PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
XXXX, XX(XX): 1~16
www.pibb.ac.cn

RMDfold：基于残差Mamba和密集连接的端到端RNA二级结构预测方法*

郝泽洲[1,2] 羊艳玲[1,2] 周 亮[3] 姚旭峰[2]**
([1] 上海理工大学健康科学与工程学院，上海 200093；[2] 上海健康医学院医学影像学院，附属嘉定区中心医院，上海 201318；
[3] 上海健康医学院协同科研中心，上海 201318）

摘要 目的 非编码RNA（non-coding RNA，ncRNA）在多种生命过程中发挥着重要的调控作用，二级结构对于其正常功能实现至关重要。现有的方法在预测RNA二级结构时普遍存在模型复杂度高、难以捕获长程依赖关系以及在复杂结构预测上泛化能力差的问题。因此，需要提出新的方法来预测RNA二级结构。**方法** 本研究提出一种端到端预测模型RMDfold，使用 残差Mamba（residual Mamba，resMamba）与密集连接（dense connection，Dense）相结合的特征提取策略。模型采用一维和二维两阶段的学习过程进行特征建模，在一维序列空间中学习核苷酸间的依赖关系，再将其映射到二维空间进一步学习核苷酸间的配对模式，并通过配对约束确定核苷酸的唯一配对，得到最终的二级结构。在一维与二维阶段中均通过Dense捕捉短程依赖特征，resMamba建模长程依赖关系，实现信息表征的逐级增强与短长程特征的有效融合。为了验证本方法的有效性，与Ufold、REDfold、TransUfold和sincfold四种方法在RNAStra lign、ArchiveII和bpRNA-new三个公开数据集上进行了比较。**结果** 本研究所提出的RMDfold方法在结构预测、假结预测、不同长度的序列预测和模型复杂度分析上优于其他算法，并且所需参数最少且推理速度最快。在结构预测中，该方法在RNAStralign的*F1*、*MCC*、*precision*和*recall*分别达到0.973 5、0.973 1、0.975 6和0.972 3；在ArchiveII的*F1*、*MCC*、*precision*和*recall*分别达到0.854 3、0.855 6、0.874 7和0.876 2；在bpRNA-new的*F1*、*MCC*、*precision*和*recall*分别达到0.382 8、0.401 5、0.536 5和0.318 7。在假结预测中，基于ArchiveII数据集，模型的*F1*、*MCC*、*precision*和*recall*分别达到0.741 4、0.743 3、0.752 5和0.739 6。在不同长度的序列预测中，该方法在200~500 nt序列中仍保持0.70的准确率。在模型复杂度分析中，RMDfold所需参数量为2.886 7 M，推理速度为0.026 0 s。**结论** RMDfold能够准确预测RNA的二级结构，有助于揭示RNA分子在基因表达调控、分子识别和催化中的核心作用；也为疾病相关变异机制解析、RNA靶向药物设计以及进化与比较基因组学研究提供重要支持。

关键词 非编码RNA，RNA二级结构，残差Mamba，密集连接，深度学习
中图分类号 Q522，Q71 **DOI**：10.3724/j.pibb.2025.0387 **CSTR**：32369.14.pibb.20250387

RNA是重要的生物大分子，参与基因的表达和调控。根据其功能和作用的不同，RNA可分为编码RNA和非编码RNA（non-coding RNA，ncRNA），编码RNA在过去的研究中已被广泛关注。随着高通量测序技术的发展，ncRNA的重要性逐渐被揭示出来。尽管ncRNA不编码蛋白质，但在细胞内也发挥着重要的调控作用。近年来的研究表明，ncRNA在基因表达调控、细胞功能、发育以及多种疾病的发生和发展中都扮演着关键角色[1-3]。该领域的研究在过去几十年中迅速发展，成为分子生物学的一个重要方向。

RNA二级结构是RNA分子中核苷酸通过碱基配对（如A-U、G-C以及非经典配对G-U等）形成的局部折叠结构。二级结构对于RNA的正确功能至关重要，作为RNA高级结构和生物学功能的基础，其构象直接影响分子在基因表达调控、分子识

* 上海市科技委员会地方高校能力建设项目（23010502700），上海市“科技创新行动计划”优秀学术/技术带头人计划基金（22XD1401300）和国家自然科学基金（61971275，81830052，82072228，62376152）资助。
** 通讯联系人。
Tel：021-65885239，E-mail：yao6636329@hotmail.com
收稿日期：2025-08-26，接受日期：2025-10-31

别及催化等过程中的作用[4-5]。深入解析RNA二级结构不仅有助于理解疾病相关变异对RNA构象及功能的影响，为疾病的研究提供重要线索[6-7]，还为靶向RNA的药物设计提供关键结构依据[8-9]，并推动进化与比较基因组学研究[10]，因此准确预测RNA二级结构具有重要的价值。相比于编码RNA，ncRNA更多地受功能性结构选择驱动[11]，其二级结构由茎、环、假结（非嵌套）[12]等多种结构单元通过碱基配对与拓扑交叉共同构建，这种多样性与复杂性增加了准确预测其二级结构的难度。

在过去的研究中，已经产生多种用于RNA二级结构预测的方法，主要包括实验方法、传统方法和机器学习方法。最初研究人员通过实验的方式进行预测，如核磁共振[13]、低温电子显微镜[14]以及高通量化学探针技术[15-16]；尽管这些方法能够提供高分辨率的结构信息，但仍存在耗时耗力、样本需求量大、通量有限以及难以捕捉RNA在不同生理条件下动态折叠等局限性[17]。实验流程中对试剂用量和反应条件的严格要求，也使得长链或高度动态的RNA区域常出现信号微弱和背景噪声高的问题[18]。此外，化学探针的浓度过高或过低都可能导致结合效率和定位偏差，也会给数据解析带来不确定性[18]。为解决上述限制，计算预测发挥了重要作用[19]。

传统的二级结构预测的方法以热力学模型为主，例如RNAfold[20]和RNAstructure[21]等，这些方法通过动态规划算法结合Turner热力学参数体系，对所有可能的碱基配对和环结构计算最小自由能并求取全局最优折叠构象。虽然该类方法理论框架清晰、实现成熟且预测速度较快，但当碱基对包含假结构时，一些基于动态编程的方法就会失效，且随着序列长度增长计算复杂度可达到O（n^3）[22]。此外，这些方法的性能在最近几年并未显著提高，平均碱基对预测准确率仍然在50%~70%左右[11]。

机器学习方法通过从数据中自动学习序列-结构映射，已成为RNA二级结构预测的重要工具。机器学习方法包含传统机器学习和深度学习，其中传统机器学习方法如支持向量机等[23]，已被应用于RNA二级结构预测。这些方法主要通过手动提取的特征来构建模型，常用特征包括二核苷酸因子、序列长度和核苷酸比例因子等[24]。CONTRAfold[25]使用条件对数线性模型，通过判别式训练和丰富的特征表示来预测结构。这些方法虽然在一定程度上提升了预测性能，但仍存在特征工程复杂和模型易过拟合的问题[26]，平均碱基对预测准确率同样在50%~70%左右[11]。

深度学习具有良好的特征提取与分析能力，在RNA结构预测中展现出强大的优势和价值。目前，卷积神经网络（convolutional neural networks，CNN）[27]和Transformer[28]等神经网络架构被广泛用于RNA结构预测中。近年来，Ufold[29]创新性地将核苷酸序列转换成图像形式，由16个表示碱基配对和1个表示配对概率形成的一个17通道2D图像作为输入，使用UNet编码器-解码器框架预测接触得分图，随后经过一个解决约束优化问题的后处理步骤得到最佳接触图。TransUfold[30]模型在此基础上引入更复杂的Vision Transformer结构，从而在RNA结构预测中提高了准确性。REDfold[22]模型在Ufold模型的基础上将RNA序列转换为包含146个通道的二维矩阵形式，包括二核苷酸（A-U等）和四核苷酸（AU-GC等），使用基于CNN的残差编码器-解码器网络架构来预测接触得分图，随后同样经过一个解决约束优化问题的后处理步骤得到最佳接触图。这三个模型均对原始的序列数据进行手动特征提取，且最后通过求解优化问题得到最终结构，sincfold模型[31]使用一维和二维残差网络打破了这些步骤，建立了一个端到端的预测模型，性能有一定的提升。尽管这些模型取得一定进展，但CNN和Transformer在处理长程依赖方面仍有不足。CNN的感受野有限，增加卷积深度或增大卷积核大小时会大大增加运行成本和模型参数量[32]；而Transformer虽然能够通过自注意力机制处理长程依赖，但在处理长序列时存在计算复杂度高的问题[33]。最新提出的Mamba在多个领域处理长程依赖任务中均展现出较好的学习能力[34-36]，为RNA结构预测提供新的可能性。

为解决现有方法在捕获长程依赖关系的模型复杂度高、泛化能力较弱的问题，本研究提出基于残差Mamba（residual Mamba，resMamba）和密集连接（dense connection，dense）的端到端预测模型RMDfold。模型使用一维和二维建模策略，在一维序列空间中学习核苷酸间的依赖关系，再将其映射到二维空间以进一步学习核苷酸之间的配对模式，并通过配对约束确定每个核苷酸的唯一配对，得到最终的二级结构。在一维与二维阶段中均通过Dense捕捉短程依赖特征，resMamba建模长程依赖关系，实现信息表征的逐级增强与短长程特征的有效融合。所提出的RMDfold方法与Ufold、

REDfold、TransUfold和sincfold经典模型在RNAStralign、ArchiveII和bpRNA-new三个公开数据集上进行对比，结果表明RMDfold在结构预测、假结预测、不同长度的序列预测和模型复杂度分析上均优于经典模型。

1 材料和方法

1.1 RNA数据及预处理

1.1.1 RNA数据

本研究使用三个公开数据集RNAStralign、ArchiveII和bpRNA-new，详细信息如下：

a. RNAStralign数据集[37]是一个广泛应用于RNA二级结构预测研究的基准数据库，包含来自8种不同RNA家族（16S、5S、Group I Intron、RNaseP、SRP、telomerase、tmRNA和tRNA）的37 149个RNA序列，是最全面的RNA结构集合之一，序列长度为30~1 851 nt。

b. ArchiveII数据集[38]是检验RNA结构预测性能的最广泛使用的数据集，较RNAStrAlign数据集多了23S和Group II Intron两个RNA家族，共3 975个序列，序列长度为28~2 968 nt。本研究选择与RNAStralign中同种类的RNA序列。

c. bpRNA-new数据集[39]来自Rfam 14.2[40]，包括1 500个任何其他数据集中都不存在的新RNA家族，共5 401个序列，序列长度为33~489 nt。

1.1.2 数据预处理

预处理流程包含去冗余、过滤、数据清洗和编码4个步骤。a. 去冗余：以序列为标准，将完全相同的样本归为一类，并在每类中仅保留一个代表样本，最终得到无重复的样本集合[28]。本研究去除了数据集内与数据集间的冗余。b. 数据清洗：剔除包含非“A”、“C”、“G”和“U”字符的序列，以及碱基对缺失的无效样本。c. 过滤：对长度超过600 nt的RNA序列进行过滤，以限制整体实验的运行时间[41]。d. 编码：对序列进行one-hot编码，4种碱基对应的编码分别为“A”：[1，0，0，0]、“C”：[0，1，0，0]、“G”：[0，0，1，0]和“U”：[0，0，0，1]。经过预处理，RNAStralign数据集保留了18 554个RNA序列样本，ArchiveII数据集保留了1 860个样本，bpRNA-new数据集保留了5 388个样本，详细信息如表1所示。

Table 1 Statistical information of the three datasets

Dataset	RNA families	Sequence length	Sample *n*
RNAStralign	8	30－600 nt	18 554
ArchiveII	8	28－593 nt	1 860
bpRNA-new	1 500	33－489 nt	5 388

1.2 RMDfold模型

为解决现有方法在捕获长程依赖关系的模型复杂度高、泛化能力较弱的问题，本研究提出RMDfold二级结构预测模型，使用resMamba和Dense结合以实现长短程特征提取与融合。如图1所示，模型框架依次由一维建模、特征映射和二维建模三部分组成。在一维建模中，模型从RNA序列中学习短长程依赖关系，为结构预测提供基础；在特征映射阶段，将一维特征映射到二维空间以形成约束矩阵，从而表示潜在的碱基配对关系；在二维建模中进一步学习核苷酸之间的配对模式，并通过配对约束确定每个核苷酸的唯一配对，得到最终的二级结构。RMDfold的流程如表2所示。

1.2.1 一维建模

一维建模从RNA序列中提取上下文依赖的序列特征，并捕捉潜在的配对关系，为结构预测提供基础。结构由一维卷积层、一维Dense模块和一维resMamba模块组成。首先，将大小为4×L的RNA编码序列通过一维卷积层提取每个核苷酸的低级特征；其次，通过Dense模块将提取的特征映射到更高维空间，进一步强化局部特征；最后，使用resMamba模块捕捉长距离的上下文依赖及远程碱基间的潜在配对关系。Dense和resMamba结构如下：

Dense模块：该模块使用4层密集连接架构，每一层的输入不仅包含前一层的输出，还融合所有先前层的特征图，从而实现信息的充分传递与累积，缓解梯度消失和信息丢失问题。每层均由一个批量归一化层、一个ReLU激活层和一个一维卷积组成，将通道数从初始的F_1逐步增加到F_1'。为减少模型的计算负担，同时保留重要的特征信息，再经过一个1×1卷积压缩回F_1。在二维建模中，该模块采用二维卷积以适应空间特征的表示需求。

resMamba模块：该模块使用基于状态空间模型（state space model，SSM）的双分支残差结构提取特征。首先，将Dense模块提取的形状为（N，L）的一维特征展平并转置为（L，N），经层归一化后输入Mamba块；其次，特征在Mamba内部分为两个并行分支：在第一分支中，特征经一个线性层扩展到（2L，N），再依次通过一维卷积层、

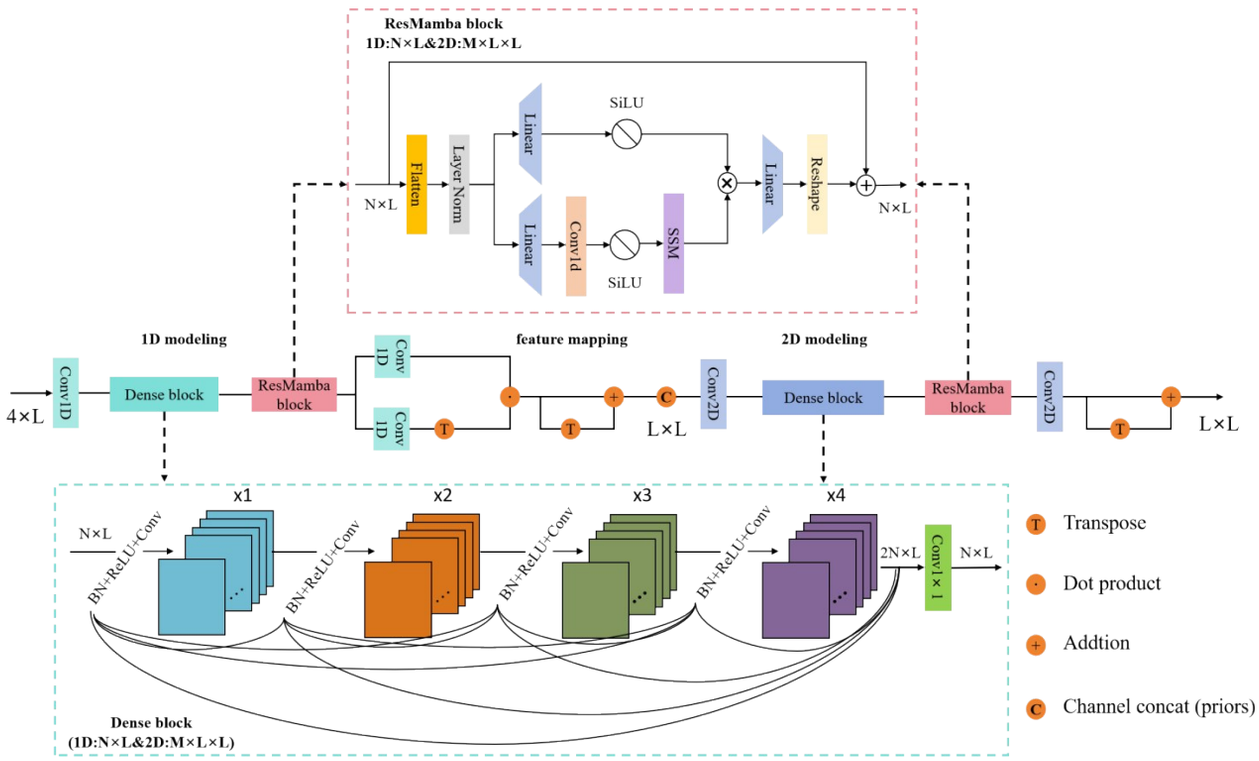


Fig. 1　The proposed RMDfold model

The input is a 4×L one-hot encoded RNA sequence， and the output is an L×L symmetric contact matrix， L is the sequence length， N is the dimension of the feature vector for each nucleotide in the one-dimensional stage， and M is the dimension of the feature vector for each nucleotide in the two-dimensional stage.

Table 2　RMDfold algorithm

RMDfold algorithm
Input： training dataset D= {（x_1， y_1），（x_2， y_2），…，（x_m， y_m）}
Steps：
1. for i=1，…，m： % Obtaining RNA sequence encodings from dataset D
2. F_{1D}^{0}= Conv1D（x_i）；
3. F_{1D}^{1} = Dense1D（F_{1D}^{0}）；
4. F_{1D}^{2} = resMamba（F_{1D}^{1}）；
5. l = *matmul*（Conv1D（F_{1D}^{2}），Conv1D(F_{1D}^{2})^{T}）； % Obtaining L×L mapping matrix
6. $l_{symmetric}$ =(l + l^{T})/2； % Obtaining symmetric mapping matrix
7. l_{final} = $l_{symmetric}$+l_{prior} ； % Integrating prior matrix
8. F_{2D}^{0}= Conv2D（l_{final}）；
9. F_{2D}^{1} = Dense2D（F_{2D}^{0}）；
10. F_{2D}^{2} = resMamba（F_{2D}^{1}）；
11. L = Conv2D（ReLU（F_{2D}^{2}））； % Obtaining L×L contact matrix
12. $L_{symmetric}$ =(L + L^{T})/2； % Obtaining symmetric contact matrix
13. end
14. Y = *postprocessing*（$L_{symmetric}$）； % Finding the maximum activation in each row and column to ensure unique pairings
Output： the final RNA secondary structure contact matrix Y.

SiLU激活函数和SSM；在第二分支中，特征同样通过一个线性层扩展到（2L， N），后经SiLU激活函数；两个分支的结果在逐元素相乘后融合，并通过投影映射到（L， N），再转置回（N， L）；最后，输出特征通过残差连接与输入特征相加，实现信息保留与梯度稳定传递。该模块在一维和二维建模中均适用，对于（M， H， W）的二维矩阵，展平后的$L = H \times W$。

1.2.2 特征映射

特征映射将一维序列信息转化为能够表征核苷酸间相互作用的二维映射矩阵，为后续结构预测提供基础表征。首先，将一维建模提取的特征通过两个一维卷积层，生成两个大小为$E \times L$的压缩编码矩阵；其次，将一个$E \times L$矩阵与另一个$E \times L$矩阵的转置相乘，得到一个初步的$L \times L$的二维映射矩阵；再次，为确保接触矩阵的对称性，该矩阵与其转置矩阵相加，从而构建对称映射矩阵；最后，加入一个通道，以编码不同碱基对（如C-G、U-A和G-U）之间的相互作用先验信息，从而在映射矩阵中注入这些碱基对的不同键合强度。

1.2.3 二维建模

二维建模进一步捕获核苷酸间的空间配对关系，从而生成更加准确的接触矩阵。二维阶段和一维阶段类似，结构由二维卷积层、二维Dense模块和二维resMamba模块组成。首先，将映射矩阵经过一个二维卷积提取低级空间特征；，通过一个4层的Dense模块提取高维特征，每层由1个批量归一化层、1个ReLU激活层和1个二维卷积组成，将通道数由F_2翻倍增加到$F_2^{'}$，再通过1×1卷积降为F_2；再次，使用resMamba模块进一步提取全局范围的配对特征并进行特征融合，经过一个二维卷积得到$L \times L$的接触矩阵，并将该矩阵与其转置相加得到对称接触矩阵。最后，通过配对约束得到最终的二级结构。

1.3 实验设计

为了验证方法的有效性，本研究进行数据集分布差异分析、消融实验、对比实验和预测二级结构对比显示四个实验，本研究将RNAStralign数据以8：2的比例随机分为训练集和内部测试集；ArchiveII作为家族内测试集；bpRNA-new作为跨家族测试集。

数据集分布差异分析：本研究提出各数据集样本均为独立同分布的假设，使用最大均值差异（maximum mean discrepancy，MMD）分别对训练集与内部测试集之间、训练集与家族内测试集之间、训练集与跨家族测试集之间，以及家族内测试集与跨家族测试集进行统计分析[42]，并迭代1 000次以获得稳定结果（P<0.05时表示结果具有统计学差异）。

消融实验：本研究通过设计两个消融实验来验证模型有效性。首先，分别移除二维resMamaba模块、二维建模模块和一维resMamba模块验证对预测结果的影响。其次，分别移除二维建模模块和resMamba模块，以及将Mamba分别平行替换为CNN和Transformer编码器验证对长序列预测结果的影响。

对比试验：为了验证RMDfold在RNA二级结构预测的有效性，本方法与Ufold、REDfold、TransUfold和sincfold模型进行了RNAStralign测试、模型复杂度分析、家族内测试、不同序列长度测试、假结预测和跨家族测试。

预测二级结构对比显示：为了观察核苷酸的配对细节，以长度为141 nt的粪肠球菌AE016830.1_3 201 675和长度为401 nt的普雷沃氏菌TRW-264731为例，本研究使用Forna[43]对不同模型的预测结果进行二级结构对比分析。

为了进行定量分析，本研究使用精确度、召回率、F1分数（F1 score，$F1$）[44]和马修斯相关系数（matthews correlation coefficient，MCC）[45]评估模型性能，使用Friedman检验方法分析模型的性能显著性（P<0.05），使用参数量和推理时间评估模型复杂度。4个性能指标公式如下：

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recal} \tag{3}$$

$$MCC = \frac{TP \times TN + FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

其中，研究基于真实接触矩阵和预测接触矩阵计算指标，矩阵中的每个元素表示该位置的两个碱基配对情况，碱基配对则值为1，反之为0。具体的，TP指真实为配对且预测也为配对的个数；FP指真实为不配对但预测为配对的个数；TN指真实为不配对且预测也为不配对的个数；FN指真实为配对但预测为不配对的个数。

训练主要超参数如下：epoch为100、批次大小为1、累计步数为4、使用学习率为1×10^{-4}的Adam优化器、损失函数使用L1正则化的交叉熵损失函数[46]，正则化系数为0.01、使用混合精度训练加速，种子数设置为0。本研究所有实验均在同一平台上完成，基于Windows系统，硬件环境使用NVIDIA GeForce RTX 3090 GPU作为计算加速器；软件环境使用Python 3.10编程语言、PyTorch 2.4深度学习框架和CUDA 12.4的并行计算平台。

2　结 果

2.1　数据集分布差异性分析

如图2所示，训练集与内部测试集之间的*P*值为0.798，置换后的*MMD*值包含实际*MMD*值，表明二者在分布上无显著差异，可适用于评估模型的拟合能力与稳定性。而其余三组之间的*P*值均为0.00，且置换后的*MMD*值远小于实际*MMD*值，表明存在显著的分布差异，可用于检验模型对不同分布数据预测的泛化能力与鲁棒性。

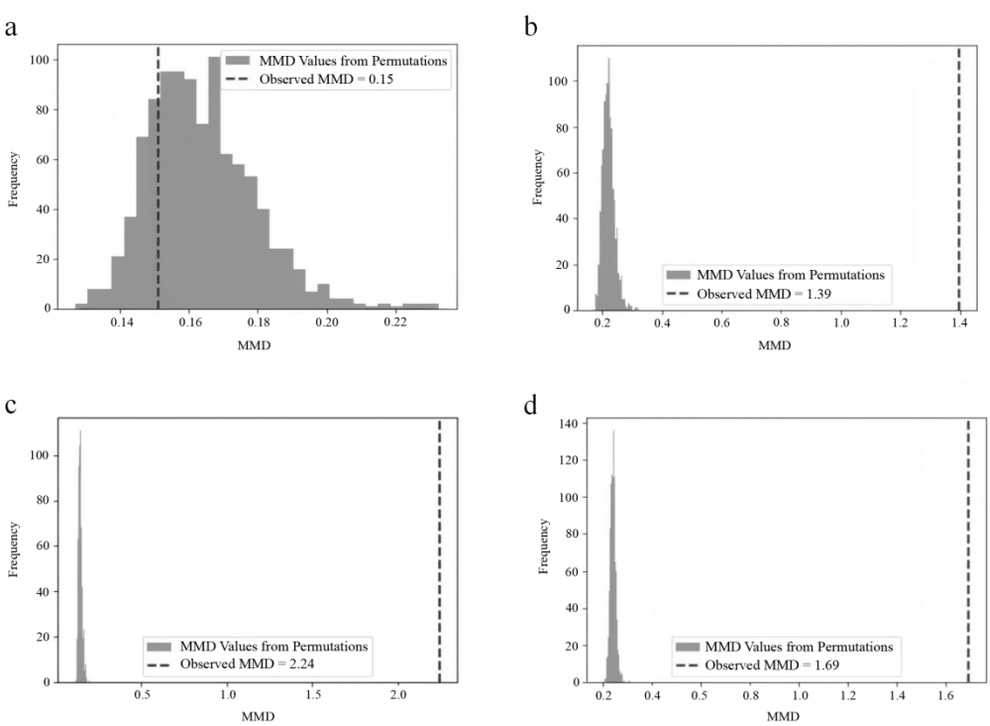


Fig. 2　Analysis of differences in dataset distribution

（a） Training set and internal test set，（b） training set and in-family test set，（c） training set and cross-family test set，（d） in-family test set and cross-family test set. MDD： maximum mean discrepancy.

2.2　消融实验

表3显示本模型在RNAStralign数据集上消融实验的结果，当移除模块时，RMDfold预测性能下降。RMDfold模型的*F1*为0.973 5，移除二维resMamba模块时，*F1*降低0.009 8；移除二维建模模块时，*F1*降低0.095 2；移除二维建模模块和一维resMamba模块时，*F1*降低0.129 9。

如图3所示，去除二维建模模块或resMamba模块均会使预测长序列（200~600 nt）的准确率下降。RMDfold在6个不同长度区间预测的平均*F1*为0.993 8、0.989 0、0.815 8、0.849 4、0.880 8和0.928 7，移除二维建模模块时，*F1*分别降低0.021 6、0.017 3、0.522 4、0.564 4、0.595 4和0.621 4；移除resMamba模块时，*F1*分别降低0.002 8、0.001 3、0.097 0、0.101 3、0.081 9和0.043 4。此外，将Mamba平行替换为CNN或Transformer会降低长序列预测能力。替换为CNN时，*F1*分别降低0.004 2、0.002 4、0.112 5、0.114 1、0.072 3、0.025 6；替换为Transformer时，在100~200 nt的序列中*F1*提高了0.000 3，其余序列区间*F1*分别降低

0.002 7、0.008 1、0.014 5、0.011 3和0.011 0。显著性分析结果表明，模型间性能差异显著（均为P=0.00）。

Table 3 Results of the model ablation experiments

Model	*precision*	*recall*	*F1*	*MCC*
RMDfold	0.975 6	**0.972 3**	**0.973 5**	**0.973 1**
w/o 2D resMamba module	**0.986 4**	0.951 0	0.963 7	0.965 7
w/o 2D modeling module	0.872 1	0.878 3	0.873 7	0.874 0
w/o 2D modeling and 1D resMamba modules	0.846 4	0.844 1	0.843 6	0.843 9

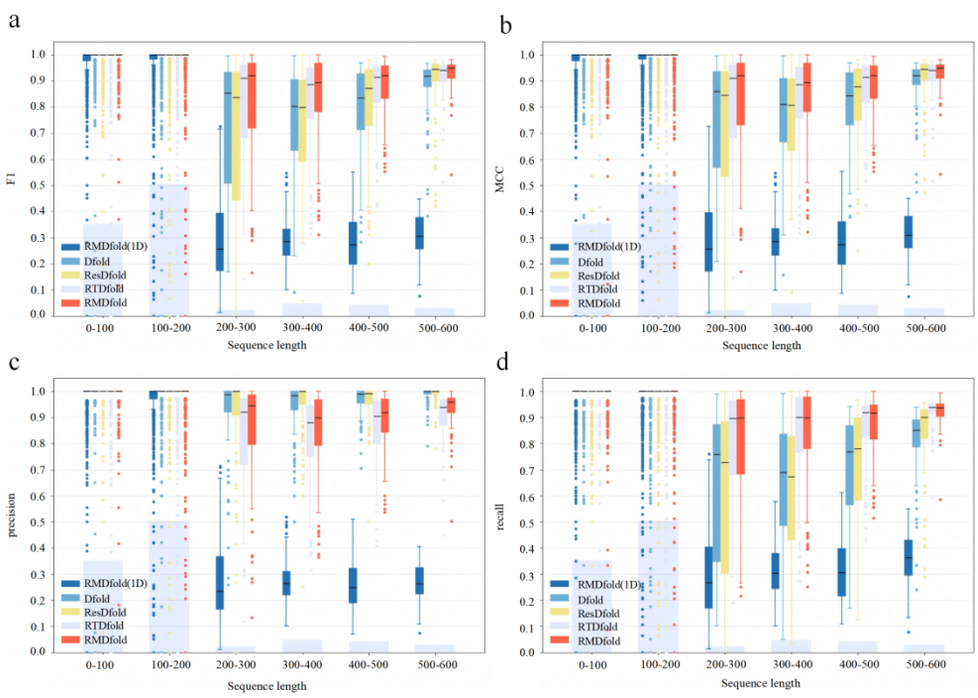


Fig. 3 Visualization of prediction metrics for different average sequence lengths ranges

（a） F1，（b） MCC，（c） precision，（d） recall. RMDfold（1D）：removed 2D modeling module；Dfold：removed resMamba module；ResDfold：parallel replacement of Mamba with CNN；RTDfold：parallel replacement of Mamba with Transformer；RMDfold：the proposed model.

2.3 RNAStralign数据集预测

2.3.1 RNA结构预测

表4所示，RMDfold模型的平均*F1*与*MCC*最高，分别达到0.973 5和0.973 1，显示出更高的RNA二级结构预测准确性。图4显示出各个模型样本预测的*F1*、*MCC*、*precision*和*rercall*分布情况，其中，RMDfold的样本分布更为集中，在多个样本上的表现较为一致，具有良好的稳定性。显著性分析结果表明，模型间性能差异显著（均为P=0.00）。

2.3.2 模型复杂度分析

如表5所示，RMDfold所需参数最少且推理速度最快，其参数量为2.886 7 M，推理速度为0.026 0 s。

Table 4 Comparison of the performance of different methods on the RNAStralign dataset

Model	*precision*	*recall*	*F1*	*MCC*
Ufold	0.968 7	0.973 2	0.970 3	0.938 5
REDfold	0.970 5	**0.975 2**	0.972 2	0.972 3
TransUfold	0.959 0	0.969 2	0.963 3	0.920 7
sincfold	**0.982 5**	0.957 7	0.967 1	0.968 5
RMDfold	0.975 6	0.972 3	**0.973 5**	**0.973 1**

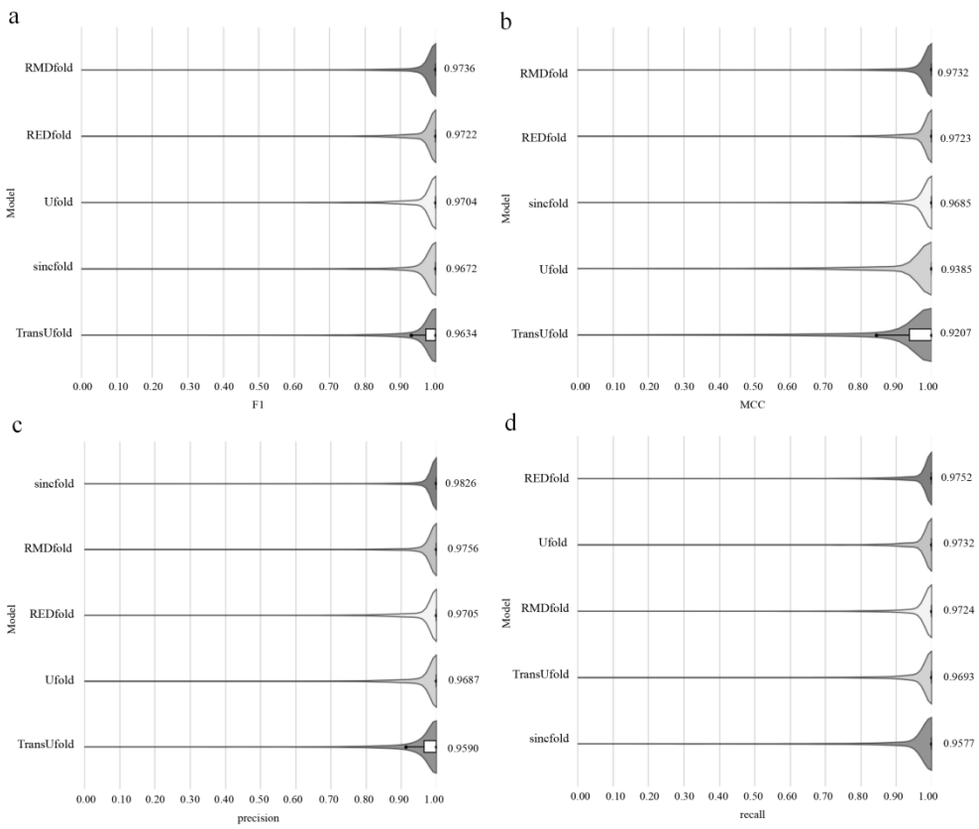


Fig. 4　Visualization of prediction metrics of different methods on the RNAStralign dataset

（a）F1，（b）MCC，（c）precision，（d）recall.

Table 5　Parameters and inference time on the RNAStralign test set

Model	Parameters/M	Time per seq/s
Ufold	8.641 3	0.124 9
REDfold	3.971 6	0.123 9
TransUfold	29.654 0	0.099 8
sincfold	9.691 5	0.098 0
RMDfold	**2.886 7**	**0.026 0**

2.4　家族内数据集预测

2.4.1　RNA结构预测

表6所示，家族内RNA二级结构预测优于经典模型，RMDfold的平均*F1*与*MCC*最高，分别达到0.854 3和0.855 6，显示出更好的泛化能力；*precision*位于第二，仅次于sincfold，达到了0.874 7；*recall*位于第二，仅次于REDfold，达到了0.846 2。图5补充说明RMDfold模型具有最高的中位数和最小的分布差异，除*precision*外，*F1*、*MCC*和*recall*均最高，分别为0.942 5、0.942 8和0.969 6，具有较高的稳定性，*precision*为0.960 7，仅次于sincfold。显著性分析结果表明，模型间性能差异显著（均为*P*=0.00）。

Table 6　Comparison of the performance of different methods on the ArchiveII dataset

Model	*precision*	*recall*	*F1*	*MCC*
Ufold	0.847 0	0.839 3	0.839 3	0.691 5
REDfold	0.854 5	**0.851 9**	0.849 6	0.850 7
TransUfold	0.815 6	0.830 6	0.819 8	0.621 5
sincfold	**0.908 0**	0.791 5	0.827 7	0.835 9
RMDfold	0.874 7	0.846 2	**0.854 3**	**0.855 6**

2.4.2　不同序列长度预测

图6所示，提出的RMDfold展现出良好的预测能力。在不同长度区间（0~100 nt，100~200 nt，200~300 nt，300~400 nt，400~500 nt，500~600 nt），RMDfold在200~600 nt区间明显优于其他模型。在较短长度区间（0~100 nt，100~200 nt），RMDfold的平均*F1*、*MCC*和*recall*仅次于

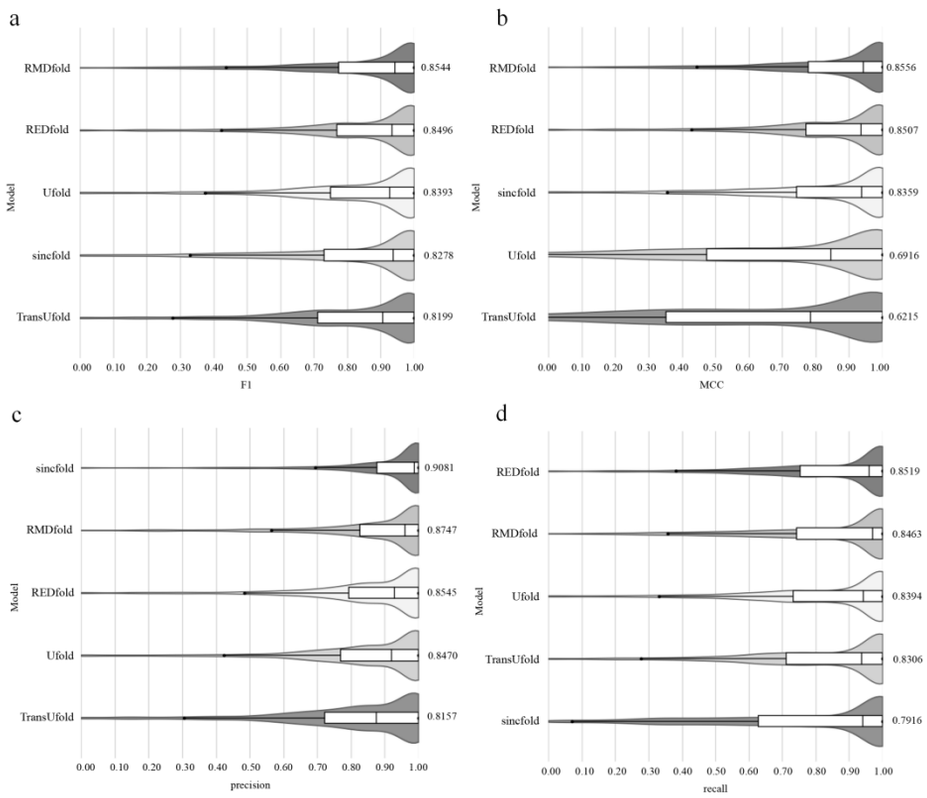


Fig. 5　Visualization of prediction metrics of different methods on the ArchiveII dataset

（a） F1，（b） MCC，（c） precision，（d） recall.

REDfold，分别为：（0.896 7，0.896 5，0.886 0）和（0.914 6、0.916 7、0.904 8）；在中长度区间（200~300 nt，300~400 nt，400~500 nt，500~600 nt），RMDfold的平均*F1*、*MCC*和*recall*最高，分别为（0.726 3，0.729 3，0.706 0），（0.769 7，0.770 8，0.763 2），（0.722 9，0.724 3，0.742 6）和（0.582 9，0.587 2，0.628 8）；在6个长度区间平均*presicion*仅次于sincfold，分别为0.924 3，0.935 7，0.760 4，0.782 1，0.710 2，0.556 0。显著性分析表明除*F1*在500~600 nt长度区间外，均具有显著性差异（P=0.00）。在500~600 nt区间内P值为0.092 4，这可能由于该区间样本量为36，受样本影响较大，因此未能观察到差异。综上表明，RMDfold的结构预测性能整体优于其他方法，尤其在中长度区间表现稳定，具有较高的准确率。

2.4.3　假结预测

2.5　跨家族数据集预测

如表8所示，RMDfold在跨家族预测显示出良好的结果，*F1*和*MCC*分别为0.382 8和0.401 5，但*recall*低于REDfold和TransUfold。

2.6　预测二级结构对比显示

如图7所示，RMDfold在茎环结构和非典型碱基对G-U预测中表现出较高的预测准确率。RMDfold在预测AE016830.1_3 201 675样本二级结

表7　所示，对筛选后的480条含假结的RNA序列预测时，RMDfold性能最佳，*F1*和*MCC*分别达到0.741 4和0.743 3。

Table 7　Comparison of the performance of different methods on pseudoknot

Model	*precision*	*recall*	*F1*	*MCC*
Ufold	0.704 8	0.663 0	0.679 3	0.372 5
REDfold	0.710 2	0.681 2	0.691 8	0.693 3
TransUfold	0.667 3	0.642 1	0.651 3	0.285 0
sincfold	**0.838 3**	0.595 4	0.677 0	0.695 0
RMDfold	0.752 5	**0.739 6**	**0.741 4**	**0.743 3**

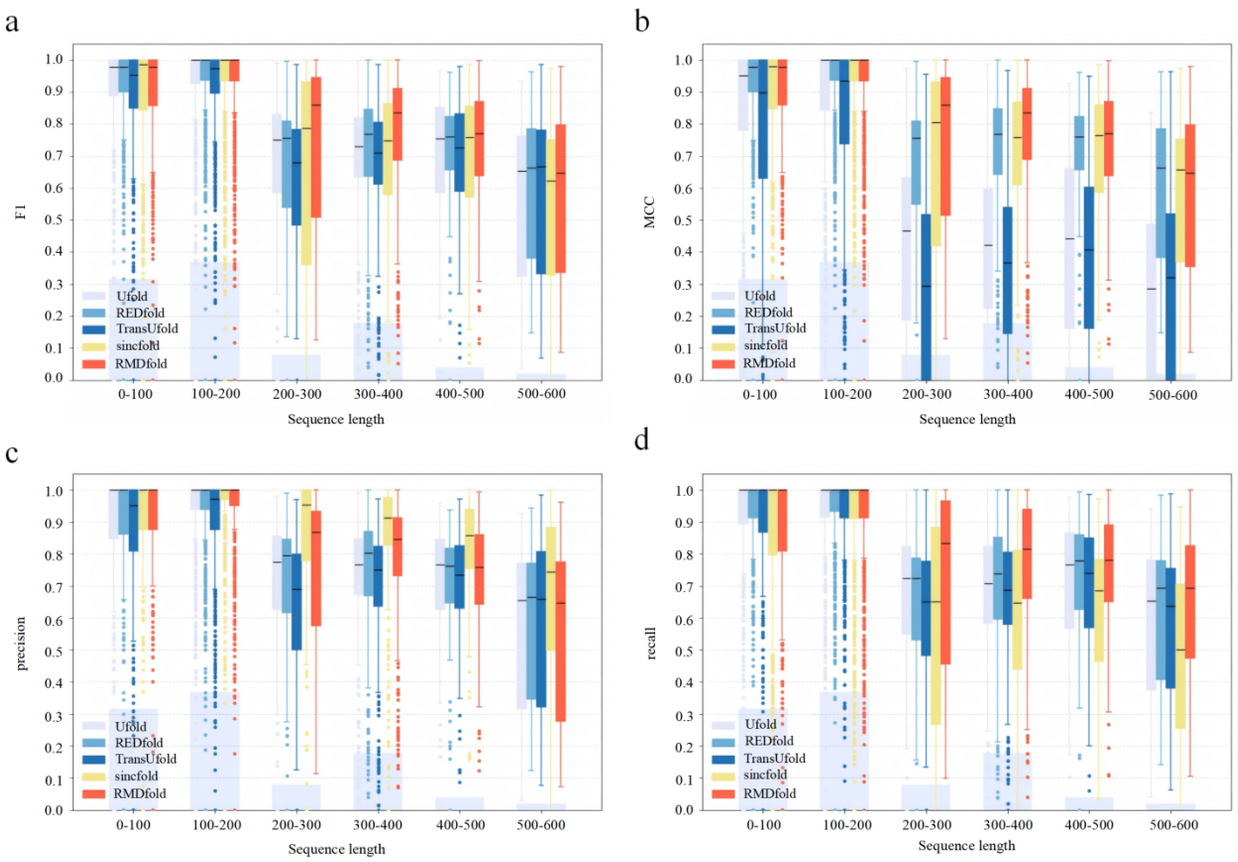


Fig. 6　Visualization of prediction metrics of different methods across different mean sequence lengths ranges

（a）F1，（b）MCC，（c）precision，（d）recall.

Table 8　Performance of different methods on the bprna−new dataset

Model	*precision*	*recall*	*F1*	*MCC*
Ufold	0.427 4	0.287 2	0.325 7	0.052
REDfold	0.433 3	**0.352 8**	0.374 4	0.379 9
TransUfold	0.311 4	0.320 6	0.309 4	-0.225 7
sincfold	0.466 1	0.203 2	0.264 8	0.293 3
RMDfold	**0.536 5**	0.318 7	**0.382 8**	**0.401 5**

构上效果最佳，*F1* 值可达0.977 2，仅出现一处预测配对错误；REDfold有多个配对未预测出，*F1* 为0.860 7；sincfold、TransUfold和Ufold表现出明显差异，*F1* 分别为0.676 4、0.425 0和0.392 9（预测错误如箭头所示）。

如图8所示，RMDfold在长序列预测中具有更高的准确率，在环状结构、茎环结构和非典型碱基对G-U预测也表现出更优的性能。RMDfold在预测Prev.rumi._TRW-264731样本二级结构上效果最佳，*F1* 值可达0.967 7，仅出现两处预测配对错误；REDfold出现多处预测错误，导致其在茎环结构预测方面出现严重失真，*F1* 为0.874 4；sincfold预测的环状结构长度高于真实值，*F1* 为0.790 9；而TransUfold未准确预测出真实的环状结构，并且在茎环结构预测方面也出现失真情况，*F1* 为0.840 7；Ufold与REDfold预测相似，但出现更多的预测错误，*F1* 为0.779 2（预测错误如箭头所示）。

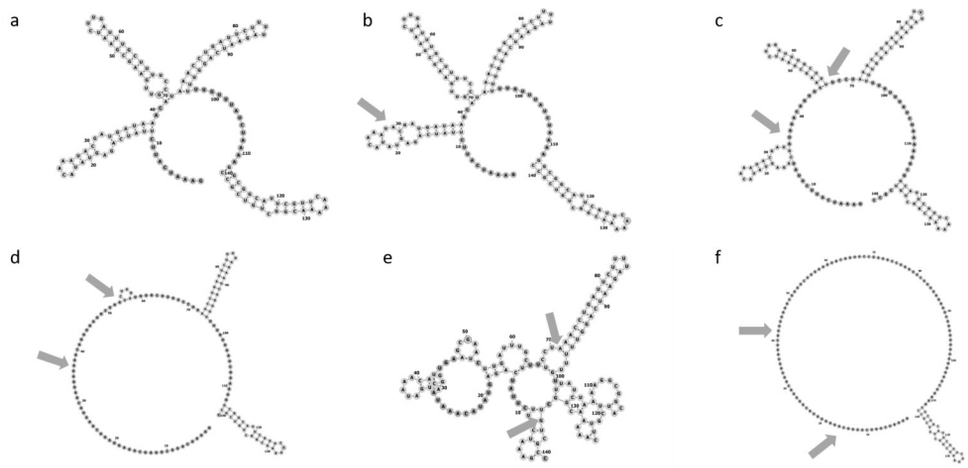


Fig. 7　Secondary structure comparison for sample AE016830.1_3 201 675

（a） Ground truth，（b） RMDfold *F1* = 0.977 2，（c） REDfold= 0.860 7，（d） sincfold = 0.676 4，（e） TransUfold= 0.425 0，（f） Ufold= 0.392 9.

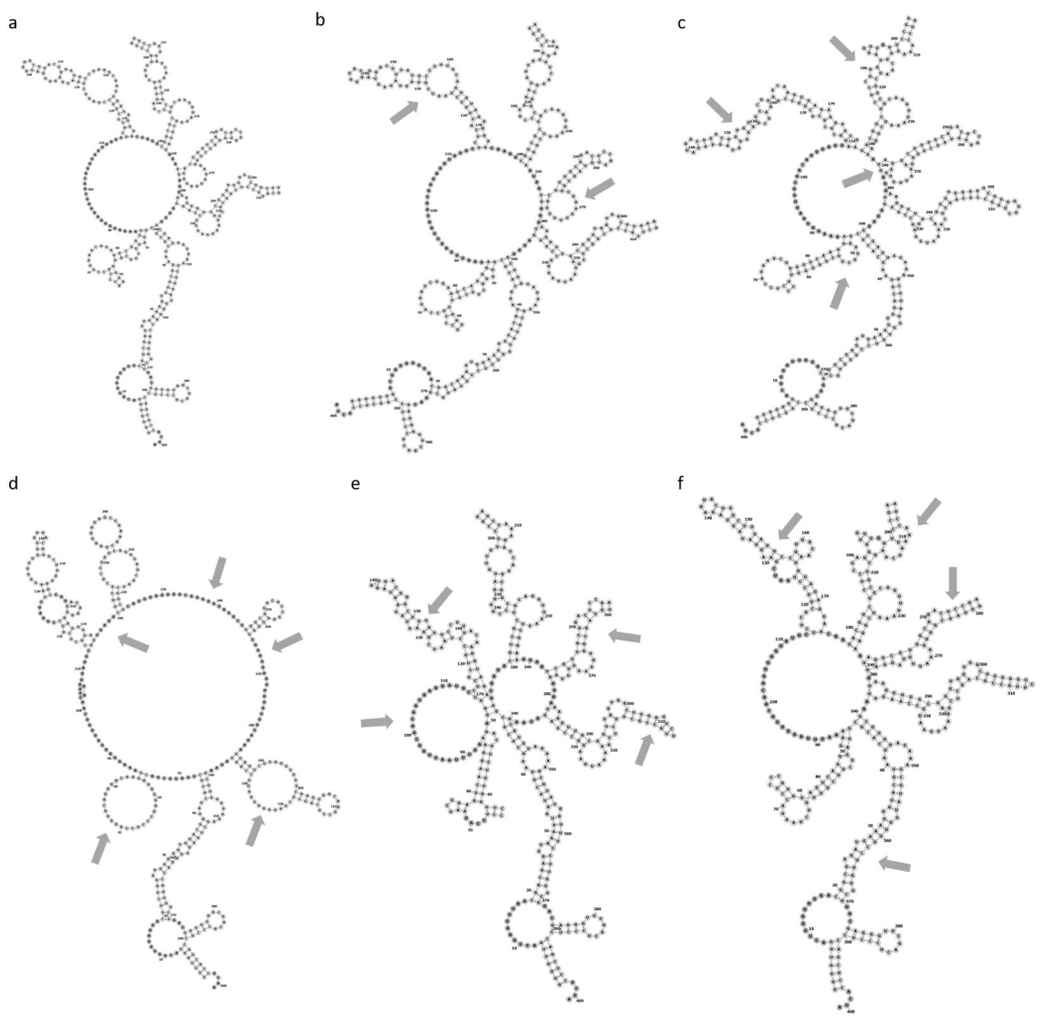


Fig. 8　Secondary structure comparison for sample Prev.rumi._TRW−264731

（a） Ground truth，（b） RMDfold= 0.967 7，（c） REDfold= 0.874 4，（d） sincfold= 0.790 9，（e） TransUfold= 0.840 7，（f） Ufold = 0.779 2.

3 讨论

本研究提出RMDfold RNA二级结构预测模型，在多个公开数据集上的预测准确率均优于经典模型，实现出更高的预测精度。尤其是在假结预测和长序列预测任务中，该模型同样展现出较强的泛化能力和稳定性。此外该模型所需参数最少且推理速度最快，有效提升模型的实用性。

本研究在RNAStralign、ArchiveII和bpRNA-new三个公开数据集上对模型进行系统评估。RNAStralign是目前被广泛使用的基准数据集，提供高质量的序列和结构信息，常被用于模型的训练与性能评估；ArchiveII作为一个包含多种RNA家族和复杂结构的数据集，在检验模型泛化能力和复杂结构预测能力方面具有重要意义，基于RNAStralign-ArchiveII的评估方案已被广泛使用[29-31]；bpRNA-new包含其他数据集中未覆盖的RNA家族序列，被广泛用于验证模型在未知家族预测的稳定性与鲁棒性。因此，通过这三个数据集的综合评估，可全面分析RMDfold在RNA二级结构预测中的预测能力。

RMDfold模型基于一维-二维建模框架[31]，不同于UFold和REDfold等方法需要在一维-二维转换中依赖预处理或人工设定，RMDfold可在训练过程中完成从一维表示到二维表示的学习。此外，相比于需要通过约束优化来修正预测结果，RMDfold只需简单的激活选择即可得到最终接触矩阵。因此，RMDfold只需最少的物理假设就可准确预测结构。

RMDfold模型基于Dense和resMamba模块进行特征提取。Dense在特征融合方面优于常规CNN和残差结构。与普通CNN相比，Dense引入跨层密集连接，使得浅层提取的局部特征能够直接传递到深层，有效缓解梯度消失问题并增强特征的多样性[47]；与残差结构相比，Dense不只是进行简单的跳跃连接，而是将所有前层特征级联用于当前层输入，从而实现更充分的信息重用与融合，提高模型在捕捉局部序列模式时的能力[48]。另一方面，resMamba结构中Mamba在处理长程依赖方面优于CNN和Transformer。CNN通过扩大感受野提取长程信息，但在一定程度上会增加模型运行成本；Transformer依赖全局自注意力机制建模任意位置间的依赖关系，但其计算复杂度较高[33]，资源开销大，特别是在长序列任务中容易受限。而Mamba采用高效的状态空间建模策略，能够以线性复杂度捕捉长距离序列信息，在保持全局感知能力的同时降低计算负担[49]。通过将Mamba替换为CNN或Transformer进行结构预测，替换后的模型在长序列预测中准确率均低于Mamba模型，CNN模型的参数量为5 723 137，约为Mamba模型的两倍，预测速度为0.027 0 s；Transformer模型的参数量为2 840 417，略低于Mamba模型，但预测速度为0.197 8 s，远高于Mamba模型。

本研究仍有一些局限性。a. 本研究主要使用了比对同源序列生成的数据集，可能会导致RNA结构预测存在一定的偏差。后续将使用高通量数据集进行研究，进一步提高模型的预测准确性。b. 本研究仅根据序列预测RNA结构，未考虑温度、盐浓度等其他因素对RNA结构的影响。后续将使用多模态数据，进一步提升模型的鲁棒性和可靠性。c. 本研究提出的模型在面对训练集中未覆盖的RNA家族时，其泛化能力显著下降。这种局限源于模型主要学习的是训练集中的特征，对于陌生的序列分布缺乏足够的建模能力。后续将考虑对比学习等自监督策略[50]，通过最大化相似结构之间的表示一致性、最小化不同结构的距离，提升模型对结构通用特征的提取能力，进而增强其对未知RNA家族的泛化能力。

4 结论

本研究提出的RMDfold模型在二级结构预测中显示出良好的预测性能。同时，模型所需参数最少且推理速度最快，表现出较好的计算效率与实用性。该模型为揭示RNA分子的正常功能提供了有力支持；并为疾病相关变异机制解析、RNA靶向药物设计以及进化与比较基因组学研究奠定了重要结构基础。

数据可用性声明 本研究主要源代码可在https：//github.com/haozezhou98/RMDfold_code获取。

参考文献

[1] Mattick J S, Amaral P P, Carninci P, *et al*. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol, 2023, **24**(6): 430-447

[2] Razin S V, Gavrilov A A. Non-coding RNAs in chromatin folding and nuclear organization. Cell Mol Life Sci, 2021, **78**(14): 5489-5504

[3] Toden S, Goel A. Non-coding RNAs as liquid biopsy biomarkers in cancer. Br J Cancer, 2022, **126**(3): 351-360

[4] Fricke M, Gerst R, Ibrahim B, *et al*. Global importance of RNA secondary structures in protein-coding sequences. Bioinformatics, 2019, **35**(4): 579-583

[5] Sanchez de Groot N, Armaos A, Graña-Montes R, *et al*. RNA structure drives interaction with proteins. Nat Commun, 2019, **10** (1): 3246

[6] Wieder N, D'Souza E N, Dawes R, *et al*. The role of untranslated region variants in Mendelian disease: a review. Eur J Hum Genet, 2025, **33**(9): 1096-1105

[7] Bohnsack K E, Yi S, Venus S, *et al*. Cellular functions of eukaryotic RNA helicases and their links to human diseases. Nat Rev Mol Cell Biol, 2023, **24**(10): 749-769

[8] Bose R, Saleem I, Mustoe A M. Causes, functions, and therapeutic possibilities of RNA secondary structure ensembles and alternative states. Cell Chem Biol, 2024, **31**(1): 17-35

[9] Shao Y, Zhang Q C. Targeting RNA structures in diseases with small molecules. Essays Biochem, 2020, **64**(6): 955-966

[10] Rivas E. Evolutionary conservation of RNA sequence and structure. Wiley Interdiscip Rev RNA, 2021, **12**(5): e1649

[11] Bugnon L A, Edera A A, Prochetto S, *et al*. Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches. Brief Bioinform, 2022, **23**(4): bbac205

[12] Hollar A, Bursey H, Jabbari H. Pseudoknots in RNA structure prediction. Curr Protoc, 2023, **3**(2): e661

[13] Grossman B D, Marchant J, Summers M F. Application of NMR to large RNAs. J Mol Biol, 2025: 169406

[14] Kretsch R C, Wu Y, Shabalina S A, *et al*. Naturally ornate RNA-only complexes revealed by cryo-EM. Nature, 2025, **643**(8073): 1135-1142

[15] Bohn P, Gribling-Burrer A S, Ambi U B, *et al*. Nano-DMS-MaP allows isoform-specific RNA structure determination. Nat Methods, 2023, **20**(6): 849-859

[16] Boerneke M A, Ehrhardt J E, Weeks K M. Physical and functional analysis of viral RNA genomes by SHAPE. Annu Rev Virol, 2019, **6**(1): 93-117

[17] Bose E, Xiong S, Jones A N. Probing RNA structure and dynamics using nanopore and next generation sequencing. J Biol Chem, 2024, **300**(6): 107317

[18] Arnold E B, Cohn D, Bose E, *et al*. Investigating the interplay between RNA structural dynamics and RNA chemical probing experiments. Nucleic Acids Res, 2025, **53**(7): gkaf290

[19] Sapoval N, Aghazadeh A, Nute M G, *et al*. Current progress and open challenges for applying deep learning across biosciences. Nat Commun, 2022, **13**(1): 1728

[20] Lorenz R, Bernhart S H, Höner Zu Siederdissen C, *et al*. ViennaRNA package 2.0. Algorithms Mol Biol, 2011, **6**: 26

[21] Reuter J S, Mathews D H. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics, 2010, **11**: 129

[22] Chen C C, Chan Y M. REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network. BMC Bioinformatics, 2023, **24**(1): 122

[23] Justyna M, Antczak M, Szachniuk M. Machine learning for RNA 2D structure prediction benchmarked on experimental data. Brief Bioinform, 2023, **24**(3): bbad153

[24] Hor C Y, Yang C B, Chang C H, *et al*. A tool preference choice method for RNA secondary structure prediction by SVM with statistical tests. Evol Bioinform Online, 2013, 9: 163-184

[25] Do C B, Woods D A, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics, 2006, **22**(14): e90-8

[26] Sato K, Hamada M. Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. Brief Bioinform, 2023, **24**(4): bbad186

[27] Zhang H, Zhang C, Li Z, *et al*. A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. Front Genet, 2019, **10**: 467

[28] Chen X, Li Y, Umarov R, *et al*. RNA Secondary Structure Prediction By Learning Unrolled Algorithms// ICLR. The 8th International Conference on Learning Representations (ICLR 2020). Red Hook: Curran Associates, Inc., 2020: 12831-12849

[29] Fu L, Cao Y, Wu J, *et al*. UFold: fast and accurate RNA secondary structure prediction with deep learning. Nucleic Acids Res, 2022, **50**(3): e14

[30] Wang Y, Zhang H, Xu Z, *et al*. TransUFold: unlocking the structural complexity of short and long RNA with pseudoknots. Math Biosci Eng, 2023, **20**(11): 19320-19340

[31] Bugnon L A, Di Persia L, Gerard M, *et al*. sincFold: end-to-end learning of short- and long-range interactions in RNA secondary structure. Brief Bioinform, 2024, **25**(4): bbae271

[32] Rangel G, Cuevas-Tello J C, Nunez-Varela J, *et al*. A survey on convolutional neural networks and their performance limitations in image recognition tasks. J Sens, 2024, **2024**(1): 2797320

[33] Farina M, Ahmad U, Taha A, *et al*. Sparsity in transformers: a systematic literature review. Neurocomputing, 2024, **582**: 127468

[34] Zhao Q, Li Z, Mao Q, *et al*. MambaCpG: an accurate model for single-cell DNA methylation status imputation using mamba. Brief Bioinform, 2025, **26**(4): bbaf360

[35] Liu X, Fu H, Yang Y, *et al*. Bio-inspired mamba for antibody-antigen interaction prediction. Biomolecules, 2025, **15**(6): 764

[36] Peng F Z, Wang C, Chen T, *et al*. PTM-Mamba: a PTM-aware protein language model with bidirectional gated Mamba blocks. Nat Methods, 2025, **22**(5): 945-949

[37] Tan Z, Fu Y, Sharma G, *et al*. TurboFold II : RNA structural alignment and secondary structure prediction informed by multiple homologs. Nucleic Acids Res, 2017, **45**(20): 11570-11581

[38] Sloma M F, Mathews D H. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. RNA, 2016, **22**(12): 1808-1818

[39] Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration.

Nat Commun, 2021, **12**(1): 941

[40] Kalvari I, Nawrocki E P, Ontiveros-Palacios N, *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res, 2021, **49**(D1): D192-D200

[41] Szikszai M, Wise M, Datta A, *et al*. Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. Bioinformatics, 2022, **38**(16): 3892-3899

[42] Papangelou C, Kyriakidis K, Natsiavas P, *et al*. Reliable machine learning models in genomic medicine using conformal prediction. Front Bioinform, 2025, **5**: 1507448

[43] Kerpedjiev P, Hammer S, Hofacker I L. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. Bioinformatics, 2015, **31**(20): 3377-3379

[44] Terven J, Cordova-Esparza D M, Romero-González J A, *et al*. A comprehensive survey of loss functions and metrics in deep learning. Artif Intell Rev, 2025, **58**(7): 195

[45] de la Cruz Huayanay A, Bazán J L, Russo C M. Performance of evaluation metrics for classification in imbalanced data. Comput Stat, 2025, **40**(3): 1447-1473

[46] Li C, Liu K, Liu S. A survey of loss functions in deep learning. Mathematics, 2025, **13**(15): 2417

[47] Huang G, Liu Z, Pleiss G, *et al*. Convolutional networks with dense connectivity. IEEE Trans Pattern Anal Mach Intell, 2022, **44**(12): 8704-8716

[48] Alzubaidi L, Zhang J, Humaidi A J, *et al*. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data, 2021, **8**(1): 53

[49] Salam A, Mahmud R, Islam T, *et al*. A comprehensive survey on mamba: architectures, challenges, and opportunities. Computer, 2025, **58**(8): 64-76

[50] Jaiswal A, Babu A R, Zadeh M Z, *et al*. A survey on contrastive self-supervised learning. Technologies, 2021, **9**(1): 2

RMDfold: an End-to-end RNA Secondary Structure Prediction Method Based on Residual Mamba and Dense Connections*

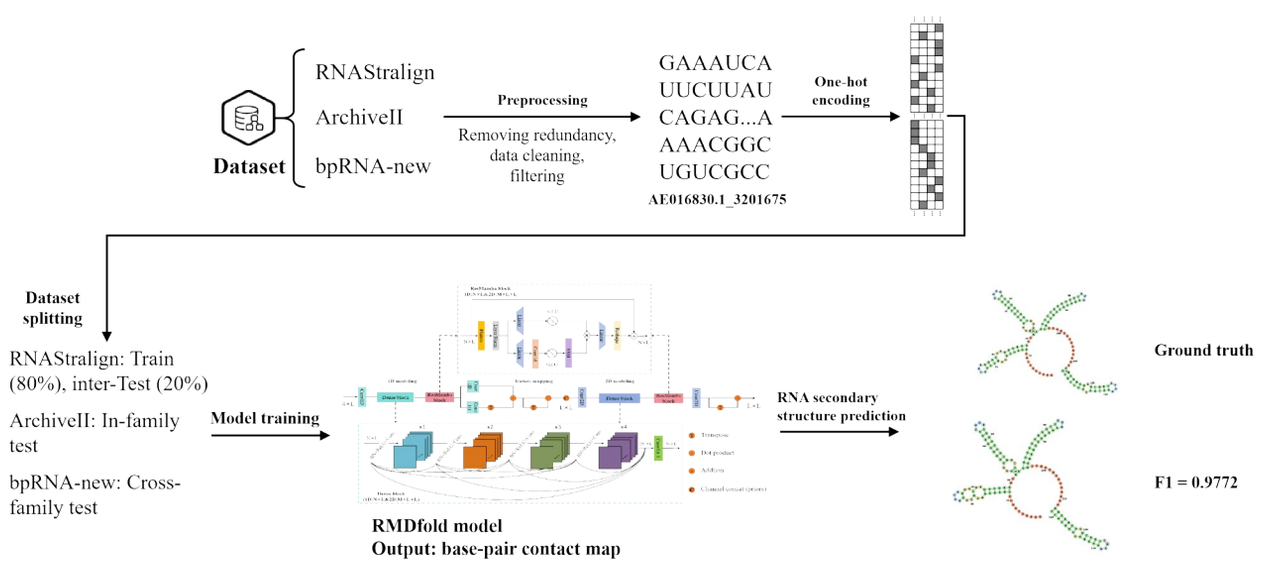
HAO Ze-Zhou[1,2], YANG Yan-Ling[1,2], ZHOU Liang[3], YAO Xu-Feng[2)**]

([1)]*School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*;

[2)]*College of Medical Imaging, Jiading District Central Hospital Affiliated Shanghai University of Medicine and Health Sciences, Shanghai* 201318, *China*;

[3)]*Collaborative Innovation Center, Shanghai University of Medicine and Health Sciences, Shanghai* 201318, *China*)

Graphical abstract



Abstract Objective Non-coding RNA (ncRNA) plays a crucial regulatory role in various biological processes. Numerous studies have shown that the functions of ncRNAs depend not only on their nucleotide sequences but also closely on their spatial conformations, particularly their secondary structures. Traditional methods for predicting RNA secondary structures often have low accuracy, usually around 50 to 70 percent, and their performance further declines when dealing with complex structures or long sequences. Although deep learning methods have improved prediction performance to some extent, they still face challenges such as high model complexity, difficulty in capturing long-range dependencies, and poor generalization in predicting complex structures. Therefore, it is necessary for RNA secondary structure prediction to develop new methods. Methods This study proposed an end-to-end prediction model RMDfold, which employed a feature extraction strategy combining residual Mamba (resMamba) and dense connections (Dense). The model framework consisted of three modules: one-dimensional (1D) modeling, feature mapping, and two-dimensional (2D) modeling. For the 1D modeling module, the model learned contextual dependencies among nucleotides from RNA sequences, providing the foundation for possible base-pairing; for the feature mapping module, the 1D features were projected into a 2D space to form a constraint matrix that represented potential base-pairing relationships; for the 2D modeling module, the model further learned the pairing patterns between nucleotides, and determined the unique pairing of

each nucleotide through pairing constraints, to obtain the final secondary structure. For both 1D and 2D modeling modules, a four-layer Dense block composed of batch normalization, ReLU activation, and convolution was used to extract short-range features; and a dual-branch residual structure resMamba based on a state-space model was used to model long-range dependencies, thereby achieving effective integration of short- and long-range features. To validate the effectiveness of the proposed method, it was compared with Ufold, REDfold, TransUfold, and sincfold on the three public datasets RNAStralign, ArchiveII, and bpRNA-new. Results The proposed RMDfold method demonstrates superior performance compared with existing algorithms in structure prediction, pseudoknot prediction, sequence prediction across varying lengths, and model complexity analysis, while requiring fewer parameters and achieving faster inference. For the structure prediction, the method achieved *F1*, Matthews correlation coefficient (*MCC*), *precision* and *recall* of 0.973 5, 0.973 1, 0.975 6 and 0.972 3 on RNAStralign, 0.854 3, 0.855 6, 0.874 7 and 0.876 2 on ArchiveII, and 0.382 8, 0.401 5, 0.536 5 and 0.318 7 on bpRNA-new, respectively. For the pseudoknot prediction based on ArchiveII, the model achieved *F1*, *MCC*, *precision* and *recall* of 0.741 4, 0.743 3, 0.752 5 and 0.739 6. For the sequence prediction across different lengths, RMDfold maintained an accuracy of 0.70 for sequences ranging from 200 to 500 nt. In terms of model complexity, RMDfold required 2.886 7 M parameters and achieves an inference speed of 0.026 0 s. Conclusion RMDfold enables highly accurate prediction of RNA secondary structures. It helps to deeply and comprehensively reveal the central roles of RNA molecules in gene expression regulation, molecular recognition, and catalysis. and also provides important structure support for elucidating disease-related variant mechanisms, designing RNA-targeted drugs, and advancing research in evolution and comparative genomics.

Key words non-coding RNA, RNA secondary structure, residual Mamba, dense connections, deep learning
DOI：10.3724/j.pibb.2025.0387 **CSTR**：32369.14.pibb.20250387

* The work was supported by grants from Shanghai Municipal Commission of Science and Technology for Capacity Building for Local Universities (23010502700), the Foundation of the Program of Shanghai Academic/Technology Research Leader under the Science and Technology Innovation Action Plan (22XD1401300), and The National Natural Science Foundation of China (61971275, 81830052, 82072228, 62376152).
** Corresponding author.
Tel：86-21-65885239，E-mail：yao6636329@hotmail.com
Received：August 26，2025 Accepted：October 31，2025