科学通报
黎乐民先生90寿辰专辑 论 文
《中国科学》杂志社
SCIENCE CHINA PRESS

基于物理启发式的深度学习方法对蛋白质三维结构的修复与扩展采样

陈振宇[1†], 林潇涵[1†], 李彦衡[1], 马子程[2], 张骏[2*], 高毅勤[1,2*]

1. 北京大学化学与分子工程学院, 北京分子科学国家研究中心, 北京 100871
2. 昌平实验室, 北京 102200
† 同等贡献
* 联系人, E-mail: jzhang@cpl.ac.cn; gaoyq@pku.edu.cn

2024-10-30 收稿, 2024-12-29 修回, 2025-02-06 接受, 2025-02-25 网络版发表

摘要 AlphaFold2作为首个达到实验精度的深度学习蛋白质结构预测模型, 已成为结构生物学领域的核心工具. 与此同时, 由此衍生的深度学习方法在蛋白质构象系综的高效采样及蛋白序列设计、配体扩散生成等一系列从蛋白质结构出发的下游结构生物学方向中也取得了长足进步. 在多构象采样方面, 传统分子动力学和蒙特卡罗方法因计算成本和采样效率限制, 难以在较大体系中广泛应用, 现有生成式深度学习方法能快速得到蛋白质多构象结构, 但难以加入适当约束以控制构象采样的范围. 此外, 蛋白质初始结构的鲁棒性对基于蛋白质结构的下游模型表现和分子动力学模拟的效率至关重要, 但传统的蛋白质结构修复工具在修复蛋白结构缺失区域方面效果有限. 针对这些挑战, 本研究提出FoldCopilot套件, 基于AlphaFold2将目标序列和多序列比对结果作为 Evoformer 模块的提示词输入, 结合同源序列扰动机制, 无需重新训练即可生成多样化的局部构象. 此外, FoldCopilot利用模板与结构初始化帧, 通过结构预测模块的迭代过程实现了对模型生成结构的演化与控制. 实验结果表明, 该方法在蛋白质结构修复和局部多构象生成等任务中表现出色, 为蛋白质-分子相互作用预测和结构建模提供了新的思路, 并为后续的分子动力学模拟和生物学研究提供了强有力的支持.

关键词 AF2, 提示词工程, 多构象采样, 动力系统控制, 同源序列扰动

蛋白质是生物体内生命活动的主要承担者. 为了更深入地了解蛋白质执行生理功能的复杂过程, 阐明其背后的机制, 研究者们一直致力于通过各种理论计算和建模方法对蛋白质自身理化性质及其与其他分子的相互作用机制进行研究和预测.

长久以来, 研究者采用经典物理学的方法对蛋白质结构的多种构象进行建模和模拟. 然而, 基于物理方法如分子动力学(Molecular Dynamics)或蒙特卡罗方法(Monte Carlo Method)的模拟采样效率较低、计算成本较高[1~6]. 近年来, 随着大量实验测定的蛋白质结构数据的积累以及深度学习技术的快速发展, 端到端的蛋白质结构建模方法取得了巨大的进展. 其中, 许多深度神经网络如AlphaFold2(下文简称AF2), AlphaFold3(下文简称AF3), RoseTTAFold和Protenix(bytedance/Protenix. 2024)[7~10]均能够实现与实验方法精度接近的蛋白质结构预测. 这些方法大大缩短了结构建模时间, 使得蛋白构象的高效采样成为可能. 此外, 这些基于深度学习的蛋白质结构预测方法能够作为实验蛋白结构解析的补充手段, 帮助我们发现新的折叠家族[11], 为分子动力学模拟、序列设计、错义突变预测、蛋白-分子和蛋白-蛋白对接提供可靠的结构初猜. 与此同时, 蛋白质-蛋白质或蛋白质-小分子的结构或相互作用数据也

引用格式: 陈振宇, 林潇涵, 李彦衡, 等. 基于物理启发式的深度学习方法对蛋白质三维结构的修复与扩展采样. 科学通报
Chen Z, Lin X, Li Y, et al. Physics-informed deep learning approach for fixing and sampling protein 3D structures (in Chinese). Chin Sci Bull, doi: 10.1360/TB-2024-1149

能够被深度学习模型所建模，用以生成与输入蛋白质结构相互作用的蛋白质序列、结构、口袋小分子，或预测蛋白-蛋白相互作用结构、亲和力等[12~18].

实验结构里的一些缺陷(见下文), 需要有方法来弥补. 在实际研究中, 基于经典物理学的蛋白质构象的理论模拟和采样需要输入蛋白质起始构象，从结构出发的深度学习模型也需要蛋白质构象作为初始输入，这些方法高度依赖蛋白质的已知构象. 在RCSB数据库中已解析的实验结构常常被视为蛋白质的已知构象，但这些结构往往因为实验解析方法的限制，存在结构缺失、分辨率有限、难以捕捉多构象和动态变化的缺点[19]. 因此，当已知构象不能涵盖待研究的蛋白质序列、结构中存在未解析的残基、研究者们需要增强下游模型结果的鲁棒性和多样性时，常常需要对已知的蛋白质结构进行结构修复和扰动，结构修复的质量直接关系到下游分子动力学模拟、口袋分子设计、蛋白-蛋白相互作用预测等任务的效果，结构扰动的效果也直观地影响深度生成式模型生成序列的质量和多样性. 目前流行的结构修复方法包括PDBFixer、Modeller[20,21]等软件: PDBFixer利用已有的蛋白质结构对缺失结构进行合理的初始化，而后采用力场进行简单的优化, 通常只能合理地修复较短的缺失结构域, 对于长序列的缺失修复效果较差; Modeller则通过蛋白质序列搜索相似模版进行“结构嫁接”，对于无法搜索到相似模版的序列则无能为力. AF2或AF3等深度学习模型可以利用已知序列对结构进行端到端建模，但一方面不能保证待预测的结构与已知的部分真实结构一致，为下游任务带来干扰. 另一方面，其搜索MSA的过程耗时较长, 不使用MSA时其建模效果较差.

通过传统分子动力学等方法采样和扰动蛋白质的结构成本极高，因此研究者们基于深度学习方法开发了AlphaFlow、AF-Cluster[22,23]等构象采样方法，端到端地生成蛋白质的构象集合. 然而, 这些基于深度学习的蛋白质多构象采样方法受限于其模型设计和有限的训练数据, 无法真正得到蛋白质的构象系综, 同时，此类方法无法采样蛋白质的局部区域，只能完成端到端的蛋白质整体构象生成.

由DeepMind公司训练的AlphaFold系列深度学习模型[7,9,24]作为最具代表性的深度学习蛋白质模型，推动了蛋白质单体、多体结构预测、蛋白与其他生物大分子、小分子的结合结构预测领域的巨大进步. AF2和AF3均从大量的序列和结构数据中训练, 通过注意力机制编码序列和结构中广泛存在的高维关联. 其中，AF2经过Invariant Point Attention机制(下文简称IPA)的迭代解码后，实现了与经典物理模拟方法相当的实验预测精度和高度可靠的结构置信度打分. 而AF3在AF2的基础上增添了扩散模块，从而能够生成包括小分子、核酸、氨基酸修饰在内的更加多样的结构.

对于AF2的底层机制，研究者进行了深入的探索: 2021年，Ovchinnikov和Huang[25]提出了蛋白质折叠的“迷宫模型”，指出同源序列在引导蛋白质折叠方面可能具有积极作用. 2022年, Roney和Ovchinnikov[26]利用AF2中的“能量函数”构建了AF2Rank, 很好地反映了不同蛋白质构象到稳定结构的差距. 2024年, Li等人[27]将物理能量景观信息通过多序列比对(下文简称MSA)形式加入 AF2, 成功增强了AF2对于蛋白质多构象的生成能力. 这些研究在一定程度上揭示了AF2的“黑箱”，从传统蛋白质物理的角度解释了蛋白质折叠神经网络.

蛋白质的结构不仅仅是序列的一个函数，而是基于序列和相互作用环境的概率分布，蛋白质折叠可以使用能量景观理论来描述. 来自MSA的信息在一定程度上表征了蛋白质折叠能量景观中的信息[28]: David Baker团队[29]开发的GREMLIN，通过直接耦合分析(DCA)从MSA中提取协同进化信号，以指导目标序列的折叠能量景观; Yang等人[30]引入了“变换约束的Rosetta”，能够直接从MSA中提取残基间的方向和距离约束, 补充经典的Rosetta能量函数. 这些方法综合展示了进化信息在构建平滑的基于知识的能量超表面中的实用性，从而更好地近似蛋白质在生物发生过程中所探索的折叠漏斗的物理相关区域. Zhang等人[31]提出的EvoGen阐述了蛋白质折叠网络AF2中存在的参数化能量景观. Bryant和Noé[32]和Dorothee等人[23]的基于AF2的方法也相继证明，改变MSA可以提升结构预测质量, 或将单链蛋白质优化为不同的折叠状态. 改变MSA的输入信息能够扰动模型的折叠能量景观，从而生成蛋白质折叠的不同构象. 但直接删改输入的MSA序列无法确保MSA中保留信息的正确性，同时会改变原本平滑的能量景观，这通常导致模型的预测结果存在较大的不合理性，甚至无法得到正确的蛋白质整体折叠结构.

与此同时，Evoformer的模板输入信息和结构模块的IPA初始化信息为结构模块中的单一表示和IPA的演化过程提供了起始点. Meiler等人[33]通过添加不同构象的模板, 并严格控制MSA的深度, 从而使得在非平滑能

量景观的不同起点开始，结构演化至不同的局部极小值，获得了多构象结构. Wallner等人[34]的AF2 sampler通过在结构模块中加入Dropout扰动，使其演化至不同的局部极小值，获得了与实验解析结构更加一致的结果. 这些工作共同说明，在结构模块的循环更新过程中存在一个对应能量景观的迭代演化系统. 通过扰动或控制该演化系统，模型可以获得不同的结果. 然而，目前尚无任何研究工作能够系统和定向地扰动该系统，直接控制其演化结果.

本文提出了一种在生物结构相关研究领域中使用AF2的新视角. 我们将高质量MSA和初始化结构帧视作AF2不同模块的最优提示词. 一方面，输入高质量MSA以确保序列折叠能量景观的基本形态，另一方面，基于MSA的隐性控制机制增加了同源序列扰动模块，不改变整体模型结构和重新训练的情况下，通过操控模型“能量函数”的变化获得一组具有不同局部构象的蛋白质结构. 此外，我们将结构模块的解码过程视为类似动力学系统的迭代演化过程，重新设计了AF2结构模块的演化逻辑，使得我们可以通过模板输入信息和IPA初始化信息的诱导，控制结构模块的演化终点停留在我们希望的固定点. 将上述改动集成到AF2框架中，形成FoldCopilot套件，套件包含Fixer和Conformer两个工具，分别适用于蛋白质结构修复和局部多构象生成两种下游场景. 测试结果表明，经过提示词增强和上下文信息诱导的AF2在与蛋白质结构相关的各种生物学挑战中，包括蛋白质结构修复和局部多构象生成等方面，取得了显著领先于其他先进方法的结果.

1 模型基本架构

AF2充分结合了模版法构建蛋白质结构和蛋白质深度学习预测接触残基任务的特点，能够对人类已知的绝大部分蛋白质进行准确预测结构，与此同时，AF2作为基础模型，能够完成一系列除了蛋白质单链结构预测之外的任务，如蛋白质-蛋白质相互作用预测、单位点突变预测、多构象预测等，为基于提示词工程的上下文学习的应用奠定了深厚的基础.

1.1 AF2模型输入

AF2的模型输入分为三部分: 待预测的蛋白质序列(Target Sequence Profile)、多序列比对文件(MSA Profile)和模版文件(Template Profile)(图1). 多序列比对文件指在序列库中检索序列相似度较高的序列并与待预测序列进行对齐(Alignment)，该输入文件又被称为MSA[35]. MSA常常被用于预测蛋白质的二级结构、残基对距离矩阵等任务，其比对中位点的保守程度通常反映了对蛋白质保持折叠或执行功能的重要性，蛋白质的两个氨基酸残基的共同进化通常意味着这些氨基酸之间存在可能的相互作用. AF2和MSA Transformer等工作通过 Attention机制学习MSA中不同残基位点之间的潜在关联，从而较好地反映残基对之间的相互作用和空间距离特征. 模型的另一部分输入是模版文件，通过序列比对，AF2将相似序列的蛋白质三维结构片段作为模版提取其特征，合理的蛋白质结构模版往往能够大大增强预测结构的正确率.

1.2 AF2模型结构

AF2的模型结构可分为Evoformer模块和结构模块: Evoformer模块是基于Attention机制和Transformer架构的编码器[36]，该编码器共有48层，层间不共享参数. 庞大的参数量和先进的架构使其能够整合原始序列、MSA和模版结构的信息，提取共进化信号，生成一维序列的高维表示(MSA表示)以及二维残基对间的高维配对表示(pair representation)，其中MSA表示的首行即为与蛋白序列长度相对应的单序列表示(single representation).

结构模块作为解码器接收Evoformer模块编码的单序列表示和配对表示，基于黑洞初始化[9](“Black Hole” Initialization)的骨架结构帧预测新的骨架结构帧并最终生成全原子结构. 该模块拥有8层共享参数的网络循环结构，每个循环迭代利用不变点注意力(Invariant Point Attention, IPA)机制来更新残基表示，并预测所需的旋转和平移向量. IPA机制具有3D等变性，确保了对旋转和平移的不变性.

1.3 基于结构“上下文”的AF2提示词工程

上下文学习(In-contextual Learning)由Brown等人[37]于2020年提出，旨在特定结构的大模型中，利用输入数据提供的上下文来生成适当的响应或预测，与传统方法需要在标记数据集上进行明确的任务特定训练和微调形成对比. 上下文学习使大型语言模型能够利用大量数据，并以灵活且动态的方式适应各种任务[38,39]. 上下文学习有多种类别，根据是否给模型提供案例样本，可将其分为零样本、单样本和少样本学习[37,40,41]. 而提示工程则指对于指定任务利用上下文

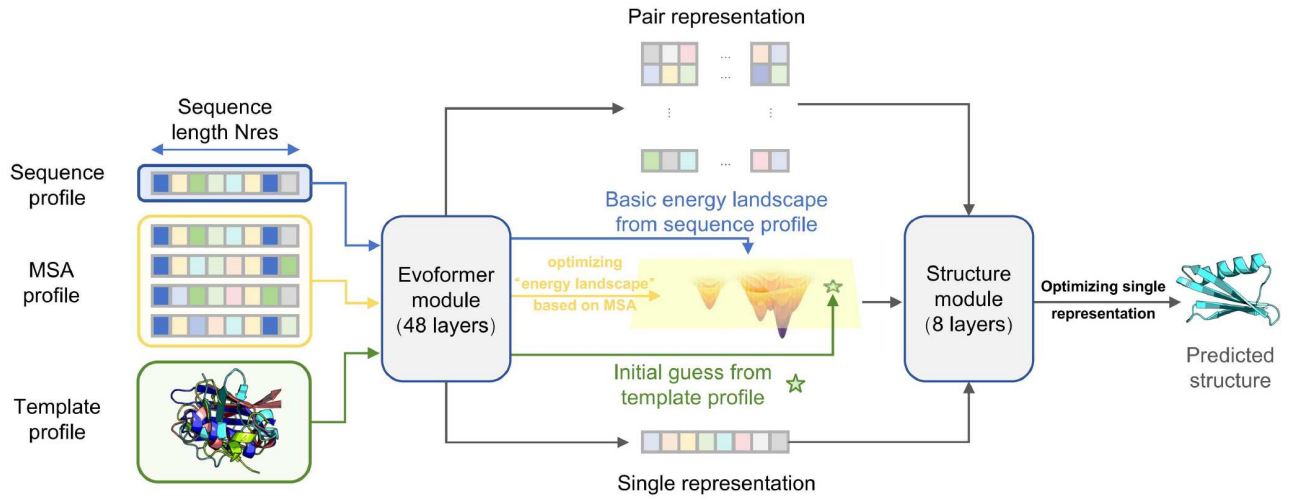

图 1 能量景观理论指导的AF2模型架构示意图
Figure 1 AF2 model architecture guided by the energy landscape theory

学习方法, 寻找最优的提示词上下文[42,43]. Li[44]于2023年提出:

$$P^{\star}=\arg\max_{P} E_{(x_i,y_i)\in\mathbb{D}}\left[S\left(f_{\theta}(P,x_i),y_i\right)\right], \tag{1}$$

即需要对于给定的大模型f和训练得到的参数θ, 输入x_i和标签y_i, 优化提示词P使得得分$S(\cdot)$最高. 在蛋白相关任务中, 研究者通常需要基于已知部分的蛋白质结构$\mathrm{x}_{\mathrm{known}}$和搜索得到的MSA信号得到一个或多个完整、符合物理、尽可能准确的蛋白质构象$\mathrm{x}_{\mathrm{all}}$, 因此, 已有部分的蛋白质结构和搜索得到的MSA输入文件即为大模型需要理解的"结构上下文". 但AF2能够理解的数据模态为氨基酸序列、MSA信息以及模版信息, 如何将"结构上下文"转化为三个模态信息(之一), 即寻找$g(\cdot)$使得$P^{\star}=g(\mathrm{x}_{\mathrm{known}})$, 成为了解决问题的关键.

在本工作中, 高质量MSA和基于已有结构初始化的初始化结构帧被我们视作AF2的最优提示词. 在广泛验证的MSA采样策略的基础上, 我们将AF2的Evoformer模块改造为同源能量景观扰动装置(图2(a)). 为了得到输入的高质量MSA, 我们使用莱文斯坦距离(Levenshtein Distance[45])将搜索得到的MSA进行过滤、层次聚类, 最终采用最远点采样(Farthest Point Sampling)策略[46]进行下采样. 在进行最后一轮Evoformer模块循环前, 装置将随机采样一行MSA表示作为单序列表示, 而后进行最后一轮Evoformer模块的推理, 最后一轮推理所用的MSA表示仅为上一轮采样得到的单序列表示. 经过48轮Evoformer模块的循环后, 装置最终得到同源能量景观扰动后的单序列表示和配对表示.

为了将已知的部分结构$\mathrm{x}_{\mathrm{known}}$输入模型, 且对结构模块进行约束, 结构模块的输入序列为完整的蛋白质序列, 不完整的蛋白质已知结构将作为模版输入模型(图2(b)). 为了让模型充分理解上下文, 蛋白质的待预测局域结构$\mathrm{x}_{\mathrm{unknown}}$将被根据已知部分结构$\mathrm{x}_{\mathrm{known}}$进行插值初始化. 插值初始化后的蛋白质结构被作为初始结构帧传入AF2的结构模块中进行更新, 特别地, 结构模块中的已知结构部分将被"冻结", 停止接受来自单序列表示的更新. 最终生成的整体构象将会按照标准AF2的流程, 通过OpenMM弛豫以消除原子碰撞等不符合物理的结构细节.

2 模型原理

2.1 诱导不动点

AF2的结构模块可以视为一个关于蛋白骨架结构$\{x_i\}$和残基表示特征$\{s_i\}$的动力系统, 其迭代的"运动方程"由IPA决定, 即$\left\{\mathrm{s}_i^{k+1}\right\},\left\{\mathrm{x}_i^{k+1}\right\}=f_{\mathrm{IPA}}\left(\left\{\mathrm{s}_i^{k}\right\},\left\{\mathrm{x}_i^{k}\right\}\right)$. 实验表明, IPA所诱导的运动方程最后将会收敛到不动点, 即$\left\{\mathrm{s}_i^{*}\right\},\left\{\mathrm{x}_i^{*}\right\}=f_{\mathrm{IPA}}\left(\left\{\mathrm{s}_i^{*}\right\},\left\{\mathrm{x}_i^{*}\right\}\right)$. AF2的训练目标保证了, 以随机初始化的结构$\left\{\mathrm{x}_i^{0}\right\}$和由MSA以及结构模板导出的残基表示特征$\left\{\mathrm{s}_i^{0}\right\}$作为初始迭代点, IPA将收敛至该序列对应的"天然结构". 然而很多与蛋白质结构有关

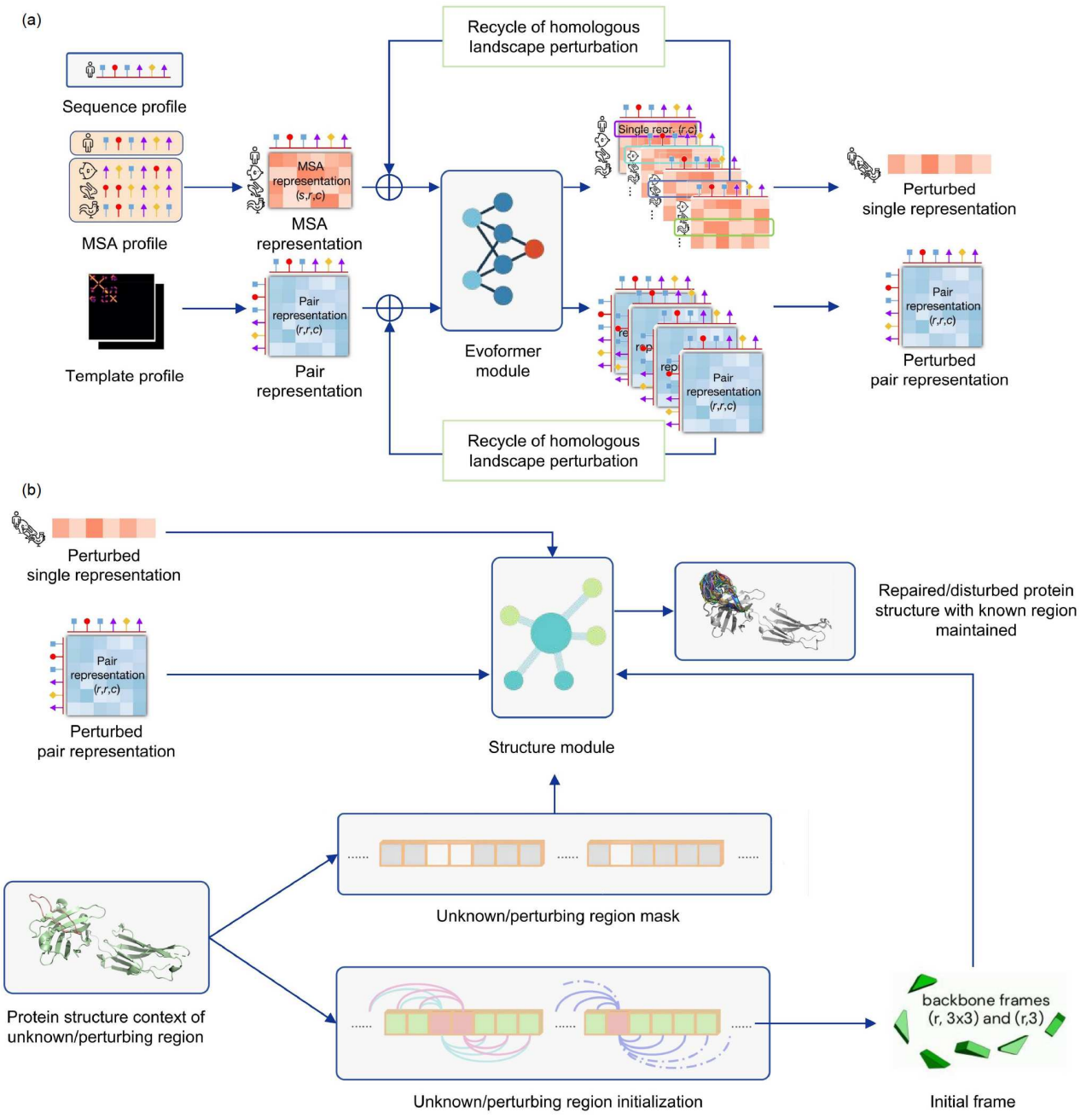

图 2 同源能量景观微扰和诱导不动点机制示意图. (a) Evoformer模块中的不同MSA序列将会产生不同的扰动序列和扰动循环; (b) 从部分区域未知/待扰动的蛋白质结构"上下文"出发完成对结构模块动力系统演化的精确控制

Figure 2 Diagram of the homogenous energy landscape perturbation and induced fixed-point mechanism: (a) Different MSA sequences in the Evoformer module generate varying perturbation sequences and cycles. (b) Precise control of the structural module's dynamical evolution is achieved from the "context" of partially unknown/unperturbed regions of the protein structure

的任务, 包括逆折叠、侧链组装、结构评估和结构补全等[47~49]往往需要模型感知和处理一个偏离天然蛋白结构分布的结构, 例如布居数较少的功能构象, 包含热涨落的动态结构等, 而AF2很难在IPA迭代中稳定这种结构, 因此不便用于这些任务当中.

对关心的骨架结构$\{\mathrm{x}_i^D\}$, 如果我们能找到一个合适的$\{\mathrm{s}_i^D\}$, 使其成为一对IPA迭代的不动点, 即满

足$\{s_i^D\}, \{x_i^D\} = f_{\mathrm{IPA}}\left(\{s_i^D\}, \{x_i^D\}\right)$，那么就可以将原本的折叠模型应用于结构感知的相关任务当中. 那么AF2是如何计算不动点的? 在传统的迭代优化算法中, 一个常用的停止准则是$|x^{k+1}-x^k|<\varepsilon$，即判断两个连续迭代点是否足够接近. 因此, 我们有理由假设AF2的结构模块的动态过程也遵循着相似的机制. 基于这个假设, 从一个初始的$\{s_i^0\}$出发, 在迭代中固定骨架结构为$\{x_i^D\}$不变, 即$\{s_i^{k+1}\}, \{x_i^D\} = f_{\mathrm{IPA}}\left(\{s_i^k\}, \{x_i^D\}\right)$, “诱导”AF2的动力学过程, 直到$\{s_i^k\}$收敛至$\{s_i^*\}$. 我们将这种方法称为“诱导不动点(induced fixed point, IFP)”算法. 根据迭代优化算法收敛判据的一般逻辑, AF2会倾向于认为$\{x_i^D\}$是一个不动点结构，从而自发地迭代出与之近似匹配的$\{s_i^*\} \approx \{s_i^D\}$. 因此, $\{s_i^*\}$可以被用于与骨架结构$\{x_i^D\}$相关的下游任务当中, 例如对$\{x_i^D\}$结构稳定性的评估, 或预测与$\{x_i^D\}$匹配的侧链结构.

除去完整的骨架结构$\{x_i^D\}$，IFP算法还可以被拓展应用于构建对部分骨架结构的感知. 例如, 在蛋白骨架结构补全的任务当中，模型需要感知给定结构的部分残基骨架$\{x_i^D\}_{i\in I}$，并补全出剩余的残基骨架$\{x_j^D\}_{j\in J}$, 按照相似的思路, 只需要将IFP的迭代更改为

$$\{s_i^{k+1}\}, \left\{\{x_j^D\}_{j\in J} \mid \{x_i^D\}_{i\in I}\right\} = f_{\mathrm{IPA}}\left(\{s_i^k\}, \left\{\{x_j^D\}_{j\in J} \mid \{x_i^D\}_{i\in I}\right\}\right).$$

2.2 同源能量景观微扰

有研究指出，AF2利用MSA辅助折叠的逻辑在于通过多序列平均平滑蛋白折叠的能量景观[31]. 具体而言，蛋白质的每一条同源序列都代表了一个粗糙的、存在多个局部极小点的折叠能量景观，AF2的Evoformer模块则通过将同源序列对应的多个折叠能量景观“积分”至目标序列的能量景观中，保留其中共同的局部极小点，抑制其他区域的涨落，构建一个更平滑的、利于折叠的能量景观. 因此, 通过操纵MSA可以获得不同的折叠能量景观, 从而诱导AF2获得不同的结构, 然而现有的操纵MSA的方法，包括对MSA进行下采样、聚类、突变都减少了MSA的条数，或降低了MSA的质量, 导致其能量景观的平滑程度有所下降, 最终影响折叠结构的置信度. 此外, 针对不同的MSA集合, 这些操纵MSA的方法都需要重复AF2的整个推理流程，在计算上相对耗时.

基于上面的思考，我们提出了一种新的操纵MSA的方式——同源能量景观微扰(Homologous Landscape Perturbation, HLP). 在HLP中, 为了不影响MSA的质量, 我们不直接微扰未经处理的MSA信息，而是选择微扰经由Evoformer提取得到的MSA信息. 在经由有限轮次的Evoformer迭代之后，每条同源序列所对应的能量景观已经充分交互, 变得平滑, 但仍然保留一部分原有的特征，包含不同的局部极小点. 此时我们使用AF2在每条同源序列对应的能量景观上进行折叠，就能得到多样且置信度高的构象系综. 此外，HLP策略可以与IFP算法结合, 在刚性固定部分残基骨架结构的情形下, 对蛋白局部的折叠能量景观进行微扰，采样局部的构象系综, 这种采样方式有利于对一些整体呈现刚性, 局部柔性较大的蛋白进行采样.

3 实验结果分析

3.1 应用1: 蛋白质稳态构象的局域修复

RCSB数据库中目前有接近22万蛋白质结构(截至2024年9月2日)，其中由于各种实验解析方法的局限性或其他原因, 蛋白质的结构会出现“缺失”部分, 即并非每个氨基酸都有对应的全原子结构，但在应用理论计算模拟方法如分子动力学模拟时，通常需要蛋白质的良好的起始构象, 如果不加以处理, 实验得出的结构的常见问题可能会导致时间和资源的浪费. 与此同时, 在蛋白质参与的分子对接、片段优化、功能区域设计等应用中，部分蛋白质构象可能通过理论或实验方法提前测得, 且可能不同于蛋白质在RCSB中解析得到的构象, 这使得下游应用的起始结构有较大误差, 影响下游方法的效果.

通常情况下，研究者选择使用商业软件如Schrödinger公司的Protein Preparation Wizard或免费软件如PDBFixer, Modeller, 以及Xponge等[50]进行修复，由于Protein Preparation Wizard需要购买正版软件许可才能使用，在此不列入讨论. Modeller首先对待补齐结构进行基于序列的模版搜索, 在提供序列相近的模版后, 模型会按照模版进行修补. 而PDBFixer和Xponge属于同

类方法，利用已有的蛋白质结构对缺失结构进行启发式初始化, 并采用力场进行简单的优化. 因此本工作选取Modeller和PDBFixer作为对照说明FoldCopilot-Fixer在给定已知部分结构时的蛋白结构修复能力. 并选用CASP14中的所有单链蛋白质作为测试集合以避免AF2过拟合，在后续测试中，FoldCopilot-Fixer采用AF2的model_1_ptm参数进行测试.

在CASP14的所有单链蛋白上，本工作按照序列长度的10%、20%、30%、40%、50%为每个蛋白随机采样了5套跨距掩码，跨距掩码并不连续，而是采样残基位点以后前后扩增，保证其与真实结构“破损”情况相类似. 从图3(a)中可以看出，在所有比例的跨距掩码上, FoldCopilot-Fixer的重构蛋白质结构与真实蛋白质结构的TM-score的中位数均大于0.97, 而PDBFixer仅能在跨距掩码为10%的情况下保证修复得到的蛋白质结构与真实结构的TM-score中位数大于0.95, 在跨距掩码大于等于序列总长的30%时，其均值迅速下降至0.9以下. FoldCopilot-Fixer得到的结构有可能出现较高的结构违反损失，这是由于PDBFixer等方法在补齐模版后会进行OpenMM优化, 降低了结构的不合理性. 随后本工作加入与AF2默认设置相同的OpenMM弛豫缓解, 结构合理性大幅提升(图3(b)). 统计修补残基的CA-RMSD可以发现，FoldCopilot-Fixer修补部分的残基在蛋白质的整体结构中处在较为合理的构象区域，与其真实结构最为接近，平均CA-RMSD最小(图3(c)). FoldCopilot-Fixer能够修复蛋白质结构的无规则区域、α螺旋、β折叠等不同的二级结构区域，并维持已知部分结构不变(图3(d)), 并且在测试中可以观察到, 对于AF2默认流程预测结果好或不好的蛋白质(图3(e)~(g)), FoldCopilot-Fixer均能够利用“结构上下文”信息，完成对待修复部分结构较高质量的修复.

通过以上实验，我们验证了FoldCopilot-Fixer能较高精度地修复包括α螺旋、β折叠、无规区域和链端结构等的绝大多数残基位点，相较于寻找缺失肽段的模版结构，基于上下文学习的AF2架构使得结构的修复质量大大提升，并节约了后续基于结构的模拟或设计工作的成本. 这充分展示了FoldCopilot-Fixer对各类蛋白质结构的修复能力，也体现了基于上下文学习的提示工程对AF2潜力的释放.

3.2 应用2: 蛋白质Loop区域构象的端到端采样

在蛋白质的识别、结合、诱导、反应催化和信号传导等过程中，蛋白质的Loop区域发挥着极为重要的作用. Loop区功能的多样性源自其结构灵活多变的特点，而这导致往往蛋白质的Loop区很难通过实验解析得到全原子三维结构, 这为蛋白质抗体的CDR3、信号受体等关键区域的精细结构、调控机理乃至进一步的理性设计带来较大的困难. 为此, 在FoldCopilot-Conformer套件中，本工作将同源能量景观微扰策略运用于Loop区采样中, 同时, 诱导不动点机制使得FoldCopilot能够将已知部分构象的约束施加在预测结果中，采样目标区域的多种构象.

本工作在CASP14蛋白质集合上测试了全局同源能量景观扰动策略. 与Ruff等人[51]的观察类似, AF2提供的每个残基的置信评分pLDDT可能与相应残基的构象柔性密切相关(图4(a)). 从采样得到的全局结构看, 蛋白质在无结构区域(如转角和环)表现出较大的柔性, 而在高度有序的区域(如α螺旋和β折叠)则趋于保守. 我们将采样的124个结构的主链二面角分别进行了层次聚类和UMAP降维可视化(图4(d), (e)), 可以明显观察到采样得到的蛋白质主链二面角分成了4类簇状结构. 为了获得更全面的结果，本工作通过对主链二面角(包括φ和ψ)的熵估计方法[52]，计算了每个残基的结构多样性. 如预期所示，φ和ψ的熵与pLDDT之间表现出较好的相关性(图4(f), (g)), 这一结果表明, 同源能量景观微扰策略能够有效探索蛋白质柔性区域的构象空间.

作为实际应用的例子，本工作利用 FoldCopilot-Conformer针对人副流感病毒3型的抗体和绿色荧光蛋白进行了构象采样(图4(b), (c)). 对于抗体蛋白而言, 其互补决定域CDR1、CDR2、CDR3，特别是CDR3呈现高度的结构柔性, 而其余部分则较为刚性, 如何采样抗体CDR3的构象是一个极具挑战的问题[53,54]. 因此我们使用局部同源能量景观微扰策略，将抗体的其余结构刚性固定，只采样CDR3的结构. 而对于绿色荧光蛋白, 其链接不同sheet结构的Loop区对其功能也有较大的影响, 因此我们选择采样其Loop区结构, 保持主要骨架区域不变. 实验表明，FoldCopilot-Conformer能够保持其余部分结构基本不变的同时，端到端地采样Loop区域的多种构象.

4 总结

综上所述，本研究基于深度学习模型AF2，提出了FoldCopilot套件, 包括Fixer和Conformer两个工具, 用于蛋白质结构的修复和局部多构象生成. 在传统的蛋白

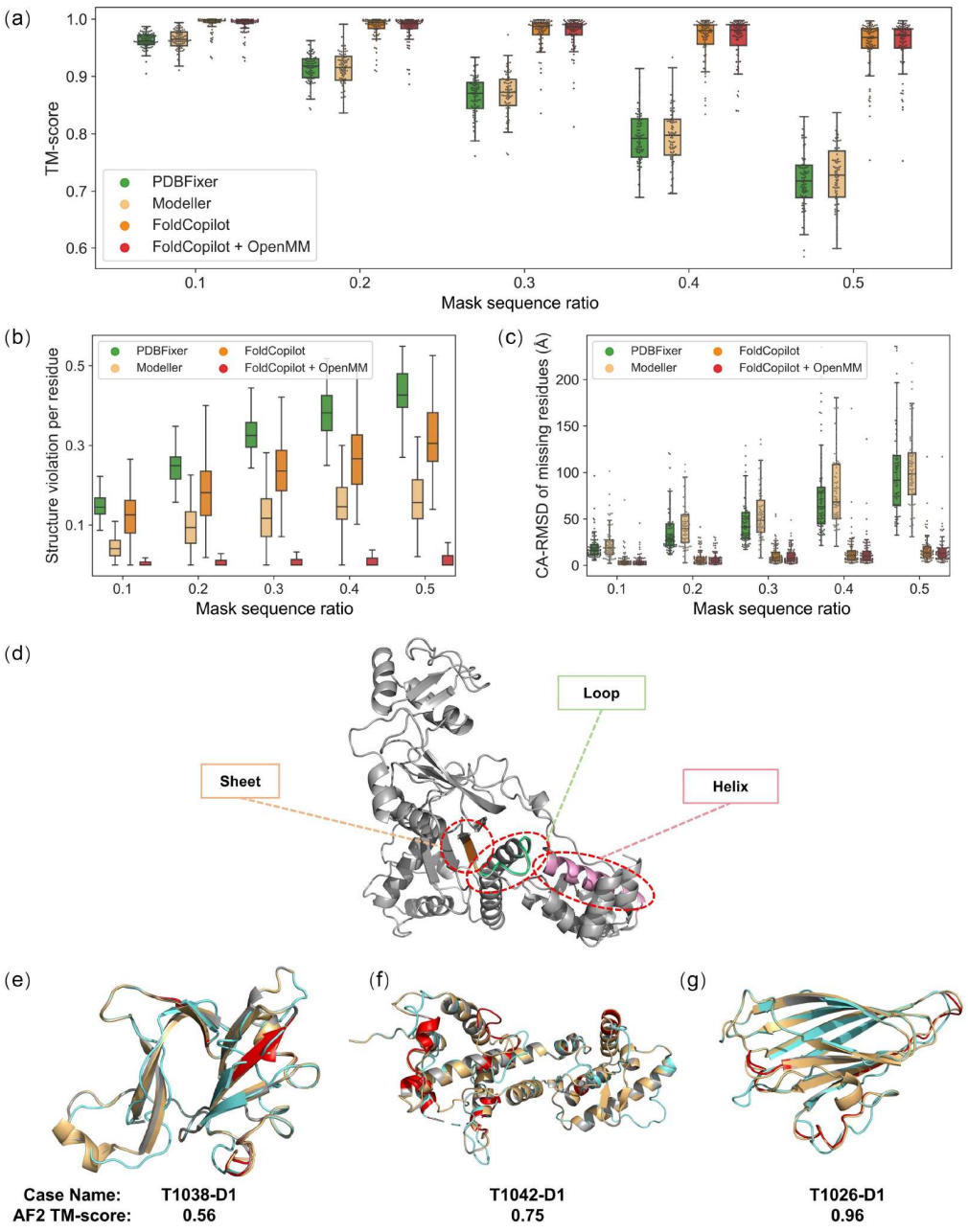

图 3 FoldCopilot-Fixer修复局域蛋白质构象能力的基准测试. (a) 不同跨距掩码时FoldCopilot-Fixer与PDBFixer的修复蛋白质与正确蛋白结构的TM-score分布. (b) 不同跨距掩码时FoldCopilot-Fixer与PDBFixer的修复蛋白质的结构违反损失, 计算方法为AF2中的Structure Violation Loss. (c) 不同跨距掩码时FoldCopilot-Fixer与PDBFixer的修复部分的CA-RMSD, 计算方法为将蛋白质已知残基CA原子对齐, 计算未知残基的CA原子均方根误差. (d) 蛋白质的修复区域示例. (e)~(g) CASP14中的三组蛋白质的FoldCopilot-Fixer修复结构图, 红色为待修复部分结构的正确结构, 灰色为已知部分结构, 金色为修复后结构, AF2 TM-score为AF2默认流程的结构与真实结构的得分; 对于AF2不能准确预测的蛋白质结构如(e)和(f), 给定上下文后FoldCopilot-Fixer也较好地完成了结构修复任务

Figure 3 Benchmark test of FoldCopilot-Fixer's ability to repair local protein conformations. (a) TM-score distribution of repaired protein structures compared to the correct structures, using FoldCopilot-Fixer and PDBFixer under different span masks. (b) Structure violation loss of repaired protein structures under different span masks, calculated based on the Structure Violation Loss metric in AF2. (c) CA-RMSD of repaired protein structures under different span masks, protein structures are aligned by CA atoms of known residues of structures, RMSD is calculated on repaired residues' CA atoms. (d) Example of the repaired regions in proteins. (e)–(g) FoldCopilot-Fixer repaired structures for three sets of proteins from CASP14, with red representing the correct structure of the region to be repaired, gray indicating the known part of the structure, and gold showing the repaired structure. The AF2 TM-score refers to the score between AF2's default predictions and the true structure. For protein structures where AF2's default approach was less accurate, such as in (e) and (f), FoldCopilot-Fixer effectively completed the repair when context was provided

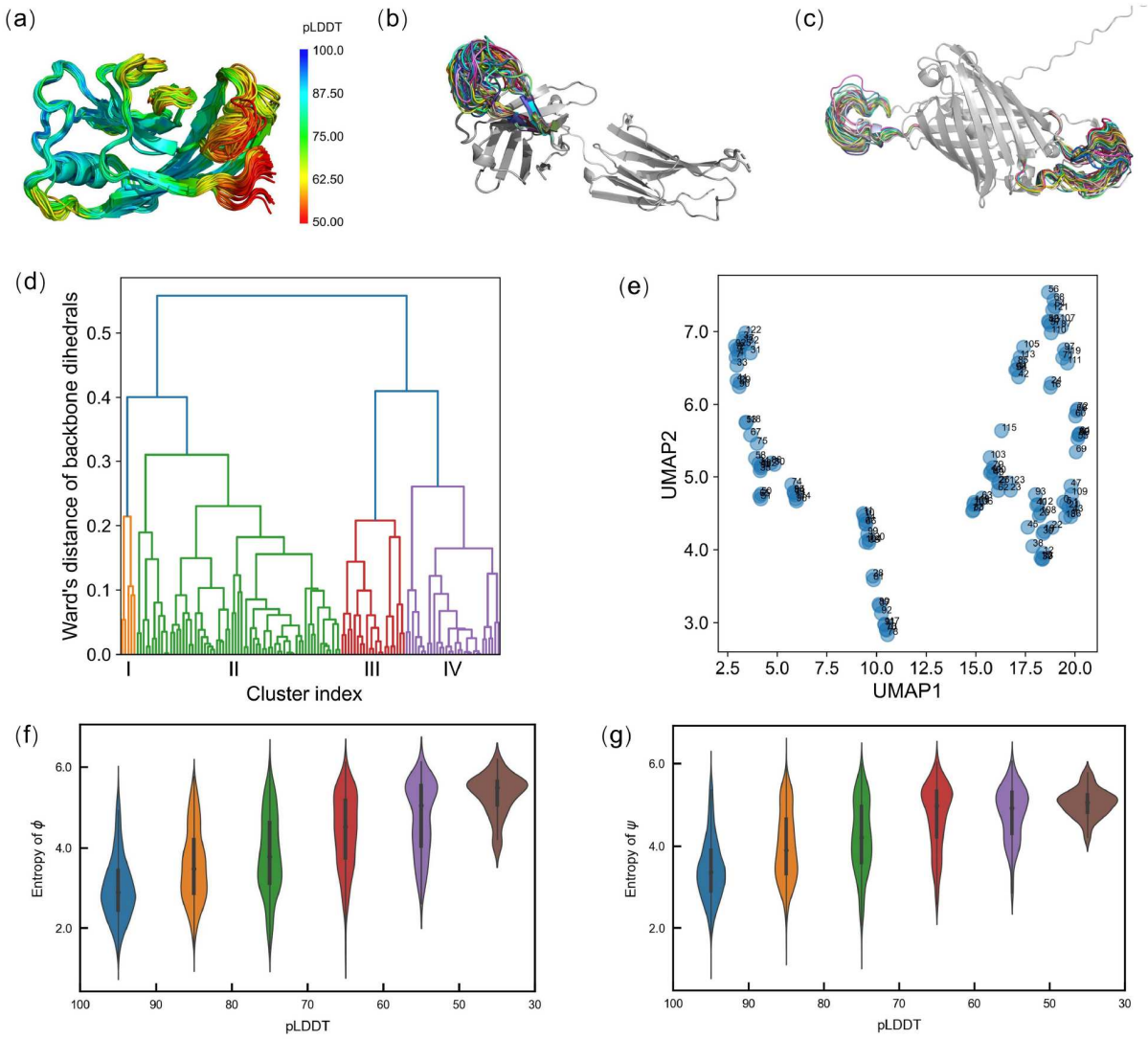

图 4 FoldCopilot-Conformer采样蛋白质构象能力的基准测试. (a) FoldCopilot-Conformer对样例蛋白(PDBID: 6Y4F)进行全局构象采样得到的构象系综, 残基按预测得到的pLDDT染色. (b) FoldCopilot-Conformer对人副流感病毒3型抗体(PDBID: 6WRP)的CDR3进行局部构象采样得到的结构系综. (c) FoldCopilot-Conformer对绿色荧光蛋白(PDBID: 2AWJ)的Loop区域进行局部构象采样得到的结构系综. FoldCopilot-Conformer对样例蛋白(PDBID: 6Y4F)采样得到的124个构象的主链二面角分别进行了层次聚类(d)和UMAP降维(e)可视化结果. (f), (g) 在CASP14蛋白集合上使用FoldCopilot-Conformer进行构象采样, 构象系综的主链二面角(φ和ψ)熵与对应残基pLDDT的关联

Figure 4 Benchmarking the conformational sampling capability of FoldCopilot-Conformer. (a) Conformational ensemble obtained by the global conformational sampling of the sample protein (PDBID: 6Y4F) using FoldCopilot-Conformer, with residues colored according to the predicted pLDDT. (b) Structural ensemble obtained by local conformational sampling of the CDR3 region of the human parainfluenza virus type 3 antibody (PDBID: 6WRP) using FoldCopilot-Conformer. (c) Structural ensemble obtained by the local conformational sampling of the loop region of green fluorescent protein (PDBID: 2AWJ) using FoldCopilot-Conformer. Hierarchical clustering (d) and UMAP dimensionality reduction (e) visualization results of the main chain dihedral angles of 124 conformations sampled for the sample protein (PDBID: 6Y4F) using FoldCopilot-Conformer. (f), (g) Correlation between the entropy of main chain dihedral angles (φ and ψ) and the corresponding residue pLDDT in the conformational ensembles sampled using FoldCopilot-Conformer on the CASP14 protein set

质结构修复和多构象采样方法中, 主要依赖于物理模拟和模板匹配. 然而, 这些方法存在计算成本高、效率低以及难以局域控制采样范围等问题. FoldCopilot通过引入基于上下文学习的提示词工程, 有效解决了这些局限性.

在蛋白质结构修复方面, FoldCopilot-Fixer能够利用已知的部分结构信息, 通过AF2模型的上下文输入, 精准修复包括α螺旋、β折叠和无规区域在内的绝大多

数残基位点. 相比传统的PDBFixer工具, FoldCopilot-Fixer在多种跨距掩码情况下均表现出更高的修复精度和稳定性, 并节省了后续模拟和设计工作的成本. 这表明, 基于上下文学习的提示词工程能够有效释放AF2在蛋白质结构修复方面的潜力. 在多构象采样方面, FoldCopilot-Conformer针对蛋白质的Loop区域, 通过改进的MSA采样策略, 能够在不改变已知部分构象的基础上, 有效探索Loop区域的多种构象. 实验结果显示, Conformer在多种蛋白质结构, 特别是复杂的抗体CDR3区域中, 能够生成不同的结构亚稳态, 为蛋白质分子设计和功能预测提供了新的思路.

目前该套件已开源以用于建模蛋白质的初始构象和采样局域多构象结构. 未来, 该套件将作为插件集成至分子动力学软件SPONGE[55]中, 高效地辅助建模蛋白质初始构象, 并对蛋白质-蛋白质相互作用、药物设计等基础科学研究和应用领域提供强有力的工具支持.

参考文献

1 Andersen H C. Molecular dynamics simulations at constant pressure and/or temperature. J Chem Phys, 1980, 72: 2384–2393
2 Arcon J P, Defelipe L A, Modenutti C P, et al. Molecular dynamics in mixed solvents reveals protein–ligand interactions, improves docking, and allows accurate binding free energy predictions. J Chem Inf Model, 2017, 57: 846–863
3 Heo L, Arbour C F, Janson G, et al. Improved sampling strategies for protein model refinement based on molecular dynamics simulation. J Chem Theor Comput, 2021, 17: 1931–1943
4 Leonov H, Mitchell J S B, Arkin I T. Monte Carlo estimation of the number of possible protein folds: Effects of sampling bias and folds distributions. Proteins, 2003, 51: 352–359
5 Metropolis N, Ulam S. The Monte Carlo Method. J Am Statistical Assoc, 1949, 44: 335–341
6 Sabbadin D, Moro S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR–ligand recognition pathway in a nanosecond time scale. J Chem Inf Model, 2014, 54: 372–376
7 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630: 493–500
8 Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, 373: 871–876
9 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
10 Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, 379: 1123–1130
11 Durairaj J, Waterhouse A M, Mets T, et al. Uncovering new families and folds in the natural protein universe. Nature, 2023, 622: 646–653
12 Dauparas J, Anishchenko I, Bennett N, et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science, 2022, 378: 49–56
13 Feng S, Chen Z, Zhang C, et al. Integrated structure prediction of protein–protein docking with experimental restraints using ColabDock. Nat Mach Intell, 2024, 6: 924–935
14 Huang Y P, Zhang H, Jiang S, et al. DSDP: a blind docking strategy accelerated by GPUs. J Chem Inf Model, 2023, 63: 4355–4363
15 Lisanza S L, Gershon J M, Tipps S W K, et al. Multistate and functional protein design using RoseTTAFold sequence space diffusion. Nat Biotechnol, 2024, doi: 10.1038/s41587-024-02395-w
16 Ren M, Yu C, Bu D, et al. Accurate and robust protein sequence design with CarbonDesign. Nat Mach Intell, 2024, 6: 536–547
17 Watson J L, Juergens D, Bennett N R, et al. De novo design of protein structure and function with RFdiffusion. Nature, 2023, 620: 1089–1100
18 Wee J J, Xia K. Persistent spectral based ensemble learning (PerSpect-EL) for protein–protein binding affinity prediction. Brief BioInf, 2022, 23: bbac024
19 Owens R J. Structural Proteomics: High-Throughput Methods. New York: Springer, 2015
20 Eastman P, Swails J, Chodera J D, et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol, 2017, 13: e1005659
21 Webb B, Sali A. Comparative protein structure modeling using Modeller. CP BioInf, 2016, 54: 5.6.1–5.6.37
22 Jing B, Berger B, Jaakkola T. AlphaFold meets flow matching for generating protein ensembles. 2024
23 Wayment-Steele H K, Ojoawo A, Otten R, et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature, 2024, 625: 832–839
24 Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature, 2020, 577: 706–710
25 Ovchinnikov S, Huang P S. Structure-based protein design with deep learning. Curr Opin Chem Biol, 2021, 65: 136–144
26 Roney J P, Ovchinnikov S. State-of-the-art estimation of protein model accuracy using AlphaFold. Phys Rev Lett, 2022, 129: 238101
27 Guan X, Tang Q Y, Ren W, et al. Predicting protein conformational motions using energetic frustration analysis and AlphaFold2. Proc Natl Acad Sci USA, 2024, 121: e2410662121

28 Smith R F, Smmith T F. Pattern-induced multi-sequence alignment (PUMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. Protein Eng Sel, 1992, 5: 35–41

29 Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci USA, 2013, 110: 15674–15679

30 Du Z, Su H, Wang W, et al. The trRosetta server for fast and accurate protein structure prediction. Nat Protoc, 2021, 16: 5634–5651

31 Zhang J, Liu S, Chen M, et al. Unsupervisedly prompting AlphaFold2 for accurate few-shot protein structure prediction. J Chem Theor Comput, 2023, 19: 8460–8471

32 Bryant P, Noé F, Schneidman D. Improved protein complex prediction with AlphaFold-multimer by denoising the MSA profile. PLoS Comput Biol, 2024, 20: e1012253

33 del Alamo D, Sala D, Mchaourab H S, et al. Sampling alternative conformational states of transporters and receptors with AlphaFold2. eLife, 2022, 11: e75751

34 Kalakoti Y, Wallner B. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. Commun Biol, 2025, 8: 373

35 Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol, 2017, 35: 1026–1028

36 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. Vol. 30. Curran Associates, Inc., 2017

37 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, et al., eds. Advances in Neural Information Processing Systems 33, NEURIPS 2020. La Jolla: Neural Information Processing Systems (NIPS), 2020

38 Liu J, Shen D, Zhang Y, et al. What makes good in-context examples for GPT-3? 2021,

39 Wang X, Zhu W, Saxon M, et al. Large language models are latent variable models: explaining and finding good demonstrations for in-context learning. In: Advances in Neural Information Processing Systems 36 (NEURIPS 2023). La Jolla: Neural Information Processing Systems (NIPS), 2023

40 Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning. 2017

41 Xian Y, Lampert C H, Schiele B, et al. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 2251–2265

42 Chen B, Zhang Z, Langrené N, et al. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. 2024,

43 Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv, 2023, 55: 1–35

44 Li Y. A practical survey on zero-shot prompt design for in-context learning. 2023

45 Berger B, Waterman M S, Yu Y W. Levenshtein distance, sequence comparison and biological database search. IEEE Trans Inform Theor, 2021, 67: 3287–3294

46 Eldar Y, Lindenbaum M, Porat M, et al. The farthest point strategy for progressive image sampling. IEEE Trans Image Process, 1997, 6: 1305–1315

47 Hsu C, Verkuil R, Liu J, et al. Learning inverse folding from millions of predicted structures. In: Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022, 162: 8946–8970

48 Sánchez R, Šali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins: Struct Funct Bioinf, 1997, 29: 50–58

49 Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Sci, 2005, 14: 1328–1339

50 Xia Y, Gao Y Q. Xponge: a Python package to perform pre- and post-processing of molecular simulations. JOSS, 2022, 7: 4467

51 Ruff K M, Pappu R V. AlphaFold and implications for intrinsically disordered proteins. J Mol Biol, 2021, 433: 167208

52 Preto J, Clementi C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. Phys Chem Chem Phys, 2014, 16: 19181–19191

53 Fernández-Quintero M L, Kokot J, Waibl F, et al. Challenges in antibody structure prediction. mAbs, 2023, 15: 2175319

54 Giulini M, Schneider C, Cutting D, et al. Towards the accurate modelling of antibody−antigen complexes from sequence using machine learning and information-driven docking. Bioinformatics, 2024, 40: btae583

55 Huang Y, Xia Y, Yang L, et al. SPONGE: a GPU-accelerated molecular dynamics package with enhanced sampling and AI-driven algorithms. Chin J Chem, 2022, 40: 160–168

Summary for “基于物理启发式的深度学习方法对蛋白质三维结构的修复与扩展采样”

Physics-informed deep learning approach for fixing and sampling protein 3D structures

Zhenyu Chen[1†], Xiaohan Lin[1†], Yanheng Li[1], Zicheng Ma[2], Jun Zhang[2*] & Yi Qin Gao[1,2*]

[1] *Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*
[2] *Changping Laboratory, Beijing 102200, China*
† Equally contributed to this work
* Corresponding authors, E-mail: jzhang@cpl.ac.cn; gaoyq@pku.edu.cn

AlphaFold2 (AF2), the first deep learning-based protein structure prediction model to achieve near-experimental accuracy, has revolutionized the field of structural biology. With its ability to predict protein structures at unprecedented levels of accuracy, AF2 has become an essential tool in structural biology. In parallel, the deep learning methodologies derived from AF2 have made significant advancements in a range of downstream applications, such as efficient sampling of protein conformational ensembles, protein sequence design, and ligand generation. These applications are underpinned by an accurate description of the protein-free energy landscape, which is critical for modeling protein behavior in various biological contexts.

Despite these advancements, challenges remain, particularly in the area of multi-conformation sampling. Traditional methods such as molecular dynamics (MD) and Monte Carlo (MC) simulations, which are based on Boltzmann statistics, can provide detailed insights into conformational distributions. However, the application of these methods is hindered by their high computational cost and limited efficiency, especially for large or complex systems. In contrast, existing generative deep learning models can rapidly generate protein conformations, but often struggle to incorporate biophysical constraints, and are limited in the power of predicting the accurate energy minima, avoiding steric clashes, or understanding physics such as electrostatic interactions. This lack of control over the physical space makes it difficult to generate conformations that are both physically realistic and biologically relevant.

Furthermore, the robustness of the initial protein structure is crucial for downstream applications, including protein-protein interaction predictions, molecular docking, and virtual screening, as well as reliable and efficient MD simulations. However, conventional structure repair tools often fail to accurately model missing regions, such as loops or partially resolved regions, which may lead to unrealistic geometries or energetically unfavorable conformations.

To address these limitations, we present FoldCopilot, an innovative suite based on AF2. FoldCopilot uses both target sequences and multiple sequence alignments (MSAs) as input prompts for the Evoformer module, leveraging a homology-based sequence perturbation mechanism that generates diverse local conformations without the need for retraining. This method allows for the efficient generation of protein conformations and at the same time maintains high fidelity with known structural templates and evolutionary constraints. In addition, FoldCopilot integrates structural templates and initial reference frames into the iterative refinement process, enabling the directed evolution and control over generated protein structures. This iterative process ensures that the generated conformations not only adhere to biophysical principles, such as realistic energy states and steric compatibility but also meet specific functional requirements dictated by the target application.

Our experimental results demonstrate that FoldCopilot excels in key tasks such as protein structure repair and localized multi-conformation generation. The method significantly outperforms traditional structure repair techniques, yielding more accurate and biologically relevant models for regions of missing data. Additionally, FoldCopilot shows a high capability to generate diverse conformations, making it an ideal tool for protein engineering, protein-ligand interaction predictions, and other structure-based drug design applications.

In summary, FoldCopilot bridges the gap between high-fidelity conformational sampling and the ability to control conformational space, offering a powerful approach to facilitate protein-molecule interaction predictions, structural modeling, and molecular dynamics simulations.

AF2, prompt engineering, conformation sampling, dynamical system control, homologous landscape perturbation

doi: 10.1360/TB-2024-1149