科学家发布开源RNA序列基础模型

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/42941.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

科学家发布开源RNA序列基础模型。
近日，中国科学技术大学
医学影像智能与机器人（MIRACLE）实验室
发布了一款开源RNA（核糖核酸）序列基础模型structRFM，有效提高AI模型在RNA结构预测与功能推断一系列下游任务中的性能，有望助力人工智能驱动的科学研究、生物学机制理解、药物发现等。

RNA是重要的生物大分子，存在于生物细胞以及部分病毒、类病毒中，广泛参与蛋白质合成、基因表达调控和RNA剪接等生命过程。
解析RNA参与序列、折叠结构相关的联系，是揭示其生物学机制的重要基础，也是药物发现的支撑要素。

随着测序技术发展，此前研究积累了海量RNA序列数据。借鉴AI自然语言处理中的预训练方法，RNA序列基础模型能够从这些数据中学习序列规律，提取可用于不同任务的通用特征，为结构预测和功能分析提供支持。

研究团队
将结构知识融入大规模序列学习，并保持模型在不同结构与功能任务中的适应性和泛化能力，研发出RNA序列基础模型structRFM。

团队首先构建了约2100万条经计算注释的序列-二级结构配对数据。在此基础上，设计算法同时预训练学习序列上下文与结构依赖。预训练
后的模型
仅需输入序列，即可提取序列级、核苷酸级和核苷酸对级特征，用于多种结构与功能推断任务，并在多项任务中达到国际前沿水平。

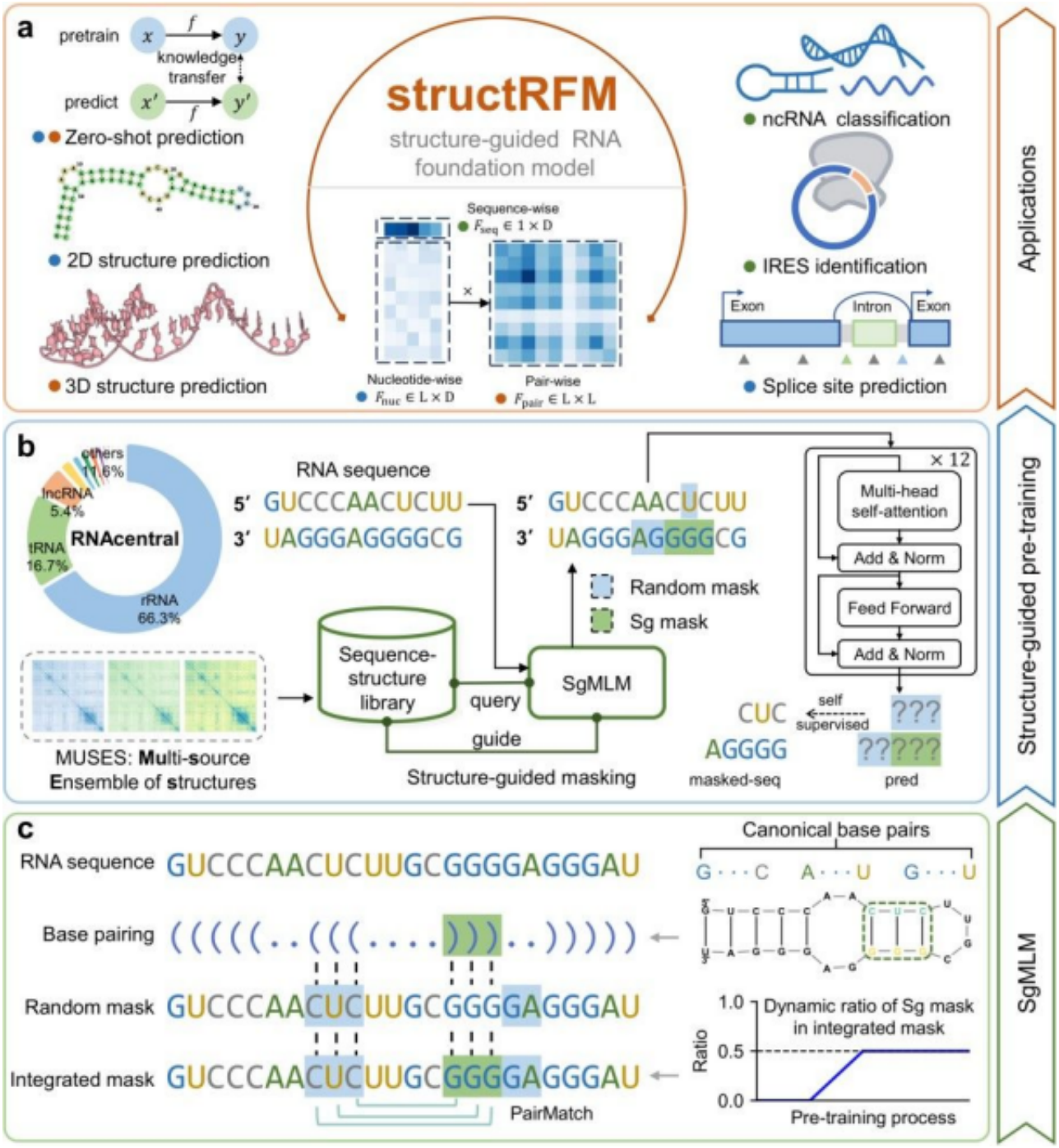
实验显示，
structRFM在零样本同源性分类中表现良好，并提高了二级结构预测性能；将其特征融入三级结构预测模型后，也改善了三维结构及其对应二级结构的预测结果。在功能推断方面，模型支持剪接位点预测、内部核糖体进入位点识别和非编码RNA分类。

此外，模型在新发布的RNA家族上保持较好的二级结构预测表现，并通过分段策略扩展至长非编码RNA分类。消融实验进一步验证了结构信息的引入及掩码策略对模型性能提升的作用，为算法设计提供了实验支持。

目前，团队已开放完整配对数据、模型权重和源代码，为研究结果的复现和相关模型的开发提供了可复用资源。

相关研究成果发表在《自然》（*Nature*）上。

论文链接



RNA序列基础模型

研究团队单位：中国科学技术大学

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发