

DOI: 10.13376/j.cbbs/2025089

文章编号: 1004-0374(2025)07-0894-10

· 蛋白质设施与产业界的合作 ·



李兆达, 毕业于中国农业大学生物系, 从事病毒学研究工作。先后在国内股权投资基金、二级市场投资基金从事医药方向研究投资工作。对国内医疗产业中早期投资、企业并购、上市公司投资均有较为深刻的理解。



贺明, 博士毕业于清华大学药学院, 研究领域主要包括小分子药物设计、PROTAC 及分子胶设计、分子动力学模拟及在药物开发中的应用等, 共发表 10 余篇高水平论文, 其中包括 *Nature Chemical Biology* 等。

蛋白质动态结构解析研究进展

蔡幸初¹, 李兆达^{2*}, 张冬妮², 肖凡玥², 贺明^{2*}

(1 湖北思朗万维计算装备制造有限公司, 孝感 432012; 2 长江3D科学计算中心, 孝感 432015)

摘要: 蛋白质结构解析是理解生命机制与药物研发的核心。本文系统回顾了人工智能 (artificial intelligence, AI) 驱动下的蛋白质结构预测的进展、分子动力学 (molecular dynamics, MD) 模拟在揭示动态构象与分子互动中的发展, 以及高性能计算 (如专用科学计算芯片 MHPC512) 对长时程模拟发挥的作用。研究表明, AI 预测模型已实现从静态单体到多分子复合体 (蛋白质 - 核酸 - 配体) 的高精度解析, 但在动态行为模拟和侧链预测上仍存局限性。经典 MD 模拟通过改进增强采样策略, 已能捕捉微秒级蛋白质构象变化, 成为揭示药物结合机制与蛋白质功能动态的关键工具。高性能计算 (如 Anton 系列、国产 MHPC512) 已经突破算力瓶颈, 使百万原子体系毫秒级模拟成为可能, 为构建蛋白质动态结构库奠定基础。未来需深度融合 “AI 预测初筛 + MD 动态精修 + 实验验证” 的范式, 推动靶向动态构象的创新药物设计。大规模的 MD 蛋白质动态结构数据库, 将在未来的科研和工业端发挥巨大的应用价值。

关键词: 蛋白质结构预测; 分子动力学模拟; 人工智能; 高性能计算; 药物设计

中图分类号: Q518; TP18 **文献标志码:** A

收稿日期: 2025-06-24

*通信作者: E-mail: zhaoda.li@smartlogitech.com (李兆达); ming.he@smartlogitech.com (贺明)

Research progress of protein dynamic structure analysis

CAI Xing-Chu¹, LI Zhao-Da^{2*}, ZHANG Dong-Ni², XIAO Fan-Yue², HE Ming^{2*}

(1 Hubei Universal Logic Calculating Equipment Manufacturing Co., Ltd, Xiaogan 432012, China; 2 Yangtze River 3D Scientific Computing Center, Xiaogan 432015, China)

Abstract: Protein structure determination serves as the cornerstone for understanding biological mechanisms and drug discovery. This review systematically examines the revolutionary advances in artificial intelligence (AI)-driven protein structure prediction, the indispensable role of molecular dynamics (MD) simulations in elucidating dynamic conformations and molecular interactions, and the breakthrough support provided by high-performance computing (HPC), exemplified by specialized supercomputing chips like MHPC512, for enabling long-timescale simulations. Key findings indicate: AI prediction models have achieved high-precision resolution from static monomers to multi-molecular complexes (protein-nucleic acid-ligand), yet remain limited in simulating dynamic behavior and accurately predicting side-chain conformations. Classical MD simulations, enhanced by refined sampling strategies, can now capture microsecond-scale protein conformational changes, establishing themselves as critical tools for revealing drug binding mechanisms and functional protein dynamics. High-performance computing breakthroughs (e.g., Anton series, domestic MHPC512) overcome computational bottlenecks, enabling millisecond-scale simulations of million-atom systems and laying the foundation for constructing dynamic protein structure libraries. Future progress necessitates the deep integration of an "AI Initial Screening + MD Dynamic Refinement + Experimental Validation" paradigm to propel the design of innovative drugs targeting dynamic conformations. Crucially, large-scale MD protein dynamic structure databases are expected to deliver immense value in both research and industrial applications.

Key words: protein structure prediction; molecular dynamics simulation; artificial intelligence; high performance computing; drug design

蛋白质作为生命功能的核心载体, 其动态构象变化与疾病机制、药物作用等关键生物过程密切相关。自 1958 年首例蛋白质晶体结构解析以来, 冷冻电镜、核磁共振与 X-ray 等技术已将结构生物学推进至原子分辨率时代。然而, 生理条件下的蛋白质始终处于运动状态, 国际人类蛋白质组计划 (Human Proteome Project, HPP) 也揭示超过 80% 的疾病相关蛋白功能调控存在动态构象依赖, 单一静态结构难以全面阐释其功能机制, 这驱动了计算模拟技术的创新发展。其中最为人所熟知的研究成果是 2018 年推出的 AlphaFold, 一个开放获取的、包含大量高精度蛋白质结构预测数据的数据库, 它使得已知蛋白质序列空间结构覆盖率达到前所未有的扩张。目前 AlphaFold 预测数据和已知实验得到的蛋白质结构相比有极高的准确度, 但在蛋白质动态结构预测领域尚未有所突破。

分子动力学 (molecular dynamics, MD) 模拟通过数值求解牛顿运动方程, 可追踪蛋白质原子级动态轨迹, 为揭示变构效应、药物结合等过程提供理论依据, 也是目前解析蛋白质动态结构的金标准。随着技术的不断完善, 利用分子动力学方法解析蛋

白质的动态结构已经成为科研领域较为成熟的方法。但传统 MD 受限于微秒级时间尺度与经典力学精度, 难以捕获功能性慢速构象变化。而以高性能计算机为基础的较长时间 MD——如美国 Anton 系列科学计算与国产 MHPC512 (5~10 μ s/d)——正逐步突破这一瓶颈, 让较长时间的 MD 模拟成为现实。

本文聚焦人工智能 (artificial intelligence, AI) 预测、MD 模拟与科学计算技术在蛋白质动态结构预测中的进展, 探讨三者协同驱动下的蛋白质动态结构解析新范式及其在药物研发中的潜力。

1 人工智能模型在蛋白质结构解析中的进展

20 世纪 50 年代, X 射线晶体衍射技术的突破性发展为蛋白质三维结构解析奠定了方法学基础。1958 年, John Kendrew 与 Max Perutz 团队首次成功解析肌红蛋白与血红蛋白的空间构象, 这项里程碑式成果使其荣获 1962 年诺贝尔化学奖。此后, 核磁共振波谱 (nuclear magnetic resonance, NMR) 与冷冻电子显微镜 (cryogenic electron microscopy, Cryo-EM) 等技术的迭代创新, 共同构成了现代结构生物学的多模态研究体系。截至当前, 全球科研界已累

计解析约 20 万种蛋白质的高精度三维结构。

1972 年, 美国生物化学家 Christian Anfinsen 因其关于蛋白质折叠规律的革命性发现被授予诺贝尔化学奖。他通过核糖核酸酶复性实验证实: 蛋白质的氨基酸序列本身蕴含其空间折叠的全部信息, 无须依赖额外遗传指令。这一理论突破不仅确立了“序列决定结构”的结构生物学基本范式, 更开启了长达半个世纪的科学探索——如何仅凭氨基酸序列精确预测蛋白质三维构象。然而, 由于蛋白质构象空间的自由度呈指数级增长 (N^{3L} , 其中 N 为残基数, L 为键长变量), 其理论可能的构型组合可达 10^{300} 量级, 这种“维度灾难”使得结构预测长期被视为计算生物学领域的重大挑战。

为系统性评估结构预测技术的有效性, 学界于 1994 年建立了具有里程碑意义的蛋白质结构预测关键评估计划 (Critical Assessment of Structure Prediction, CASP)。该计划由 Moult 和 Fidelis 等学者发起, 采用双盲评估机制: 组织者定期发布系列未公开的氨基酸序列, 其对应蛋白质结构虽已被实验解析但尚未公开发表; 全球研究团队基于序列信息, 采用各类计算建模方法 (包括同源建模、从头预测等) 进行三维结构预测; 最终由独立评审委员会通过计算模型与实验解析结构的空叠合度、局部构象匹配度等指标进行量化评估, 构建结构预测技术的客观评价基准。

CASP 后来演变为竞赛, 旨在加速该领域的发现。然而, 直到多年后才取得重大突破。AlphaFold 在 2018 年的 CASP 横空出世, 其可以根据蛋白质的氨基酸序列预测蛋白质的结构, 但在当时的预测准确性也不到 60 分, 2020 年 11 月份则突破性地提高到 92.4 分, 成为人工智能深入到生物领域的一大突破。AlphaFold 前所未有的准确性和速度使得大规模结构预测数据库得以建立, Deepmind 于 2021 年 7 月在 *Nature* 杂志发表文章称 AlphaFold 已经预测出了 35 万种蛋白质结构^[1], 涵盖了 98.5% 的人类蛋白质组以及 20 种生物的蛋白质, 并开源了它的数据库。AlphaFold 使得已知蛋白质序列空间结构覆盖率达到前所未有的扩张, 生物学家能够获得几乎任何蛋白质序列的结构模型, 这改变了他们解决研究问题的方式, 很大程度上提高了科研人员的工作效率, 加速了生物领域创新成果的进展。AlphaFold 给蛋白质预测领域打开了新的大门, 许多工作基于这个重要的工具得以展开。例如, 2024 年 Monteiro da Silva 等^[2]使用 AlphaFold2 预测了蛋

白质构象的相对分布: 通过对多序列比对进行子抽样, 直接预测了不同蛋白质构象的相对分布; 与核磁共振实验进行对比测试, 其测试结果正确率超过 80%。

在 DeepMind 发布 AlphaFold2 突破性成果的同时, 美国华盛顿大学 David Baker 团队^[3]于 *Science* 杂志报道了新型深度学习框架 RoseTTAFold 的研发进展。该模型创新性地构建了三级协同处理通道: 第一通道解析氨基酸序列特征, 第二通道建模残基间空间几何关系 (包括距离与方向角), 第三通道直接作用于三维笛卡尔坐标空间的原子级构象优化。相较于 AlphaFold2 的双轨架构, 其新增的坐标空间优化模块显著强化了序列-结构-空间拓扑的多维度耦合建模能力, 使得蛋白质三维结构重建效率达到分钟级运算水平。在 CASP14 国际评估中, 该模型全局拓扑相似性评分 (GDT_TS) 达到 73.2 分, 虽略低于 AlphaFold2 的 92.4 分, 但仍超越传统计算方法两个数量级, 充分验证了多轨协同神经网络在构象空间搜索中的方法论优势。

2024 年 3 月, David Baker 团队^[4]在 *Science* 发表的 RoseTTAFold All-Atom (RFAA) 模型突破了传统蛋白质建模的局限。该创新性框架通过整合氨基酸残基的抽象表示与原子级精度的分子描述, 实现了对包含核酸链、金属离子及共价修饰的复杂生物组件的三维建模。在 PDB 数据库的训练支持下, RFAA 展现出与 AlphaFold2 (AF2) 相当的蛋白质结构预测能力, 尤其在柔性骨架小分子对接方面表现突出。值得关注的是, RFAA 对共价修饰蛋白质复合物的建模精度达到可应用水平, 为药物设计提供了新工具。

同年 5 月, DeepMind 团队^[5]在 *Nature* 发布的 AlphaFold3 标志着结构预测技术的又一里程碑。该模型通过改进算法架构, 将预测范围扩展到蛋白质与核酸、配体间的相互作用网络, 在药物靶点结合预测中实现了突破性进展。实验数据显示, 其预测抗体-抗原结合界面的准确度较前代提升显著, 这对免疫机制研究和抗体工程具有重要价值。

目前, 已有多个科研机构积极尝试基于深度学习框架, 使用蛋白质 MD 模拟数据训练, 生成蛋白质多样构象的新研究范式, 解决了传统计算无法克服的难题, 取得了一定的成果。例如, 2024 年德国亥姆霍兹研究中心发布的 MISATO 数据集^[6], 系统整合了 20 000 个蛋白质-配体复合物的量子力学修正结构与 170 μ s 分子动力学轨迹。在 2024 年 2 月

的 NeurIPS 上, Jaakkola 团队^[7]开发了一种基于流的生成模型方法 AlphaFlow, 用于学习和采样蛋白质的构象景观, 该方法基于 MD 模拟数据库中 1 522 条蛋白质链生成的 13.2 TB 的原始数据。2022 年 12 月, 西湖大学联合团队开发了 ProtMD 模型, 该模型的训练使用了大量的 MD 柔性数据^[8]。

基于物理模型的人工智能, 在蛋白质动态结构预测领域也有明显的突破。浙江大学团队开发的 EquiScore 模型通过异构图神经网络, 在等变几何空间实现了物理先验与数据驱动的有机融合^[9]。该模型在先导化合物优化阶段的成功率较传统方法提高 30%。在构象分布建模方面, ByteDance 团队提出的 CONFDIFF 模型创新性地将玻尔兹曼分布原理与扩散模型结合^[10]。通过物理场引导的生成策略, 该模型可高效产生符合热力学规律的蛋白质构象集合。与传统采样方法相比, 其构象生成速度提升百倍, 且能量分布与实验数据相关性达 0.89。复旦大学团队研发的 AlphaFolding4D 模型创新性地将时间维度引入扩散框架, 实现了对蛋白质运动轨迹的连续预测^[11]。该模型可处理长达 256 个氨基酸、32 个时间步长的动力学模拟, 对局部柔性运动和全局构象变化的预测精度分别达到 0.8Å 和 2.3Å。通过整合分子动力学先验知识, 其预测轨迹与实验观测的时间一致性成分 (TIC) 相关系数达 0.92。

上述人工智能模型大大加速了人类对蛋白质结构的预测能力, 是计算生物学近几年最为重要的研究成果, 但是像 AlphaFold 这样的模型也并不能完全模拟蛋白质的真实物理运动。例如, 2022 年 *Bioinformatics* 发表的文章指出, AlphaFold2 产生的“折叠轨迹”没有概括真实的蛋白质折叠动力学^[12]。AlphaFold 官方也公开表述其存在明显的短板。一方面, AlphaFold 只针对蛋白质“静态”结构的预测有较高的准确率, 无法预测蛋白质“动态”结构。具体表现为在直接使用靶蛋白与分子对接后, 其预测结果可能产生严重偏差。另一方面, 供 AlphaFold 训练的静态蛋白质结构库已无法支撑 AI 模型的快速发展。其根本原因在于 AI 算法根本上还是依赖于多年湿实验研究积累的大量的高质量的结构生物学数据供 AI 算法使用。这些实验解析的结构数据来自 X 射线衍射、冷冻电镜等方法, 均为极低温度下静态的蛋白质结构。而蛋白质的结构本质上是一个动态模型^[13], 而非实验得到的一个静态结果。另一个重要的事实是, 当前蛋白质结构数据很大程度上依赖体外表达蛋白的结构解析, 缺少部分翻译后

修饰的结构数据也对理解真实的蛋白质结构具有误导性。

室温下动态的蛋白质结构可由液态核磁共振光谱得到, 但目前数据较少, 尚未被 AI 算法使用。因而, 目前的 AI 方法对于时刻处于变化过程中的蛋白质结构, 描绘能力还比较有限。蛋白质的动态结构, 是继 AlphaFold 之后的重大科学问题。目前的瓶颈在于, 与现有的结构数据库中大量的蛋白质静态结构数据相比, 蛋白质结构的动态数据以及蛋白质与小分子相互作用细节的数据非常缺乏。在现有的蛋白质数据库中, 能够用于 AI 模型训练的氨基酸序列和蛋白质三维结构数据有限且基本已被充分使用, 导致生物模型的片面化和局限性。

动态 AI 模型将比静态 AI 模型发挥出更大的作用, 解决生命科学领域更为重大的问题, 比如静态模型无法解决的蛋白质构象分布、构象转化时间和路径、蛋白质 - 小分子结合 / 解离的快慢等。基于动态 MD 仿真模拟的 AI 模型可引入基于物理的能量函数作为先验知识引导构象采样, 而 AlphaFold3 强烈依赖于多序列比对 (MSA) 中的进化和协同进化信息, 对于找不到太多同源序列的蛋白质 (例如抗体、孤儿蛋白、病毒蛋白、复合物等), AlphaFold3 预测的精确度就会大幅度下降。因此, 与静态 AI 模型 AlphaFold3 相比, 动态 AI 模型将能具有更强的泛化能力。动态 AI 模型的另一技术特点是, 与静态 AI 模型 AlphaFold3 相同, 两者在可预见的未来不会被 AI 通用大模型涵盖, 因此与当下如火如荼开展的 AI 通用大模型研发不构成竞争关系, 而是并行关系。具体地说, 蛋白质科学领域的 AI 模型因其数据的特殊性、专业性、不易获取性, 在可预见的未来不会被纳入 AI 通用大模型的训练中。因此, 蛋白质科学领域的 AI 模型开发有其重要的意义与价值。此外, 相比生命科学其他领域的数据 (如细胞实验数据等), 包括 MD 仿真模拟数据在内的蛋白质相关数据与 AI 技术天然地更易于结合。而建立海量的蛋白质动态结构 MD 数据库, 对于训练出超越 AlphaFold3 的蛋白质动态结构预测模型具有至关重要的作用。

2 分子动力学模拟在蛋白质动态结构解析中的应用

蛋白质作为生命功能的核心执行者, 其结构与人类多数重大疾病的病理机制密切相关。研究表明, 人体内存在超过两万种蛋白质, 这些分子不仅是病

毒侵染和药物作用的直接靶点，也为疾病机制解析及创新药物研发提供了关键基础。在此背景下，“国际人类蛋白质组计划”(HPP)作为“人类基因组计划”的延续性工程，已成为全球范围内最具影响力的科研项目之一。该计划的核心技术路线依赖于质谱分析、冷冻电镜等实验方法，通过分离并解析蛋白质的三维结构以揭示其生物学功能。借助冷冻电镜等尖端技术，科研人员现已能够获得原子分辨率级别的蛋白质静态三维结构。然而，蛋白质在生理环境中始终处于动态变化状态，单一静态结构难以全面揭示其复杂的生物学功能。尤其对于蛋白质-药物结合、蛋白质-蛋白质相互作用 (protein-protein interaction, PPI) 等关键生物过程，传统生化实验方法往往存在观测局限性。在此背景下，MD 模拟技术展现出独特优势，该计算方法能够动态追踪蛋白质复合物体系中原子尺度的相互作用机制，为理解生命活动分子基础提供全新视角。

MD 模拟的核心原理基于经典牛顿力学，通过数值求解系统中各原子的运动方程，从而实现原子时空轨迹的精确预测。该方法通过迭代计算每个时间步长内原子的受力情况与位移变化，最终构建出系统的动态演化过程。其显著优势在于能够生成包含完整时空信息的轨迹文件，这些文件不仅记录特定时刻的构象状态，更重要的是揭示了蛋白质分子中各个氨基酸残基的连续动态行为。作为一种计算研究方法，MD 模拟能够在实验开展前提供关键的分子机制预测，这种“计算先行”策略可显著提高实验研究的针对性和成功率。正因如此，整合计算模拟与实验验证的研究范式已成为当前解析蛋白质构象变化和分子识别过程的方法之一。

从发展历程来看，MD 模拟技术的科学应用最早可追溯至 1957 年 Alder 和 Wainwright 对气体系统的开创性研究。到 20 世纪 80 年代，该技术成功拓展至蛋白质体系，首次从理论计算层面展现了生物大分子的动态特征，这一突破性进展促使学界重新审视蛋白质动态结构与功能的关系。对于现代生物医学研究而言，深入理解蛋白质的动态行为特征已成为揭示生命活动分子机制、指导理性药物设计的重要理论基础。2006 年，研究发现通过靶向从 A β 构象转变 MD 模拟中提取的 A β 中间结构的 C 末端 β -折叠区域，使用虚拟筛选发现了一种新型抑制剂，可以抑制 A β 纤维化^[14]。2009 年，也有研究报道使用分子动力学方法模拟了 ATP 结合过程的构象转变路径，揭示了 ATP 结合猿猴病毒大肿瘤

抗原的动态过程^[15]。如今，MD 模拟不仅可以模拟单个蛋白质，还可以模拟蛋白质与其他生物大分子间的相互作用，如蛋白质-蛋白质、蛋白质-核酸复合物等。

2009 年，DE Shaw 研究所的单一兵研究员^[16]通过对 Abl 激酶进行长时间的 MD 模拟，以原子级别的细节可视化了特征性的构象变化 (“DFG 翻转”)，这种构象变化在催化活性和非活性构象之间转换；该研究还提出了一个能量模型，预测 DFG 天冬氨酸的质子化控制着翻转。该研究团队通过实验证明伊马替尼与 Abl 激酶结合的动力学具有 pH 依赖性，这种依赖性在 DFG 天冬氨酸突变后消失。2011 年，单一兵研究团队^[17]首次提出 Swimming 的 MD 模拟策略，即对蛋白质-小分子体系进行相对较长的无指导 MD 模拟，Swimming 模拟不需要任何关于结合位点位置的先验知识，且不施加额外偏置力，尽可能模拟真实生理条件。其中配体 (癌症药物达沙替尼或激酶抑制剂 PPI) 被初始随机放置在包含一个已知与该配体结合的蛋白质 (Src 激酶) 的盒子内。在几次模拟中，配体正确识别了其目标结合位点，形成了与晶体学确定的结合结构几乎完全相同的复合物。这些模拟轨迹提供了整个结合过程的连续、原子级视角，揭示了持久且值得注意的中间构象，并阐明了水分子的作用。2022 年，Mingione 等^[18]通过核磁共振光谱化学位移微扰 (chemical shift perturbation) 的实验方法，证实了单一兵团队于 2011 年用 Swimming 的 MD 模拟寻找到的 Src 蛋白上的新结合位点 G-loop 位点为真实的结合位点。

MD 模拟在激酶类蛋白调节剂筛选和检测方面被广泛应用。2011 年，Cardoso 等^[19]使用分子建模工具，分析了由芳香醛和丙二腈设计的曲酸衍生物对酪氨酸酶活性的抑制作用。所有衍生物均显示出与酶活性位点的构象亲和力，以及与螯合铜离子的有利距离，这对酶功能至关重要。通过 MD 模拟揭示了衍生物和激酶形成的稳定复合物，部分衍生物显示了有利的结合自由能，表明其有成为激酶竞争性抑制剂的潜能。Hernández-Ochoa 等^[20]在 55 种化合物中筛选出 5 种具有最佳抑制效果的幽门螺杆菌葡萄糖-6-磷酸脱氢酶 (*H. pylori* glucose-6-phosphate dehydrogenase, HpG6PD) 抑制剂，并通过分子动力学的方法，模拟了化合物与 HpG6PD 的结合方式和结合部位，并合理推测这些化合物可能与 NADP⁺ 产生竞争性抑制作用。EGFR 激酶的小分子抑制剂

也有相关文献报道^[21], 研究人员通过 MD 模拟技术观察到小分子 Lapatinib 结合在 EGFR 激酶上与晶体结构一致的结合位点上(两次结合事件分别发生在模拟时间约 7 μ s 与 8 μ s 时), 且结合姿态也与晶体结构一致, 这表明了 MD 模拟确定蛋白质-小分子结合位点与结合姿态的正确性。此外, 该研究还发现 EGFR 激酶的 L834R 突变(也称为 L858R), 通过抑制 EGFR 的局部无序性促进了 EGFR 的二聚化, 这为 EGFR 激酶的 L834R 突变(也称为 L858R)的致癌表现与异常高的活性提供了机理解释。

2020 年, Yan 等^[22]对小分子 AZD9291(奥希替尼)与 EGFR 激酶(野生型和 T790M 突变型)的结合进行了若干个 10 μ s 的 Swimming 模拟, MD 模拟生成的蛋白质-小分子复合物结构模型通过实验确定的 EGFR/T790M 复合物晶体结构得到了证实, 这也证明了 MD 模拟确定蛋白质-小分子结合位点与结合姿态的正确性。同时, 该 Swimming 模拟解释了为什么 AZD9291 与 T790M 的亲和力更高: 模拟生成的结构模型表明, AZD9291 与 T790M 的结合姿态与其与 WT 的结合姿态不同, 并且 AZD9291 与 T790M 中的门控残基(MET 790)发生广泛的相互作用, 而不是与 WT 中的 THR 790 相互作用。这项工作可能有助于理性设计能够克服 AZD9291 耐药突变的药物, 并且更广泛地表明了使用 Swimming 模拟来阐明小分子结合姿态的潜力。

蛋白质-蛋白质相互作用(PPI)是普遍存在的生物分子过程, 对几乎所有细胞功能的各个方面都至关重要。识别能够调节特定疾病相关 PPI 的小分子是一种极具前景的药物发现策略。然而, 设计破坏 PPI 的药物是具有挑战性的, 因为在 PPI 界面上的许多潜在药物结合位点是“隐匿”的: 当没有配体占据时, 隐匿位点通常是平坦且无特征的, 因此在晶体结构中不易识别, 只有在配体结合时才会显示出典型小分子结合位点的几何和化学特性。已有文献报道 MD 在抑制蛋白质互作药物研究中的应用。2022 年, 单一兵研究团队^[23]进行了四种小分子抑制剂(SP4206 和三个 SP4206 类似物)与白细胞介素 2(IL2)结合的 Swimming 模拟, 未结合任何关于配体结合的先前结构信息。在多次结合事件中, 小分子稳定地嵌入到 IL2 的 PPI 界面, 形成的蛋白质-小分子结合位点和姿势几乎与现有的 IL2-SP4206 复合物晶体结构中观察到的完全相同。小分子的结合稳定了 IL2 的结合槽, 而当小分子未结合时, 该结合槽仅短暂且不完全地出现。此外, 自

由能微扰(free energy perturbation, FEP)计算成功区分了模拟中发现的 IL2-小分子结合的本征和非本征结合姿势, 表明结合模拟结合 FEP 计算可能为识别隐匿结合位点和确定小分子破坏 PPI 界面设计的结合姿势提供了一种有效工具。

MD 模拟对小分子诱导蛋白质变构导致的蛋白质-小分子结合的模拟能力也得到了文献的支持。2023 年, 来自 DE Shaw 研究所的 Ayaz 等^[24]进行了 Abl 激酶与抗癌药物伊马替尼结合的 Swimming 模拟。在模拟中伊马替尼首先选择性地与 Abl 激酶的自抑制构象结合, 随后伊马替尼引发了蛋白质的大构象变化, 形成与已发表晶体结构非常相似的结合复合物。

MD 前瞻性地模拟确定片段在先前未识别的口袋中的结合构象过程也有相关文献的报道。Greisman 等^[25]用长时间尺度 MD 模拟, 前瞻性地发现了两个系列的片段在一个重要且具有挑战性的药物靶点——蛋白酪氨酸磷酸酶 1b(PTP1b)的变构口袋中的结合构象结果。该模拟可逆地采样了片段结合和解离的过程, 发现一个结合口袋未见有片段结合, 另一个口袋则采用了非常罕见的构象。该研究团队通过实验的方法验证了上述结果的真实性。

MD 模拟广泛应用于 G 蛋白偶联受体(G protein-coupled receptor, GPCR)与各种药物分子结合过程的揭示。早在 2011 年, Dror 等^[26]利用无偏的 MD 模拟捕捉了多种药物与两种典型 GPCR 结合的全过程。尽管模拟过程中并未给予药物任何有关结合位点的“信息”, 药物仍需自发“发现”结合口袋, 但模拟所得的结合构象与晶体结构中的姿态高度一致, 精确至亚埃(subangstrom)级别。该研究团队在 2013 年的另一篇文章^[27]中通过 MD 模拟, 确定了几种 M2 毒蕈碱乙酰胆碱受体(M2 受体)变构调节剂的结合位点、结合构象及其与受体的特异性相互作用。M2 受体是 A 类 GPCR 的代表性成员。在模拟过程中, 这些调节剂能够自发地与受体结合。这些调节剂能在受体的胞外前庭区域(距离经典的“正构”配体结合位点约 15Å)与芳香残基簇形成阳离子- π 相互作用。在 GPCR 的偶联受体人类多巴胺 D2 受体(D2R)的研究中, Mansoor 等^[28]通过 MD 模拟, 在全原子分辨率下研究了内源性激动剂多巴胺、部分激动剂阿立哌唑(aripiprazole)以及拮抗剂舒必利(sulpiride)与 D2R 结合后的结构决定因素。模拟发现多巴胺结合复合物维持了在 TM3、

TM5 和 TM6 上的关键相互作用点,从而引发离子锁的部分打开,这是 TM6 向外移动所必需的。而舒必利的结合方式则破坏了切换开关 (toggle switch),从而全面改变了跨膜螺旋的动态行为,这种动态在另外两种配体中是保守的。

常规 MD 能够从原子尺度研究体系的热力学和动力学出发,为实验获得的宏观描述提供必要补充;但受限于飞秒级积分步长,难以模拟复杂体系的长期演化。而在生物体系中,受配体相互作用、蛋白质折叠和功能构象转变等动力学过程需要长时的 MD 模拟,且存在难以跳出局部稳定构象的可能,主要原因是不同低能构象之间存在较高的自由能垒,跨越这些能垒的构象转变过程需要花费毫秒或更长时间。

在过去几十年中,各种增强采样方法通过克服能量势垒扩展了模拟的时间尺度,实现了对体系热力学和动力学的研究,促进了对体系演化过程的深入探索。增强采样算法被广泛用于研究受体-配体相互作用、蛋白质折叠、功能构象转变等。典型的增强采样算法包括温度加速动力学 (Temperature Accelerated MD)、元动力学 (Metadynamics MD)、副本交换动力学 (Replica Exchange MD)、伞形采样 (Umbrella Sampling) 等。从是否需要特定反应坐标 (CV) 的角度可将其分为两类,分别适用于不同的系统。

近年来,基于人工智能的预测方法在蛋白质构象发现领域也有非常多的进展,常见方法是通过在深度学习网络中引入分子力场的先验知识以及在 MD 轨迹数据上训练等方式,使得模型具备从单一结构预测到蛋白质多构象预测的拓展能力,能够快速产生类似 MD 轨迹的多构象分布。

MD 模拟依赖于经典力场参数化模型与牛顿运动方程的数值求解,其计算资源消耗巨大且存在显著的时间尺度限制(通常局限于微秒量级)。尽管增强采样方法的引入在一定程度上缓解了构象空间采样不足的问题,但在实际应用中仍难以完整捕获蛋白质功能相关的慢速动态过程(如毫秒级结构重排或变构效应)。蛋白质的种类繁多、结构复杂,如果建立完备的蛋白质 MD 模拟库意味着天文级的计算量,强大的能够实现长时间 MD 模拟的科学计算机是必不可少的投入。

表 1 对预测蛋白质与功能相关新构象的计算方法进行了比较。

3 高性能计算技术在蛋白质动态结构解析中的应用

MD 模拟是“观测”蛋白质构象演化并演绎其生理功能的唯一有效的科学手段,短时(纳秒级)的蛋白质分子动力学仿真模拟难以看到真正有意义的生命现象,更不能指导新药设计现实需求;针对重要蛋白质生理现象的观察,都要在微秒乃至毫秒级的仿真模拟以上才有意义。针对蛋白质进行长时间尺度的 MD 模拟需要机器庞大的计算量和通信量,现有的常规计算很难高效率地完成,这也导致该领域的科技创新工作难以得到真正实施。

微秒级长时动态模拟需要天文量级的计算量、通信量(百万原子体系百微秒仿真计算量需要 1 021 次算术运算,1 021 次非线性运算,1 023 bit 邻域内及全局数据通信),这使得传统计算机需要超长时间(需要数百年乃至上万年)的计算方能完成或是根本无法满足。这需要借助极其强大的算力,可以说科学计算已成为当前人类认识研究蛋白质微观世界动态结构变化的核心工具,也是打开 21 世纪生命科学奥秘大门的钥匙。事实上,当前以高性能计算为核心的研究方法已逐渐被广泛采用,其推动诸多领域研究范式的演进——由过去传统的“实验发现+理论推演”向“科学计算+实验验证”的方式转变。蛋白质作为生命活动的主要承担者,对于蛋白质的研究,只有从分子乃至原子水平观察到酶催化反应、蛋白质构象变化及折叠等重要的生命活动现象,才能揭露更加深层的生理、病理的本质规律,进而指导医药研发及疾病诊治,最终服务于人类生命健康。

在过去的近二十年间,全球范围内仅有美国的安腾 (ANTON) 能够高效地完成该领域的任务。2003 年美国 Anton 1 问世,其分子动力学计算性能世界第一(约 1 μ s/d),可以提供对蛋白质、细胞膜、核酸和其他分子在原子尺度上的动态模拟。2013 年美国又推出了 Anton 2,其性能达到 10 μ s/d,在分子动力学领域其性能较传统科学计算具备 4 个数量级加速。相比于 Anton 2,2022 年推出的 Anton 3 的计算性能再次提升了一个数量级^[29],在 512 个节点并行处理下,能够对达 100 万个原子的大体系,每天进行超过 100 μ s 量级的 MD。模拟发展至今,分子动力学已经占美国全部科学计算资源的 30% 以上,ANTON 系列也在全世界垄断了 MD 模拟领域近二十年,并成功孵化了新药上市公司 Relay

表1 预测蛋白质与功能相关新构象的计算方法比较

方法	主要特点	优势	局限
常规MD (如Swimming MD)	·通过长时间的模拟增大系统跨越能垒访问其他构象的概率	·不需要先验知识 ·能够解释持久且值得关注的中间构象 ·未引入人工偏置, 结果更具说服力	·需要长时的MD模拟
元动力学 Metadynamics	·在已访问能量极小值区域添加势能以推动新的构象空间探索	·适用于探索复杂能量景观, 加速对高能构象的访问 ·无需特定反应坐标CV	·势能添加可能导致自由能景观过度平滑, 选择合适的集体变量挑战性大
温度加速动力学 Temperature Accelerated MD	·在高温下进行模拟以增加状态空间探索	·增加复杂系统状态空间的探索效率 ·无需特定反应坐标CV	·高温模拟可能导致非物理的构象变化
副本交换动力学 Replica Exchange MD	·运行多个不同参数设置的模拟副本并交换构象	·适合广泛温度范围的探索, 增强能量障碍区域的采样。基于一些广义的物理量(如温度和体系势能等)作为增强采样的参数 ·无需特定反应坐标CV	·对计算资源需求高, 设置交换参数挑战性大
伞形采样 Umbrella Sampling	·通过偏置势能增加稀有事件发生概率, 用于自由能计算	·能够准确计算过程的自由能差 ·需要特定的反应坐标CV	·需要精心选择偏置势能, 数据后处理复杂
自适应偏置动力学 Adaptive Biasing Force	·实时计算并抵消特定反应坐标上的平均力	·精确计算自由能障碍, 适用于复杂能量景观 ·需要特定的反应坐标CV	·实现复杂, 反应坐标选择挑战性大
Steered MD	·通过施加外部力引导系统沿特定路径移动	·直观模拟分子机械过程, 理解分子相互作用细节 ·需要特定的反应坐标CV	·依赖预定路径, 可能导致非自然动态行为
AI预测	·在深度学习网络中引入分子力场或在MD轨迹数据上训练使得模型能够学习构象分布	·速度快 ·能够从序列直接预测多构象分布	·准确度和适用广度不如MD模拟 ·难以处理复杂的蛋白质动态变化 ·缺少实际检验

Therapeutics。依托 ANTON 的强大性能, 借助高性能计算的计算生物学成本更低、速度更快, 能够大幅度缩短新药研发的周期及成本, 让研发效率得到数倍的提升, 进而推动生物医药技术的快速发展。但遗憾的是, ANTON 仅面向美国科研人员使用, 对全球其他国家科学家严格禁用, 长久以来中国的科学家没有机会使用这类强大的工具。

目前我国已经成功研制出 100% 自主原创的高性能科学计算领域微处理器——MaPU 芯片, 并基于此款芯片打造了 MHPC512 科学计算机, 其科学计算性能达到了世界一流的水平, 在分子动力学领域更是媲美 Anton 2, 打破了美国在该领域长期以来的绝对垄断地位。首台样机位于上海科技大学, 是世界第一台完全可编程的 3D 科学计算服务器, 在该服务器上实测 100 万原子体系的 MD 模拟每天可达 5~10 μ s, 达到了国际领先水平, 较国内传统

科学计算系统快 3~4 个数量级。目前该机型计算效率对标 Anton 2, 远远高于世界上其他主流超级计算机的计算效率。MHPC 512 科学计算机由 512 颗基于 MaPU 架构的 MDP 1.0 芯片构成, 并且这些芯片通过自研的 Superlink 高速通讯端口相互连接, 形成了一个具有六面体通信能力的处理单元网络, 并被组织成 8×8×8 的立方体架构, 通过 3D 并行物理子空间遍历计算和通信机制, 使系统跨越处理器形成逻辑上无子空间边界的整体三维物理空间。在 MD 模拟中, 这种物理架构的设计在以下多个方面发挥了关键作用。

(1) 并行计算效率提升。MaPU 架构的灵活性和高度可编程性使得 MDP 芯片能够更高效地执行 MD 模拟中的计算任务, 如力场计算和位置速度更新。立方体架构的并行性能保证了系统在处理这些任务时能够同时进行多个计算, 显著提升了整体的

并行计算效率。

(2) 通信机制的优越性。通过 Superlink 高速通讯端口的应用,MDP 芯片之间的通信得以迅速而高效地进行。这在 MD 模拟中特别重要,因为模拟需要在不同处理单元之间频繁地交换信息。六面体通信能力的设计使得系统具备更为灵活的通信方式,有助于优化数据交换和提高整体通信效率。

(3) 三维物理空间的整合。立方体架构和 3D 并行物理子空间遍历计算机制的应用,使得系统能更好地模拟三维物理空间中的对象相互作用和运动问题。这对于 MD 模拟来说至关重要,因为系统需要在三维空间内准确模拟粒子的运动和相互作用。

综合而言,MHPC 512 为 MD 模拟提供了高效的计算平台。通过在整体三维物理空间中无缝整合处理单元,系统在三维仿真领域取得了超高的计算效率,为科学计算和仿真领域带来了卓越的性能,为蛋白质动态结构库的构建提供了坚实的算力基础。目前,MHPC512 已投入商用,主要用户包括某国家重大任务、上海科技大学、复旦大学、上海交通大学、北京大学、中国科学院大连化学物理研究所等。MHPC512 的出世,使得国内具备了独立自主开展大规模蛋白质长时动态仿真模拟的能力。

目前位于我国湖北孝感的长江 3D 科学计算中心,依托 MHPC512 整机和自主 3D 科学计算软件平台,可全面支持分子动力学、量子化学计算、计算流体动力学仿真等各领域的商业化应用,是目前全球最大规模的通用型 3D 科学计算算力集群。该中心面向学术界和工业界提供病理研发平台、药物研制平台、材料仿真研发平台、计算流体仿真平台的计算服务。中心承接完成的某部委委托的新冠病毒复盘研究国家级重大课题,是中国迄今为止最大的算力课题。长江 3D 科学计算中心的落地,为我国规模化的蛋白质动态结构模拟提供了有效的算力基础。

在模拟蛋白质动态结构领域,强大的计算资源需要与顶级的研究平台紧密结合。作为“十一五”国家重大科技基础设施,国家蛋白质科学研究(上海)设施(简称“蛋白质设施”)是全球生命科学领域首个综合性的大科学装置,集成性地建设了具有跨尺度和时间分辨能力的蛋白质科学研究体系,是国际一流的蛋白质科学研究中心、蛋白质研究技术与方法创新中心、蛋白质药靶的转化研究中心和蛋白质研究高层次科技人才培养中心,是我国蛋白质科学研究和技术创新基地。

在蛋白质研究领域,“结构决定功能”是主流思想。遗憾的是,蛋白质结构研究仍以静态为主,而蛋白质在人体内是不断运动的,最真实的蛋白质还未呈现在我们的眼前。世界领先的蛋白质结构预测模型——AlphaFold 也仅能够实现以 90% 以上的精度预测蛋白质静态结构。可以说,蛋白质结构研究正由“静态”走向“动态”。

近日,蛋白质设施依托 MHPC512 整机建立“高性能蛋白质动态计算中心”,计划开展蛋白质动态结构研究,结合蛋白质设施的实验平台进行验证,“干湿结合”为蛋白质关键领域的重要科学问题提供解决方案。依托“高性能蛋白质动态计算中心”,蛋白质设施及多家机构联合发起“人源蛋白动态组学”科学计划,旨在借助 MHPC 超算,完成人体已知的两万余种蛋白质的动态结构库及其与重要小分子的动态相互作用图谱;“蛋白动态结构库”的构建将积累起海量动态蛋白数据,训练出世界上独一无二的、超越 AlphaFold 的蛋白质动态结构预测模型。上述高性能计算技术与蛋白质结构研究体系等底层创新,将有助于我国更好地应对全球挑战,保障国家安全稳定。

4 结论与展望

目前,以 AlphaFold 为代表的人工智能对于蛋白质的静态结构解析已经取得了非常好的效果,但动态结构的预测难以突破的重点在于缺乏足够的蛋白质动态结构数据。未来以分子动力学数据为基础,结合 AI 的蛋白质动态结构预测将成为该领域的重要趋势,这更加凸显了蛋白质动态结构数据库的重要性。未来的蛋白质研究将依托“AI 预测提供初始构象→科学计算 MD 探索动态景观→实验数据验证关键靶点”的干湿闭环,推动药物研发从“静态靶点”向“动态构象”范式的根本转变。随着国产科学计算的规模化应用,中国有望在动态结构驱动的创新药物研发中占据战略高地。

[参 考 文 献]

- [1] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583-9
- [2] Monteiro da Silva G, Cui JY, Dalgarno DC, et al. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nat Commun*, 2024, 15: 2464
- [3] Humphreys IR, Pei J, Baek M, et al. Computed structures of core eukaryotic protein complexes. *Science*, 2021, 374:

- eabm4805
- [4] Krishna R, Wang J, Ahern W, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 2024, 384: eadl2528
- [5] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024, 630: 493-500
- [6] Siebenmorgen T, Menezes F, Benassou S, et al. MISATO: machine learning dataset of protein-ligand complexes for structure-based drug discovery. *Nat Comput Sci*, 2024, 4: 367-78
- [7] Jing B, Berger B, Jaakkola T, et al. AlphaFold meets flow matching for generating protein ensembles. *arXiv*, 2024, <https://doi.org/10.48550/arXiv.2402.04845>
- [8] Wu F, Jin S, Jiang Y, et al. Pre-training of equivariant graph matching networks with conformation flexibility for drug binding. *Adv Sci (Weinh)*, 2022, 9: e2203796
- [9] Cao D, Chen G, Jiang J, et al. Generic protein-ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nat Mach Intell*, 2024, 6: 688-700
- [10] Wang Y, Wang L, Shen Y, et al. Protein conformation generation via force-guided SE(3) diffusion models. *arXiv*, 2024, <https://doi.org/10.48550/arXiv.2403.14088>
- [11] Kaihui C, Liu C, Su Q, et al. AlphaFolding: 4D diffusion for dynamic protein structure prediction with reference and motion guidance. *arXiv*, 2024, <https://doi.org/10.48550/arXiv.2408.12419>
- [12] Outeiral C, Nissley DA, Deane CM. Current structure predictors are not learning the physics of protein folding. *Bioinformatics*, 2022, 38: 1881-7
- [13] Fraser JS, Murcko MA. Structure is beauty, but not always truth. *Cell*, 2024, 187: 517-20
- [14] Liu D, Xu Y, Feng Y, et al. Inhibitor discovery targeting the intermediate structure of β -amyloid peptide on the conformational transition pathway: implications in the aggregation mechanism of β -amyloid peptide. *Biochemistry*, 2006, 45: 10963-72
- [15] Shi Y, Liu H, Gai D, et al. A computational analysis of ATP binding of SV40 large tumor antigen helicase motor. *PLoS Comput Biol*, 2009, 5: e1000514
- [16] Shan Y, Seeliger MA, Eastwood MP, et al. A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proc Natl Acad Sci U S A*, 2009, 106: 139-44
- [17] Shan Y, Kim ET, Eastwood MP, et al. How does a drug molecule find its target binding site? *J Am Chem Soc*, 2011, 133: 9181-3
- [18] Mingione VR, Foda ZH, Paung Y, et al. Validation of an allosteric binding site of Src kinase identified by unbiased ligand binding simulations. *J Mol Biol*, 2022, 434: 167628
- [19] Cardoso R, Valente R, Souza da Costa CH, et al. Analysis of kojic acid derivatives as competitive inhibitors of tyrosinase: a molecular modeling approach. *Molecules*, 2021, 26: 2875
- [20] Hernández-Ochoa B, Navarrete-Vázquez G, Aguayo-Ortiz R, et al. Identification and *in silico* characterization of novel *Helicobacter pylori* glucose-6-phosphate dehydrogenase inhibitors. *Molecules*, 2021, 26: 4955
- [21] Shan Y, Eastwood MP, Zhang X, et al. Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell*, 2012, 149: 860-70
- [22] Yan XE, Ayaz P, Zhu SJ, et al. Structural basis of AZD9291 selectivity for EGFR T790M. *J Med Chem*, 2020, 63: 8502-11
- [23] Shan Y, Mysore VP, Leffler AE, et al. How does a small molecule bind at a cryptic binding site? *PLoS Comput Biol*, 2022, 18: e1009817
- [24] Ayaz P, Lyczek A, Paung Y, et al. Structural mechanism of a drug-binding process involving a large conformational change of the protein target. *Nat Commun*, 2023, 14: 1885
- [25] Greisman JB, Willmore L, Yeh CY, et al. Discovery and validation of the binding poses of allosteric fragment hits to protein tyrosine phosphatase 1b: from molecular dynamics simulations to X-ray crystallography. *J Chem Inf Model*, 2023, 63: 2644-50
- [26] Dror RO, Pan AC, Arlow DH, et al. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc Natl Acad Sci U S A*, 2011, 108: 13118-23
- [27] Dror RO, Green HF, Valant C, et al. Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. *Nature*, 2013, 503: 295-9
- [28] Mansoor S, Morra G. Insights into ligand-specific activation dynamics of dopamine D2 receptor explored by MD simulations. *J Chem Inf Model*, 2025, 65: 5050-61
- [29] Shaw DE, Adams PJ, Azaria A, et al. Anton 3: twenty microseconds of molecular dynamics simulation before lunch[C]. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, St. Louis, MO, USA, 2021: 1-11