

文章

<https://doi.org/10.1038/s41467-019-12920-0>

OPEN

预测 RNA 成分在蛋白质表面的结合偏好的深度学习框架

林浩明、李宇

^{1,13}朱丽哲

^{2,3*}, Ramzan Umarov, Hanlun Jiang, Amélie Héliou ,

Fu Kit Sheong, Tianyun Liu, Yongkang Long

^{1,7}, Yunfei Li , Liang Fang, Russ B. Altman, Wei Chen *

黄旭辉

^{2,8,9,10,11,12} & Xin Gao ^{1*}

蛋白质-RNA相互作用在转录后调控中发挥重要作用。然而，在给定蛋白质结构的情况下预测这些相互作用的任务很困难。在这里，我们表明，通过利用深度学习模型 NucleicNet，可以根据蛋白质结构表面的局部理化特征来预测 RNA 主链成分和不同碱基的结合偏好等属性。在一系列具有挑战性的 RNA 结合蛋白上，包括 Fem-3 结合因子 2、Argonaute 2 和核糖核酸酶 III，NucleicNet 可以准确地恢复结构生物学实验发现的相互作用模式。此外，我们表明，在没有看到任何体外或体内测定数据的情况下，NucleicNet 仍然可以实现与实验的一致性，包括 RNAcompete、免疫沉淀测定和 siRNA Knockdown Benchmark。因此，NucleicNet 可以为给定的结合口袋提供 RNA 序列的定量适应性，或预测潜在的结合口袋和先前未知的 RNA 结合蛋白的结合 RNA。

¹ 阿卜杜拉国王科技大学计算机、电气和数学科学与工程部计算生物科学研究中心

和技术 (KAUST), Thuwal 23955-6900, 沙特阿拉伯。香港科技大学化学系, 中国香港。香港中文大学 (深圳) 生命与健康科学学院沃谢尔计算生物学研究所, 广东深圳 518172 生物化学系和蛋白质设计研究所, 华盛顿大学, 西雅图, 美国。法国帕莱索综合理工学院计算机科学系信息实验室。斯坦福大学医学、遗传学和生物工程系, 美国加利福尼亚州斯坦福。南方科技大学生物系, 广东深圳 518055 香港科技大学生物医学工程系, 中国香港。香港科技大学分子神经科学国家重点实验室, 中国香港。香港科技大学国家组织修复与重建工程研究中心香港分中心, 中国香港。香港科技大学高等研究院, 中国香港。香港科技大学深圳研究院, 深圳市南山区高科技园区, 邮编: 518057 这些作者做出了同等贡献: Jordy Homing Lam, Yu Li * 电子邮件: zhulizhe@cuhk.edu.cn; chenw@sustech.edu.cn; xuhui Huang@ust.hk; xin.gao@kaust.edu.sa

转录后，mRNA 经历一系列相互交织的过程，最后被翻译成功能蛋白。这些转录后调控为细胞提供了微调其蛋白质组的扩展选择，通常是通过 RNA 和 RNA 结合蛋白 (RBP) 之间的相互作用介导的。在细胞中，RNA 在很大程度上受到两种特定相互作用模式的调节——要么通过直接识别 RBP 表面上的 RNA 基序，要么通过间接的 RNA 引导方式。在前一种情况下，RBP 与 RNA 碱基直接接触。例如，Pumilio/FBF (PUF) 家族可以通过直接碱基-蛋白质接触来控制翻译，例如，使用 RNA 转录物上的 UGUR 基序。在后一种情况下，RBP 与碱基的主干或非沃森-克里克 (WC) 边缘相互作用，留下 WC 边缘用于目标识别。例如，在 RNA 干扰 (RNAi，例如 Argonautes) 和基因编辑复合物 (例如 CRISPR-Cas) 的核心酶中，将指导 RNA (gRNA) 选择性加载到 RBP 中是激活酶的先决条件；然后，目标 D/RNA 识别通过 gRNA 的 WC 边缘介导，而 gRNA 的其他部分仍与 RBP 接触。因此，破译 RNA-蛋白质相互作用的特异性和机制对于理解 RBP 的功能、识别 RBP 以及设计用于 RBP 识别和调节的 RNA 至关重要。为了系统地绘制这些相互作用，已经开发了各种实验和计算技术。在实验类型中，体内紫外交联免疫沉淀测定 (例如 CLIP-HITS) 和体外选择测定 (例如 HT-SELEX 和 RNAcompete) 是最成功的技术。一般来说，从这些方法获得的特异性模式可以表示为每个 RBP 的徽标图或单个 RNA 序列的分析分数。通过结构阐明技术，许多这些特征性 RBP (例如 hnRNP、Nova 和 PAZ) 的结合机制也已得到阐明。然而，尽管取得了如此显著的成就，实验分析仍受到反应性、检测和可扩展性的限制。例如，UV 交联测定更喜欢富含尿苷的序列，因为嘧啶比嘌呤更容易被光激活。虽然可以说这些测定特异性的化学起源可以通过核糖核蛋白共晶来验证，但单个或几个这样的共晶很难解释标志图上真正模糊的模式 (例如，对同一位置上的 U 和 A 都具有特异性)。为此，计算方法可以增强实验结果。在这种类型中，可以对采样的实验知识、分析和结构的主体进行改进，以发现以前错误/未承认的特异性模式。示例性的基于测定的计算方法，例如 DeepBind 和变体，可以集成并学习为 RBP 收集的测定数据，以推断与大规模测定一致的特异性模式。基于结构和序列的计算方法也较少被探索。通常，在后面的方法中，给定三维蛋白质结构或其氨基酸序列，可以以残基为单位提取其他结构信息 (例如溶剂可及性、二级结构、疏水性和静电斑块) 中的局部蛋白质序列背景并用于参考蛋白质数据库 (PDB) 中的 RNA-RBP 结构来训练模型。因此，基于测定的方法放宽了对实验数据的要求。然而，由于可用特征的数量非常有限，它们的预测能力仅限于区分 RNA 结合位点与非位点，即对蛋白质残基的位置或索引进行二元预测，而不建议首选碱基/序列，也不建议任何建议。信息交互模式 (例如，通过主干或基础)。

然而，计算方法具有可扩展性和成本效益，因此是实验技术的重要补充。

在这项工作中，我们介绍了 NucleicNet，一种基于结构的计算框架，它解决了上面提出的主题挑战：(i) 我们开发了从 PDB 中有效学习的方法，以便我们可以预测不同 RNA 成分 - 磷酸盐 (P) 的相互作用模式、核糖 (R)、腺嘌呤 (A)、鸟嘌呤 (G)、胞嘧啶 (C)、尿嘧啶 (U) 和非位点 - 并在任何蛋白质表面上可视化它们；(ii) NucleicNet 不需要外部分析输入即可得出与分析数据一致的徽标图，包括 RNAcompete、免疫沉淀分析和 siRNA Knockdown Benchmark；(iii) 从 NucleicNet 获得的标志图或位置权重矩阵 (PWM) 可用于对单个 RNA 序列的结合潜力进行评分；(iv) NucleicNet 可以推广到不同的 RBP 家族，并有可能用于识别新的 RBP 及其结合口袋/偏好。我们的管道建立在特征向量框架的基础上，它将蛋白质表面的物理化学特性编码为高维特征向量。这种丰富的向量空间不仅涵盖了其他程序中开发的大多数功能，而且还可以通过其离散径向分布设置来解释局部拓扑的细微差异。重要的是，从这些高维特征向量空间中学习是很重要的，因此为此目的提出并训练了深度残差网络。

我们从三个不同的数据源 (结构、体外和体内实验) 对 NucleicNet 进行基准测试。对于结构数据，进行了两项测试：

(i) 参考外部基准，我们表明 NucleicNet 在区分蛋白质表面上的 RNA 结合位点和非位点方面可以有效优于所有可用的基于序列的方法；(ii) 与我们自己精心构建的非冗余 7 类数据集相比，我们表明 NucleicNet 可以解析 RNA 成分，对于所有 6 个 RNA 成分和非位点，类平均 AUROC 为 0.77，对于 6 个 RNA 成分和非位点，类平均 AUROC 为 0.66 到 4 个基地。对于体外数据，采用 RNAcompete (RNAC) 测定来评估我们的 NucleicNet PWM 在处理直接识别其表面 RNA 基序的 RBP 时的准确性。在所有八个可用示例中，我们表明，在没有对分析数据进行任何训练的情况下，NucleicNet PWM 在从所有可能的 7 聚体序列中识别最佳结合 7 聚体方面与 RNAC PWM 相当。最后，我们还探索了与体内 RNAi 实验相关的下游应用。我们表明 NucleicNet 评分能够解释 Argonaute 2 (hAgo2) 引导链加载的体内不对称性以及不同 siRNA 序列中不同的敲低水平。

结果

NucleicNet 概述。在 NucleicNet 中，我们的目标是预测蛋白质表面的每个位置 (网格点)，现场呈现的理化环境是否适合与 RNA 结合，如果是，则预测对每种类型 RNA 成分的结合偏好 - 磷酸 (P)、核糖 (R)、腺嘌呤 (A)、鸟嘌呤 (G)、胞嘧啶 (C) 和尿嘧啶 (U) - 与该位置结合。在计算上，我们将该问题视为有监督的七类分类问题。因此，我们将 NucleicNet 的端到端训练制定如下 (图 1 上图)。首先，从 PDB 中检索核糖核蛋白复合物的表面位置，并将其典型为 7 个类别，对应于结合的 RNA 成分和非 RNA 结合位点 (X)。然后使用 FEATURE 程序对每个位置相应的物理化学环境进行表征 (方法，图 1 中图)。接下来，训练深度残差网络，将每个物理化学环境与 7 个环境之一关联起来。

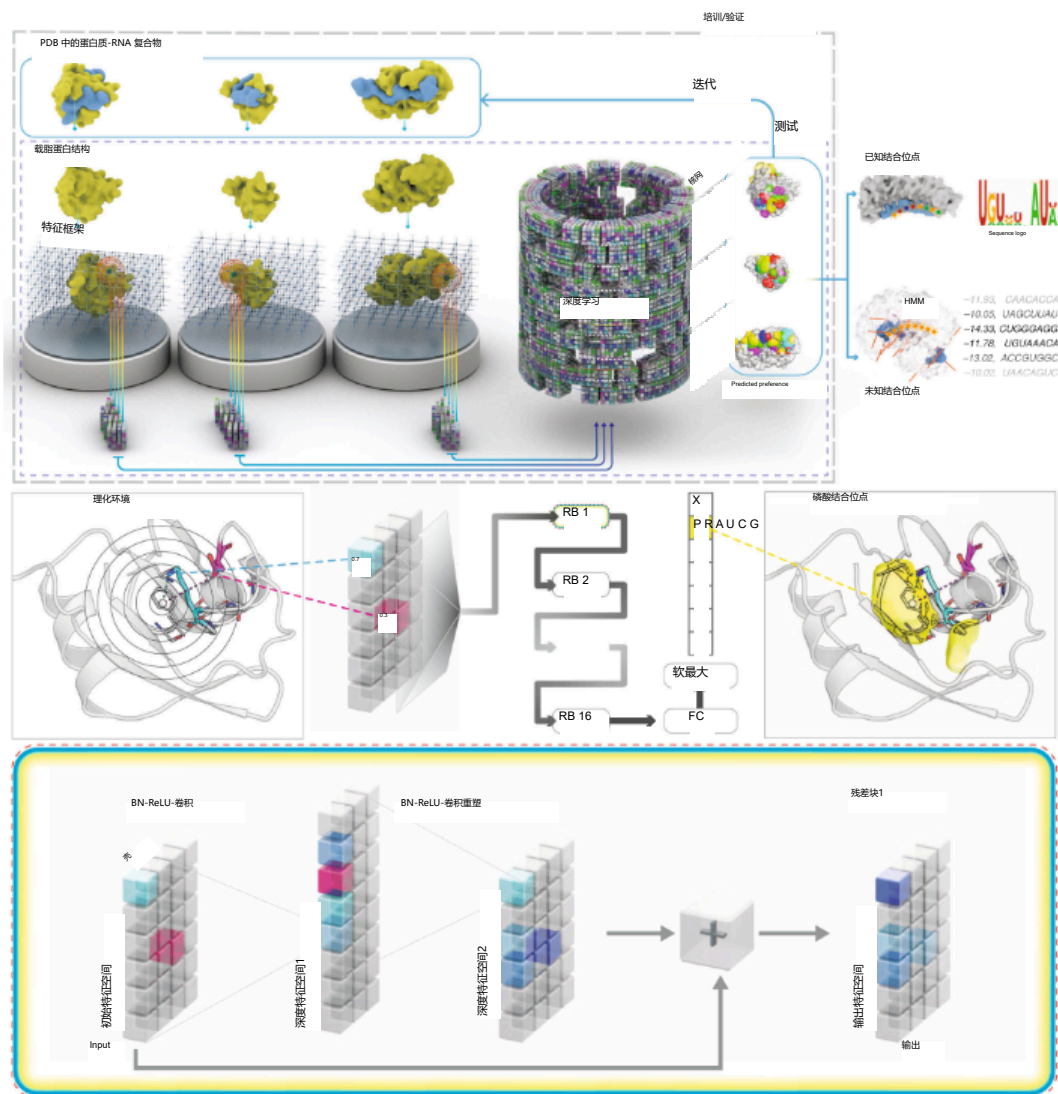


图 1 NucleicNet 概述。顶部面板：NucleicNet 的训练策略和实用程序。PDB 中的核糖核蛋白结构与其结合的 RNA 被剥离。网格点放置在蛋白质的表面上。FEATURE 程序会分析每个网格点（位置）的周围生理化学环境，并将其编码为特征向量。所有六类 RNA 成分的结合位置和非 RNA 结合位点均进行了相应标记。标签及其各自的向量被编译；这制定了深度残差网络的训练输入。

网络中的参数通过误差反向传播迭代更新，并经过训练以区分七个类别。训练完成后，学习的模型可用于预测任何查询蛋白质结构表面的 RNA 成分的结合位点。预测结果的下游应用包括生成 RBP 微标图 and 任何查询 RNA 序列的评分界面。中图：物理化学环境的加入和残差网络的介绍。在特征向量框架中，物理化学环境是通过径向分布设置中网格点 7.5 Å 范围内蛋白质原子的计算特性。因此，每个网格点周围的空间被分为六个同心球壳，对于每个球壳，有 80 种物理化学特性（例如，负/正电荷、部分电荷、原子类型、残基类型、所拥有残基的二级结构）、疏水性、溶剂可及性等）被考虑，产生尺寸为 6×80 的张量。然后，张量由具有 16 个连续残差块的深度残差网络进行变换。之后，最终的残差块通过 softmax 操作连接到全连接层，以评估该位置上每个类的结合概率。底部面板：我们说明了残差网络中的原理操作，即批量归一化（BN）、修正线性单元（ReLU）、局部连接网络和典型的跳跃连接（将初始输入添加回倒数第二个输出层）

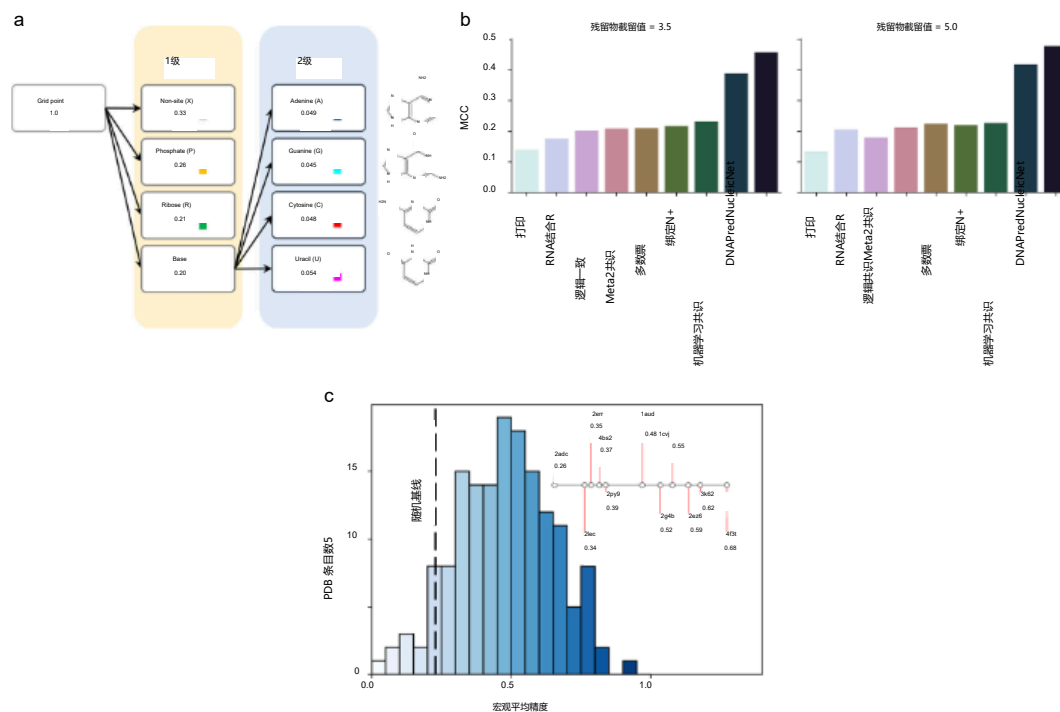


图2 NucleicNet的数据统计和性能。a 7 类分类的统计数据和层次结构。每个类别的可用数据比率显示在每个框的第二行中。每个网格点首先分为 4 个粗标签——非位点、磷酸盐、核糖和碱基——然后分类为 4 个细基标签——腺嘌呤 (A)/鸟嘌呤 (G)/胞嘧啶 (C)/尿嘧啶 (U)。每个成分的颜色代码显示为正方形。b NucleicNet 区分站点和非站点的基准。最近的综述论文中列出的所有方法都使用马修相关系数 (MCC) 在两个截止值方面进行了比较, 以埃为单位, 如标题所示。c NucleicNet 的基准, 用于在 PDB 中蛋白质-RNA 复合物结构的 3 倍交叉验证中区分六种 RNA 成分和非位点。提供了宏观平均准确度的直方图。参考随机 7 类预测变量的基线精度 (0.23) 用虚线表示

以分层方式划分类（方法，图 2a）。最后，通过分类交叉熵损失的标准反传传播来优化网络参数。请注意，训练数据完全源自 PDB 中的三维结构，即我们并没有使用来自外部部分的训练数据。一旦 NucleicNet 的训练完成，就可以将查询蛋白质表面位置上的 FEATURE 提取的原始表面特征向前馈送，以逐个位置推断每个类别的结合偏好。

我们的方法与相关工作的区别之一是，不仅可以预测所有 6 类 RNA 成分的结合位点并在蛋白质表面上进行可视化，而且同时，这些详细结果可以被同化为徽标图或 RNA 序列的评分界面。因此，前馈模块的结果被打包成三个实用模块——一个可视化模块，以曲面图的形式指示最预测的 RNA 成分（图 3a-c）；一个徽标图模块，当 RNA 被预测时生成徽标图。蛋白质表面上的结合口袋是已知的（图 4a-h），并且有一个评分接口模块来获取查询 RNA 序列的结合分数（图 4a-h 和 5a、b），它可以预测最可能的 RNA 序列以及任何查询蛋白上相应的结合口袋（图 3a-c）。后两个模块可以概括为隐马尔可夫模型（HMM）

它结合了碱基的位置和可行 RNA 序列的几何约束 (方法, 补充图 4 和 5)。可视化模块用于将我们的预测与结构生物学实验进行比较。徽标图和评分模块用于将我们的预测与体内或体外测定数据进行比较。

从结构角度验证 NucleicNet。可以从 PDB 中存储的已知核糖核蛋白结构中提取各种可靠的基本事实。首先,我们首先区分 RNA 结合残基和非 RNA 结合残基,即二元分类。这是大多数蛋白质-RNA 相互作用计算预测器解决的经典问题。一般来说,如果蛋白质残基的至少一个原子与RNA原子在一定距离内,则认为该蛋白质残基在共晶中与RNA结合。在最近的一篇综述中,考虑了 3.5 Å 和 5.0 Å 的截止值。其中提出的基准RNA_t数据集由175条RNA结合蛋白链组成,是通过根据序列和结构相似性对蛋白链进行聚类而生成的,其中RNA结合残基的注释在相似的链之间转移以减轻链截断的影响。基于这一基本事实,我们使用基于序列信息的各种最先进的预测器对 NucleicNet 进行了基准测试(图 2b)。到

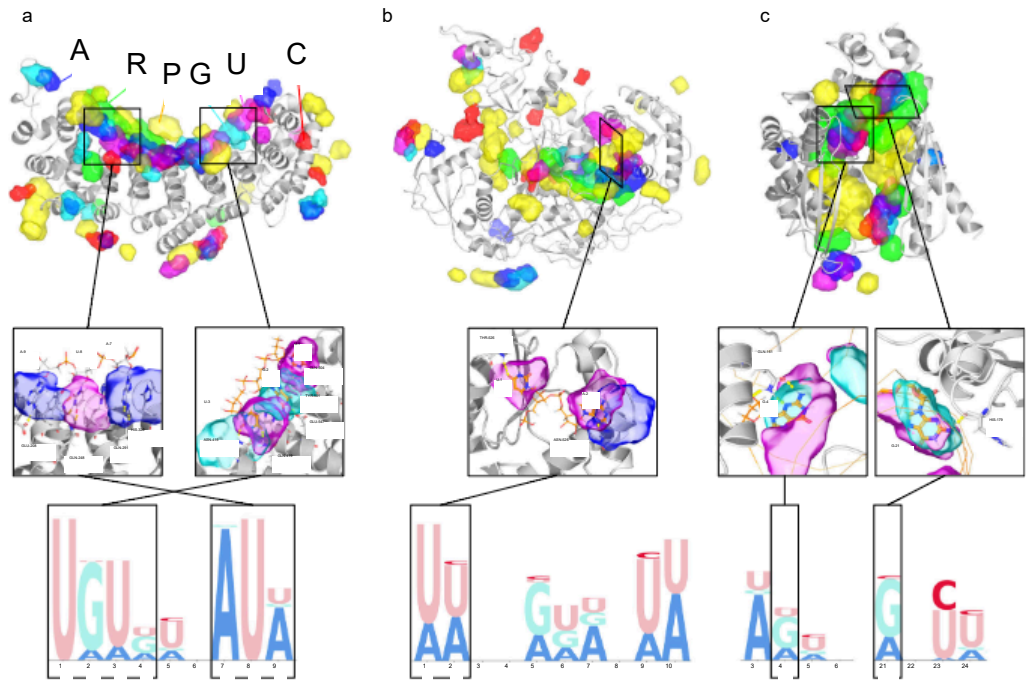


图 3 NucleicNet 预测捕获了结构生物学实验确定的详细结合运动。a FBF2、b hAgo2 和 c Aa-RNase III。上图：NucleicNet 对查询 RBP 的预测。每个类别前 10% 的结合位点都绘制在透明表面上。颜色代码：磷酸黄色、核糖绿色、腺嘌呤蓝色、胞嘧啶红色、鸟嘌呤青色、尿嘧啶紫色和蛋白灰色卡通。中间面板：化学相互作用的详细视图。下图：各个 RBP 的预测序列标识图

使用我们的 NucleicNet 预测器在每个蛋白质残基上分配二元标签（位点或非位点），该预测器在蛋白质表面上的网格点上工作，最接近蛋白质残基的 30 个网格点上的得分向量被用来投票给 2 个粗类，即“RNA 结合位点”和“非位点”；6 个更细的类别（对应于单个 RNA 成分）被认为是“RNA 结合位点”。训练中省略了测试基准蛋白质。在上述两个距离截止范围内，NucleicNet 均优于所有可用方法（图 2b）。因此，这证明了 NucleicNet 作为预测一般 RNA 结合位点的工具的基本实用性。

接下来，我们评估 NucleicNet 检索所提议的六种详细 RNA 成分的结合位点的能力；其中包括磷酸盐 (P)、核糖 (R)、腺嘌呤 (A)、鸟嘌呤 (G)、胞嘧啶 (C) 和尿嘧啶 (U)。对来自 PDB 的所有蛋白质-RNA 复杂结构（参见方法）精心挑选和整理的非冗余数据集进行了 3 倍交叉验证，该数据集由 158 个复杂结构组成，在数据集中产生了约 280,000 个网格点。我们将 158 个蛋白质分为三部分。每次，将其中的两折用于训练，另一折用于测试。折叠之间，不允许 BLASTclust 序列同源性 $\geq 90\%$ （参见方法）。请注意，这种交叉验证的粒度是单个蛋白质，而不是网格点，这消除了蛋白质大小的偏差。表 1 报告了每个类别的 AUROC、F1 分数、精确度和召回率方面的性能（指标在 SI 中进行了解释）。对于碱基 (A/U/C/G)，平均 AUROC 可以达到 0.66。值得注意的是，区分站点和非站点的能在 AUROC 中概括为 0.97。还计算了每种蛋白质的分类准确性

准确率得分的分布如图 2c 所示；案例研究中涵盖的蛋白质（图 3a-c 和 4a-h）在插图线图上标出了它们的 PDBID 以表明它们的性能，这表明案例研究的准确性分布广泛。一般来说，非冗余 3 倍交叉验证的中位准确度达到 49%（参见随机基线 23%，补充说明 1）。因此，这一原理验证分析表明，NucleicNet 可以从理化环境的多样化结构数据库中学习，并推广到看不见的 RBP，以回忆潜在的结合 RNA 成分，前提是所阐明的蛋白质的结构基本完整并包含相关的 RNA 结合域。

NucleicNet 再现 RNA 结合位点的复杂空间模式。基于结构的方法的优势之一是它们能够揭示和可视化蛋白质表面的结合位点。虽然以前基于结构的方法仅涉及二元分类（位点或非位点），但我们的方法可以进一步说明所有六种常见的 RNA 成分 - “磷酸盐” (P)、“核糖” (R)、“腺嘌呤” (A)、“鸟嘌呤” (G)、“胞嘧啶” (C) 和“尿嘧啶” (U)。我们通过三个示例性 RBP 展示了我们方法的这种独特功能：Fem-3 结合因子 2 (FBF2, PDB Entry 3k62, 图 3a)、Human Argonaute 2 (hAgo2, PDB Entry 4f3t, 图 3b) 和 Aquifex aeolicus 核糖核酸酶 III (Aa-RNase III, PDB 条目 2ez6, 图 3c)。FBF2 是 RBP 的一个例子，它通过碱基接触直接与单链 RNA (ssRNA) 基序相互作用，而 hAgo2 是 RBP 的一个例子，它通过主干或非 WC 边缘接触以 RNA 引导的方式发挥作用。第三个例子，AaRNase III，涉及双链 RNA 结合域

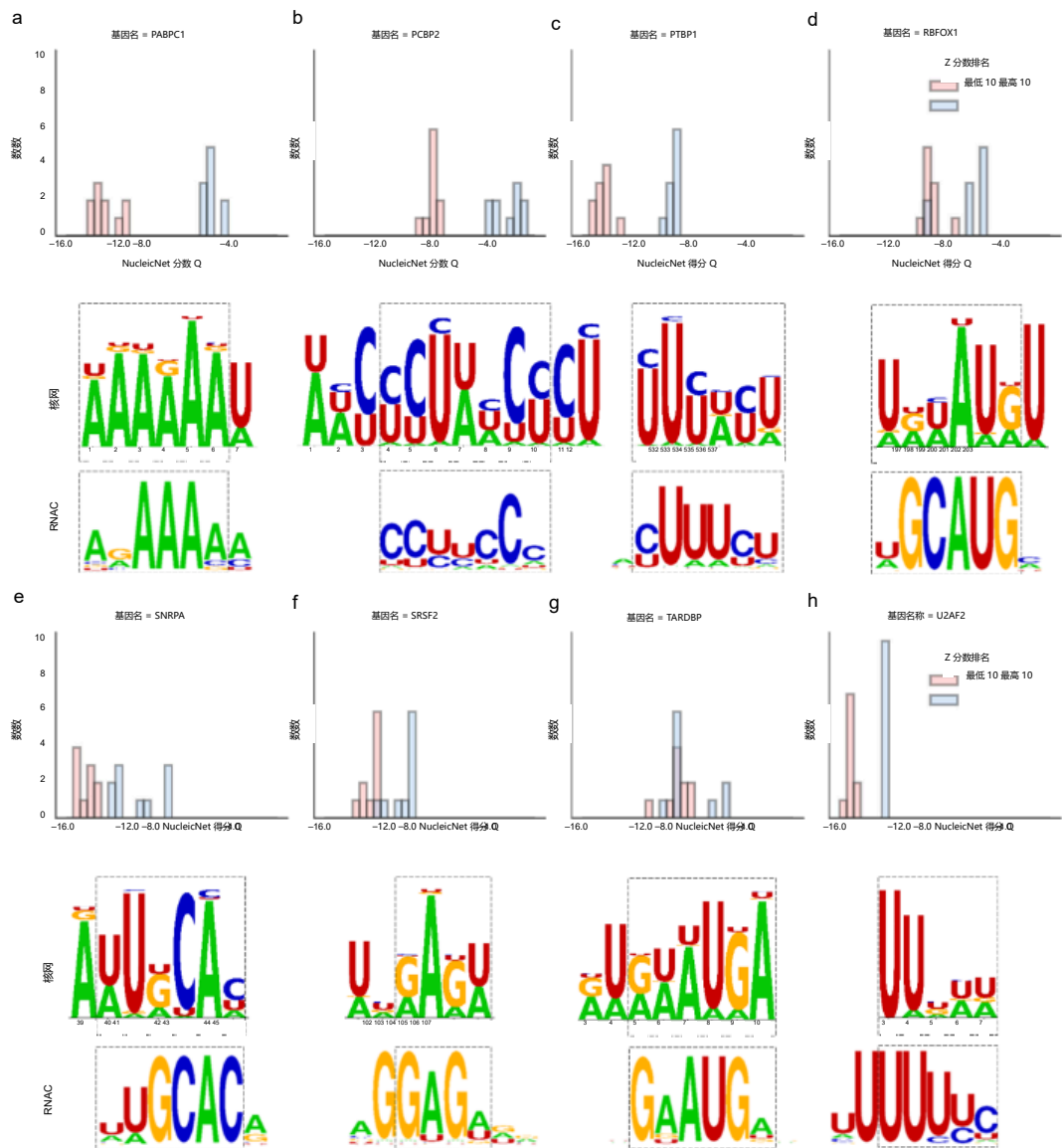


图 4 将 NucleicNet 的分数和徽标图与 RNAcompete (RNAC) 测定获得的分数和徽标图进行比较。a PABPC1、b PCBP2、c PTBP1、d RBFox1、e SNRPA、f SRSF2、g TARDBP 和 h U2AF2。上图：我们表明 NucleicNet 分数能够区分 RNAC Z 分数指示的顶部和底部 10 个序列。下图：我们将 NucleicNet 生成的序列标识与 RNAC 生成的序列标识进行比较。字母的最佳匹配部分用破折号框突出显示

(dsRBD)。在图 3 中，我们使用可视化模块指示了这些蛋白质上每个结合类别的最高预测结合位点。在所有情况下，预测都是在从核糖核蛋白复合物中去除 RNA 后对蛋白质结构进行的。这些蛋白质及其同源物都被排除在训练过程之外。在图 3 中图中，我们表明，对核碱基的强烈偏好主要出现在核苷酸叠加在核糖核蛋白结构上时与蛋白质残基明确相互作用的位置。在图中

下图 3 中，序列标识图是通过长天然 RNA 链上核碱基位置处的 NucleicNet 得分进行平均而生成的（方法）。在所有情况下，我们都表明 NucleicNet 再现了结构生物学实验捕获的详细结合特异性。

Fem-3 结合因子 2。RBP 的 PUMILIO/Fem-3 结合因子 (PUF) 家族是重要的转录后因子

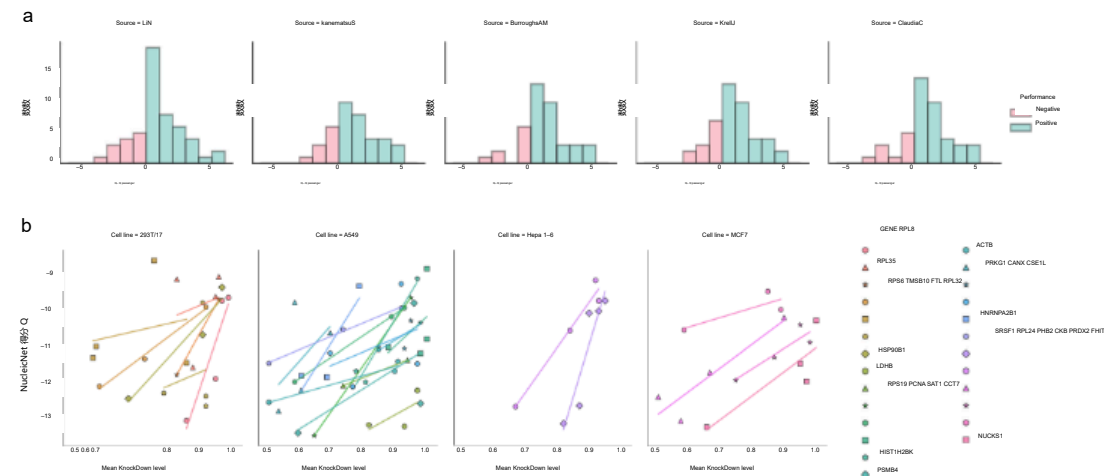


图 5 NucleicNet 预测与 Ago2 的体内实验一致。a 来自五种不同细胞系（即四种来自人类）的引导链和过客链之间 NucleicNet 得分差异的直方图（BurroughsAM：急性单核细胞白血病 THP-1、KanematsuS：结肠癌 DLD、KrellJ：结肠癌 HCT116 和 ClaudiaC：T 细胞白血病）和一种来自小鼠（LiN：神经母细胞瘤 N2a）。具有正差值的条形以绿色突出显示，否则如果为负值则以红色突出显示。b 将 NucleicNet 得分与四种不同细胞系中针对 26 个基因的 shRNA 的基准平均敲低水平进行比较

表1 PDB非冗余数据集交叉验证性能统计							
指标	非现场	磷酸盐	核糖	腺嘌呤	鸟嘌呤	尿嘧啶	胞嘧啶
奥罗克	0.97	0.93	0.84	0.67	0.67	0.65	0.66
F1-分数（宏观）	0.90	0.70	0.63	0.47	0.38	0.48	0.32
回忆（宏）	0.88	0.82	0.63	0.46	0.38	0.45	0.37
精度（宏观）	0.92	0.61	0.63	0.48	0.38	0.51	0.29
F1-分数（微）	0.90	0.70	0.64	0.47	0.41	0.48	0.32
召回（微）	0.88	0.81	0.64	0.47	0.40	0.46	0.36
精度（微米）	0.92	0.61	0.64	0.48	0.41	0.51	0.29

监管机构。在典型的 PUF-mRNA 相互作用中，所有 PUF 中常见的 PUM-HD 结构域将与包含保守 UGUR 序列基序的 mRNA 3' 非翻译区结合。强序列特异性是通过蛋白质表面残基和 RNA 核碱基之间的直接相互作用（芳香堆积和氢键）介导的。Fem-3 结合因子 2 (FBF2) 是最具特征的 PUF 家族蛋白之一。在图3a中图，我们显示 NucleicNet在FBF2的Q504/Q419/N415/E542/Y501和 Q248/Q291/E208/H326处指示的相互作用表面主要涉及PUM-HD重复上的氢键供体或受体。从这些位置得出的相应序列标识图（图3a下图）表明碱基1-4和7-8处有很强的序列偏好，这与之前报道的5'-UGUR和下游A7-U8模式一致。此外，NucleicNet 还正确捕获了碱基 9 处对 A 或 U (A > U > G) 的适度偏好，这与酵母三杂交检测报告的共识一致，即使该位置处的晶体结合天然碱基是 C因此，这表明 NucleicNet 能够揭示晶体结构中未见的、在缺乏第三方分析数据的情况下的潜在序列特异性模式。

其中引导RNA链可以是小干扰RNA (siRNA)或微小RNA (miRNA)。在细胞中，这两种 RNA 预先以互补单链的双链体形式存在。然而，在 RNA 诱导沉默复合物 (RISC) 的组装过程中，通常会将其其中一条链优先加载到 hAgo2 中，以引导靶 RNA 的裂解。这种不对称行为很大程度上受到前体双链体 RNA 序列微小变化的影响。确定了两个归因因素 - (1) 5' 端之一的碱基对减弱，这决定哪条链将在其 5' 端解旋并随后进入 RISC 复合体；(2) 引导RNA-hAgo2在碱基1处的相互作用（图3b中图）和碱基2-8的非Watson-Crick边缘（种子区域）（图3b），这些相互作用被假设为降低RISC 组装成本。然而，与对 RISC 复合物识别靶标 RNA 的深入研究相比，第二个因素（将加载和敲低效率与指导 RNA 蛋白相互作用相关联的 RISC 组装）的研究要少得多。

在图 3b 上图中，我们显示了引导链的结合位点，包括 PAZ 和 N 结构域周围的磷酸核糖主链，被 NucleicNet 正确捕获。具体来说，在图 3b 中图和下图中，我们重点关注 Mid 结构域上的 5' 端结合口袋，并表明 NucleicNet 正确预测了碱基 1 处的强 U 结合口袋 (U > A >> C/G)，并且基数 2 处有一个 U/A 绑定口袋 (U = A)。基数 1 上的第一偏好及其顺序得到了结构的良好支持

Human Argonaute 2. Human Argonaute 2 (hAgo2) 是一个以 RNA 引导方式运行的示例性 RBP

使用单磷酸核苷 (5' 端的模拟物, UMP (0.12 mM) > AMP (0.26 mM) > CMP (3.6 mM)/GMP (3.3 mM)) 进行证据和 NMR 滴定实验。对于种子区域中的其他结合偏好, 仅提供结构证据, 并且它分散在包含不同种子序列的不同 PDB 条目中。例如, 在PDB条目4f3t中, A2和G5分别与N562和Q757交互; 在 PDB 条目 5js1/5t7b 中, U2 与 N562 交互。这些结果与 NucleicNet 提供的微标图一致 (图 3b 下图)。我们稍后在图 5a、b 中显示, 这些 NucleicNet 预测得到了免疫沉淀实验和最低测定的支持, 证实了引导加载效率和序列-蛋白质相互作用是相关的。

Aquifex aeolicus 核糖核酸酶 III。双链 RNA 结合结构域 (dsRBD) 是广泛存在于双链 RNA 特异性核糖核酸内切酶 (包括此处介绍的 Aa-RNase III) 中的结构域。最初, dsRBD 中 RNA 的识别被认为是形状依赖性的, 而不是序列特异性的。然而, 最近的结构证据证实该结构域可以通过与小沟相互作用来识别碱基。在图3c 中图和下图中, 我们显示NucleicNet正确预测了集中在H179和Q161 周围的两个强G结合位点, 对应于dsRBD的第一个 α 螺旋和 β 链1和2 之间的环, 两者非常吻合与现有的共晶。

使用体外 RNACompete 测定数据进行验证。为了在直接识别其表面 RNA 基序的 RBP 上验证 NucleicNet, 我们将 NucleicNet 得分与从 RNACompete 测定 (RNAC) 获得的得分进行比较。RNAC 是一项大规模体外实验, 它使用表位标记的 RBP 从设计池中竞争性地选择 RNA 序列。对于每个 RBP, 获得的 7 聚体 RNA 结合谱可以总结为单个 RNA 序列的 Z 分数, 或通过比对前 10 个评分序列总结为 PWM。Z 分数越高表示结合越好。我们在所有 RNAC 数据和 PDB 结构均可用的 RBP (PABPC1、PCBP2、PTBP1、RBF0X1、SNRPA、SRSF2、TARDBP 和 U2AF2) 上测试了 NucleicNet。在所有情况下 (图 4a-h, 表 2), 执行 Welch t 检验并显示 NucleicNet 能够区分 RNAC Zscore 指示的顶部和底部 10 个序列, 并具有正检验统计量和 p 值 < 0.005, TARDBP 除外, 其 RNAC 结合谱特定于单个序列。在补充表 2 中, 我们进一步将 NucleicNet 分数与 RNAC Z 分数不同排名范围 (顶部/底部 10、50 和 100) 中的 RNAC PWM 分数进行比较。在所有情况下, NucleicNet 都能够区分序列, 尽管它从未接受过任何检测数据的训练。因此, 这表明 NucleicNet 评分具有预测性, 适合补充选择分析。

有趣的是, NucleicNet 能够预测 PDB 中结构生物学信息之外的结合偏好。例如, 蛋白质 PTBP1 的所有三个 PDB 条目 (PDBID: 2adc、2adc 和 2ad9) 均与 RNA 序列 CUCUCU 结合, 该序列偏离 RNAC 建议的序列 YUUUYU (表 2)。这表明单个或少数 PDB 共晶结构可能无法全面了解 RNA 结合偏好。然而, 通过训练与其他 PDB 数据集成, NucleicNet 预测了与 RNAC 序列合理一致的 UUUUYU 的建议序列 (图 4c), 这表明它能够做出训练数据中不存在的预测。因此, 在这些情况下, NucleicNet 相对于 PDB 的准确度分数可能较低

表 2 NucleicNet 和 RNACompete (RNAC) 的性能统计和建议序列图 4

基因名称	a	b	c	d	e	f	g	h
‘采样 PDBID	PABPC1	PCBP2	PTBP1	RBOX1	国家自然资源保护局	SRSF2	TARDBP	U2AF2
RNAC ID	1ovj	2py9	2adc	2err	1aud	2lec	4bs2	2g4b
RNAC建议序列	阿拉姆	CCYYCCH	悠悠悠	世界GCAUGM	WUGCAOR	GGGWD	GAUGD	79
NucleicNet 建议序列	啊啊啊啊	WHCYCUWHCYCU	乌号乌号	乌尔蒙古	惊人的	旺加格夫	鲁沃加	乌号乌号
PDB存放序列	啊啊啊啊	AACCUAACCCU	布谷鸟	考古	AUUGCAC	乌加占	古嘎嘎嘎	呜呜呜
Pearson 相关性 (RNAC PWM 分数与 NucleicNet 分数)	0.70	0.73	0.27	0.27	0.74	0.32	0.77	0.72
哪个 ‘s t- 测试统计 (最高 10 – 最低 10)	20.7	16	25.3	5.2	6.2	7	1.7	20.2
哪个 ‘s t- 测试 P- 值	6.10E-13	1.90E-09	6.70E-13	2.40E-04	4.90E-05	3.90E-06	1.10E-01	8.30E-09

RNAC 和 NucleicNet 之间的最佳匹配建议序列
带有下划线。R: A/G, M: A/C, Y: G/T, H:
A/C/T, W: A/T, D: A/G/T, N: A/C/G/U

结构数据（如图 2c 插图 2adc 所示，精度为 0.26）。另一个例子是蛋白质 RBF0X1，它只有两个沉积的 PDB 条目 2err（具有 RNA 序列 UGCAUGU）和 2n82（具有 RNA 序列 GGCAUGA）。即便如此，NucleicNet 仍可以正确预测第一个位置上具有主导 U 的 U/A，这与 RNAC 建议的序列一致（图 4d）。

hAgo2 引导链加载的偏好。如上所述，引导 RNA 5' 端（碱基 1–8）的微小序列变化可能会导致 RISC 组装产生不同的结果，从而影响了 siRNA 敲低效率。因此，了解 guide-hAgo2 相互作用和加载效率如何相关对于开发有效的 RNA 诱导沉默工具至关重要。为了评估 NucleicNet 预测 gRNA 负载不对称性的能力，我们将 NucleicNet 得分 Q 与两种类型的体内实验（免疫沉淀测定和 siRNA 敲低）的定量结果进行了比较：Q 来自 hAgo2 结构（PDBID: 4f3t）的分析以及和三核苷酸构象文库的比对（补充图 4 和 5 以及方法）。

对 Ago2-RIP-Seq 实验的评估。我们表明，Q 可以区分来自相同前体 miRNA 双链体的引导序列和过客序列，这些序列是通过 Ago2 IP 确定的，然后是来自不同细胞系的小 RNA 测序，即来自人类的四种细胞系（急性单核细胞白血病 THP-1、结肠癌 DLD、结肠癌 HCT116），和 T 细胞白血病）和一种来自小鼠（神经母细胞瘤 N2a）。在每个数据集中，当 Ago2-RIP-Seq 实验（Ago2-RNA 免疫沉淀和测序）中一条链的每百万读数（RPM）取代其互补链至少 2 个数量级时，该链被视为双链体中的指导。引导链速度低于 25 RPM 的双链体也被丢弃，导致总共 222 个双链体处于评估状态（补充表 4）。对于每个数据集，生成每个双链体的引导链和乘客链之间的 NucleicNet 得分差 Q-Qpassenger 的直方图（图 5a）。正差异意味着根据 NucleicNet 分析，导游在结合方面的预测比乘客更有利，这是期望的结果。总之，76% 的测试双链体显示出积极的差异。为了量化这些差异的统计显著性，进行了配对 T 检验和 Wilcoxon 符号秩检验。两项测试在所有数据集中均满足 p 值 < 0.005 标准，证实了 NucleicNet 预测体内设置定义的小 RNA 不对称性的能力（补充说明 3、补充表 3、补充图 7 和 8）。

siRNA 敲低实验的评估。在 siRNA 敲低实验中，不同的引导序列具有不同的加载效率，会影响 RISC 组装，因此它们的沉默效率可能不同。在这里，我们评估 NucleicNet 预测的 Guide-hAgo2 相互作用如何解释这些差异。在这方面，我们从经销商的网站 (<http://www.sigmaldrich.com/life-science/functiongenomics-and-rnai.html>) 收集了在 Broad Institute RNAi Consortium 注册的 shRNA 的击倒基准，并对其进行了测试与 NucleicNet 评分的相关性（补充表 5、补充说明 4 中提供的数据）。为了适应细胞系和靶基因的异质性，对每个实体分别进行回归分析，并且仅限于包含多个数据点的实体（即碱基 1–8 处的不同 shRNA 序列）（图 5b）。击倒等级范围小于 0 的实体

由于看不到趋势，1 被排除。综上，使用 127 个数据点进行评估，总共覆盖 37 个基因；90 个数据点（26 个基因）显示与 NucleicNet 评分呈正相关（图 5b），而 37 个数据点（11 个基因）显示负相关（补充图 9）。尽管许多因素会影响敲低效率，但我们的结果表明，引导链加载的序列偏好是其中之一，因此在未来的 siRNA 设计中应予以考虑。

讨论

实验测定和基于测定的计算方法是了解蛋白质 RNA 结合特性的典型起点。然而，除了鉴定 RNA 序列基序之外，几乎无法推断出碱基-蛋白质相互作用的化学性质，即特异性的起源，因为 RBP 的原子和拓补细节被排除在分析之外。可以说，这种理解上的差距可以通过阐明更多的核糖核蛋白共晶体来填补。然而，即使结构阐明技术变得更加标准化并且共晶集合不断积累，利用这种巨大的抽象结构知识的有效方法尚未实现。在这项工作中，通过深度残差网络感知局部物理化学环境，我们表明可以在基于纯结构的计算框架中推断出关于 RNA 结合位点和 RNA 成分相互作用模式的有意义的预测。更重要的是，我们的结果表明，这些结构方面的知识可用于与最先进的体外和体内实验分析数据进行比较，这表明我们能够捕获真正的 RNA 结合相互作用，并具有可验证的生物学意义。然而，几乎没有什么限制可以掩盖基于结构的范例。首先，没有考虑通过 RNA-RNA 相互作用进一步稳定的特异性；一个极端的例子是在核糖体中，RNA 含量比蛋白质含量多出数倍，因此 RNA-蛋白质相互作用中的不匹配可以通过 RNA-RNA 相互作用来补偿。在我们的数据集中，这些蛋白质被排除在分析之外。其次，在某些情况下，RNA-蛋白质相互作用模式是通过碱基堆积、碱基配对和凸出来辅助的，例如，在 FBF2 和 RNase III 中。尽管这些部分远离蛋白质表面，但它们可能会在整个结合机制中产生焓/熵成本，因此理想情况下不应被忽视。未来，基于结构的方法可能会扩展到涵盖核糖核蛋白复合物的 RNA 结构注释和 RNA 相关理化特征的训练；这可以用于理解 RNA 引导机器（例如 Argonautes 和 CRISPR/Cas）中的靶标 D/RNA 结合。最后，基于结构的方法不了解 RNA 结合机制中的蛋白质动力学。例如，Argonaute 和 RNase III 都需要较大的构象变化才能整合 RNA。此外，蛋白质在与不同的 RNA 序列结合后可能会发生构象变化。为此，如果蛋白质同系物中存在共晶，则可以通过马尔可夫状态模型、正常模式甚至大规模同源建模来增强相关蛋白质构象异构体的采样。尽管如此，基于结构的方法在恢复化学结合特异性模式方面的潜力非常引人注目，并且这种类型可能在不久的将来成为主流。

方法

NucleicNet 框架概述。在 NucleicNet 中，我们的目标是预测蛋白质表面的每个位置，现场呈现的理化环境是否适合与 RNA 结合，如果是，则预测最可能的 RNA 成分类型 - 磷酸盐 (P)、核糖 (R)、腺嘌呤 (A)、鸟嘌呤 (G)、胞嘧啶 (C) 和尿嘧啶 (U) - 与位置结合（图 1）。在下面的

小节中，我们将总结我们的方法。首先，我们定义如何从 PDB 中标记和提取相关的非冗余位置（对应于 RNA 结合位点的正例和负例）。然后，我们描述 FEATURE 程序如何感知这些位置的物理化学环境，该程序为我们的深度学习网络制定输入。接下来，我们检查 NucleicNet 的学习策略和模型架构，以预测这些物理化学环境中的结合类别。最后，我们解释如何从 NucleicNet 预测中推断字母 RNA 序列。

蛋白质-RNA 复合物上的相关非冗余位置。表面距任何蛋白质残基 2.5–5.0 Å 的位置是通过立方晶格上间隔为 1 Å 的网格点的三维坐标确定的。通过考虑局部蛋白质表面的拓扑结构及其结合的 RNA 成分标签，从表面位置中选择相关位置。通过从确定的相关位置去除与同源蛋白质相关的网格点来检索非冗余位置。这种收集相关非冗余数据集的严格策略确保了训练和测试数据集在交叉验证下是不相交的（图2），并且数据集不携带有关蛋白质的先验信息。

为了确定相关的表面位置，所有核糖核蛋白结构均从 NPDB 检索，这是一个托管 RCSB 蛋白质数据库 (PDB) 结构的最新服务器，按其结合核苷酸（例如 RNA-、DNA- 或 D/RNA）进行分类（选择SI 中涵盖了 PDB 结构的标准）。为了定义这些 PDB 结构上的表面位置（表面网格点），采用基于 alpha 球的外部程序 Fpocket 来标记埋藏和溶剂暴露的蛋白质表面上的网格点。提供RNA成分结合位点的正面例子，每个重原子的几何质心

成分被标记。这些标记质心 3 Å 以内且距蛋白质最多 5 Å 的表面网格点被认为是正相关位置；每个正相关位置都由结合的 RNA 成分标记。接下来，我们考虑 RNA 不可能结合的位置。这些负相关位置由从任何 RNA 原子 3 Å 以内的体积以及 Fpocket 中的 alpha 球体排除的空间中随机选择的表面网格点提供。注意，去除冗余位置后，正相关位置和负相关位置的数量以2:1的比例平衡。

为了删除冗余位置，从 PDB 收集的数据充满了冗余。由于同源或异源多聚体复合物的形成，同一 RNA 结合蛋白链的多个拷贝可以存在于同一 PDB 条目中。同源链也可以在不同的 PDB 条目之间共享，这些 PDB 条目专用于不同的结合 RNA 序列、解析蛋白质的质量和突变体等。我们将前一种情况定义为内部冗余，后者定义为外部冗余。通常，这些同源链可以在很大程度上共享共同的 RNA 结合构型和物理化学环境。使用包含冗余的数据进行训练和测试可能会评估带来很大的偏差，并夸大模型的泛化能力。所以，

必须从数据中删除冗余。为了消除外部冗余，PDB 条目被聚类成组，其中每个条目都与另一个共享至少一条 RNA 结合链且 BLASTClust 序列同源性≥90% 的条目链接（补充图 1–3）；对于每个集群，选择具有最佳全局分辨率的 PDB 条目。这样，483 个有效的 PDB 条目就变成了 158 个簇，每个簇仅向数据集贡献一个条目。此外，如果所选条目包含相同蛋白质/RNA链的多个副本（即内部冗余），则仅保留附着于最佳局部解析RNA的网格点；附着于同源蛋白质的网格点也被丢弃。局部分辨率由 RNA 原子上的 B 因子的平均值定义；网格点被分配以粘附最接近的RNA/蛋白质残基。请注意，剩余的非冗余网格点是其特征在于存在内部冗余蛋白链以保持完整的理化环境。总共约 280k 个数据点是从有效的 PDB 条目中编译而来的；其中三分之二是正面例子。数据点被随机分成三个不相交的折叠，不允许外部和内部冗余，即使同一 BLAST 组的成员（补充图 1–3）可以存在于同一训练折叠中，以最大限度地提高训练数据的可用性。请注意，测试是在数据点上执行的仅由每个 BLAST 组的代表成员贡献，在所有三个折叠中，有 80k 个这样的数据点。

使用 FEATURE 捕捉物理化学环境。RNA-蛋白质间-作用是通过物理力和性质（例如静电、疏水性、溶剂可及性等）来维持的，但这些相互作用的起源和强度是由蛋白质表面上化学成分和原子的不同空间排列决定的（例如，带电的）残基、氢键供体/受体等）。这些复杂的拓扑特征（我们将其概括为物理化学环境）可以转化为特征向量，并利用深度学习的力量来预测蛋白质表面位置上的 RNA 结合伙伴——这是我们 NucleicNet 方法的基础。

在这项工作中，采用我们一些人开发的 FEATURE 向量框架来感知三维蛋白质表面上的物理化学环境。此前，该框架已应用于预测阳离子和

配体/片段结合位点。在这些研究中，它已被证明是一种有效的方法来描述结构或序列相似性很少的蛋白质共享的相似结合位点。与之前基于结构/序列的研究使用的其他矢量框架相比，其中物理/结构特征（总共最多 60 个）以残基为单位，无论其空间分布如何，我们的物理化学特征以原子为单位以及它们在某个位置上的离散径向分布（图 1 中图）。因此，这些特征总共 480 个，比任何其他特征保留了更广泛的细节（包括原子类型、元素、残基、官能团、二级结构、电荷、疏水性、溶剂可及性等以及它们的径向分布）矢量框架。为了完整起见，补充表 1 中复制了 Halperin 等人正在考虑的功能列表；仅使用与蛋白质相关的特征，不相关的特征设置为零。这种关于物理化学环境的全面信息对于解决微妙的问题是必不可少的。

RNA 碱基和骨架结合位点之间的差异（图 2a 和 3a–c）。它使我们不仅能够像之前的其他研究一样分辨出 RNA 结合的空间区域，还能将这些结合位点分为六种不同的 RNA 成分，并推断出对 RNA 碱基的特异性。

总而言之，在获得一组标记的相关非冗余位置后，在不存在核酸、溶剂、底物和离子的情况下，在 FEATURE 框架下对它们的蛋白质相关理化环境进行表征。因此，每个位置都由一个 FEATURE 向量和一个指示绑定类的标签进行注释，并且我们的 NucleicNet 经过训练以根据 FEATURE 向量预测标签。

物理化学环境的层次分类。在 NucleicNet 中，我们的目标是预测蛋白质表面的每个位置，现场呈现的理化环境是否适合与 RNA 结合，如果是，则预测与该位置结合的最可能的 RNA 成分类型。这是一个多类分类问题，可以进行端到端训练，其中七个可达到的类别是磷酸盐 (P)、核糖 (R)、腺嘌呤 (A)、鸟嘌呤 (G)、胞嘧啶 (C)、尿嘧啶 (U) 和非结合位点 (X)。然而，由于主干成分 (P/R/R) 的正例比核碱基 (A、U、C、G) 丰富4-5倍，直接的深度学习模型训练会遇到严重的类别不平衡问题（图1）。2a）。为了解释这种情况，我们因此采用了平衡数据的分层分类方案（图2a）。在第一级中，网格点通过 4 类粗略模型进行分类，其中可达到的类是碱基、核糖、磷酸盐和非位点，产生归一化的多标签 4 类得分向量。该模型的训练需要将 A/U/C/G 注释的数据点合并到 Base 中。这缓解了类别不平衡的问题。为了区分 A/U/C/G 四个基数，编译了第二级分类器，该分类器不会遇到类别不平衡问题。最终的归一化多标签 7 类得分向量是通过将第二级结果（也已归一化）与第一级的先验基础相乘而生成的。因此，基于这样的层次结构，为整个问题构建了两个模型：一个用于预测四个粗类，另一个用于区分四个基。下一小节将介绍两个级别的学习者通用的模型架构。

模型架构。神经网络的架构随着深度学习领域的发展而不断发展。从传说中的AlexNet到残差网络 (ResNet) 和生成对抗网络 (GAN) 等前沿架构，每一种架构都是为了突破预测精度的极限，解决特定类别训练中遇到的具体问题。数据。在这项工作中，考虑到问题的复杂性和模型的收敛速度，选择ResNet作为我们的基本单元架构，因为它能够处理梯度消失问题，这阻碍了基线多层卷积神经网络模型的广泛训练当深度网络被编译时。我们的模型由 16 个残差块、一个全连接 (FC) 层和一个最终 Softmax 层组成，用于进行 4 类概率预测。残差块被视为特征提取器，FC-Softmax 被视为分类器。总共编译了 32 个卷积层，其中每个残差块包含 2 个卷积层。FEATURE 程序的输入张量的形状为 1 × 6 壳 × 80 物理化学属性。在卷积运算中，大小为 1 × 2 × 80 的共享滤波器在输入上滑动，在每个位置生成内积作为中间输出，然后经过批量归一化 (BN) 和逐元素非线性激活，在我们的例子中，使用修正线性单元 (ReLU) 来产生中间输出。BN 层的使用缓解了内部协变量偏移问题。总共使用了 80 个过滤器。请注意，为了使输出张量的大小保持一致 (1 × 6 × 80)，输入张量是零填充的。在每个残差块中，添加相同的快捷方式以允许学习输入和第二中间输出之间的残差。最终残差块的输出随后被展平并馈送到全连接层，以在最终 Softmax 层中进行四类概率预测。在 Adam 的指导下，使用分类交叉熵作为损失函数，网络中的所有参数都经过权重衰减的优化。训练是通过 TensorFlow 实现的。一般来说，在 Titan X GPU 上训练各个级别的模型需要 4 天的时间。在补充讨论中，我们还比较了 NucleicNet 预测器中 ResNet 的替代方案，例如、不考虑空间的浅层机器学习方法和神经网络

信息。我们发现，在网络级预测上，所提出的模型 NucleicNet 在相同的实验设置下优于所有浅层方法以及其他深度学习架构。还考虑了替代机器学习策略，例如 MAX-AUC 和带有数据采样的集成学习，尽管遇到了运行时和过度拟合的问题。

获得具有预定碱基位置的序列标志。NucleicNet 的前馈模块用归一化得分向量注释每个网格点，该向量指示相对于该位置上七个可达到的类别的预测结合概率。对于具有预定核糖核蛋白结构的 RBP（例如，与图 4 中的 RNACompete 测定进行比较的那些），通过考虑相应核碱基的质心位置 i 可以轻松生成序列标志图。因此，对每个碱基质心 3 Å 以内的网格点预测的 NucleicNet 得分向量进行平均，以产生平均结合概率 p （参见补充图 4 中的说明过程）。然后根据以下等式计算每个碱基位置 i 上的信息内容 Ξ ，其中 p 是类 c 的碱基位置 i 上的平均结合概率：

$$\Xi = -\log_2 \sum_{c \in \text{AUGCPRNG}} p(c) \log_2 p(c)$$

然后可以通过根据 $P(c)$ 对信息内容 Ξ 进行比例来生成序列标志图。徽标图中省略了对应于磷酸盐、核糖和非 RNA 结合位点的 P、R 和 X 类。请注意，当位置 i 距蛋白质 ≥ 5 Å 时，会自动分配间隙。

对 Ago2 的 RNA 字母序列进行评分。与应用位置权重矩阵分数（PWM 分数）来研究转录因子的 DNA 序列特异性结合的想法类似，NucleicNet 对于各个蛋白质表面的结果可以总结为方程 Q，以对任意 RNA 字母序列输入进行评分：

$$Q = \sum_i \log_2 \frac{p(\text{observed})}{p(\text{background})}$$

这个方程，我们称为固定隐马尔可夫模型（HMM），由发射概率 p 和转移概率 T 组成。我们的目标是通过 p 和 T 来同化 NucleicNet 输出，以考虑共价键提出的几何约束网络和真正的 RNA 链的扭转空间。隐藏状态是由 i 个碱基索引的位置，该碱基与 RBP 结合的连续 RNA 链相关，长度为 N 的字母序列。发射概率 $p(b)$ 是指碱基与 RNA 序列的结合概率。通过对碱基位置 i 的 3 Å 范围内的 NucleicNet 输出进行平均（参见补充图 4 和图 5 中的说明过程）。请注意，这里的 $p(b)$ 是在碱基之间标准化的。转移概率 T 是指从 5' 端到 3' 端的连续 RNA 链上第 i 个碱基和第 $i+1$ 个碱基之间的转移概率。如果碱基位置是由核糖核蛋白共晶预先确定的，连续碱基之间的转换及其位置 i 是确定的，则 $T = 1$ 和 $p(b)$ 可以通过对位置上的 NucleicNet 输出进行平均来推导，就像我们生成徽标一样图表。然后方程 Q 简化为普通 PWM 评分函数；然后通过滑过共晶天然 RNA 链位置来评估长度为 N 的 RNA 串列，以获得 Q 的最大值，这用于计算 NucleicNet 分数以与 RNAC 分数进行比较（图 4a–h，补充笔记 2）。

我们还研究了涉及连续 RNA 链的碱基位置未知但 NucleicNet 预测的 RNA 结合位点明确由磷酸核糖主链引导的情况（例如，在 RNA 引导的情况下，例如图 5 中的 Ago2，补充图 4）。在这种情况下，分数 Q 不能像共晶的情况那样轻松生成，因为这些隐藏位置及其转移概率未知，尽管一旦位置近似，仍然可以根据位置 i 周围的 NucleicNet 输出计算 $p(b)$ 。在下一节中，我们概述了如何通过将 RNA 成分的顶部预测结合位点与 RNA 三核苷酸构象库进行比对来有效地估计这些未知数。在这种情况下，当查询长度为 N 的 RNA 字母序列时，可以通过最大化所有可能的 $i, i+1$ 条转换路径来获得分数 Q 。

连续的 RNA 链可以被认为是连续碱基位置之间的转换图，其中碱基身份可以通过每个节点上的发射概率 $p(b)$ 来表示，该节点由位置 i 索引，指的是 RNA 上碱基 b 的位置与 RBP 结合的链。如果这些位置被隐藏，则转移概率 T 是未知的。然而，无论链长如何，这些转变肯定会受到共价键和 RNA 扭转空间的限制。因此，可以通过筛选 RBP 表面上一系列预测的 RNA 结合位点所耐受的 RNA 几何结构数据库来估计它们。特别是，对于预测的 RNA 结合位点由磷酸核糖主链明确引导的情况（例如，在 Ago2 中，已知 RBP 以 RNA 引导的方式工作），主链结合位点和间歇性碱基的轨迹结合位点在视觉上指示出连续的 RNA 链（补充图 4）。在这种情况下，为了有效筛选与连续 RNA 链相关的结合位点，NucleicNet 报告的前 10% 的结合位点与 Humphris-Narayanan 等人采用的非冗余三核苷酸构象库进行比对。

该文库是通过在 RNA 主链的伪扭转空间上分箱，从 PDB 中的核糖核蛋白复合物编译而来的，从中选择包含 296 个构象异构体的 15°-bin 文库。为了编译全面的三核苷酸构象异构体库，15°-bin 库被排列以覆盖每个构象异构体的原子细节中的所有 4° 可能的三核苷酸序列；所得的 18944 个三核苷酸构象异构体在

AMBER99SB-ILDN 力场确保正确的几何形状。最后，这些三核苷酸被简化为其 RNA 成分（节点）的质心，从而形成一些 9 节点粗粒度模型，准备与顶部结合位点对齐。团对齐过程是通过 Bron-Kerbosch 算法完成的，其中仅保留 ≥ 7 个与蛋白质没有原子冲突的团。选择 7 派系，使得连续 $i, i+1$ 个碱基（即 9 节点模型上的 2 个碱基节点）之间的转换必须由至少 5 个骨干成分引导。这些标准确保所提出的结合位点在几何上是可行的。为了系统地评估这些对齐的 3 聚体如何形成连续链，我们将问题表述为固定 HMM。假设的基本位置是隐藏状态。为了提出这些位置，由对齐的基本节点覆盖的欧几里德空间被划分为多个由 k 均值中心播种的 Voronoi 单元。为了表达碱基的身份，这些 Voronoi 单元（每个代表一个假设的碱基位置）的特征是从 k 均值中心 3 Å 内的网格点平均的发射概率 p 。然后，对于同一对准集团内的连续碱基，不同 Voronoi 单元之间的转移被计数并对称化，作为转移概率 T 的估计。在 Ago2 的情况下，由于确定 5' 位置位于 Mid 域中，所以将一定的起始概率 1 分配给位于 Mid 域中距离任何其他小区最远的小区。然后通过到该起始 Voronoi 单元的排序距离来确定 5' 到 3' 的过渡方向；转换图上的边的方向仅允许从高等级转换到低等级。HMM 的详细信息显示在补充图中。5 和 6。使用 p 和 T 拟合，然后可以使用上面给出的等式计算分数 Q 。

报告摘要。有关研究设计的更多信息，请参阅本文链接的《自然研究报告摘要》。

数据可用性

支持本研究结果的数据可根据合理要求从通讯作者处获得。作者声明，支持本研究结果的所有其他数据均可在论文及其补充信息文件中找到。

代码可用性

NucleicNet 托管在我们的网络服务器 <http://www.cbrc.kaust.edu.sa/NucleicNet/> 上。NucleicNet 工作版本的源代码可在 <https://github.com/NucleicNet/NucleicNet> 上获取。

收稿日期：2019 年 5 月 10 日；接受日期：2019 年 10 月 8 日；

Published online: 30 October 2019

参考

1. Quenault, T., Lithgow, T. and Traven, A. PUF 蛋白：抑制、激活和 mRNA 定位。趋势细胞生物学。21, 104–112 (2011)。
2. Darnell, R. B. HITS-CLIP：活细胞中蛋白质–RNA 调节的全景图。威利跨学科。修订版 RNA 1, 266–286 (2010)。
3. Roulet, E. 等人。用于转录因子结合位点定量建模的高通量 SELEX-SAGE 方法。纳特。生物技术。20, 831–835 (2002)。
4. 雷, D. 等人。用于解码基因调控的 RNA 结合基序概要。自然 499, 172–177 (2013)。
5. Burd, C. G. & Dreyfuss, G. RNA 结合蛋白的保守结构和功能多样性。科学 265, 615–621 (1994)。
6. Lunde, B. M., Moore, C. 和 Varani, G. RNA 结合蛋白：实现高效功能的模块化设计。纳特。莫尔牧师。细胞生物学。8, 479–490 (2007)。
7. Hudson, W. H. & Ortlund, E. A. 结合 DNA 和 RNA 的蛋白质的结构、功能和进化。纳特。莫尔牧师。细胞生物学。15, 749–760 (2014)。
8. 杉本, Y. 等人。分析用于蛋白质–RNA 相互作用的核苷酸解析研究的 CLIP 和 iCLIP 方法。基因组生物学。13, R67 (2012)。
9. Alipanahi, B., DeLong, A., Weirauch, M. T. 和 Frey, B. J. 通过深度学习预测 DNA 和 RNA 结合蛋白的序列特异性。纳特。生物技术。33, 831–838 (2015)。
10. 赵 H., 杨 Y. 和周 Y. RNA 结合域和 RNA 结合位点的基于结构的预测及其在结构基因组学目标中的应用。核酸研究。39, 3017–3025 (2011)。

11. Yan, J., Friedrich, S. and Kurgan, L. 基于序列的 DNA 和 RNA 结合残基预测因子的全面比较综述。简短的。生物信息。 17, 88–105 (2016)。

12. Halperin, I., Glazer, D. S., Wu, S. and Altman, R. B. 蛋白质功能注释的 FEATURE 框架：建模新功能、提高性能并扩展到新应用。BMC 基因组学 9, S2 (2008)。

13. Yan, J. 和 Kurgan, L. DRNApred, 一种基于序列的快速方法, 可准确预测和区分 DNA 和 RNA 结合残基。核酸研究。 45, e84–e84 (2017)。

14. Kumar, M., Gromiha, M. M. & Raghava, G. P. S. 使用 SVM 和 PSSM 配置文件预测蛋白质中的 RNA 结合位点。蛋白质结构。功能。生物信息。 71, 189–194 (2008)。

15. Wang, L., Huang, C., Yang, M. Q. and Yang, J. Y. BindN+, 用于根据蛋白质序列特征准确预测 DNA 和 RNA 结合残基。BMC 系统。生物。 4, S3 (2010)。

16. 瓦利亚, R.R. 等人。使用机器学习进行蛋白质-RNA 界面残基预测：对现有技术的评估。BMC 生物信息。 13, 89 (2012)。

17. Wang, Y., Opperman, L., Wickens, M. & Hall, T. M. T. PUF 调节蛋白特异性识别多个 mRNA 靶标的结构基础。过程。国家科学院。科学。美国 106, 20186–20191 (2009)。

18. Bernstein, D., Hook, B., Hajarnavis, A., Opperman, L. 和 Wickens, M. 线虫 PUF 蛋白 FBF-1 的结合特异性和 mRNA 靶标。RNA 11, 447–458 (2005)。

19. 施瓦茨, D.S. 等人。RNAi 酶复合物组装的不对称性。细胞 115, 199–208 (2003)。

20. Frank, F., Sonenberg, N. 和 Nagar, B. 人类 AGO2 对引导 RNA 的 5'-核苷酸碱基特异性识别的结构基础。自然 465, 818–822 (2010)。

21. Elkayam, E. 等人。人 argonaute-2 与 miR20a 复合物的结构。细胞 150, 100–110 (2012)。

22. Schirle, N.T. 等人。与修饰的 siRNA 向导结合的人 argonaute-2 的结构分析。J. Am. 化学。苏克。 138, 8694–8697 (2016)。

23. 甘, J. 等人。核糖核酸酶 III 加工双链 RNA 机制的结构洞察。细胞 124, 355–366 (2006)。

24. 雷, D. 等人。RNAcompete 方法和应用来确定非常规 RNA 结合蛋白的序列偏好。方法 118–119, 3–15 (2017)。

25. 雷, D. 等人。快速、系统地分析 RNA 结合蛋白的 RNA 识别特异性。纳特。生物技术。 27, 667–670 (2009)。

26. 巴勒斯, A.M. 等人。对人类 Argonaute 相关小 RNA 进行深度测序可深入了解 miRNA 分选, 并揭示 Argonaute 与不同来源的 RNA 片段的关联。RNA 生物学。 8, 158–177 (2011)。

27. Kanematsu, S., Tanimoto, K., Suzuki, Y. 和 Sugano, S. 筛选结肠癌细胞系中可能的 miRNA-mRNA 关联。基因 533, 520–531 (2014)。

28. 克雷尔, J. 等人。TP53 调节 miRNA 与 AGO2 的关联, 以重塑 miRNA-mRNA 相互作用网络。基因组研究。 <https://doi.org/10.1101/gr.191759>. 115 (2015)。

29. 卡里西米, C. 等人。Jurkat 细胞中 miR-21 过表达后 AGO2 相关 RNA 的综合 RNA 数据集。数据简介。 7, 604–606 (2016)。

30. 李, N. 等人。miRNA 和发夹前体的全局分析：深入了解 miRNA 加工和新的 miRNA 发现。核酸研究。 41, 3619–3634 (2013)。

31. Petri, R. & Jakobsson, J. mRNA 衰变：方法和方案 (Lamandé, S. R. 编辑) 131–140 (Springer, 纽约, 2018 年)。 https://doi.org/10.1007/978-1-49397540-2_9。

32. Paddison, P. J., Caudy, A. A., Bernstein, E., Hannon, G. J. 和 Conklin, D. S. 短发夹 RNA (shRNA) 在哺乳动物细胞中诱导序列特异性沉默。基因开发。 16, 948–958 (2002)。

33. 愤怒, A.M. 等人。人类和果蝇 80S 核糖体的结构。自然 497, 80–85 (2013)。

34. 朱, L. 等人。灵活的域-域铰链促进了诱导拟合显性机制, 将向导 DNA 加载到嗜热栖热菌的 argonaute 蛋白中。J. Phys. 化学。 B 120, 2709–2720 (2016)。

35. Bowman, G. R. & Geissler, P. L. 单折叠蛋白质的平衡波动揭示了许多潜在的神秘变构位点。过程。国家科学院。科学美国 109, 11681–11686 (2012)。

36. Parton, D. L., Grinaway, P. B., Hanson, S. M., Beauchamp, K. A. 和 Chodera, J. D. Ensembler: 实现超家族规模的高通量分子模拟。公共科学图书馆计算。生物。 12, e1004728 (2016)。

37. 基尔萨诺夫, D.D. 等人。NPIDB: 核酸-蛋白质相互作用数据库。核酸研究。 41, D517–D523 (2013)。

38. Le Guilloux, V., Schmidtke, P. 和 Tuffery, P. Fpocket: 用于配体口袋检测的开源平台。BMC 生物信息。 10, 168 (2009)。

39. Glazer, D. S., Radmer, R. J. 和 Altman, R. B. 使用分子动力学改进基于结构的功能预测。结构 17, 919–929 (2009)。

40. Liu, T. & Altman, R. B. 通过将循环建模与机器学习相结合来预测钙结合位点。BMC 结构。生物。 9, 72 (2009)。

41. Wu, S., Liu, T. 和 Altman, R. B. 重复蛋白质结构微环境的识别和 CYS 残基周围新功能位点的发现。BMC 结构。生物。 10, 4 (2010)。

42. Zhou, W., Tang, G. W. 和 Altman, R. B. 使用 FEATURE 对 3D 蛋白质结构中的钙结合位点进行高分辨率预测。J. 化学。信息。模型。 55, 1663–1672 (2015)。

43. Tang, G. W. 和 Altman, R. B. 基于知识的片段结合预测。公共科学图书馆计算。生物。 10, e1003589 (2014)。

44. Tang, G. W. 和 Altman, R. B. 使用进化守恒和结构动力学进行远程硫氧还蛋白识别。结构 19, 461–470 (2011)。

45. Liu, T. & Altman, R. B. 使用多个微环境寻找相似的配体结合位点：应用于激酶抑制剂结合。公共科学图书馆计算。生物。 7, e1002326 (2011)。

46. Ren, H. & Shen, Y. 使用结构特征预测 RNA 结合残基。BMC 生物信息。 16, 249 (2015)。

47. Jaapkowicz, N. & Stephen, S. 阶级不平衡问题：系统研究。英特尔。数据分析。 6, 429–449 (2002)。

48. Krizhevsky, A., Sutskever, I. 和 Hinton, G. E. 深度 ImageNet 分类卷积神经网络。在过程中。第 25 届神经信息处理系统国际会议, 卷。1, 1097–1105 (Curran Associates Inc., 2012)。

49. He, K., Zhang, X., Ren, S. 和 Sun, J. 深度残差网络中的身份映射。ArXiv160305027 Cs (2016)。

50. 古德费洛, I. 等人。神经信息处理系统进展 27 (编者: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. 和 Weinberger, K. Q.) 2672–2680 (Curran Associates, Inc., 2014)。

51. Ioffe, S. & Szegedy, C. 批量归一化：通过减少内部协变量偏移来加速深度网络训练。在过程中。第 32 届国际机器学习会议, 卷。 37, 448–456 (JMLR.org, 2015)。

52. Wang, S., Sun, S. 和 Xu, J., 《数据库中的机器学习和知识发现》(编辑: Frasconi, P., Landwehr, N., Manco, G. 和 Vreeken, J.) 1–16 (Springer) 国际出版, 2016)。

53. Van Hulse, J., Khoshgoftaar, T. M. 和 Napolitano, A. 从不平衡数据中学习的实验观点。在过程中。第 24 届国际机器学习会议。 935–942, <https://doi.org/10.1145/1273496.1273614> (ACM, 2007)。

54. Schmidhuber, J. 神经网络中的深度学习：概述。神经网络。 61, 85–117 (2015)。

55. Wasserman, W.W. 和 Sandelin, A. 应用生物信息学识别调控元件。纳特。牧师。热内特。 5, 276–287 (2004)。

56. Humphris-Narayanan, E. & Pyle, A. M. 来自质扭转空间的离散 RNA 文库。J. 莫尔。生物。 421, 6–26 (2012)。

57. Schneider, B., Moróvek, Z. 和 Berman, H. M. RNA 构象类别。核酸研究。 32, 1666–1677 (2004)。

58. Lindorff-Larsen, K. 等人。改进了 Amber ff99SB 蛋白质力场的侧链扭转电位。蛋白质 78, 1950–1958 (2010)。

59. Bron, C. & Kerbosch, J. 算法 457：查找无向图的所有派系。交流。ACM 16, 575–577 (1973)。

致谢

我们感谢王伟的有益讨论。图 1 由阿卜杜拉国王科技大学 (KAUST) 的科学插画家 Heno Hwang 创作。这项工作得到了 KAUST 向 X.G. 的资助。(BAS/1/1624-01、FCC/1/1976-18-01、FCC/1/1976-23-01、FCC/1/1976-25-01、FCC/1/1976-26-01 和 FPCS/ 1/4102-02-01) 以及 KAUST 向 X.G. 提供的资金和 X.H. (URF/1/3007-01)。香港研究资助局 (HKUST C6009-15G、AoE/M-09/12 和 AoE/P705/16) 及创新科技署 (ITCPD/17-9 和 ITC-CNERC14SC01) 至 X.H.; L.F., Y.F.L. 和 W.C. 深圳市科学技术创新委员会科研经费 (No. RQTD20180411143432337 和 JCYJ20170307105752508) 的资助。部分生物信息学分析得到南方科技大学计算科学与工程中心的支持。

作者贡献

X.G., X.H. 和 R.B.A. 构思了这项研究。H.J., A.H. 和 T.L. 发起了这项研究。J.H.M.L. 从 PDB 中提取数据集。Y.L. 和 J.H.M.L. 实施深度学习模型。J.H.M.L. 和 F.K.S. 设计了字母序列的评分界面。Y.K.L., Y.F.L., L.F. 收集了有关 hAgo2 的实验数据; W.C., X.G. 和 J.H.M.L. 为这些数据设计了统计检验。R.U., J.H.M.L. 和 Y.L. 设计了网络服务器。J.H.M.L., L.Z. 和 Y.L. 手稿是在 X.G., X.H. 和 L.Z. 的监督下撰写的。和 W.C. 所有作者都参与了手稿的讨论和定稿。

利益争夺

作者声明没有竞争利益。


附加信息

本文的补充信息可在 <https://doi.org/10.1038/s41467019-12920-0> 上找到。
信件和材料请求应发送至 L.Z.、W.C.、X.H. 或 X.G.

同行评审信息 Nature Communications 感谢邓雷和其他匿名审稿人对这项工作的同行评审所做的贡献。同行评审报告可供使用。

重印和许可信息可在 <http://www.nature.com/reprints> 上找到

出版商说明施普林格·自然对于已出版地图和机构隶属关系中的管辖权主张保持中立。

 开放获取 本文根据知识共享署名 4.0 国际许可证获得许可。该许可证允许以任何媒介或格式使用、分享、传播、分发和复制，只要您对原作者和来源给予适当的认可，提供知识共享许可证的链接，并指出是否进行了更改。除非材料的出处另有说明，本文中的图像或其他第三方材料均包含在文章的知识共享许可中。如果文章的知识共享许可中未包含材料，并且您的预期用途不受法律法规允许或超出了允许的用途，则需要直接获得版权所有者的许可。要查看此许可证的副本，请访问 <http://creativecommons.org/licenses/by/4.0/>。

© 作者 2019