

杨建益：基于冷冻电镜的蛋白复合物结构建模算法CryoAtom

一、研究背景与核心动机

在蛋白质结构预测领域，AlphaFold系列算法取得显著突破，其中AlphaFold2已基本解决蛋白质单体结构预测问题，相关成果斩获诺贝尔化学奖。然而，该系列算法并非万能，在复合物结构预测与核酸结构解析方面仍存在重大挑战。例如，包含十几万氨基酸的超大蛋白复合物，以及RNA纳米笼（Nanocage）等这类核酸结构，无论是AlphaFold1、AlphaFold2乃至AlphaFold3都无法实现精准预测，这些结构的解析仍需探索新的技术路径。

冷冻电镜（Cryo-TEM）是当前解析复合物结构的主要工具，但在实际应用中面临“最后一公里”难题。从电镜获取的密度图中解析出原子结构，往往需要研究人员进行繁琐的手动反复迭代操作，不仅耗费大量时间，在部分情况下甚至无法成功解析。同时，冷冻电镜对密度图分辨率要求极高，原子半径约为1.7-2Å，当密度图分辨率处于3-7Å范围时，难以直接观测到原子，这一局限性严重制约了冷冻电镜在复合物结构解析中的广泛应用。

针对上述问题，本研究以“计算+实验”的交叉融合为核心思路，旨在将结构预测的算法优势与冷冻电镜的实验数据相结合，开发出能够自动构建全原子结构的方法。通过该方法，降低对电镜密度图分辨率的要求，突破现有技术瓶颈，最终解决复合物及核酸结构解析的难题，为结构生物学研究提供更高效的技术支撑。

二、核心研究成果：CryoAtom算法

（一）算法核心思路

1. 整合建模框架

CryoAtom算法采用整合建模框架，将结构预测（AI）与冷冻电镜实验数据紧密结合。该框架的输入包含两部分，一是蛋白质或核酸的一维序列，二是冷冻电镜获取的三维密度图（呈实心体结构，仅能观测到轮廓，细节模糊）；输出则为三维全原子结构。在整个建模过程中，结构预测为算法提供支撑，冷冻电镜提供实验约束，二者相互配合，实现“计算辅助实验、实验优化计算”的共赢局面，有效打破单一技术的局限性。

2. 两阶段核心流程

CryoAtom算法通过两阶段流程实现从密度图到全原子结构的构建：

第一阶段：从密度图解析Cα原子位置。该阶段采用三维U-Net神经网络，结合“均值漂移原理”筛选最大值，去除无效假设，从而精准定位Cα原子。此阶段主要借鉴领域内现有成熟方法，未进行过多创新性设计，但其为后续全原子结构的构建奠定了坚实基础。

第二阶段：从Cα原子构建全原子结构。这是CryoAtom算法的核心创新阶段，算法基于AlphaFold2的EvoFormer和Structure Module框架，针对如何有效融入密度图信息进行了关键改进，最终实现从Cα原子到全原子结构的精准构建。

（二）关键技术创新

1. Local Attention（局部注意力机制）

AlphaFold2采用Global Attention（全局注意力机制），需构建 $r \times r$ 矩阵（其中 r 为氨基酸数量），当 R 数值较大时，计算与存储成本极高。而CryoAtom算法的Local Attention机制则做出优化，仅考虑每个原子的 K 个邻居，构建 $k \times r$ 矩阵。这种改进充分利用了 $C\alpha$ 原子已知的邻居空间信息，无需进行全局比对，大幅降低了矩阵维度。例如，当 $r=10$ 万时，矩阵规模从 $10\text{万} \times 10\text{万}$ 降至 $20 \times 10\text{万}$ ，显著节省了计算时间与存储空间，有效提升了算法效率。

2. 3D RoPE（三维旋转位置编码）

CryoAtom算法参考自然语言处理领域的一维RoPE，并将其拓展至三维空间。在Transformer结构中，算法将嵌入向量按 x 、 y 、 z 三维分量进行三等分（嵌入维度设定为3的倍数）。对每个分量，根据预设的 θ 值进行旋转编码，使得query与key的注意力得分与原子间距离呈现正相关关系，即原子间距离越近，注意力得分越高，这一特性完全符合真实的分子结构规律。通过3D RoPE创新，模型对三维空间位置信息的捕捉能力显著提升，训练过程中损失更稳定、数值更低，有效优化了学习效率，让网络能够更好地理解原子间的空间距离关系。

（三）算法性能与优势

1. 低分辨率适应性强

在中低分辨率的密度图解析任务中，CryoAtom算法表现出优异的适应性。相较于2024年发表于《Nature》的ModelAngelo方法，CryoAtom算法的完成度更高。例如，在4-7Å分辨率的密度图解析中，部分案例的完成度从ModelAngelo方法的20%—30%提升至40%。即便在密度图较为模糊的情况下，该算法仍能生成更完整的结构框架，为后续的结构

分析与研究提供可靠基础。

2. 突破人工建模极限

CryoAtom算法能够解析人工建模未完成的结构区域，即密度图存在约束但人工未覆盖的区域，且算法解析出的结构并非随机生成，而是严格遵循密度图信息。为验证解析结构的可靠性，研究通过AlphaFold3进行验证，结果显示解析结构与预测结构的TM-score（结构相似性评分，取值范围0—1，分数越高表明结构相似性越强）达到0.86，同时pLDDT（预测局部距离差异测试，用于评估结构可靠性）评分在90分以上。从密度图约束和结构预测两个维度，双重验证了解析结构的可靠性，突破了人工建模的极限。

3. 支持超大结构解析

针对超大复合物结构，CryoAtom算法展现出强大的解析能力。研究中，算法成功解析了包含15万个氨基酸的结构（使用A100显卡运行1—2天，完成度比Model Angelo方法高10%），以及包含20万个氨基酸的结构（完成度较Model Angelo方法提升30%）。对于此类高复杂度结构，单纯依靠结构预测难以实现精准解析，而CryoAtom算法通过“计算+实验”结合的方式，高效完成了超大结构的解析任务，拓展了结构解析的应用范围。

4. 可扩展至核酸结构

通过修改原子体系与算法流程，CryoAtom算法可扩展应用于核酸及蛋白-核酸复合物结构的解析。在针对162个非冗余核酸冷冻电镜密度图的测试中，该算法的完成度比ModelAngelo等方法高15%，有效解决了核酸结构解析中常见的结构断裂、碎片化问题；对RNA纳米笼结构的解析仅需2小时，远超人工搭建效率，且解析结构与天然结构的吻合度较高，充分证明了算法在核酸结构解析领域的应用价值。

（四）应用场景

1. 助力结构解析与新蛋白发现

CryoAtom算法已在实际研究中展现出重要应用价值，能够帮助结构生物学家解析此前无法完成的结构。例如，西湖大学Alexey团队通过该工具成功解析出3个未发表密度图中的新结构。同时，该算法还可用于发现未知蛋白，无需预先提供完整序列，仅通过密度图解析结构后，与物种数据库进行比对，即可识别样品中未预期的蛋白链，为新蛋白的研究提供了有力支持。

2. 支撑RNA药物研发

当前RNA结构预测精度较低，制约了RNA相关研究及药物研发进程。CryoAtom算法为RNA结构解析提供了高效工具，能够精准解析RNA结构，助力研究人员挖掘RNA的未知功能，为RNA药物研发奠定基础，推动RNA药物研发领域的发展。

3. 拓展至核酸结构建模

蛋白-核酸复合物及纯RNA结构的解析仍是结构生物学领域的难点问题。CryoAtom算法已初步拓展至核酸建模领域，在162个测试样本中，其表现优于现有方法，尤其在避免“结构断裂”问题上取得了明显改进。如前文所述，RNA nanocage的结构在2小时内即可通过该算法自动解析完成，而人工建模则需要更长时间，充分体现了算法在核酸结构建模中的高效性与优越性。

三、未解决问题与未来研究方向

（一）核酸结构的纯算法预测

目前，核酸（尤其是RNA）结构预测精度较差，仍需依赖冷冻电镜实验数据。未来，研究需探索不借助实验数据、仅通过算法改进实现RNA结构精准预测的方法。通过这种纯算法预测，挖掘基因组中潜在的稳定RNA结构，为RNA相关研究及药物研发提供更有力的技术支撑，进一步释放RNA在生命科学与医学领域的应用价值。

（二）从静态到动态结构预测

现有结构预测研究主要聚焦于静态结构，而蛋白质、核酸的功能实现依赖于其动态构象变化。因此，未来需克服动态结构建模的技术瓶颈，开发能够实现生物大分子动态结构精准预测的方法。通过动态结构预测，深入揭示生物大分子的功能机制，为理解生命活动规律、疾病发生发展机制提供更全面的视角。

（三）小分子结构解析

在生物体系中，小分子（如药物分子）具有重要作用，其与生物大分子的相互作用对生命活动及疾病治疗至关重要。然而，目前冷冻电镜结构解析尚未涵盖小分子，无法完整呈现生物大分子与小分子复合物的结构信息。未来，需扩展CryoAtom算法，实现蛋白-核酸-小分子复合物的全结构自动解析，完善生物大分子结构解析体系，为药物研发、疾病治疗等领域提供更全面的技术支持。

四、其他关键观点

（一）模型与数据相关

在CryoAtom算法的开发过程中，训练集初始包含3000多个密度图，且以分辨率3Å以下的密度图为主；测试集则涵盖4-7Å分辨率的密度图，共103个样本。研究发现，少量数据即可验证算法效果，例如仅用几百个样本就能实现算法的快速迭代。此外，现有EMDB数据库中约有几万个密度图，但仅一半已解析出PDB结构，这表明CryoAtom算法仍有大量的应用空间，未来可通过解析更多未完成的密度图，为结构生物学研究积累更多数据与成果。

（二）与通用AI模型的关系

CryoAtom算法的底层算法与OpenAI、DeepSeek等通用大模型存在相似之处，但二者在训练数据与应用场景上存在显著差异。CryoAtom算法的训练数据聚焦于生物结构数

据，应用场景为生物大分子结构解析；而通用大模型的训练数据更为广泛，应用场景涵盖自然语言处理、图像识别等多个领域。尽管如此，通用大模型仍能为CryoAtom算法的开发提供辅助，例如辅助完成格式转换、代码生成等基础工作，有效提升研究效率。

（三）现存局限

尽管CryoAtom算法取得了显著成果，但

仍存在一些局限。当密度图分辨率过低时，算法解析出的结构仍会出现断裂、碎片化问题；在算法性能评估中，尚未严格测试氨基酸链长度对解析完整性的影响；此外，对于多肽及无序蛋白的结构解析，算法仍面临较大挑战。这些局限为未来算法的优化与改进指明了方向，需通过进一步的研究与开发，不断提升算法的性能与适用性。