统综述了 AI 在

收稿日期：2026-01-05；**修回日期**：2026-06-03
基金项目：上海市科技创新行动计划（24JS2820200）；国家重点研发计划（2023YFC3403200）
This work was supported by Shanghai Action Plan for Science, Technology and Innovation (24JS2820200) and the National Key Research and Development Program of China (2023YFC3403200).
通信作者：温翰（wenh@aisi.ac.cn）

RNA研究中的核心应用，涵盖RNA序列元件优化、RNA修饰位点预测、RNA二级与三级结构预测，以及RNA-蛋白质相互作用建模等关键方向。进一步剖析了当前AI驱动RNA研究面临的核心挑战，包括高质量标注数据资源匮乏、模型可解释性不足等；并在此基础上，从数据资源整合、模型与算法优化、干湿实验闭环转化体系构建等维度明确未来发展路径，以期推动AI与RNA研究的深度融合，加速基础科学发现与创新疗法的研发进程。

关键词 核糖核酸；人工智能；序列优化；结构预测；RNA修饰

中图法分类号 TP399

DOI: 10.7544/issn1000-1239.202660024 **CSTR:** 32373.14.issn1000-1239.202660024

1 RNA分子的性质与功能

核糖核酸(ribonucleic acid, RNA)是一类由核苷酸通过磷酸二酯键连接形成的生物分子，其基础结构如图1所示。这类分子通常以单链形式存在，依赖碱基互补配对形成稳定的二级结构及更为复杂的高级结构。核苷酸序列、空间构象、化学修饰状态及其动态变化过程共同决定了RNA的生物学功能。

细胞内存在多种功能特异的RNA分子，主要的RNA分子类型如图2所示。不同类型的RNA分工协作，保障生命活动有序进行。信使RNA(messenger RNA, mRNA)是最主要的编码RNA，负责将DNA中储存的遗传信息转录并传递至核糖体，作为蛋白质翻译的模板指导具有特定生物学功能的蛋白质的合成。不编码蛋白质的RNA统称为非编码RNA(non-coding RNA, ncRNA)。其中，核糖体RNA(ribosomal RNA, rRNA)是核糖体的重要结构与功能组分，并在蛋白质合成过程中发挥催化作用；转运RNA(transfer RNA, tRNA)通过反密码子与mRNA密码子的特异性识别，将相应氨基酸准确递送至核糖体；小核RNA(small nuclear RNA, snRNA)主要参与真核生物前体mRNA的剪接过程，促进内含子的切除及外显子的正确连接。

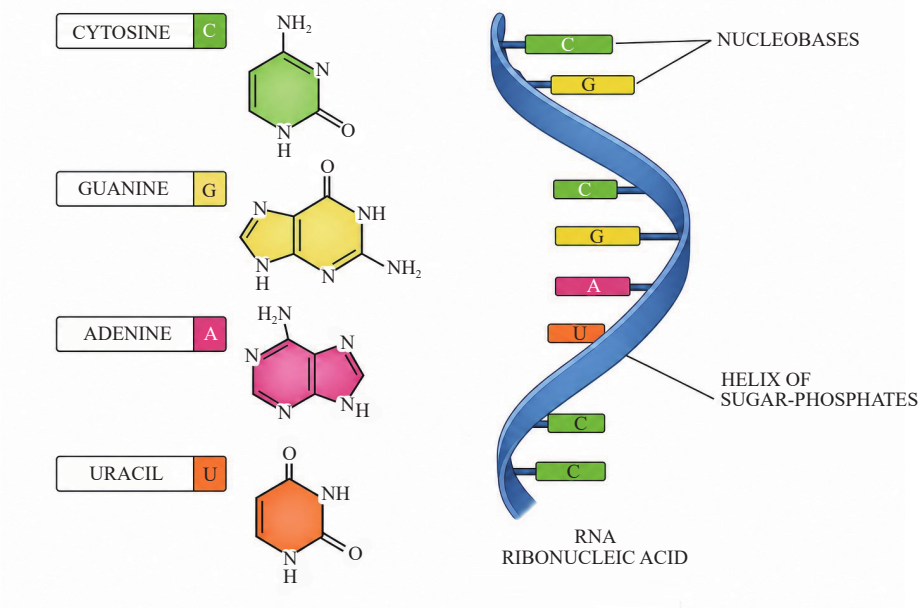


Fig. 1 Basic structure of RNA molecule

图1 RNA分子基础结构

RNA在基因表达调控、细胞命运决定乃至机体稳态维持过程中发挥重要作用。近年来大量研究表明，RNA异常与多种重大疾病的发生发展密切相关。RNA表达失衡、剪接异常、结构稳态破坏及化学修饰紊乱，均可参与肿瘤、遗传性疾病、神经退行性疾病、代谢性疾病及感染性疾病等病理过程[1]。因此，RNA已成为揭示生命活动基本规律和阐明疾病分子机制的重要研究对象。

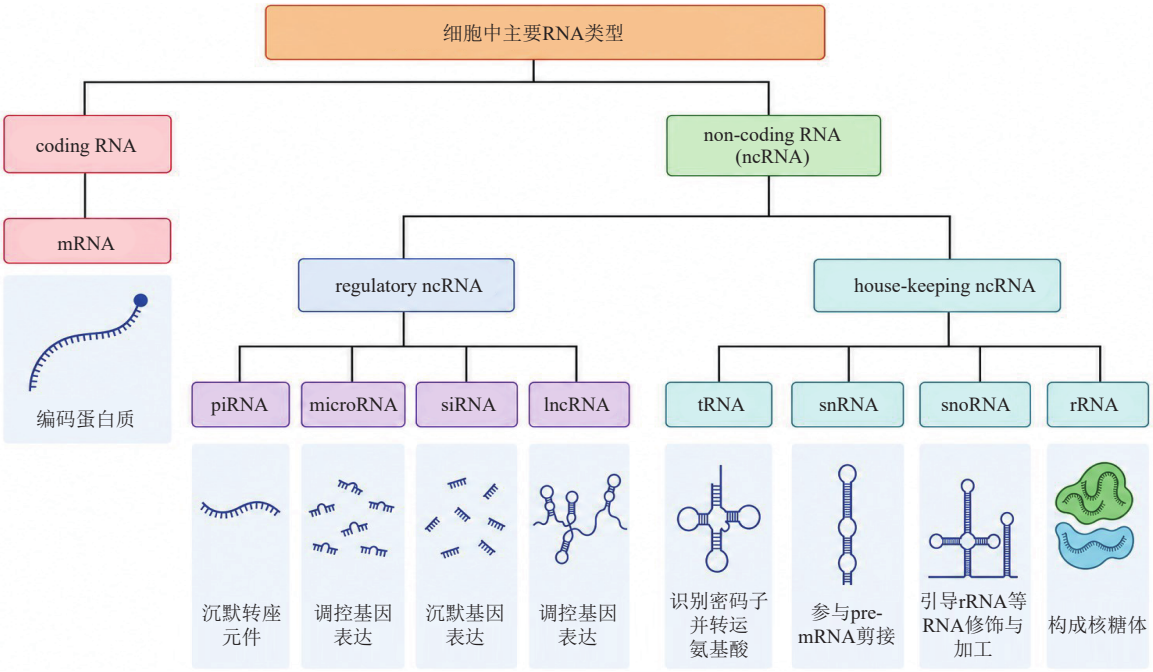


Fig. 2 Different types of RNA

图 2 不同种类的 RNA

与此同时，传统药物研发长期面临部分致病靶点难以成药的挑战。部分致病蛋白由于缺乏稳定的结合口袋、构象变化复杂，或在维持正常生理功能中发挥关键作用，因而难以通过传统小分子药物或抗体进行有效调控。相较于直接靶向蛋白质，RNA 位于基因表达调控的关键中间环节，可在转录后加工、翻译调控等多个层面实现干预，从而为绕开不可靶向蛋白、抑制致病通路提供新的策略。以反义寡核苷酸（antisense oligonucleotide, ASO）和小干扰 RNA（small interfering RNA, siRNA）为代表的 RNA 靶向技术，可通过特异性识别靶 RNA 并促进其降解或抑制其翻译，进而下调致病蛋白的表达水平。

此外，凭借高度可设计与可编程的分子特征，RNA 可被改造以编码肿瘤新抗原及个体化特异性抗原。递送编码特定抗原的 mRNA 至宿主细胞可促使其表达相应抗原并诱导免疫应答，这为肿瘤疫苗研发和精准治疗开辟了新的技术路径。

当前，RNA 已成为基础研究、临床转化和生物医药产业共同关注的重要方向。然而，随着 RNA 研究的不断深入，其内在复杂性所带来的技术与方法学挑战也日益凸显。RNA 并非静态的一维序列分子，其生物学功能通常由序列信息、二级与三级结构、化学修饰状态，以及与蛋白质、小分子和其他核酸之间的相互作用等共同决定。更为重要的是，这些因素具有显著的情境依赖性，可随细胞类型、发育阶段、疾病状态及外界刺激的变化而发生动态改变。因此，RNA 研究本质上面对的是一个高维、多尺度、动态耦合的复杂体系。

近年来，高通量测序、单细胞组学、空间转录组学等技术快速发展，推动 RNA 相关数据以前所未有的速度积累。在此背景下，如何从海量、异质且噪声复杂的数据中提取稳健、可靠且具有生物学解释意义的规律，已成为当前 RNA 研究亟待解决的关键问题。传统依赖实验验证和规则驱动分析的研究模式，普遍存在成本较高、周期较长、通量有限及跨模态整合能力不足等局限，已难以满足 RNA 功能解析和疾病机制研究以及药物筛选对规模化、系统化及高精度分析的现实需求。而发展面向 RNA 研究的人工智能（artificial intelligence, AI）方法是突破现有研究瓶颈的关键举措。

在 RNA 研究中，AI 方法有以下突出优势：

1）AI 在刻画 RNA 复杂生物学特性方面具有显著优势。传统建模方式往往依赖较强的手工特征或显式假设，难以适配 RNA 所呈现的高维、跨尺度与动态耦合特性。深度学习模型减少对人工特征工程的依赖，可利用非线性表征学习与多尺度特征提取能力，从数据中挖掘传统方法难以捕获的关联模式，从而为 RNA 调控网络解析与致病机制研究提供新的技术路径。

2）AI 与 RNA 领域数据多源化、海量化的增长趋

势具有良好的契合性。机器学习与深度学习模型可辅助开展数据质量评估、异常检测、表征学习与多源信息融合等任务，提升分析流程的自动化水平，从而显著减少人工处理与重复计算所需的时间与人力投入，提升研究效率。

凭借在大规模数据处理与复杂特征自动学习方面的优势，AI 方法能够学习并表征 RNA 序列、结构与功能之间的内在关系，支撑生物标志物发现、mRNA 疫苗候选序列设计，以及 ASO 等寡核苷酸药物的序列优化与候选筛选，在 RNA 药物研发与精准诊疗领域展现出重要的应用潜力。AI 驱动 RNA 研究的核心任务与代表性方法系统梳理如表 1 所示。

Table 1　Core Tasks and Typical Methods in AI for RNA Research

表 1　人工智能在 RNA 研究中的核心任务与典型方法

子领域	任务	典型模型与方法
RNA 表征	序列表征	独热编码、概率矩阵、MSA、PSSM、预训练嵌入
	结构表征	邻接矩阵、连通性表、点括号标注法
RNA 序列设计	元件优化	CDS 区域优化：CDSfold[2]、LinearDesign[3]、RNop[4]、CodonBERT[5] 等；UTR 区域优化：UTRGAN[6]；多区域整合优化：iDRO[7]、LinearDesign2[8] 等
	修饰位点预测	单一位点预测：传统机器学习模型（SRAMP[9]、iRNA-Methyl[10] 等）、深度学习模型（EDLm(6)APred[11] 等）；多位点联合预测：RNA-ModX[12] 等
RNA 结构预测	二级结构预测	一维方法（建模为多分类问题）：DMfold[13] 等；二维方法（建模为语义分隔问题）：UFold[14] 等基于 U-Net 的模型
	三级结构预测	AI 辅助分子模拟：预测接触图引导构象采样（SPOT-RNA-2D[15] 等）、候选构象打分（ARES[16] 等）
RNA-其他分子相互作用预测	RNA-蛋白质相互作用预测	DeepBind[17]、iDeep[18]、HydRA[19]、RPITER[20] 等

综合来说，RNA 相关研究的兴起正在推动生物学发展，拓宽现代药物研发的技术版图，而与 AI 的结合有望进一步释放 RNA 研究与 RNA 相关药物开发的潜力，推动医药领域迈入继小分子药物和抗体药物之后的“第三次浪潮”，并为多种重大疾病的精准治疗提供全新的解决方案。

2　RNA 序列分析与优化

RNA 作为连接遗传信息与蛋白质功能的核心分子，其生物学功能的精准实现，依赖于序列功能元件的精细调控，以及多种转录后修饰的协同配合。现有实验技术在成本、操作周期及检测分辨率等方面仍存在固有局限。在此背景下，AI 方法凭借其在复杂模式识别、高维数据挖掘方面的独特优势，已逐渐成为功能元件优化及修饰位点精准解析的核心技术工具，为系统阐释 RNA 调控的分子机制、加速 RNA 药物的研发进程提供了创新性技术路径。

针对 RNA 序列的特征表征，研究者已经提出并发展出多种较为成熟的编码策略以满足不同计算任务和建模需求，常用的表征方式如表 2 所示。

表 2 中，最基础且应用最广泛的方法是独热编码，该方法对 RNA 序列中的 4 种标准碱基 A、U、C、G 分别分配唯一的 4 维二进制向量，即仅在对应碱基维度取值为 1，其余维度为 0。通过遍历 RNA 序列并逐位将碱基符号映射为相应的独热向量，可将原始符号序列转换为一个维度为“序列长度 × 4”的数值矩阵，为后续机器学习和深度学习模型提供结构清晰、无歧义的输入表示。

Table 2　Representation of RNA Sequence

表 2　RNA 序列表征方式

特征类型	说明
独热编码	在欧式空间中的离散特征编码
概率矩阵	基于热力学原理计算所得碱基配对概率矩阵
MSA	多物种序列间的相似性关系
PSSM	序列各位置的碱基频率矩阵
预训练嵌入	大规模预训练模型得到的潜在序列表示

此外，在实际测序数据和生物学研究中，RNA 序列中常包含表示碱基不确定性或多重可能性的符号。为此，国际纯粹与应用化学联合会（International Union of Pure and Applied Chemistry, IUPAC）制定了统一的核苷酸编码标准，用以规范描述单一碱基或多种碱基组合的情形，例如用特定字符表示“2 种或多种碱基之一”或“任意碱基”，具体编码规则如表 3 所示。该标准在处理高通量测序数据、变异位点标注及复杂 RNA 序列建模中具有重要意义，并为更灵活、信息更丰富的 RNA 特征编码提供了通用框架。

基于多序列比对（multiple sequence alignment, MSA）构建的二维特征是另一常用的 RNA 序列表征方法，

Table 3 IUPAC Codes of DNA/RNA Bases

表 3 DNA / RNA 碱基的 IUPAC 编码

IUPAC 碱基代码	所代表的碱基
A	腺嘌呤（Adenine，A）
C	胞嘧啶（Cytosine，C）
G	鸟嘌呤（Guanine，G）
T	胸腺嘧啶（Thymine，T）
U	尿嘧啶（Uracil，U）
R	A 或 G
Y	C 或 T（U）
S	G 或 C
W	A 或 T（U）
K	G 或 T（U）
M	A 或 C
B	C 或 G 或 T（U）
D	A 或 G 或 T（U）
H	A 或 C 或 T（U）
V	A 或 C 或 G
N	任意碱基
. 或 –	缺口（gap）

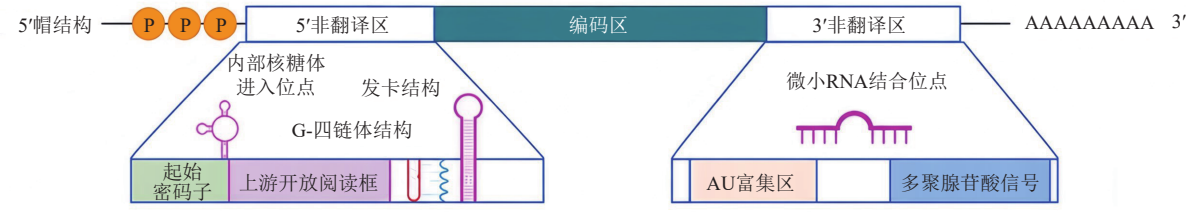
其核心优势在于能捕捉序列的进化保守性与残基间协同变异信息，典型代表为位置特异性计分矩阵（position-specific scoring matrix，PSSM）。此外，基于自由能热力学模型的算法（如 LinearPartition[21]）计算得到的碱基配对概率矩阵，也能有效表征 RNA 的二级结构潜在特征，为后续结构预测提供关键先验信息。近年来，随着预训练模型的发展，预训练模型也被用于将 RNA 序列编码为高维嵌入向量[22]。预训练模型学习到的高阶潜在表示，能够整合序列上下文依赖、进化规律及潜在结构模式等多维度信息，在下游任务中具有优异的表现。

2.1 RNA 序列元件优化

RNA 序列元件优化是指基于特定生物学功能目标，通过理性设计、计算建模或实验筛选等手段，对 RNA 分子的特定结构域或功能元件的核苷酸序列进行改造与优化，以提升其在目标场景下的活性、稳定性、特异性或调控效率。

mRNA 元件优化是 RNA 序列元件优化的核心分支，更是 mRNA 疫苗研发中的关键核心环节之一。其核心逻辑在于通过精准优化 mRNA 分子的核心功能元件，包括编码区的密码子序列、非编码区的调控序列等，使疫苗携带的外源抗原基因在宿主细胞内获得更高效的翻译效率，增强 mRNA 分子的体内稳定性，并降低免疫原性，最终实现疫苗免疫保护效果的提升。mRNA 各元件的传统优化目标如图 3 所示。

传统 mRNA 元件优化方法存在显著局限性：不仅实验周期冗长、操作成本高昂，还多依赖研究者的经验驱动进行定向改造，缺乏系统性的科学设计支撑。更关键的是，传统策略常采用单指标优化或分步优化模式，难以兼顾翻译效率提升、分子稳定性增强、免疫原性降低等多目标的动态平衡。而 AI 的融入为 mRNA 元件优化带来了突破性变革，其能够基于海量生物数据自主学习 RNA 序列与功能的内在关联特征，构建多目标协同优化框架，实现对多个元



(a) 合成mRNA的关键元件

帽结构	5′非翻译区	编码区	3′非翻译区	多聚腺苷酸尾
采用Cap 1、Cap 2等甲基化帽结构并结合高效加帽策略，通常有助于提高翻译效率，并降低外源RNA诱导的非预期先天免疫激活。	通过影响核糖体募集与5′端扫描过程调控翻译起始；Kozak序列优化通常有助于增强起始效率，而复杂二级结构或上游开放阅读框（uORF）可能产生干扰。	密码子优化可协同提升翻译效率并改善mRNA胞内稳定性。	终止密码子邻近序列以及AU富集元件（ARE）、miRNA结合位点等顺式调控元件，可影响mRNA稳定性与表达效率；对潜在不利元件进行筛选与优化通常有助于提升表达量。	尾长影响mRNA脱腺苷化降解速率与翻译环化效率，常用设计长度约为100～150 nt。
	核苷酸修饰通常可降低先天免疫激活			

(b) 主要设计特征

Fig. 3 mRNA sequence optimization

图 3 mRNA 序列优化

件的同步精准优化，大幅提升优化效率与设计的科学性。

2.1.1 编码序列优化

编码序列（coding sequence, CDS）是基因序列中能够编码蛋白质的核心区域，该区域的优化通常以提升表达效率和RNA稳定性作为核心目标。其中，RNA稳定性常用最小自由能（minimum free energy, *MFE*）和平均未配对概率（average unpaired probability, *AUP*）进行定量表征，2个指标的定义如下。

1）*MFE*

*MFE*是RNA分子折叠形成二级结构时，自发趋向的热力学能量最低状态所对应的自由能值。*MFE*对应的结构通常认为是最可能的天然折叠构象。*MFE*是衡量mRNA二级结构稳定性的常用指标。该值越低，代表mRNA形成的二级结构越稳定。

2）*AUP*

*AUP*代表序列中核苷酸的平均未配对概率，是评估mRNA稳定性的重要指标。*AUP*值越低，代表核苷酸配对越充分，形成的mRNA结构越稳定。

RNA表达效率则以相对同义密码子使用（relative synonymous codon usage, *RSCU*）频率、密码子适应指数（codon adaptation index, *CAI*）、tRNA适应指数（tRNA adaptation index, *tAI*）或密码子稳定系数（codon stability coefficient, *CSC*）作为指标。各个指标的定义如下。

1）*RSCU*

*RSCU*是某一同义密码子的实际使用频率与理论预期频率（每种密码子均匀使用）的比值。其计算公式为：

$$RSCU_{ij}=\frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij}},$$

其中，X_{ij}代表第i种氨基酸的第j种密码子的实际出现次数，n_i代表第i种氨基酸对应的同义密码子个数。*RSCU*＞1表示该密码子被偏好使用；*RSCU*=1表示密码子无偏好使用；*RSCU*＜1表示该密码子被避免使用。在序列优化时，应让目标基因的密码子的*RSCU*分布贴近宿主表达系统的偏好，使得密码子使用频率的与宿主细胞内对应tRNA的丰度匹配，以此减少核糖体在翻译过程中的“等待时间”，降低翻译延宕概率，提升蛋白质合成效率。

2）*CAI*

*CAI*是衡量序列整体优化程度的核心指标，其计算公式为：

$$CAI=\frac{CAI_{\mathrm{obs}}}{CAI_{\max}}=\frac{\left(\prod_{k=1}^{L}RSCU_k\right)^{1/L}}{\left(\prod_{k=1}^{L}RSCU_{k\max}\right)^{1/L}},$$

其中，$RSCU_k$代表基因中第k个密码子的实际*RSCU*值，$RSCU_{k\max}$代表该密码子对应的最大*RSCU*值，L代表基因中的密码子总数；CAI_{obs}为基因的实际*CAI*值，$CAI_{\max}$为理论最大*CAI*值。*CAI*取值范围介于0至1之间，取值越接近1代表密码子优化效果越优。

3）*tAI*

*tAI*以tRNA丰度为核心，评估密码子与细胞内tRNA供应的匹配度，进而预测翻译效率。由于直接测量tRNA丰度难度较大，研究中通常用tRNA基因数量或tRNA测序获得的tRNA水平作为替代指标。

需注意的是，*tAI*可能存在局部偏差。例如，内源性mRNA的前30~50个密码子中常富集稀有密码子。这一现象可通过翻译斜坡假说解释，即初始阶段缓慢翻译有助于核糖体正确组装。因此，对CDS的起始区域进行密码子优化时，建议降低*tAI*的权重。

4）*CSC*

*CSC*是mRNA序列优化中用于量化密码子对翻译过程“稳定性”贡献的关键指标。其计算逻辑为：首先对收集到的所有CDS进行密码子拆分，分别统计61种有义密码子（排除3种终止密码子）在每个基因mRNA序列中的出现频率；采用相关性分析方法（常用皮尔逊相关系数），逐一计算每种密码子的出现频率与对应mRNA半衰期之间的相关系数，这个系数就是该密码子的*CSC*值。

AI方法通过构建多目标优化框架，在解空间中高效搜索最优解。CDSfold模型基于热力学参数和动态规划算法[2]，使用*MFE*与*CAI*结合的打分函数评估序列优劣。在满足氨基酸序列编码约束的前提下，搜索二级结构稳定且具有高表达量的编码序列。但该模型计算复杂度为序列长度的三次方，因此只适用于短序列。而LinearDesign算法结合动态规划和语法解析[3]。首先将蛋白质序列对应的众多mRNA候选序列用确定有限自动机进行转化。随后根据RNA折叠语法规则和密码子使用偏好，采用计算语言学中的网格解析技术在mRNA网络图上进行最优序列搜索。该算法可在序列长度的二次方时间内找到*MFE*最小的最优序列，若采用启发式波束搜索，甚至能在“线性时间”内获得近优解，在平衡mRNA的稳定性和翻译效率上有较好的表现。尽管

如此，基于动态规划的方法仍然固有地受限于无法捕获同义位点之间的高阶上位性相互作用。深度学习架构可以超越动态规划方法的限制，RNop 是一种视觉 Transformer 模型，将 CDS 编码为二维位置频率图，通过设计 L_{GP}（GC 含量惩罚）、L_{CAI}、L_{tAI} 和 L_{MFE} 四种损失函数，实现对于序列保真度、物种特异性密码子适应性、翻译效率和二级结构稳定性多指标的同时优化[4]。

与此同时，深度学习模型可以通过整合大规模实验数据，挖掘序列中隐藏的规律，从全局视角对序列进行优化。CodonBERT 是一个基于 ProteinBERT 架构（BERT 变体），用于 mRNA 密码子优化的大语言模型[5]。研究团队首先从人类蛋白质图谱（human protein atlas, HPA）数据库提取 RNA 测序数据，并筛选出其中高表达、高 *CAI*（≥0.7）、低 *MFE*（≤−200）且长度处于特定范围内的转录本。将筛选后的 HPA 序列与 Java 密码子适配工具这一经典密码子优化工具处理后的序列按不同比例混合构建训练集，让模型同时学习“自然高表达特征”和“人工优化特征”。模型采用多头注意力 Transformer 架构，以氨基酸序列为查询及密码子序列为键和值，借助跨注意力机制捕捉两者间的长程依赖；同时在训练过程中采用逐步掩码策略，从初始不掩码密码子序列，逐步增加掩码比例，最终实现完全掩码，迫使模型逐步摆脱对原始密码子序列的依赖，仅输入氨基酸序列就能生成 *CAI* 适配、二级结构稳定的优化密码子序列。

2.1.2 非翻译区优化

非翻译区（untranslated region, UTR）是 mRNA 上不直接参与蛋白质编码的重要调控区域，通常分布于编码序列的 5′ 端和 3′ 端。尽管 UTR 本身不决定蛋白质的氨基酸序列，但其序列和结构特征在转录后调控过程中发挥至关重要的作用。例如，5′UTR 中的核糖体结合位点、上游开放阅读框以及二级结构特征，会显著影响核糖体的招募与起始效率；3′UTR 中的微小 RNA（microRNA, miRNA）结合位点、RNA 结合蛋白识别序列等，则与 mRNA 的稳定性、定位和降解速率密切相关。这些因素共同决定了基因的表达水平和翻译效率，使 UTR 成为合成生物学和基因工程中一个重要且复杂的优化对象。

在 UTR 优化任务中，传统的基于规则或局部突变搜索的方法往往难以在庞大的序列空间中高效探索全局最优解。近年来，研究者引入了一种结合生成模型与功能预测模型的新范式，其核心思想是：由生成模型负责提出候选 UTR 序列，通过功能预测模型评估这些序列在表达量、翻译效率等功能指标上的表现，从而为生成模型提供优化方向和约束信号。

UTRGAN 是该类方法中的代表性工作之一[6]。该模型以生成对抗网络（generative adversarial network, GAN）作为核心生成框架，通过在大量人类天然 5′UTR 序列上进行训练，使生成器学习真实 UTR 的序列特征，从而能够生成在整体结构和碱基组成上接近天然序列的候选 UTR。这种基于数据驱动的方式避免了手工规则设计的局限，为探索更丰富的功能序列提供了可能。此基础上，UTRGAN 进一步引入多个功能预测模型构成多目标优化体系。例如，Xpresso 用于预测基因表达量，FramePool 用于估计平均核糖体负载（mean ribosome load, MRL），MTtrans 用于预测翻译效率（translation efficiency, TE）。在优化阶段，UTRGAN 采用梯度上升策略，对生成器的输入噪声向量进行迭代更新，使生成的 UTR 序列在保持天然性的同时，朝着更高的表达量、更大的 MRL 或更优的 TE 方向演化。通过这种方式，模型实现了在连续潜在空间中的定向搜索，将原本离散的序列优化问题转化为可微的优化问题。整体而言，这种生成模型与功能预测模型协同互补的策略，为 UTR 的理性设计和定向优化提供了一条高效且可扩展的技术路线，也为基因表达调控元件的计算设计奠定了重要基础。

2.1.3 多区域整合优化

RNA 分子的功能实现依赖 5′UTR、CDS、3′UTR 等不同区域的协同作用，各区域的结构与功能紧密关联、相互制约。传统仅针对单一区域（CDS 或 UTR）的优化策略，往往因忽视区域间的适配性而难以达成 mRNA 整体性质最优，制约了其在疫苗研发、基因治疗等领域的应用效能。因此，近年来基于 AI 的 mRNA 序列优化研究逐渐由单一区域优化转向多区域整合优化，以实现 mRNA 全长序列的性能提升。

在多区域整合优化的技术探索中，整合深度学习 mRNA 优化算法（integrated deep-learning-based mRNA optimization, iDRO）是通过深度学习方法实现 mRNA 全长序列（涵盖 5′UTR、CDS、3′UTR）整合式优化的代表性研究[7]。该算法针对不同区域的功能特性采用差异化优化模型：对于 CDS，引入双向长短期记忆-条件随机场（bidirectional long short-term memory-conditional random field, BiLSTM-CRF）模型，通过该模型强大的序列特征提取与依赖关系建模能力，精准优化密码子，保障翻译效率；针对 UTR 则采用基于 Transformer 架构的 RNA-Bart 模型，借助其对

RNA 序列上下文信息的深度挖掘能力生成最优 UTR 序列，以强化核糖体结合效率与 mRNA 稳定性。为验证优化效果，研究团队将优化后的 CDS 与 UTR 进行组合，构建完整 mRNA 序列，并通过增强型绿色荧光蛋白（enhanced green fluorescent protein, GFP）基因的体外细胞实验进行验证筛选，最终证实该整合优化策略可显著提升 mRNA 的表达效率与功能稳定性。

LinearDesign2 算法则在整合优化的逻辑上实现了进一步升级，通过循环迭代机制强化区域间的协同适配[8]。其核心流程为：首先利用初代 LinearDesign[3] 算法生成初始 CDS 序列；随后固定该 CDS 序列，通过进化算法对 5′UTR 区域进行针对性优化，确保 5′UTR 与初始 CDS 的初步适配；在此基础上，固定优化后的 5′UTR 序列，对 CDS 区域进行二次优化。通过多轮循环迭代，算法可动态协调 CDS 与 5′UTR 的结构与功能关系，每一轮优化均以另一区域的当前最优状态为约束条件，从根本上规避了“分别优化”模式下易出现的区域间适配性差、功能冲突等问题。相较于 iDRO 算法的分区域优化后组合验证模式，LinearDesign2 通过动态反馈机制实现了区域间的深度适配，更充分地诠释了多区域整合优化的核心思想，进一步提升了优化序列的整体性能。

除上述专用优化算法外，预训练大语言模型的兴起为 RNA 序列整合优化提供了全新技术路径。以 Uni-RNA[23] 为代表的 RNA 预训练大语言模型，通过对海量全类型 RNA 序列进行大规模预训练，具备同时对 CDS 与 UTR 等多个功能元件进行高效表征和深度解析的能力。这类模型的出现有望打破传统算法对不同区域的割裂式处理模式，可实现从 RNA 序列输入到优化序列输出的端到端调控，无需人工拆分区域进行分步优化。凭借强大的泛化能力与特征提取能力，类似的预训练大语言模型有望成为新一代 RNA 序列端到端整合优化的基础工具，为突破现有技术瓶颈、推动 RNA 设计技术的标准化与智能化发展提供核心支撑。

2.2 RNA 修饰预测

RNA 的生物学功能并不仅由其一级序列决定，还在很大程度上受到化学修饰的影响。作为真核生物转录后基因表达调控网络中的关键组成部分，RNA 修饰通过在特定核苷酸位点引入化学修饰基团，对 RNA 的结构稳定性、相互作用能力以及生命周期等特性进行精细调控。这些修饰广泛存在于 mRNA、tRNA、rRNA 及各类非编码 RNA 中，目前已鉴定的 RNA 修饰类型超过 100 种。mRNA 上常见的修饰类型如图 4 所示。其中，N^6-甲基腺苷（m^6A）与 mRNA 的输出、翻译及降解过程密切相关；假尿苷（Ψ）及其衍生修饰（如 $m^1Ψ$）则通过增强 RNA 的构象稳定性和翻译保真性，在应激响应和免疫调控中发挥重要作用。

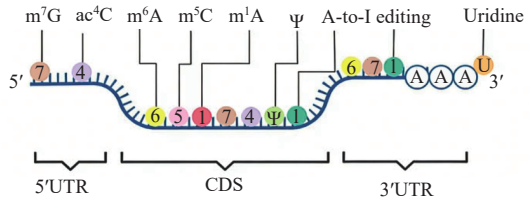


Fig. 4 Common mRNA modification types

图 4 mRNA 上常见的修饰类型

尽管 RNA 修饰在生物学功能中的重要性已得到广泛认可，但单纯依赖实验手段对修饰位点进行解析，仍然面临诸多现实挑战。一方面，高通量测序结合化学或抗体富集的实验方案往往成本较高、实验流程复杂，难以在大规模样本或多条件背景下频繁开展；另一方面，许多抗体富集型方法在空间分辨率上存在先天不足，通常只能将修饰信号定位到几十甚至上百个核苷酸的区间内，难以实现单碱基精度的准确定位。此外，对于化学性质相近的修饰类型，不同修饰在实验信号层面的区分度有限，容易产生混淆，从而影响下游功能分析的可靠性。在此背景下，借助计算模型对 RNA 修饰位点进行高效、精细的预测，逐渐成为实验研究的重要补充手段。

早期的 RNA 修饰预测方法主要基于传统机器学习框架。研究者通常从 RNA 序列中人工提取特征，例如 k-mer 组成、核苷酸频率等，再将这些特征输入支持向量机、随机森林等分类模型中，用以判定特定位点是否发生修饰[9-10,24]。这类方法在特定修饰类型和特定物种背景下取得了一定成功，验证了 RNA 修饰在序列层面存在可被学习的统计规律。然而，其性能在很大程度上依赖于人工特征工程，难以自动捕捉长程序列依赖关系和复杂的非线性调控模式，同时在跨数据集或跨物种应用时往往面临泛化能力不足的问题。

随着深度学习技术的发展，研究者开始将卷积神经网络（convolutional neural network, CNN）、循环神经网络（recurrent neural network, RNN）以及注意力机制引入 RNA 修饰预测任务中，使模型能够直接从原始序列中自动学习多尺度、高阶的序列表示，从而显著提升预测精度与稳定性。以 m^6A 修饰为例，包括 EDLm^{6}APred[11] 在内的多种深度学习模型已在公开数

据集上验证了其对修饰位点的高灵敏度和高特异性识别能力。这类模型通常在目标位点上下游构建固定长度的序列窗口，通过卷积或循环结构捕捉局部与长程上下文信息，从而学习与甲基转移酶复合物识别相关的隐含序列偏好，并在不同物种和细胞系中表现出较好的适用性。

在此基础上，研究者逐渐认识到，单一修饰类型的预测难以全面反映真实生物体系中多种 RNA 修饰共存、相互影响的复杂调控格局。为此，多修饰联合建模策略开始受到关注。近年来提出的多标签或多任务深度学习模型，通过共享底层序列表示，在同一网络框架中同时预测多种 RNA 修饰位点，不仅能够捕捉不同修饰之间潜在的协同或拮抗关系，还显著降低了为每一种修饰类型分别训练独立模型所带来的计算成本。例如，RNA-ModX[12] 等模型采用多任务学习思想，在统一的网络中联合预测多种转录后 RNA 修饰，在整体建模能力和预测一致性方面均表现出优势。

此外，越来越多的研究表明，RNA 修饰的发生并非仅由线性序列信息决定，还与局部二级结构甚至更高层次的空间构象密切相关。特定修饰酶往往偏好位于发卡结构、环区或特定配对状态下的核苷酸位点。因此，部分工作开始在序列特征之外，引入碱基配对状态、*MFE* 或结构概率分布等信息，将结构特征与深度学习模型相结合，以期进一步提升 RNA 修饰预测的生物学合理性与准确性。这一趋势也体现了 RNA 修饰计算研究从序列驱动向联合使用序列信息与结构信息进行建模。

3 RNA 结构预测

除一级序列外，RNA 的生物学功能也高度依赖于其特定的空间构象及 RNA 与其他分子的相互作用关系。其中 RNA 的二级与三级结构为 RNA 实现分子识别、催化反应及转录调控等关键功能提供了重要的物理基础。然而，传统 RNA 结构解析技术受限于高昂的实验成本、复杂的操作流程，以及对动态构象捕获能力不足等因素，严重制约了对 RNA 结构的系统性研究。AI 技术的引入为 RNA 结构研究提供了新的解决思路，借助其在复杂模式识别和高维特征学习方面的优势，AI 模型能够直接从 RNA 序列中学习其与空间结构之间的非线性映射关系，从而在一定程度上突破传统计算方法在搜索空间规模和建模精度方面的固有局限。

3.1 RNA 二级结构预测

RNA 分子可通过碱基之间的氢键互补配对自发折叠，从而形成具有特定拓扑特征的二级结构，常见的二级结构单元如图 5 所示。

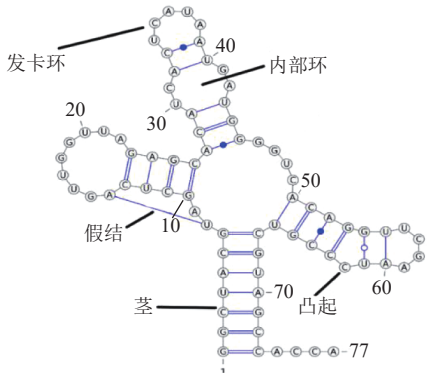


Fig. 5 Secondary structure of RNA

图 5 RNA 二级结构

在 RNA 二级结构体系中，茎结构是最基础、最核心的结构单元，由连续互补碱基配对经轴向堆叠形成局部双链螺旋，通过降低分子自由能实现构象稳定，是支撑 RNA 空间折叠的核心元件。发夹环、内环、凸环、多分支环等经典二级结构单元，均可视为茎结构经拓扑重排衍生的功能变体：发夹环由单一茎结构末端游离碱基闭合形成；内环由两组反向排列的茎结构之间未配对碱基区段构成；凸环与多分支环则由茎结构的空间连接模式变化所产生。值得注意的是，假结因其碱基配对存在拓扑交叉，无法被嵌套式茎结构模型表征，通常被视为超出传统二级结构定义范畴的特殊构象。总体而言，茎结构构成了 RNA 二级结构的骨架，其组合方式和连接模式决定了 RNA 分子整体折叠形态与潜在功能区域。

在 RNA 二级结构研究与建模中，合理的结构表示方式与特征编码策略是开展计算分析、结构预测以及 AI 建模的基础。针对 RNA 二级结构，研究者目前通常采用邻接矩阵、连通性表以及点括号标注法 3 种表示方式，它们从不同角度刻画了 RNA 分子中碱基配对关系及拓扑特征。

1）邻接矩阵

邻接矩阵是一种基于图论思想的结构表示方式，用于刻画 RNA 分子中任意 2 个碱基之间是否存在氢键配对关系。如图 6 所示，对于长度为 L 的 RNA 序列，其邻接矩阵为一个 $L\times L$ 的二值矩阵，其中矩阵元素 a_{ij} 定义为：当第 i 个碱基与第 j 个碱基形成氢键配对时，$a_{ij}=1$，否则 $a_{ij}=0$。由于碱基配对关系具有对称

性，即若 i 与 j 配对，则 j 亦与 i 配对，因此邻接矩阵是一个严格对称矩阵。该表示方式能够完整保留 RNA 二级结构中的全局配对信息，适合与图神经网络等深度学习模型结合，用于结构模式识别与结构比较分析。

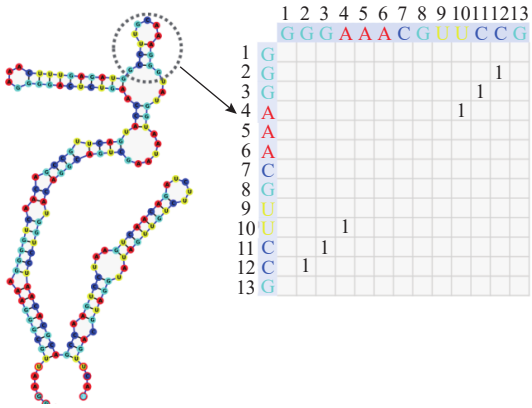


Fig. 6 Adjacency matrix of RNA secondary structure

图 6 RNA 二级结构的邻接矩阵

2）连通性表（connectivity table, CT）

连通性表是一种结构化、可读性较强的 RNA 二级结构表示格式，广泛用于传统 RNA 结构预测与分析软件中。如图 7 所示，CT 文件的首行通常包含核苷酸总数、预测或计算得到的折叠自由能以及文件名等全局信息；其后每一行对应一个核苷酸，并包含 6 列数据，分别为：核苷酸序号（从 1 至 N，N 为核苷酸总数）；核苷酸类型（A、G、U 或 C）；前一个核苷酸序号（从 0 递增至 N–1）；后一个核苷酸序号（从 2 递增至 N）；配对核苷酸序号（若为 0 则表示该核苷酸未配对）；最后一列重复第一列的核苷酸序号。

120	ENERGY=0	A00001			
1	U	0	2	0	1
2	G	1	3	0	2
3	A	2	4	0	3
4	U	3	5	0	4
5	A	4	6	0	5
6	C	5	7	115	6
7	G	6	8	114	7
8	G	7	9	113	8
9	C	8	10	112	9
10	G	9	11	111	10
11	G	10	12	110	11
12	U	11	13	109	12
13	C	12	14	0	13
14	A	13	15	0	14
15	U	14	16	0	15
16	A	15	17	0	16
17	G	16	18	69	17

Fig. 7 Connectivity table of RNA secondary structure

图 7 RNA 二级结构的连通性表

3）点括号标注法

点括号标注法是一种简洁直观的 RNA 二级结构表示方式，常用于结构可视化和算法输入输出。如图 8 所示，该方法以字符串形式描述结构：未发生配对的碱基用“.”表示；形成氢键配对的 2 个碱基分别用成对的括号“(”和“)”表示，其中靠近 5′端的碱基标记为“(”，靠近 3′ 端的碱基标记为“)”。点括号标注法能够清晰反映 RNA 二级结构中茎、环等基本结构单元，但由于其基于嵌套结构假设，难以直接表示假结等非嵌套配对关系。

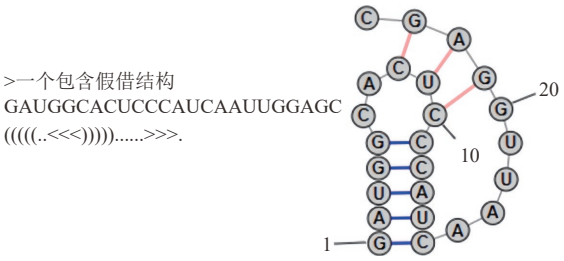


Fig. 8 Dot bracket annotation of RNA secondary structure

图 8 RNA 二级结构的点括号标注法

RNA 的生物学功能在很大程度上依赖其空间构象，而二级结构作为连接 RNA 一级序列与高级结构及功能的重要中间层级，在 RNA 分子稳定性、翻译调控及其与其他分子的相互作用中发挥着关键作用。因此，RNA 二级结构的准确解析对于理解 RNA 的功能机制及其在疾病发生发展中的作用具有重要意义。多年来，科研人员围绕 RNA 二级结构预测问题开发了大量计算工具，从早期基于热力学模型的方法到近年来引入深度学习的预测框架，均取得了阶段性成果。总体来看，现有 RNA 二级结构预测方法大致可以分为一维方法和二维方法 2 类。

一维方法将 RNA 二级结构预测视作序列分析问题，其核心思想是将 RNA 的结构信息线性化表示，从而能够直接套用成熟的序列建模框架。该类方法通常将采用点括号标注法表示 RNA 序列[25]。通过这种方式，原本具有复杂拓扑约束的结构预测问题被转化为一个逐位置多分类任务，即对序列中每个位置进行多分类判定。

从建模形式上看，该任务与自然语言处理（natural language processing, NLP）中的序列到序列建模或序列标注问题高度相似：RNA 的碱基序列可类比为输入词序列，而点括号结构则对应输出标签序

列。因此，研究者可以直接借鉴 NLP 领域中广泛使用的模型架构，如 RNN、长短期记忆网络(long short-term memory, LSTM)以及基于自注意力机制的 Transformer 模型，对 RNA 序列进行上下文建模，并输出每个位置对应的结构类别[13]。在训练过程中，通常采用交叉熵损失函数对预测标签与真实点括号标签进行监督学习。这类方法的优势在于实现简单、计算效率较高，且易于与其他序列特征进行融合。然而需要注意的是，在真实 RNA 序列中，未配对碱基所占比例往往远高于配对碱基，导致标签分布高度不均衡，这在实践中可能使模型偏向预测“.”类别。因此，研究中往往需要通过类别加权、重采样或引入辅助约束等方式缓解类别不平衡问题，以提升模型对配对碱基的识别能力。

与一维方法不同，二维方法直接以 RNA 的碱基配对关系作为建模对象，通常将二级结构表示为一个 $L\times L$ 的邻接矩阵，其中 L 为序列长度。该建模方式在形式上与计算机视觉中的语义分割任务高度相似：语义分割的目标是为图像中的每一个像素分配类别标签，而二维 RNA 结构预测则是为矩阵中的每一个“像素点”预测其是否属于“配对”类别。因此，语义分割领域中成熟的网络结构、特征提取策略以及多尺度融合思想，能够较为自然地迁移到 RNA 二级结构预测问题中。

UFold 模型是二维方法中的代表性工作[14]。该模型创新性地将经典的 U-Net 语义分割架构引入 RNA 二级结构预测任务中，将 RNA 序列编码为二维特征表示后输入网络，通过编码器-解码器结构实现对局部特征与全局依赖关系的高效融合。U-Net 的下采样路径能够捕捉长程碱基依赖和整体结构模式，而上采样路径则有助于恢复精细的局部配对信息，从而在无需显式引入热力学模型的情况下，直接学习序列与配对关系之间的映射。实验结果表明，UFold 在多种基准数据集上均能实现高精度的配对概率预测，验证了将 RNA 二级结构预测视作二维语义分割问题的可行性与有效性。

总体而言，一维方法与二维方法从 2 个角度对 RNA 二级结构预测问题进行了重构：前者强调与 NLP 模型的自然对接，具有实现简洁、计算高效的优势；后者则更直接地刻画碱基配对的全局拓扑结构，能够更充分地利用深度学习在空间特征建模方面的能力。2 类方法从不同角度对 RNA 二级结构预测问题进行建模，为该领域的持续发展提供了互补的技术路线。

3.2 RNA 三级结构预测

RNA 分子在二级结构的基础上，通过复杂的分子间相互作用进一步形成稳定的三维空间构象。这些相互作用包括非经典碱基配对形式(如 Hoogsteen 配对)、碱基与核糖或磷酸骨架之间的氢键作用、Mg^{2+} 等二价金属离子在带负电的磷酸骨架之间提供的电荷中和与桥接效应，以及碱基堆积力和疏水相互作用等远距离协同作用。在多种物理化学因素的共同驱动下，RNA 分子能够从局部稳定的茎环、内环或多分支结构出发，逐步折叠为高度有序且具有特定功能的三级结构。

RNA 三级结构是 RNA 分子具备生物学活性，成为功能执行体的核心结构基础，在介导分子识别、调控信号传导及催化生化反应等关键生物学过程中发挥不可替代的作用。例如，tRNA 特有的 L 型构象保证了反密码子环与氨基酸接受臂在核糖体中的精确空间定位；核酶通过形成特定的三维催化口袋，实现底物识别与化学反应的高效催化；而长非编码 RNA (long non-coding RNA, lncRNA) 等非编码 RNA 的三级结构则为其介导转录调控、染色质重塑以及 RNA-蛋白相互作用等复杂生理与病理过程提供了必要的结构支撑。

从计算建模角度看，RNA 三级结构预测的核心目标是获得 RNA 分子中各原子的三维空间坐标，这一点在形式上与蛋白质三级结构预测高度一致。然而，与蛋白质结构预测领域中机器学习方法的快速发展相比，当前 RNA 三级结构预测仍主要依赖分子模拟这一传统技术路线。其根本原因在于高精度 RNA 三级结构实验数据的数量严重不足：通过 X 射线晶体学、核磁共振或冷冻电镜技术解析得到的 RNA 原子级结构远少于蛋白质结构，难以为深度学习模型提供足够的训练样本，从而限制了模型在不同 RNA 类型和尺度上的泛化能力。因此，RNA 三级结构预测中的机器学习研究整体仍处于探索与方法验证阶段。

然而，单纯依赖分子模拟方法在实际应用中也面临显著挑战。RNA 分子的构象空间极其庞大，在缺乏有效先验信息或结构约束的情况下，模拟过程往往需要在高维自由度空间中进行长时间搜索，不仅计算成本高昂，而且容易陷入非物理合理的局部极值，从而影响预测精度。为缓解这一问题，研究者开始引入 AI 作为辅助工具，为分子模拟提供引导信息。其中，一种具有代表性的策略是首先利用机器学习或深度学习模型预测 RNA 的碱基接触关系[15]，

再将碱基接触图作为软约束整合进后续的分子动力学或蒙特卡洛模拟流程中，从而显著缩小搜索空间，同时加速构象收敛。

除了作为结构生成过程中的辅助约束，机器学习与深度学习模型还可在 RNA 三级结构预测中承担结构评分器的角色，用于对分子模拟生成的大量候选构象进行高效筛选与排序。传统基于物理能量的评分函数往往难以准确反映近似模型与真实结构之间的几何差异，而基于数据驱动的学习型评分方法则能够自动捕捉更复杂的空间与化学特征。例如，斯坦福大学研究团队提出的原子旋转等变评分器（atomic rotationally equivariant scorer, ARES）[16]，该模型基于旋转等变神经网络，可直接以原子三维坐标和元素类型作为输入，在不依赖 RNA 特有先验规则的前提下学习分子结构特征，对候选构象相对于真实结构的均方根偏差（root mean square deviation, RMSD）进行预测。在国际 RNA 结构预测评测任务中，此类“AI+分子模拟”融合策略已展现出优于传统方法的整体性能，能够从大量候选构象中更准确地识别接近真实结构的高质量模型。这些结果充分表明，将 AI 与分子模拟有机结合，是突破 RNA 三级结构预测精度瓶颈的一条重要发展方向，也为构建更高精度、更具泛化能力的 RNA 结构预测体系奠定基础。

3.3 RNA 与其他分子的相互作用预测

RNA 与其他生物大分子的相互作用贯穿于几乎所有关键生命过程，在基因表达调控、信号转导、细胞分化以及疾病发生发展等方面发挥核心调控作用。其中，RNA 结合蛋白（RNA-binding protein, RBP）作为连接 RNA 分子与细胞功能执行系统的重要桥梁，在复杂调控网络中占据枢纽地位。RBP 通常依赖其特有的 RNA 结合结构域，如 RNA 识别基序、锌指结构域等，与 RNA 分子中特定的序列或结构基序实现高度特异性的识别与结合，广泛参与 RNA 可变剪接、核输出、翻译调控与降解过程，其介导的多层次调控路径如图 9 所示。这种精准且动态的 RNA-蛋白相互作用，构成了转录后调控网络的重要分子基础。

在非编码 RNA 研究领域，RNA-蛋白相互作用的生物学意义尤为突出。以 lncRNA 为例，其自身通常不具备蛋白质编码能力，但可通过与染色质调控复合物、转录因子及核糖核蛋白复合物形成稳定的 RNA-蛋白复合体，参与染色质重塑、转录激活或抑制等关键生物学过程。lncRNA 的功能行使往往依赖

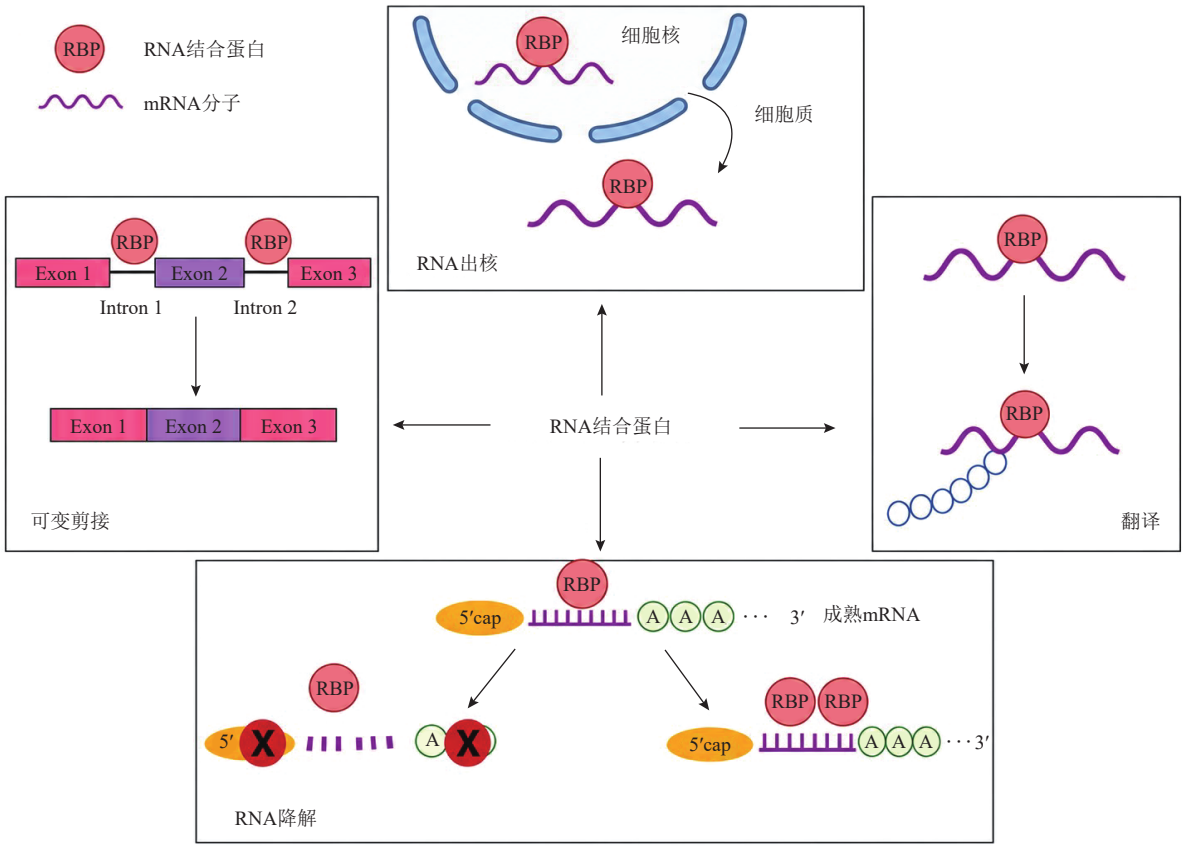


Fig. 9 RNA-binding proteins

图 9 RNA 结合蛋白

于与特定 RBP 的选择性结合，而这种结合特异性不仅受 RNA 一级序列的影响，还与其局部或全局的空间结构密切相关。因此，系统解析 RNA-蛋白相互作用网络，对于阐明 RNA 调控机制及揭示疾病相关分子通路具有重要的科学意义。

目前，RNA-蛋白相互作用的研究在很大程度上依赖于高分辨率实验技术，例如 X 射线晶体学、核磁共振波谱学以及冷冻电子显微镜等方法。这些技术能够直接解析 RNA-蛋白复合物的原子级空间结构，从而获取精准的相互作用位点与分子作用模式信息。然而，此类实验方法普遍存在成本高昂、操作流程复杂、实验周期冗长以及对样本纯度与稳定性要求苛刻等固有局限，难以支撑大规模 RBP 结合位点的系统性筛查，亦无法满足多条件、多物种背景下相互作用机制的全面解析需求。在此背景下，开发高效、精准且具备可扩展性的计算预测方法，用于识别 RBP 结合位点、刻画其结合偏好并挖掘潜在的相互作用信息，逐渐成为 RNA-蛋白相互作用研究领域的重要发展方向。

随着深度学习技术的快速发展，RNA-蛋白相互作用预测领域涌现出一系列具有代表性的模型，并在多个核心任务中取得突破性进展。DeepBind 是最早将 CNN 引入核酸-蛋白相互作用预测的经典模型之一[17]。该模型通过卷积核在核酸序列上的滑动扫描，自动学习与蛋白结合相关的局部序列模式，再结合池化层与全连接层对高维特征进行整合与降维，最终输出目标 RNA 序列与特定蛋白的结合强度预测结果。DeepBind 的成功构建验证了深度学习在自动挖掘核酸-蛋白结合基序方面的技术优势。

在此基础上，iDeep 模型进一步引入多模态深度学习框架以提升预测性能[18]。该模型不仅利用 RNA 的一级序列信息，还整合了预测得到的 RNA 二级结构特征，通过深度神经网络对多源异构特征进行联合表示学习，从而更全面地刻画 RNA 分子在特定序列与结构背景下被 RBP 识别的分子机制与结合概率。iDeep 的研究证实，单纯依赖 RNA 一级序列信息往往难以充分解释 RBP 的结合偏好，而 RNA 二级结构信息的引入，能够显著提升模型对真实生物调控过程的刻画能力与预测精度。

与上述以 RNA 为核心的预测策略不同，HydRA 模型从蛋白质视角出发，为 RBP 功能预测提供了全新的研究路径[19]。HydRA 通过整合蛋白质的外在分子交互信息（如蛋白-蛋白相互作用网络及功能关联网络中同源 RBP 的比例）与内在序列特征信息（包括短肽基序、氨基酸组成、理化性质以及保守功能域模式等），构建混合深度学习模型，实现对蛋白质 RNA 结合能力的精准预测，并进一步定位其潜在的 RNA 结合相关结构域。

此外，RPITER 等模型则致力于实现 RNA 与蛋白质的协同建模[20]，构建双通道预测框架以刻画二者之间的相互作用关系。该类模型通常分别对 RNA 核苷酸序列和 RBP 的氨基酸序列进行针对性特征编码，并送入独立的特征提取网络进行深层次特征挖掘，随后在模型高层通过特征融合模块或联合分类器，对给定的 RNA-蛋白对是否存在真实相互作用进行整体判定。这种建模方式能够同时兼顾 RNA 与蛋白质两侧的分子特征，有助于提升模型在复杂生物场景下的预测稳定性与泛化能力。

总体而言，随着深度学习方法的不断演进，RNA-蛋白相互作用计算预测正从早期的单一序列基序识别，逐步迈向多模态融合、多层次特征挖掘与系统性网络建模的新阶段。未来，通过进一步整合 RNA 三级结构信息、蛋白质三维空间特征以及细胞微环境背景数据，并结合模型可解释性分析与实验验证机制，AI 驱动的 RNA-蛋白相互作用研究，有望在基础生物学机制探索与临床疾病诊疗靶点开发中发挥更为关键的作用。

4 讨论与总结

结合现有研究进展，RNA 领域的 AI 方法可以概括为 3 类：传统机器学习方法、深度学习方法，以及机器学习/深度学习与物理建模和采样过程相结合的协同策略。

需要说明的是，这种划分并非严格互斥的分类体系。其中，深度学习方法中的预训练基础模型既可以单独作为判别任务的表征骨干，也可与机器学习中提取的特征结合，设计融合策略来完成进一步的复杂任务。在将 AI 方法应用到 RNA 相关研究中时，通常需要同时考虑序列长度与搜索空间规模、标注数据的数量、任务类型，以及对可解释性、生物物理一致性与结果可控性的要求。

对于目标函数能够较为明确表达且对可控性与可复现性要求较高的任务，物理或规则驱动方法仍具有重要基础价值。此类方法通过将 *MFE*、*CAI* 等可显式表达的目标或约束项纳入求解或优化过程，使其输出与既有生物物理假设或工程约束保持较高一致性，因此在 RNA 二级结构推断以及受约束密码

子优化任务中往往具有较好的结果稳定性。

传统机器学习主要依赖人工构建的序列、结构或进化特征，在小样本且特征定义较清晰的场景中仍可作为稳健基线，尤其适用于局部规律较明确的任务，如局部基序识别与基础功能分类等问题。然而，该类方法对表达调控、修饰背景与细胞情境等跨模态信息的整合能力相对有限；当任务涉及长序列、多目标耦合优化或复杂环境条件时，目标函数构建的完备性与计算代价往往共同成为限制因素。

相比之下，以Transformer[26]为代表的深度学习模型能够直接从原始输入中学习多尺度表征，因此在二级结构概率预测以及RNA-蛋白相互作用预测等复杂任务中通常具备更高的性能上限。然而，深度学习模型的高性能往往建立在相对一致的数据分布与标注口径之上；若缺乏跨平台、跨物种与跨实验条件验证，模型可能更容易过拟合到一些环境相关的伪特征，从而产生具有系统性偏差的虚假关联。

生成式模型是深度学习模型中的另一大类，推动AI赋能的RNA研究由属性预测进一步走向理性设计。在mRNA序列优化[27]等任务中，生成式框架能够直接输出候选序列方案，从而更契合RNA药物的工程化需求。其关键瓶颈在于，生成模型通常依赖代理指标，如预测得分、结构或能量先验、配对概率等进行条件生成以引导候选产生；但由于代理指标对应的实验数据缺失或预测模型能力不足，高评分候选序列未必能够在真实实验体系中实现相应功能。此类方法的应用价值不仅取决于生成能力本身，更高度依赖评估器质量、约束机制设置以及后续实验反馈通路的有效建立；只有在生成、评估与验证之间形成较稳定的联动关系时，才更有可能降低代理目标误差被持续放大的风险，并提升候选的可验证性与可实现性。

近年来，预训练基础模型作为新兴技术，为RNA研究提供了新的拓展方向[22-23,28]。预训练模型通过大规模未标注或弱标注RNA数据学习通用表征，有望缓解标注稀缺问题，并提升跨任务迁移与长程依赖建模能力，在小样本场景下可能发挥更显著的作用；不过，其实际表现仍受预训练语料分布、下游任务对齐方式与置信度校准质量影响，在RNA相关任务中的性能仍有待验证。

另一方面，在RNA三级结构预测等标签稀缺且构象空间巨大的任务中，将机器学习或深度学习与物理建模和采样过程相结合的策略，通常比依赖单一方法更具现实可行性，既能发挥数据驱动模型的预测优势，又能兼顾生物物理合理性。例如，可将机器学习/深度学习模型预测得到的配对概率、接触图或其他结构表征作为条件输入、约束项或候选评分依据，以缩小搜索空间并提高候选筛选效率。但其效果同样依赖先验质量、采样覆盖率与评分函数准确性。且该策略更适合被视为模型与模拟协同的工作流，而非完全独立的单一方法类别。

总体而言，未来AI在RNA研究中的进一步发展，关键不只是构建更复杂的模型，更在于面向具体生物学问题建立清晰的方法选择依据，从而实现任务需求与技术路线的合理匹配。

5 未来展望

在数据规模呈指数级增长的时代背景下，科研活动正从“假设驱动”逐步迈向“数据驱动”的新范式。生命科学领域尤为如此，随着高通量测序、单细胞组学、空间转录组学以及大规模药物筛选技术的快速发展，研究者所面对的数据不仅体量巨大，而且来源多样、结构复杂，亟需引入新的计算处理方法。RNA分子的功能机制、调控网络及其与疾病发生发展的关联信息，往往隐匿于基因表达谱、细胞异质性数据、分子互作网络以及药物-靶点关系等多源异构数据之中。传统统计分析与经验驱动的方法在应对这类高维、非线性、强噪声数据时逐渐显露出局限性，而AI凭借其在高维特征提取与复杂模式识别方面的独特优势，正在成为“从数据海洋中发现科学规律”的关键工具。

在RNA研究领域，AI的引入正在深刻改变基础研究与应用研究的范式。从RNA修饰位点解析、非编码RNA功能注释以及RNA疫苗序列设计，到RNA二级与三级结构预测、RNA-蛋白相互作用识别，AI方法已在多个方向展现出显著潜力。AI与RNA科学的交叉融合，已跃升为生命科学领域发展最为迅猛的前沿研究方向之一：相关学术成果与发文量呈爆发式增长，领域研究热度持续攀升[29]；该方向不仅是学术界的核心研究热点，更成为产业界高度聚焦的热门赛道，展现出极为广阔的临床转化与产业落地前景。然而，AI在RNA领域的应用目前仍处于快速发展但尚未成熟的阶段，其进一步突破与大规模落地仍面临一系列关键挑战。

首先，高质量数据资源的匮乏与整合度不足，仍是制约AI模型性能的基础性瓶颈[30]。一方面，尽管公共数据库中已积累了大量RNA测序数据，但在数

据规模、物种覆盖范围、实验条件一致性以及标注质量等方面仍存在明显不足，不同实验平台和处理流程带来的系统偏差直接影响模型训练的稳定性和泛化能力。多组学数据之间缺乏严格配对和系统整合，使得模型难以全面学习 RNA 在真实生物体系中的多层次调控关系。另一方面，与蛋白质相比，已解析的 RNA 三维结构数量仍然十分有限，这在一定程度上制约了 AI 方法在 RNA 结构建模、结构-功能关系解析等方向的应用潜力。

其次，模型可解释性不足是当前 AI 在 RNA 研究中亟待突破的另一核心瓶颈。深度学习模型在 RNA 序列功能预测等任务中往往能够取得优异的预测性能，但其内部决策过程通常难以直接对应明确的生物物理机制或分子作用路径。这种“黑箱”特性使研究者难以从模型中提炼可检验的生物学假设，限制了模型结果对实验设计的指导价值，也在一定程度上降低了其在临床与产业应用场景中的可信度与接受度。

此外，计算预测结果向临床与实际应用转化的效率仍然偏低。尽管 AI 能够高效生成具有潜在活性的 RNA 序列、结构及相关功能预测结果，但其能否真正转化为具有临床开发价值的药物候选物，仍高度依赖后续湿实验的系统性验证。以 mRNA 疫苗设计为例，研究者不仅需要在真实生物体系中确认其功能有效性、结构可实现性与作用机制可靠性，还需进一步评估其在递送体系中的稳定性、化学修饰后的活性保持情况、载体适配性，以及潜在的毒性、免疫原性，而上述环节通常涉及大规模筛选、多轮迭代优化及临床前安全性研究。 由于 RNA 分子的序列组成、构象稳定性、修饰策略、递送方式与成药性之间存在高度耦合，加之实验验证成本高、周期长、标准路径尚不完全成熟以及跨学科协同要求较高，从 AI 设计到成药性确认的全过程呈现出高投入、高损耗的典型特征。计算预测结果向临床与实际应用转化的效率仍然有限。

针对上述挑战，未来 AI 在 RNA 研究中的发展需要在数据、模型与应用 3 个层面协同推进、系统布局[31]。在数据层面，应充分依托单细胞测序、长读长测序和空间组学等新兴技术，不断完善非编码 RNA 与转录后调控相关的高质量参考数据库，同时加强基因组、转录组、蛋白组及表观组等多组学数据的系统性整合。此外，通过实验技术革新和国际化数据共享机制的构建，逐步扩充 RNA 结构与功能相关的数据资源，并积极探索迁移学习、小样本学习等策略，以降低模型对大规模标注数据的依赖。

在模型层面，未来研究不仅需要继续提升预测性能，更应注重引入生物学约束与可解释性机制。例如，将第一性原理的物理建模思想与数据驱动模型相结合，或通过可解释性分析方法，增强模型输出与具体分子机制之间的对应关系，从而提高模型在科学发现与临床决策中的可用性和可信度。

在应用层面，关键在于形成从计算预测到实验验证，再到应用转化的完整路径。当前，行业头部企业已率先开展实践探索，形成具有示范意义的转化路径，为领域发展提供参考。Insitro 针对 siRNA 研发中靶点筛选不精准、脱靶效应、体内稳定性不足及递送效率低等瓶颈，整合多组学与临床数据，通过机器学习平台生成和优化 siRNA 候选，涵盖序列选择、靶基因敲低效果以及多维药物样性参数的联合建模与评估；并与 N - 乙酰半乳糖胺（N-acetylgalactosamine，GalNAc）靶向递送技术相结合，推进临床前开发进程，打通基础研究至临床开发的转化通道。

在初步形成转化路径的基础上，未来更值得关注的是构建干湿闭环研究体系。Moderna 与 Merck 合作开发的个体化新抗原 mRNA 疫苗 mRNA-4157（V940）是一个具有参考意义的实践。当前个体化新抗原疗法面临的核心瓶颈在于，患者肿瘤突变高度异质，真正具有免疫原性的候选新抗原比例有限，而开发又对时间窗口和候选筛选效率提出了较高要求。为应对这些问题，Moderna 和 Merck 基于患者肿瘤及血液样本的 mRNA 测序数据，借助 AI 算法识别体细胞突变并筛选最可能诱导免疫应答的候选新抗原，提高个体化候选发现与优先级排序的效率。在此基础上，候选序列还需进入后续制造、免疫原性评估以及临床疗效观察等湿实验与临床验证环节；而这些阶段积累的免疫原性指标与临床数据，能够为后续候选选择策略的校准与优化提供支持，进而形成“数据获取—计算预测—实验/临床验证—策略迭代”的反馈闭环。mRNA-4157（V940）在高风险黑色素瘤中的随机 IIb 期研究已显示出较好的临床前景[32]，并已进一步拓展至多项后续临床项目。

更进一步地，AI 在 RNA 药物研发中的作用不应再局限于生成候选，而应深度嵌入“设计—合成—测试—学习”的闭环体系。例如结合主动学习、贝叶斯优化等策略，根据每一轮实验结果动态挑选信息增益最高的候选分子和实验条件，从而减少盲目筛选与重复试错，使 AI 进一步发展为实验决策协同系统，加快 RNA 分子从体外验证走向体内评估的整体进

程。以 Insitro 的 TherML 平台为例，该平台旨在构建一个闭环主动学习系统，利用计算预测指导实验设计并优先选择高信息增益实验，随后基于实验结果持续更新模型与优化策略，并通过对接自动化实验设施实现高通量循环迭代。

综上所述，尽管当前仍面临数据质量参差不齐、模型可解释性有待提升、转化路径尚未完全打通等挑战，但随着高质量数据资源的持续积累、算法模型的迭代更新以及计算和实验闭环体系的深入构建，AI 与 RNA 科学的交叉融合必将释放巨大潜能，为攻克重大疾病、推动精准医疗发展提供强有力的技术支撑。

作者贡献声明：顾睿初、梁雅琳负责相关文献资料的收集、分析和论文撰写；温翰构思了论文整体内容，负责了论文的撰写及修改。

参 考 文 献

[1] Cooper T A, Wan Lili, Dreyfuss G. RNA and disease[J]. Cell, 2009, 136(4): 777−793

[2] Terai G, Kamegai S, Asai K. CDSfold: An algorithm for designing a protein-coding sequence with the most stable secondary structure[J]. Bioinformatics, 2016, 32(6): 828−834

[3] Zhang He, Zhang Liang, Lin A, et al. Algorithm for optimized mRNA design improves stability and immunogenicity[J]. Nature, 2023, 621(7978): 396−403

[4] Gong Zheng, Jiang Ziyi, Gao Weihao, et al. A new deep-learning-based approach for mRNA optimization: High fidelity, computation efficiency, and multiple optimization factors[J]. arXiv preprint, arXiv: 2505.23862, 2025

[5] Ren Zilin, Jiang Lili, Di Yaxin, et al. CodonBERT: A BERT-based architecture tailored for codon optimization using the cross-attention mechanism[J]. Bioinformatics, 2024, 40(7): btae330

[6] Barazandeh S, Ozden F, Hincer A, et al. UTRGAN: Learning to generate 5′UTR sequences for optimized translation efficiency and gene expression[J]. Bioinformatics Advances, 2025, 5(1): vbaf134

[7] Gong Haoran, Wen Jianguo, Luo Ruihan, et al. Integrated mRNA sequence optimization using deep learning[J]. Briefings in Bioinformatics, 2023, 24(1): bbad001

[8] Liu Yang, Gao Jie, Zhang Xiaonan, et al. Joint design of 5′ untranslated region and coding sequence of mRNA[J]. arXiv preprint, arXiv: 2410.20781, 2024

[9] Zhou Yuan, Zeng Pan, Li Yanhui, et al. SRAMP: Prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features[J]. Nucleic Acids Research, 2016, 44(10): e91

[10] Chen Wei, Feng Pengmian, Ding Hui, et al. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition[J]. Analytical Biochemistry, 2015, 490: 26−33

[11] Zhang Lin, Li Gangshen, Li Xiuyu, et al. EDLm^{6}APred: Ensemble deep learning approach for mRNA m^{6}A site prediction[J]. BMC Bioinformatics, 2021, 22(1): 288

[12] Chen Yuge, Hang E S, Mamtha M R N, et al. RNA-ModX: A multilabel prediction and interpretation framework for RNA modifications[J]. Briefings in Bioinformatics, 2025, 26(1): bbae688

[13] Wang Linyu, Liu Yuanning, Zhong Xiaodan, et al. DMfold: A novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle[J]. Frontiers in Genetics, 2019, 10: 143

[14] Fu Laiyi, Cao Yingxin, Wu Jie, et al. UFold: Fast and accurate RNA secondary structure prediction with deep learning[J]. Nucleic Acids Research, 2022, 50(3): e14

[15] Singh J, Paliwal K, Zhang Tongchuan, et al. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning[J]. Bioinformatics, 2021, 37(17): 2589−2600

[16] Townshend R J L, Eismann S, Watkins A M, et al. Geometric deep learning of RNA structure[J]. Science, 2021, 373(6558): 1047−1051

[17] Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning[J]. Nature Biotechnology, 2015, 33(8): 831−838

[18] Pan Xiaoyong, Shen Hongbin. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach[J]. BMC Bioinformatics, 2017, 18(1): 136

[19] Jin Wenhao, Brannan K W, Kapeli K, et al. HydRA: Deep-learning models for predicting RNA-binding capacity from protein interaction association context and protein sequence[J]. Molecular Cell, 2023, 83(14): 2595−2611

[20] Peng Cheng, Han Siyu, Zhang Hui, et al. RPITER: A hierarchical deep learning framework for ncRNA–protein interaction prediction[J]. International Journal of Molecular Sciences, 2019, 20(5): 1070

[21] Zhang He, Zhang Liang, Mathews D H, et al. LinearPartition: Linear-time approximation of RNA folding partition function and base-pairing probabilities[J]. Bioinformatics, 2020, 36(Suppl_1): i258−i267

[22] Chen Jiayang, Hu Zhihang, Sun Siqi, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions[J]. arXiv preprint, arXiv: 2204.00300, 2022

[23] Wang Xi, Gu Ruichu, Chen Zhiyuan, et al. Uni-RNA: Universal pre-trained models revolutionize RNA research[EB/OL]. 2023[2026-04-26]. https://www.biorxiv.org/content/10.1101/2023.07.11.548588v1

[24] El Allali A, Elhamraoui Z, Daoud R. Machine learning applications in RNA modification sites prediction[J]. Computational and Structural Biotechnology Journal, 2021, 19: 5510−5524

[25] Hong J, Deng Dezhong, Hendrix D A. Improving transformer secondary structure predictions with secondary structure “fixing” task[C]//Proc of the 18th Machine Learning in Computational Biology Meeting. Cambridge: PMLR, 2024: 267−278

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J].

arXiv preprint, arXiv: 1706.03762, 2017

[27] Zhang He, Liu Hailong, Xu Yushan, et al. Deep generative models design mRNA sequences with enhanced translational capacity and stability[J]. Science, 2025, 390(6773): eadr8470

[28] Penić R J, Vlašić T, Huber R G, et al. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks[J]. Nature Communications, 2025, 16(1): 5671

[29] Liu Xinyu, Zhang Zhen, Tan Chao, et al. Global trends in machine learning applications for single-cell transcriptomics research[J]. Hereditas, 2025, 162(1): 164

[30] Zhang Jun, Lang Mei, Zhou Yaoqi, et al. Predicting RNA structures and functions by artificial intelligence[J]. Trends in Genetics, 2024, 40(1): 94−107

[31] Jung V, Vincent-Cuaz C, Tumescheit C, et al. Decoding the interactions and functions of non-coding RNA with artificial intelligence[J]. Nature Reviews Molecular Cell Biology, 2025, 26(10): 797−818

[32] Weber J S, Carlino M S, Khattak A, et al. Individualised neoantigen therapy mRNA-4157 (V940) plus pembrolizumab versus pembrolizumab monotherapy in resected melanoma (KEYNOTE-942): A randomised, phase 2b study[J]. Lancet, 2024, 403(10427): 632−644



Gu Ruichu, born in 2001. PhD candidate. His main research interests include artificial intelligence methods and RNA research.

顾睿初, 2001 年生。博士研究生。主要研究方向为人工智能方法、RNA 研究。



Liang Yalin, born in 2003. Master candidate. Her main research interests include artificial intelligence methods.

梁雅琳, 2003 年生。硕士研究生。主要研究方向为人工智能方法。



Wen Han, born in 1990. PhD, associate professor. His main research interests include AI for Science, omics research, and RNA research.

温　翰, 1990 年生。博士，副研究员。主要研究方向为 AI for Science、生命组学研究、RNA 研究。