



人工智能加速分子动力学模拟

钱润彤^{1,2}, 杨世悦^{1,2}, 宋子林^{1,2}, 黄晶^{1,2*}

1. 西湖实验室人工智能药物设计核心实验室, 杭州 310024

2. 西湖大学生命科学学院, 杭州 310024

*通讯作者, E-mail: huangjing@westlake.edu.cn

收稿日期: 2025-02-28; 接受日期: 2025-04-07; 网络版发表日期: 2025-05-26

国家自然科学基金(编号: 32171247)、浙江省“尖兵”“领雁”研发攻关计划(编号: 2023C03109, 2024SSYS0036)和博士后科学基金(编号: 2023M743138)资助项目

摘要 分子动力学模拟是解析复杂分子体系结构、热力学和动力学性质的重要工具, 但其应用受限于计算效率, 模拟和真实过程的时间尺度往往存在落差. 随着深度学习技术的飞速发展, 利用人工智能加速分子动力学模拟成为可能. 本文探讨了人工智能在不同层面上实现加速的潜力: 时域加速突破步长限制, 序列加速捕捉构象跃迁, 空间采样优化构象遍历. 我们也讨论了现有方法在物理一致性与跨尺度适应性方面的诸多困难与挑战. 未来, 多尺度方法与物理约束框架的有机融合, 有望协同提升精度、效率与可解释性, 推动分子模拟智能化发展, 并进一步拓展其在机制解析与筛选设计上的应用潜力.

关键词 分子动力学模拟, 人工智能, 构象系综, 构象动力学, 深度学习, 人工智能驱动的科学发现

1 引言

近年来, 随着科技的不断进步, 对复杂分子系统的研究取得了革命性的突破^[1,2]. 核磁共振^[3]、X射线晶体衍射^[4]和冷冻电子显微镜^[5]等先进实验技术, 使大分子静态结构的精确测定成为现实. 而在生物大分子动态构象的表征领域, 现有实验技术体系仍存在显著局限性. 以核磁共振为例, 该方法虽能有效解析蛋白质折叠路径、热力学稳定性及局部构象变化等特征, 但对大尺度构象运动的实时动态捕捉仍面临技术瓶颈^[6]. 此外, 复杂耗时的样品制备过程也限制了实验方法在动力学研究中的广泛应用.

与此同时, 从生物大分子序列出发预测其高精度

结构的能力也有了显著提高. AlphaFold^[7,8]等端到端的机器学习(machine learning, ML)方法, 可以高效精准地预测全部蛋白质的单体结构, 并给出生物大分子及其相互作用分子的复合物结构. 然而, 目前的结构预测方法局限于生成单个或少数几个有代表性的静态结构, 无法准确捕捉分子的动态功能构象.

生物大分子的构象动力学是揭示其功能机制的关键所在. 蛋白质在执行催化、信号传导和物质转运等重要生物学功能时, 往往需要在多个构象之间转换, 而这些构象变化也受到如配体结合等环境因素的调控^[9]. 以激酶这类重要的信号通路枢纽蛋白为例, 它们通过活化环的大尺度构象变化介导“开放-闭合”状态转换, 协同催化核心与其他调控元件的特异性空间重

引用格式: Qian R, Yang S, Song Z, Huang J. Accelerating molecular dynamics simulations with artificial intelligence. *Sci Sin Chim*, 2025, 55: 1688–1703, doi: 10.1360/SSC-2025-0055

组, 实现ATP结合与磷酸基团转移的精确时空调控^[10]. 激酶调控异常与肿瘤的发生发展密切相关, 而开发高选择性激酶抑制剂的关键挑战在于: 激酶家族成员间高度保守的核心结构域和复杂多样的调控机制导致基于ATP结合口袋静态结构的药物设计难以实现高选择性识别. 已有研究表明, 激酶靶向药物的高选择性很可能源于诱导契合的动力学过程, 而非静态结构特征^[11]. 这些发现凸显了开发动力学研究方法的迫切需求, 特别是能在原子分辨率下追踪构象转换路径与时间依赖关系的技术手段.

分子动力学(molecular dynamics, MD)模拟等基于

物理的计算方法, 为研究复杂体系的动力学和动态行为提供了强大的工具. MD模拟基于统计热力学理论, 通过数值求解牛顿运动方程模拟分子体系的时间演化过程.

根据遍历性假设 $\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T A(x(t)) dt = \int_{\Omega} A(x) \rho(x) dx$, 一个各态遍历系统的性质 $A(x)$ 的时间平均等于相空间 Ω 上的平均, 充分的MD采样能够遍历整个相空间. 这使得MD成为解析和预测分子体系热力学和动力学性质的宝贵工具(图1)^[12-16].

MD模拟的精度主要是由其使用的势能计算模型决定的, 而势能计算也是模拟效率的主要瓶颈. MD使

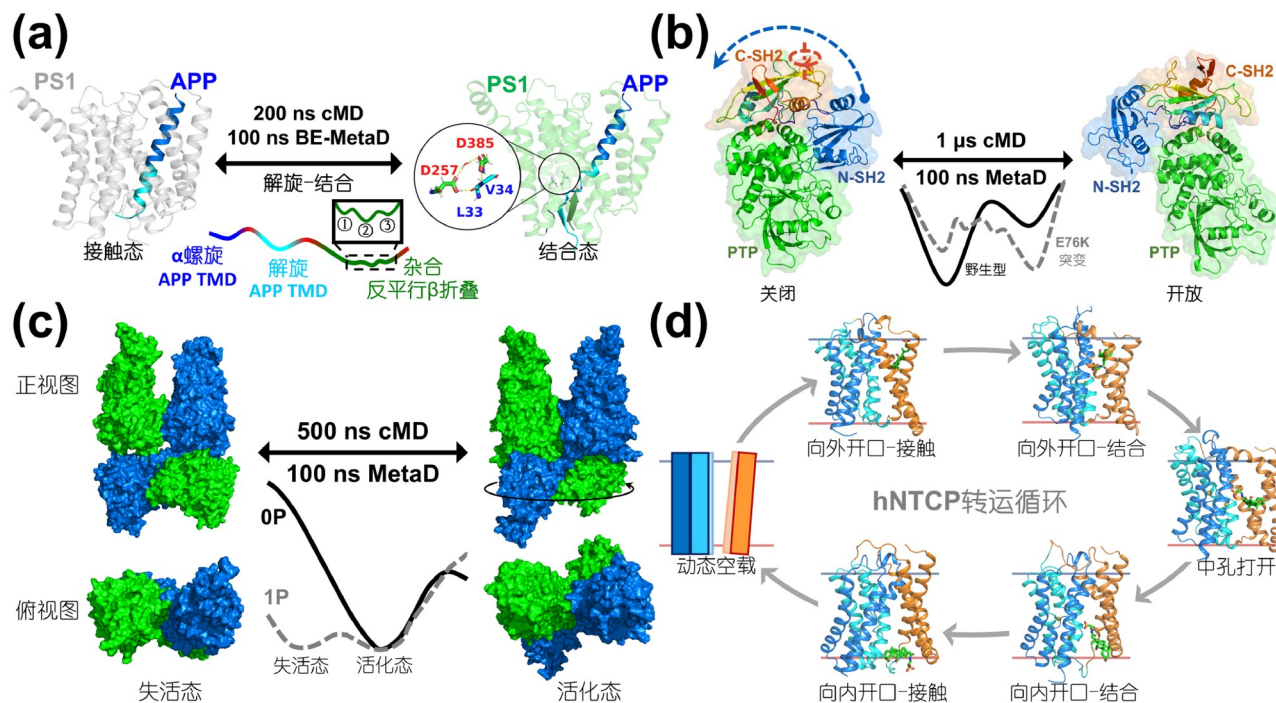


图 1 (网络版彩图)基于全原子常规MD (conventional MD, cMD)模拟和元动力学(metadynamics, MetaD)增强采样的联合策略揭示的4种复杂蛋白体系的构象变化机制. 蛋白质均采用CHARMM36m力场^[12]描述. (a) 人源 γ -分泌酶识别淀粉样前体蛋白(APP)底物的构象动力学变化^[13]. 模拟结果表明, 人源 γ -分泌酶可诱导APP跨膜区解旋, 并形成杂合反平行 β 折叠. 自由能面解析揭示了催化裂解可能生成的三类不同产物: ①失活; ②A β 49; ③A β 48. (b) E76K单一突变调控蛋白质酪氨酸磷酸酶SHP2的开闭构象动力学^[14]. (c) 溶质载体蛋白SLC12A6在去磷酸化和单磷酸化状态下不同的活化-失活态构象动力学变化^[15]. (d) 5 μ s全原子常规MD模拟和900 ns二维元动力学模拟揭示了人源钠离子-牛磺胆酸盐共转运蛋白(Na^+ -taurocholate co-transporting polypeptide, NTCp)中钠离子驱动的底物耦合转运机制^[16].

Figure 1 (Color online) Conformational transformation mechanisms of four complex protein systems revealed by a combined strategy integrating all-atom conventional molecular dynamics (cMD) simulations and enhanced sampling via metadynamics (MetaD). All proteins are described using the CHARMM36m force field [12]. (a) Conformational dynamics of human γ -secretase in recognizing amyloid precursor protein (APP) substrate [13]. The simulations indicate that human γ -secretase can induce the unwinding of the APP transmembrane region, forming a hybrid antiparallel β -sheet. Analysis on the free energy surface reveals three possible catalytic cleavage products: ① inactive, ② A β 49, and ③ A β 48. (b) The single E76K mutation modulates the open-closed conformational dynamics for SHP2 protein tyrosine phosphatase [14]. (c) Conformational dynamics of solute carrier protein SLC12A6 in dephosphorylated and monophosphorylated states, highlighting its transformation between active and inactive conformations [15]. (d) The 5 μ s all-atom cMD simulations and 900 ns two-dimensional MetaD simulations reveal the sodium-driven substrate-coupled transport mechanism in the human Na-taurocholate co-transporting polypeptide (NTCP) [16].

用的势能面通常可以分为基于经典核运动与解析势能函数的分子力场方法和基于第一性原理的从头算方法。基于第一性原理的分子模拟(ab initio MD, AIMD)在模拟的每一步依赖电子结构计算获取原子核所受的力, 对计算资源消耗巨大。近年来, 机器学习方法被应用于直接拟合量子力学计算的势能面^[17~20], 对AIMD模拟实现了大幅加速, 已有很多优秀的综述深入探讨了机器学习势或机器学习力场领域的发展^[21~23]。在本文中, 我们将重点探讨机器学习在加速经典MD模拟中的应用。

基于分子力场的MD模拟可及的时空尺度多年来不断提高, 对百万原子体系进行微秒级模拟现在已成为常规。基于GPU等高性能并行硬件的MD算法实现^[24~28]、更高效的增强采样方法^[29~32]和更精准的分子力场^[12,33~35]都极大地促进了分子动力学研究工具的发展。专为MD模拟设计的Anton等专用计算架构进一步提高了模拟效率, 最新一代Anton 3超级计算机已成功将百万原子体系的模拟速度提升至每天数百微秒^[36,37]。尽管如此, 许多涉及大分子构象变化的生理过程发生在更长的时间尺度, 如激酶活化-非活化态转换的典型时间尺度为毫秒级^[38]; 利用转运蛋白构象变化实现的跨膜物质转运则通常发生在1~10 ms的尺度^[39]。在现有的高性能计算机上执行如此长时间的模拟往往需要数月甚至数年, 而若考虑此类过程涉及的多路径竞争与过渡态捕捉, 则需要许多模拟轨迹才能获取可靠统计结果, 更凸显出计算效率的限制。因此, 开发更高效的方法来加速MD模拟对于推动我们了解复杂的生物和化学过程至关重要。

近年来, 机器学习技术的兴起为加速MD模拟提供了新思路。机器学习模型擅长拟合复杂高维函数和建模复杂高维数据分布, 在预测分子性质和生成给定性质的分子方面显示出巨大的潜力。已经有多篇优秀的综述广泛地回顾了机器学习在分子模拟中的应用, 包括势能面拟合、模型粗粒化、自由能预测、集合变量(collective variable, CV)表征、轨迹分析等方面的突破^[21,40,41]。本文将不再对这些方向做过多探讨, 而是重点关注基于经典MD算法本身的加速模型。

2 分子动力学模拟的基本原理

分子动力学模拟应用有限差分方法来求解经典多

体系统的运动方程。在这种方法中, 系统中 N 个原子或原子团簇之间的相互作用可以用经典势函数 $U: \mathbf{r} \in \mathbb{R}^{N \times 3} \rightarrow \mathbb{R}$ 来描述。对粒子 i , 其动量 $\mathbf{p}_i$ 及构型 $\mathbf{q}_i$ 随时间的演化服从哈密顿运动方程:

$$\begin{aligned}\dot{\mathbf{q}}_i &= \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} \\ \dot{\mathbf{p}}_i &= -\frac{\partial \mathcal{H}}{\partial \mathbf{q}_i}\end{aligned}\quad (1)$$

其中, $\mathcal{H}$ 是系统的哈密顿量, 等价于系统总能量, 即动能和势能之和。势能函数的选择对于模拟的准确性至关重要, 而运动方程的演化需要不断计算势能函数对坐标的一阶导数。

哈密顿运动方程主要适用于孤立系统, 而实际模拟中常需通过扩展方法引入非平衡条件或耗散效应, 以满足实际应用需求。例如, 在Langevin动力学中, 运动方程会附加阻尼项和随机噪声项以模拟环境效应, 式(1)进而扩展为

$$\dot{\mathbf{p}}_i = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}_i} - \gamma \mathbf{p}_i + \delta \dot{\mathbf{w}} \quad (2)$$

其中 γ 是摩擦系数, $\dot{\mathbf{w}}$ 是维纳过程(布朗运动), 其强度由系数 δ 控制。此外, 扩展系综方法(如Nosé-Hoover热浴、Parrinello-Rahman方法等), 均通过修改哈密顿量或引入额外自由度来实现特定系综约束。因此, 积分器的选择需综合考虑模拟体系的物理特性、计算效率与数值稳定性等要素。

模拟条件的设定直接影响MD模拟的结果和适用性。MD模拟通常在不同的统计力学系综下进行, 以再现实验条件, 反映系统与外界环境的相互作用。常见的系综包括微正则系综(NVE)、正则系综(NVT)和等压等温系综(NPT)等。在NVE系综中, 体系的分子数、体积与能量保持恒定, 适用于研究绝热过程, 常采用严格服从公式(1)的Verlet类积分器; NVT系综允许能量的热涨落, 通过Langevin动力学或Nosé-Hoover热浴保持温度恒定, 描述与恒温热浴接触的封闭系统; NPT系综则允许体积变化, 采用Andersen、Parrinello-Rahman等方法维持压力稳定, 适合研究相变等对体系压强敏感的现象。选择合适的系综和物理参数, 有助于模拟结果更准确地反映待研究的实际物理化学过程。

在给定的势能面上, MD模拟能够在多层次上提供分子体系的丰富信息, 涵盖空间尺度、时间尺度与统

计信息等不同层级. 这些层面的信息不仅各自独立存在, 更在模拟过程中相互关联, 共同揭示体系的动力学演化机制与热力学信息.

(1) 空间尺度. 模拟体系的粒子坐标是MD模拟提供的最本质微观信息. 根据力场的颗粒化精度, MD模拟提供从原子到纳米尺度的空间细节, 揭示分子体系的结构特征. 对于全原子力场而言, MD模拟中的空间信息包含了多个尺度: 分子内信息, 如键长、键角、二面角等; 分子间信息, 如水的氢键网络、苯分子晶体内苯环的堆叠; 生物大分子的构象, 如蛋白质的二级和三级结构, 反映分子在空间中的排列方式; 多体相互作用下的分子聚集行为, 如胶束、囊泡的形成. 这些空间信息对于理解分子体系在不同尺度下的组织结构以及结构背后的驱动力至关重要.

(2) 时间尺度. 模拟可以捕捉在亚皮秒级到微秒甚至更长时间尺度上的分子系统的动力学行为. 选择合适的时间窗口对于观察特定的动力学过程至关重要, 例如, 化学键振动通常需要飞秒级的时间窗口, 而蛋白质折叠则可能需要微秒至毫秒级的时间窗口. 为了保证数值积分的稳定性, 经典MD模拟的时间步长需小于体系中最快运动的周期, 通常设置为2 fs. 模拟轨迹蕴含的时间信息因此天然是多尺度的. 正则系综下, MD轨迹随模拟时间的演化, 可以观察到体系随热涨落在稳态、过渡态、高维鞍点、非稳态之间的变化路径. 此外, 通过分析分子事件发生的时序关系和时间间隔, 可以揭示动态过程中的因果关系. 时间相关函数, 包括自相关和交叉相关函数, 可用于描述系统在不同时间点之间的关联性, 直接与谱学测量相联系, 有助于通过体系对扰动的响应揭示分子间相互作用信息. 弛豫时间作为系统从非平衡态恢复到平衡态所需的平均时间, 反映了系统对外界扰动的响应速度.

(3) 统计信息. 将MD模拟视为一种采样算法, 其最主要的目标则是产生符合玻尔兹曼分布的体系状态集合. 通过对体系中的部分自由度进行积分, 可以得到剩余维度上的自由能变化, 即所谓的potential of mean force (PMF). 许多增强采样算法如伞形采样、元动力学, 基于这一统计信息进行加权, 调整采样分布. 对完整系综的统计分析可以推导分子体系的宏观性质. 例如, 通过采样统计能够获得体系的热力学平均性质, 如温度、压力、内能等. 均方位移衡量粒子随时间的位移平方平均值, 可用于计算扩散系数. 径向分布函

数统计每个粒子在特定距离内邻居粒子的数量, 用于了解局部结构和空间分布特性, 并可以通过Kirkwood-Buff积分推导体系全部热力学性质. 系综平均的自相关函数, 可用于推导体系的输运性质. 这些统计量弥补了微观模拟结果与宏观观测指标之间的鸿沟, 从而为实验结果提供了有价值的微观预测和见解.

在MD模拟中, 这三个层面的信息相互交织, 共同描绘出体系的完整图景. 分子体系的空间构型会随时间演变, 得到动力学轨迹, 最终完成对相空间的采样, 通过系综平均对宏观性质进行统计预估. 模拟的具体研究目标往往决定了需要关注系统哪些层面的信息. 例如, 理性药物设计的核心问题在于预测蛋白质与小分子之间的相互作用以及探究小分子结合如何调控蛋白质构象. 通过对靶标蛋白构象动力学、小分子与蛋白之间结合模式与空间结构的变化进行模拟, 能够辅助发现重要的相互作用位点, 从而进一步挑选和优化分子, 这需要空间层面的信息. 时间层面的信息, 如蛋白与小分子结合的动力学信息, 可以通过中间过渡态的采样, 预估结合发生的能垒从而预估结合的速率, 对于了解药物的作用机制和半衰期至关重要. 这些不同层次的信息相互关联, 共同揭示了分子相互作用的细节. 根据不同阶段的不同研究需求, 研究人员可以针对性地从MD模拟中提取所需信息用以研究(图2)^[42-44].

因此, 利用机器学习加速MD模拟可以根据具体研究目标, 针对特定层次的信息进行加速. 根据模拟的目标任务, 定向提取与目标性质高度关联的关键信息. MD模拟区别于其他采样方法的关键在于其时间信息. 根据时间信息的保留程度不同, 我们将ML在MD加速中的应用分为三个主要方向: 时域加速、序列加速和空间采样加速, 并分别在下文展开讨论.

蛋白质等生物大分子的动力学演化是典型的多尺度复杂过程. 构象动力学主要关注相对较慢的蛋白质功能性构象演化过程, 主要体现为蛋白亚稳态间的跃迁. 基于MD轨迹所提供的时间序列信息, 时域加速和序列加速通过选择不同的时间粗粒化尺度, 在不同的层面上描述构象动力学. 这种多尺度动力学建模不仅有助于解析大分子的运动特征, 还能为优化空间采样方法提供基准. 另一方面, 自由能面的解析能够提示热力学平衡下的构象分布规律——低自由能态通常占据主导地位, 而高能稀有态则需通过特殊采样策略捕

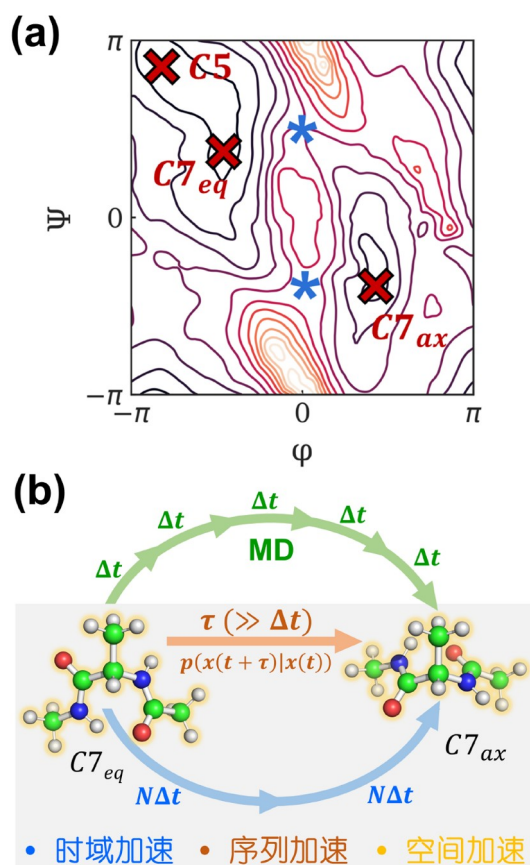


图 2 (网络版彩图)经典MD模拟系统及加速方案(以丙氨酸二肽为例). (a) 丙氨酸二肽主链构象空间(ϕ 、 ψ 二面角)的气相自由能景观. 我们以CHARMM36m^[12]力场描述丙氨酸二肽, 在真空环境中以朗之万积分器(温度350 K, 耗散系数 30 ps^{-1})对其主链上 ϕ - ψ 二面角以 $\pi/15$ 为采样窗口进行伞形采样, 并利用FastMBAR^[42]重权算法获得其自由能景观. 红色标记了三个主要稳定构象: $C7_{ax}$ 、 $C7_{eq}$ 、 $C5$ ^[43], 蓝色标记了C7异构体间两个可能的过渡态^[44]. (b) MD模拟的三种加速策略

Figure 2 (Color online) Classical MD simulation system and various acceleration schemes, taken alanine dipeptide as a model system. (a) Free energy landscape of gas-phase alanine dipeptide backbone conformational space (ϕ , ψ are dihedral angles). The system was described using the CHARMM36m force field^[12], and umbrella sampling was performed on the ϕ - ψ dihedral angles with a sampling window of $\pi/15$ in vacuum. Langevin dynamics integrator was employed at a temperature of 350 K with a damping coefficient of 30 ps^{-1} . The free energy landscape was reconstructed using the FastMBAR reweighting algorithm^[42]. Red indicates the three main stable conformations: $C7_{ax}$, $C7_{eq}$, $C5$ ^[43]; while blue indicates two possible transition states between the C7 conformers^[44]. (b) Three ML acceleration strategies for MD simulation.

获. 当研究目标主要关注自由能面所包含的机理信息时, 我们可以通过空间采样方法解耦时间维度, 直接在服从玻尔兹曼分布的构象集合中进行采样, 从而实

现平衡热力学构象系综的高效构建.

值得注意的是, 任何形式的信息剥离都可能以丢失其他信息为代价. 例如, 用机器学习模型替代积分器进行单步或多步推演可能破坏辛结构, 从而导致系统能量漂移, 并影响模拟轨迹的稳定性和热力学性质的计算准确性. 因此, 在研究和应用基于ML的MD加速方法时, 必须慎重考虑加速策略以保留研究目标所需的信息不被破坏(表1).

3 时域洞察: 高保真地提升MD推演效率

MD模拟的时间步长受限于数值积分的稳定性, 使得高保真的动力学描述需要耗费巨大的计算资源. 因此, 如何在严格保留显式物理时间信息的前提下提高计算效率, 成为高保真MD加速方法的核心问题. 时域方法旨在从历史动力学轨迹中学习并预测未来状态, 通过机器学习直接建模动力学方程的时间演化规律, 从而突破经典MD模拟中以微小时间步长为代表的限速环节与计算精度间的强耦合限制, 最终在保持物理规律约束的前提下显著减少达到特定物理观测时间所需的迭代或演化次数, 实现采样加速.

这类方法的主要挑战是如何在长时空尺度上保持稳定、避免累积误差过大. 潜在的发展方向包括引入物理信息正则化项, 开发基于注意力的自适应时间粗粒化策略, 以及结合基于强化学习的动态步长优化算法. 总体而言, 时域方法有可能在模拟效率和瞬时分子动力学机制解析方面取得突破, 是非常值得期待的发展方向.

一个最直接的思路是借鉴物理信息神经网络(physics-informed NN, PINN)的理念, 在求解微分方程时, 并不是优化神经网络预测的解与真实解间的误差, 而是直接将预测解代入微分方程作为损失函数. MD积分方法通常是以数值求解牛顿运动方程或郎之万方程的方式展开, 若使用神经网络进行拟合, 便涉及使用自动微分计算二阶微分, 成为神经网络方法效率的瓶颈^[45]. 为避免直接求解二阶微分方程, Marios等^[46]提出了将哈密顿方程作为损失函数的哈密顿神经网络, 结合简单的双层感知机, 能够得到连续、平滑的解析解, 并严格守恒正确的系统哈密顿量. 然而, 该研究在方法对比时仅与一阶欧拉方法进行精度对比, 且积分时长较短, 处理较大体系与长时间模拟尚未得到验证.

表 1 三类机器学习加速MD方法的区别

Table 1 Comparison among three ML-accelerated MD strategies

加速方法类别	时间信息	约束/假设	构象间依赖关系
时域加速	$N\Delta t$ 严格保留物理时间	系综物理约束	有
序列加速	统计时间尺度 τ 不严格对应物理时间	马尔可夫假设	有
空间采样加速	无	玻尔兹曼假设	无

使用物理信息神经网络直接求解微分方程在其他领域已经得到了广泛的应用,但在MD模拟领域还没有获得太多的发展.

另一个思路是以预测解与真实解之间的误差作为损失函数,通常使用当前构象的坐标与动量为输入,预测下一步或数步的原子坐标、动量或受力等物理量.图神经网络(graph neural network, GNN),尤其是消息传递网络(message-passing neural network, MPNN)已在预测分子性质方面展现出优异的性能,因此一些初步研究开始探索GNN在加速MD中的潜力. Zheng等^[47]基于MPNN开发了MDNet,学习MD轨迹中前后时间步之间坐标和动量的变化.通过使用消息传递方法和用于边操作的多层感知机,MDNet可以对Lennard-Jones (LJ)粒子等简单系统在大时间步长(128 fs)上进行高精度预测.该模型复杂度低,易于理解和扩展,但在处理更复杂的分子体系时表现不佳.这凸显了分子复杂性对机器学习网络架构和模型训练的更高要求.基于图结构的机器学习模型会在每一步迭代时计算节点的隐藏特征表示. Schaaf等^[48]发现这些特征随时间的变化很小,即前后时间步的特征高度相关.基于这一观察,他们提出了基于等变MPNN的BoostMD方法,利用前一时间步的节点隐藏特征和几何信息预测后续步骤的能量.该方法的模型开销较小且并行性能良好,在小分子体系的加速模拟中表现出色. Li等^[49]开发了图神经网络架构GAMD (GNN accelerated MD),用于预测输入构象的力,从而取代耗时的势能偏导计算过程以加速分子动力学. GAMD采用与SchNet^[50]类似的图卷积架构,并在边特征中加入原子间化学键的信息,从而能够处理更复杂的系统,如NVT系综下的水盒子模拟.有趣的是,随着水分子模型从简单到复杂,GAMD的预测误差也随之显著增大,尤其是在处理柔性水分子的局部作用力(如成键相互作用)时尤为明显.这可能反映了MPNN架构的一个关键缺陷,即由于高阶相邻数

据的多轮聚合,低阶邻居节点的信息会被不断“稀释”,导致模型很快“遗忘”近邻信息,表征能力下降.而在分子体系中,距离相近的原子对,尤其是参与键合作用的原子对,往往具有更紧密的相互作用.

此外,基于GNN的模型在物理一致性上高度依赖于训练数据的质量和多样性,这限制了它们在不同体系中的可迁移性.另一个关键挑战在于计算效率.在模拟系统的规模增大时,由于每个节点的嵌入递归地依赖于其所有邻居的嵌入,节点表示和梯度的计算复杂度会随着神经网络层数的增加而呈指数增长,在训练时导致“邻居爆炸”问题.这使得大尺度系统中GNN模型的效率严重下降,需要进一步优化模型构架以提高它们的可扩展性.

还有一种思路,借鉴增强采样算法中常用的CV概念,从高维轨迹数据中提取低维隐表示,并学习动力学信息与低维表征的关系.基于编码器-处理器-解码器(encoder-processor-decoder, EPD)框架的模型,通过编码器 E 将轨迹 $\mathbf{x}_t$ 压缩为隐向量 $\mathbf{s}_t$,在压缩表示的隐空间(latent space)中训练处理器 P 学习隐向量随时间演化的规律,再通过解码器 D 重构原始高维轨迹的构象.这种方法提供了一种优雅的方式来学习系统动力学内涵的规律,但在保持隐空间中动力学的物理约束和轨迹演化与重构的稳定性等方面仍存在一些挑战.

如果轨迹采样的时间间隔 Δt 足够大,使迭代过程成为一个马尔可夫过程,那么一个理想的处理器应该最小化 $\mathbf{s}_{t+\Delta t}$ 与 $P\mathbf{s}_t$ 之间的误差.通过最大似然或最小回归误差在隐空间中优化编码器比优化处理器更具挑战性.这是因为编码器更容易收敛到平凡的、无信息的结果,难以在保持泛化能力的同时捕捉到轨迹中最有意义的动力学.解码器也面临类似的挑战,因为它应当是编码器的严格逆操作,以确保能准确地重建原始高维空间.自编码器(autoencoder, AE)通过自回归方式保证编码源与解码信息的一致性,而变分自编码器

(variational autoencoder, VAE)提供了一种更强大的替代方案. 与AE的确定性映射不同, VAE可以将输入空间中变量的分布映射到一个隐分布, 并从隐分布中采样以解码生成样本, 这对于学习MD模拟构象的概率分布尤为有用.

然而, 在显式溶剂模拟体系中, 溶质和溶剂之间的相互作用存在动力学耦合效应, 从而使马尔可夫假设变得不适用. 溶剂壳层通常会引入记忆效应, 而记忆效应本质上是非马尔可夫的, 因此需要更复杂的建模策略. 循环神经网络(recurrent neural network, RNN)模型, 尤其是改进的长短期记忆(long short-term memory, LSTM)模型可作为处理器, 通过存储之前时间步骤的信息来处理序列数据, 利用上下文信息捕捉MD轨迹中的时间依赖性. 相比传统RNN, LSTM可以在更长的时间尺度上存储信息, 避免梯度消失, 并且在对序列数据建模时更加灵活. 一些研究探索了LSTM在加速MD模拟中的应用. 例如, Tsai等^[51]使用基于LSTM的自然语言模型, 通过学习高维动力学轨迹的一维编码概率, 能够重现不同时间尺度上的动力学分布. 该方法可以学习非马尔可夫的蛋白质动力学过程, 并能直接计算热力学可观测量. 其局限包括需要对状态空间进行离散化、不具备生成能力以及无法识别动力学的低维序参量表征. Kadupitiya等^[52]使用LSTM根据前几帧轨迹预测隐空间中下一个时间步的构象和动量, 为时间步长外推提供了一种很有前景的解决方案. 该方法目前在更复杂系统中的适用性方面仍然有限, 并受到RNN对训练数据量强依赖的限制. Vlachas等^[53]提出了一种基于混合密度网络自动编码器和LSTM的数据驱动生成框架(MDN-LSTM), 既可以实现细粒度构型和粗粒度隐向量之间的概率映射, 又可以有效捕捉非马尔可夫动力学特征. MDN-LSTM框架为复杂的时变系统建模提供了一种更具通用性和稳健性的方法, 但对于更复杂的分子系统仍需要进一步验证.

扩散模型作为前景广阔的生成式模型架构, 也被应用于加速MD模拟. Schreiner等^[54]提出了隐式转移算子学习框架, 通过采用条件去噪的扩散概率模型和SE(3)几何等变架构, 学习多个时间分辨率下的代理模型, 而这些模型可以在部分观察到的状态下自动生成一致的随机动力学过程. 该方法在Müller二维势、丙氨酸二肽以及Chignolin等四个快速折叠蛋白体系中都取得了很好的效果, 然而其泛化性能相对受

限. Wu等^[55]则提出了DiffMD框架, 利用扩散模型和几何自注意力机制直接预测下一时间步的原子坐标. 该模型使用3D球面傅里叶-贝塞尔表示来增强自注意力模块, 以便更好地捕捉原子之间的朝向和速度信息. 这种混合模型是时间步长外推法一个很有前景的方向.

多时间步长方法^[56,57]是突破MD积分步长限制的经典策略. 基于这一思想, Fu等^[58]和Sun等^[59]探索了将系统拆分为不同时间尺度分别处理的想法, 以更高效地模拟大体系中的短程和长程相互作用. 这些方法通过利用不同机器学习模型处理多尺度信息, 但在方法评估和整体性能方面仍面临挑战. 此外, 预报-校正方法也展示了在MD时域加速中的应用前景. 该方法通过预报和校正的迭代循环来改进初始猜测, 近似求解数学问题. Kochkov等^[60]在计算流体力学中应用学习修正(learned correction, LC)神经网络模型来预测和修正速度场残差. 在他们的实验中, LC模型在逐点准确性、统计准确性和稳定性方面与学习插值模型结果一致, 并且在多种流体类型中均优于EPD和ResNet架构. LC模型结构简单, 实现容易, 可能具备更好的泛化性. 生成式框架LiFlow^[61]应用预报-校正技术来学习锂基固态电解质的分子动力学, 实现对下一时间步原子坐标的推演. 该框架中还引入了条件流匹配模型, 使得模型能够更好地适应不同的温度条件, 并实现对不同类型原子的自适应处理. 类似地, MultiPDENet方法^[62]将流体运动分为宏观尺度和微观尺度, 在微观尺度用高精度物理方程数值求解, 而在宏观尺度则通过设计基于有限差分格式的卷积滤波器来求解偏微分方程, 再用误差校正网络修正结果, 结合两个尺度以实现加速流体模拟的效果. 这些框架结合了预测和修正的迭代过程, 在效率和自适应度方面取得了令人印象深刻的成果.

Young^[63]以阿司匹林分子为模型体系对几类网络的MD轨迹预测性能开展基准研究. 结果表明, 图卷积网络的性能优于其他架构, 等变神经网络(E3NN)次之, 略优于RNN. 扩散模型有较强的生成能力, 但需要维护物理可行性和等变性, 以确保生成的数据点是合理的; 不考虑时间依赖性的前馈神经网络性能最差. 这项研究虽然简单, 且阿司匹林分子也是一个很小的体系, 但对于不同方法和架构的测评将为推动科学探索提供宝贵的参考价值.

4 序列关联: 探索构象演变中的隐性联系

在研究大分子构象转变机制时, 如何高效表征其长时演化规律是一个亟待解决的核心问题. 传统连续积分方法在解析长时间尺度的动态过程时存在两大缺陷: 一是高维构象空间中状态遍历效率的限制, 二是显式物理时间标签对构象演化本质规律的表征冗余性. 实际上, 仅通过蛋白质的几个亚稳态结构及在这些结构间的跃迁顺序即可大致勾勒出其整体工作机制的蓝图. 序列加速方法通过建立统计动力学框架, 将研究焦点从显式物理时间标签转移到构象演化序列的统计关联性, 通过隐式马尔可夫模型或深度时序网络捕捉态跃迁的潜在模式. 这类方法通过用隐式时间尺度(implied timescale)替代显式积分时长, 将连续轨迹简化为亚稳态之间的跃迁. 这种降维策略降低了表征和计算的复杂度, 特别适用于研究蛋白质折叠、转运蛋白构象变化等涉及复杂自由能景观的长时间尺度动态过程.

一种有效的思路是通过插值在两个已知状态之间寻找有序连接的亚稳态^[64,65]. 尽管这种方法不是一种直接的加速方法, 但可以视为对小时间步长下高分辨率MD模拟的加速, 用于高效搜索转变路径并勾勒出动力学过程.

当观测的时间延迟 τ 足够大时, 可以将亚稳态之间的动力学转变近似为马尔可夫过程. 基于 τ 的构象分布概率预测可被视为对长时间尺度上系统动力学演化的粗粒化描述. 当 $\tau \gg \Delta t$ 时, 其预测结果与通过连续时间 $N\Delta t$ 积分得到的轨迹演化包含等价的机制信息. 但是, τ 本质上是用于简化和加速长期动力学描述的统计时间尺度, 而非严格对应物理时间步长累积. 因此, 与时域方法不同, 序列方法并不严格复现连续动力学的物理时间演化, 而主要关注长时间尺度下稀有构象变化的动力学机制.

马尔可夫过程变分法网络(variational approach for Markov processes networks, VAMPnets)是一个使用EPD框架的深度学习架构^[66,67], 基于经过预处理的分子动力学轨迹构建马尔可夫动力学网络. VAMPnets通过使用具有编码和解码层的神经网络学习高维状态空间的低维编码, 将高维状态空间聚类并使用转移概率矩阵来构建马尔可夫网络, 训练目标是 minimized 真实马尔可夫过程与从低维表示中估计的过程之间的差

异. 然而, VAMPnets主要关注隐空间内粗粒化构象坐标上投影动力学的马尔可夫化, 其本身并不关注全原子坐标的精细解码重构, 因此通常不作为生成式采样网络使用. 此外, 使用马尔可夫状态模型(Markov state model, MSM)来准确采样隐空间动力学并估计动态过程的时间尺度是成功运行VAMPnets工作流的前提. Hernández等^[68]提出的变分动力学编码器在VAE框架中集成了自相关损失函数与输入层加噪机制: 前者通过时序约束捕捉分子体系的长期动力学特征, 后者提升模型对轨迹噪声的鲁棒性. 通过这些改进, 该方法实现了跨时间步加速, 并能通过分析隐空间特征有效地识别体系中的慢运动.

Timewarp方法^[69]通过引入条件归一化流模型作为马尔可夫链蒙特卡罗(MCMC)的提议分布, 直接从当前帧构象对 τ 后的构象分布 $p(x(t+\tau)|x(t))$ 进行采样, 显著扩展了模拟的时间跨度至亚微秒级, 同时维持了平衡热力学的玻尔兹曼分布. 该模型可以被看作时间依赖的玻尔兹曼生成器^[70], 且有较好的迁移性. 在隐式溶剂模型下的多肽系统中, 该方法结合MCMC采样能够获得与MD采样接近的结果. 目前该方法还存在采样接受率低、较MD模拟生成样本数更少等问题, 有待进一步改进. 此外, 基于条件扩散模型的Score Dynamics方法^[71]将传统力场替换为神经网络预测的分数场, 也实现了对转移概率矩阵的预测, 可在皮秒级时间步长下生成符合玻尔兹曼分布的构象系综, 且具备捕捉稀有事件的能力, 在丙氨酸二肽和小烷烃等较为复杂的体系上表现出良好的效果.

序列方法高度依赖于对动力学马尔可夫性的先验假设. 然而, 为实现动力学降维而对状态空间系统进行粗粒化投影会引入记忆效应, 导致亚稳态转变对历史轨迹存在非马尔可夫关联. 这种现象在分析短延迟时间的构象动力学建模时尤为显著, 使得基于MSM模型对粗粒化动力学构象跃迁的预测产生系统性偏差. 为解决这一问题, 研究者们发展了多种非马尔可夫的方法. 考虑到亚稳态可与机器学习模型中的隐变量对应, 一个方案是通过构建隐马尔可夫模型(hidden Markov models, HMM)以替代MSM^[72,73], 但HMM基于的最大期望算法易陷入局部最优且收敛效率低. 上海交通大学吴昊教授^[74]基于可观测量子模型(observable operator models, OOM)开发的谱学习方法从非平衡数据学习平衡动力学, 从而重构非马尔可夫动力学过程. 黄旭

辉教授团队近年来在非马尔可夫动力学上做了一系列开创性的重要工作^[75-78], 如基于广义动力学主方程构建了准马尔可夫态模型(quasi-Markov state models, qMSM)^[79]. 该模型通过计算记忆核函数(memory kernel)量化快自由度的历史贡献, 能够在较短时间尺度上捕捉快动力学过程, 进而实现对长时间尺度动力学行为的准确预测.

序列方法的另一个核心争议在于其数学抽象与物理真实性的匹配度: 由于缺乏明确的时间动力学参数, 序列模型与真实动力学的一致性通常难以量化; 序列建模对状态驻留时间的模糊处理可能导致虚假关联的产生, 例如将环境热噪引起的构象涨落误判为真实的功能性转变. 此外, 高维构象空间的序列建模面临维度灾难, 现有解决方案多依赖粗粒化策略或运动模式分解, 可能导致丢失关键的时空尺度信息. 值得关注的是, Timewarp等方法尝试将热力学约束(如玻尔兹曼分布)嵌入序列生成过程, 通过物理引导的注意力机制提升状态转变路径的物理合理性, 这或许能有效弥合数据驱动模型与统计力学原理之间的鸿沟, 但仍有待进一步深入研究和验证. 此外, 许多现有方法基于粗粒化模型(如仅保留蛋白质主链自由度), 原子级细节的保真度仍有待提升.

5 空间搜索: 无需时间束缚的构象自由度

在许多具体应用中, 研究者关心的是体系的关键构象态及其热力学分布, 而非其具体的时间演化过程或顺序, MD模拟只是用来获得构象态和构象系综的计算方法. 例如, 在蛋白质-配体相互作用预测或构象采样任务中, 仅生成符合平衡态分布的代表性结构常常即可满足研究需求. 不依赖动力学路径的平衡态构象生成方法在这些场景下可大幅提升计算效率.

基于空间搜索的结构预测方法通过高效生成符合热力学平衡态分布的构象集合摆脱了对时间积分的依赖, 并在药物结合口袋采样、蛋白功能构象搜索等场景中展现应用前景. 近年来, 机器学习驱动的分

主链/侧链原子间的非物理重叠^[84], 或与溶剂分子形成热力学不利的接触模式, 有时甚至出现动力学不可能的拓扑缠绕. 这些现象揭示了现有算法在处理多组分复杂体系时的固有局限性.

最近, 研究者基于AlphaFold等蛋白质结构预测工具开发蛋白多构象生成器. 由于AlphaFold通过对目标序列进行多重序列比对(multiple sequence alignment, MSA)提取共进化信息来预测蛋白质结构, 从技术上, 可以通过扰动所输入的MSA来生成蛋白质及其复合物的多个构象^[85]. 然而, 现有基于MSA扰动的方法在蛋白构象多样性预测上同样被AlphaFold本身的预测精度所限制: 对于目标蛋白的固有无序片段, AlphaFold所生成的构象系综常与实验观测的二级结构多样性不符^[86]. 通过相似性聚类等手段解耦序列共进化信息以获取蛋白质多构象的预测策略曾被寄予厚望^[87], 但近期研究表明其预测效果与对MSA序列空间随机采样相比并无统计显著的改善^[88]. 这意味着所谓的序列相似度聚类实际上仅提取了PDB训练集中同源蛋白的结构特征, 而与解耦序列共进化信息关联不大, 同时也反映了AlphaFold等生物学基础大模型缺乏物理可解释性的客观现实. 许多近期研究尝试将动力学原理嵌入AlphaFold框架, 成功的方案包括在损失函数中引入先验知识约束引导结构预测^[89], 结合增强采样方法优化构象搜索^[90]、利用能量挫折分析提示序列MSA特征构建^[91]等. 尽管如此, 基于AlphaFold的构象集合与MD模拟的平衡态系综仍存在本质差异: 首先在于进化约束主导的预测偏差, 非保守位点的构象变化难以被MSA表述的共进化信号捕获; 其次在于非物理的能量景观表征, 生成的构象轨迹缺乏连续的自由能面特征, 导致其构型比例与MD结果存在系统性偏差.

为进一步突破AlphaFold的固有局限, 研究者开发了多类改进型算法. Str2Str模型^[92]受模拟退火方法启发, 将AlphaFold的结构预测模块与得分匹配(score matching)模型相结合, 在几何等变约束下对加噪扰动的蛋白结构进行降噪重构. 值得注意的是, Str2Str在推断时可以通过调节随机扰动的程度(即加噪时间步数)来实现结构多样性程度的显式控制, 而无需从随机高斯分布采样的隐空间开始完整降噪过程. 此外, Str2Str的训练集主要来源于PDB结构数据, 尽管所生成的构象系综可能无法反映严格的热力学平衡分布, 但具有

对蛋白的构象空间零样本探索能力。基于类似思路的P2DFlow^[93]以ESMFold^[94]预测的静态结构为起点, 基于SE(3)流匹配模型预测目标构象分布。近期发表的BioEmu框架^[95]基于分布式Graphormer架构, 利用AlphaFold结构数据库预训练捕捉基础构象特征, 通过微调阶段整合大规模MD轨迹及折叠自由能实验数据, 最终实现吻合自由能景观分布的多构象生成。初步测试表明, 在快速折叠蛋白等体系中, BioEmu所预测的二维自由能景观与MD结果相比已达到化学精度。

在非AlphaFold驱动的大分子系综采样领域, 学界通过融合物理原理与生成式模型, 形成了多样化的技术方案。美国南卫理公会大学陶鹏团队^[96]系统评估了VAE对蛋白构象生成任务的适用性。研究表明, 尽管VAE能很好地拟合训练集MD轨迹的构象分布, 但在隐空间作分布外解码时会不可避免地产生非物理结构冲突, 这提示了纯数据驱动的概率分布拟合方法需要显式引入物理约束来校正采样偏差。对此, Baker团队^[97]提出创新的两段式架构: 首先将蛋白质结构编码为邻接矩阵, 利用VAE在邻接矩阵空间作自回归采样, 再通过RosettaFold基于解码网络输出的邻接矩阵重建三维结构。该方法有效避免了利用VAE直接预测原子坐标时的几何失真问题。

在热力学约束嵌入方面, 玻尔兹曼生成器^[70]开创性地通过归一化流模型, 将符合玻尔兹曼分布的能量函数映射到高斯隐空间, 结合MCMC采样实现了对热力学构象系综的高效生成。Ruzmetov等^[98]开发的ICoN模型则采用基于向量编码的键-角扭转(vBAT)自编码器模型, 对二面角空间的低维隐空间进行非线性插值以生成构象系综。与传统VAE相比, ICoN在保持构象合理性的同时, 更精确捕获了蛋白质分子的关键物理特征。Wang等^[99]开发的ConfDiff则提出了一种基于全原子MD的能量和力引导策略, 使生成构象的势能分布更接近热力学平衡。数据驱动方面, Hu等^[100]构建了覆盖单/多态蛋白的大规模MD数据集, 通过自由能景观分析识别构象转换路径, 并以此训练深度学习模型预测高能态与过渡态。在膜蛋白体系中, 浙江大学王勇团队针对P型ATP酶的构象多样性这一具体问题, 提出了基于条件扩散模型的动态采样框架。通过引入膜环境特异性条件概率约束, 结合MD数据扩展训练的构象态分类器, 他们成功生成了符合实验观测的转运循环态^[101]。该工作进一步提出了生成模型与MD模拟

的协同策略: 生成模型提供初始态分布, 短轨迹MD模拟细化蛋白构象演化, 以揭示构象态的时间演变和稳定性。

固有无序蛋白(intrinsically disordered protein, IDP)因其高度动态的构象分布, 成为分子构象采样领域的特殊挑战。针对这一难题, 密歇根州立大学Feig团队系统探索了基于深度学习的IDP系综预测方法。他们开发的idpGAN模型^[102]采用基于自注意力机制的生成式对抗网络架构, 直接从序列预测的粗粒化IDP构象系综。该模型展现出良好的跨体系迁移能力, 并且可以应用于更高结构分辨率的蛋白构象上。在此基础上, 该团队进一步开发的idpSAM框架^[103]整合了SE(3)-不变自编码器与扩散模型, 在大规模MD模拟构象数据集上进行训练, 实现了对IDP构象系综的高精度拟合, 优化采样效率的同时也表现出更好的可迁移性。

当前, 机器学习驱动的大尺度构象搜索仍然面临两大核心瓶颈。首先是复杂体系构象空间的维度灾难问题。考虑一个由 N 个氨基酸残基组成的蛋白质, 其主链可旋转二面角约 $2N$ 个, 构象自由度随 N 呈指数级增长。其次是结构误差的级联放大效应, 这导致对每个自由度的预测容错极小。对各自由度的预测模型误差(如主链二面角 $\pm 1^\circ$)在折叠过程中会被逐级放大, 最终导致蛋白质全局构象显著偏离真实分布。

空间采样方法的核心革新在于摆脱对时间序列的显式依赖, 各个构象生成过程彼此独立而能够高效并行。然而, 其热力学合理性仍不明确。尽管生成的构象可满足结构化学约束(如键长、键角分布和无原子重叠), 但这些构象仍可能位于热力学或动力学不可达的构象空间区域。问题的源头在于: (1) 生成模型未严格遵循热力学分布约束; (2) 训练数据未能覆盖真实系统的构象空间。当前研究通过两条可能的路径来提升模型的物理可靠性。一是加入显式物理约束, 将能量函数嵌入整合到生成模型的损失函数中, 使生成的构象系综满足局部能量分布条件。二则在于利用后验筛选流程, 通过分子对接、增强采样等技术对生成构象进行热力学验证, 剔除异常结构。此外, 构象空间采样和生成模型对训练数据分布高度依赖性, 这导致对构象建模的泛化性能低下。前沿研究正尝试融合构象预测与增强采样技术, 在高效的构象生成的同时引入热力学可解释性。这种混合范式具有突破当前构象采样效率瓶颈的潜力, 但是需要在工程化中优化这类异构算

法对计算资源的调度效率和实际应用时的自动化程度。

6 总结与展望

本文探讨了机器学习加速分子动力学模拟的三类方法: 时域驱动、序列解析与空间采样。它们本质上反映了“精度-效率-可解释性”不可能三角的权衡关系: 时域方法以物理保真度见长, 在热力学可解释性上占据优势, 但在跨越飞秒级时间步长缩减计算成本上效率尚不可观; 序列方法实现了中等时间尺度的态间跃迁解析, 但其有效性高度依赖于强马尔可夫性假设; 空间方法能高效遍历相空间构象, 却由于完全丧失时间关联性导致动力学过程无法重构。多层次融合可能是突破这些局限的关键: 通过时域模型捕捉局域快速运动, 序列模型解析介观尺度构象跃迁, 空间模型提供全局相空间先验分布, 构建自适应采样框架。当然, 三个层面上均需取得更多的突破性进展方能实现这一融合策略。

机器学习模型生成的构象采样、构象系综或构象动力学, 始终绕不开物理约束的嵌入难题。破解难题的关键在于突破传统数据驱动范式的局限, 通过将物理原则深度整合到神经网络架构中, 使生成轨迹自发地满足能量守恒、动量连续等物理基本约束。相较于常规机器学习模型对统计相关性的被动捕捉, 这种主动编码的策略能够实现物理规律与数据特征的有机融合。技术上需同时考虑架构设计和优化目标。在架构设计层面, 通过构建等变神经网络严格保持分子体系的旋转、平移和交换对称性, 确保构象生成过程符合分

子间相互作用的几何约束。在优化目标层面, 可能需要超越PINN直接将物理方程嵌入损失函数的方案, 使神经网络参数更新过程同步优化动力学方程的残差项, 约束模型产生的平滑连续的轨迹解。

从学科范式演进的视角看, 分子模拟正从“计算观测”向“智能设计”迈进。研究者已不再满足于通过对给定分子体系进行MD模拟来解释实验观测现象和解析分子机制, 而是期望利用分子模拟开展高通量的筛选与设计, 引导实验验证和功能优化。针对高通量药物分子筛选, 将生成模型与MD或基于MD的自由能微扰方法(free energy perturbation, FEP)结合的级联流程已初步展现出工业应用潜力^[104,105]。在人工酶设计领域, 主流策略仍以数据驱动的分步优化为主。例如先用MD筛选关键突变位点, 再结合机器学习解析机制^[106]; 或通过机器学习分析MD轨迹定位关键残基以指导突变^[107]。开发AI+MD深度融合方法, 建立实时构象评估与理性设计迭代的闭环体系仍在探索中。

AI加速方法的多样化发展催生了对标准化评估体系的迫切需求。然而, 由于不同AI加速策略针对MD模拟的优化维度和方法框架存在较大异质性, 构建统一的基准测试框架极具挑战, 需要严谨全面且差异化的评价指标制定。通过将物理规律编码为神经网络的先验约束, 同时利用大数据捕捉超越理论框架的关联特征, 新一代算法有望在保持物理可解释性的前提下, 突破传统分子模拟方法的效率瓶颈。这种物理增强的混合智能范式, 或将重塑生物分子复杂性研究的方法论体系——不仅为理解蛋白质折叠等基础生物学过程提供新工具, 更有望推动定制化药物设计、人工酶理性合成等应用领域的跨越式发展。

致谢 感谢鲁晓莉博士对本文工作的贡献, 感谢西湖大学高性能计算中心对本课题组的长期支持与帮助。

参考文献

- 1 Dobson CM. *Annu Rev Biochem*, 2019, 88: 25–33
- 2 Rout MP, Sali A. *Cell*, 2019, 177: 1384–1403
- 3 Huang C, Kalodimos CG. *Annu Rev Biophys*, 2017, 46: 317–336
- 4 Moffat K. *Annu Rev Biophys Biophys Chem*, 1989, 18: 309–332
- 5 Yip KM, Fischer N, Paknia E, Chari A, Stark H. *Nature*, 2020, 587: 157–161
- 6 Bonomi M, Heller GT, Camilloni C, Vendruscolo M. *Curr Opin Struct Biol*, 2017, 42: 106–116
- 7 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M,

- Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. *Nature*, 2021, 596: 583–589
- 8 Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung CC, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Žemgulytė A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Židek A, Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM. *Nature*, 2024, 630: 493–500
- 9 Miller MD, Phillips Jr. GN. *J Biol Chem*, 2021, 296: 100749
- 10 Huse M, Kuriyan J. *Cell*, 2002, 109: 275–282
- 11 Pitsawong W, Buosi V, Otten R, Agafonov RV, Zorba A, Kern N, Kutter S, Kern G, Pádua RA, Meniche X, Kern D. *eLife*, 2018, 7: e36656
- 12 Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell Jr AD. *Nat Methods*, 2016, 14: 71–73
- 13 Lu X, Huang J. *Commun Biol*, 2022, 5: 837
- 14 Hou Y, Lu X, Xu Z, Qu J, Huang J. *RSC Adv*, 2023, 13: 4263–4274
- 15 Lu X, Xue J, Zhou Q, Huang J. *Biochemistry*, 2025, 64: 1405–1414
- 16 Lu X, Huang J. *Biophys J*, 2024, 123: 1195–1210
- 17 Behler J, Parrinello M. *Phys Rev Lett*, 2007, 98: 146401
- 18 Zhang L, Han J, Wang H, Car R, E W. *Phys Rev Lett*, 2018, 120: 143001
- 19 Devereux C, Smith JS, Huddleston KK, Barros K, Zubatyuk R, Isayev O, Roitberg AE. *J Chem Theor Comput*, 2020, 16: 4192–4202
- 20 Unke OT, Stöhr M, Ganscha S, Unterthiner T, Maennel H, Kashubin S, Ahlin D, Gastegger M, Medrano Sandonas L, Berryman JT, Tkatchenko A, Müller KR. *Sci Adv*, 2024, 10: eadn4397
- 21 Noé F, Tkatchenko A, Müller KR, Clementi C. *Annu Rev Phys Chem*, 2020, 71: 361–390
- 22 Behler J. *Angew Chem Int Ed*, 2017, 56: 12828–12840
- 23 Kocer E, Ko TW, Behler J. *Annu Rev Phys Chem*, 2022, 73: 163–186
- 24 Hwang W, Austin SL, Blondel A, Boittier ED, Boresch S, Buck M, Buckner J, Caflisch A, Chang HT, Cheng X, Choi YK, Chu JW, Crowley MF, Cui Q, Damjanovic A, Deng Y, Devereux M, Ding X, Feig MF, Gao J, Glowacki DR, Gonzales Ii JE, Hamaneh MB, Harder ED, Hayes RL, Huang J, Huang Y, Hudson PS, Im W, Islam SM, Jiang W, Jones MR, Käser S, Kearns FL, Kern NR, Klauda JB, Lazaridis T, Lee J, Lemkul JA, Liu X, Luo Y, MacKerell Jr. AD, Major DT, Meuwly M, Nam K, Nilsson L, Ovchinnikov V, Paci E, Park S, Pastor RW, Pittman AR, Post CB, Prasad S, Pu J, Qi Y, Rathinavelan T, Roe DR, Roux B, Rowley CN, Shen J, Simonett AC, Sodt AJ, Töpfer K, Upadhyay M, van der Vaart A, Vazquez-Salazar LI, Venable RM, Warrensford LC, Woodcock HL, Wu Y, Brooks Iii CL, Brooks BR, Karplus M. *J Phys Chem B*, 2024, 128: 9976–10042
- 25 Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. *SoftwareX*, 2015, 1-2: 19–25
- 26 Lee TS, Cerutti DS, Mermelstein D, Lin C, LeGrand S, Giese TJ, Roitberg A, Case DA, Walker RC, York DM. *J Chem Inf Model*, 2018, 58: 2043–2050
- 27 Phillips JC, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, Bernardi RC, Buch R, Fiorin G, Hénin J, Jiang W, McGreevy R, Melo MCR, Radak BK, Skeel RD, Singharoy A, Wang Y, Roux B, Aksimentiev A, Luthey-Schulten Z, Kalé LV, Schulten K, Chipot C, Tajkhorshid E. *J Chem Phys*, 2020, 153: 044130
- 28 Eastman P, Galvelis R, Peláez RP, Abreu CRA, Farr SE, Gallicchio E, Gorenko A, Henry MM, Hu F, Huang J, Krämer A, Michel J, Mitchell JA, Pande VS, Rodrigues JP, Rodriguez-Guerra J, Simonett AC, Singh S, Swails J, Turner P, Wang Y, Zhang I, Chodera JD, De Fabritiis G, Markland TE. *J Phys Chem B*, 2024, 128: 109–116
- 29 Sugita Y, Okamoto Y. *Chem Phys Lett*, 1999, 314: 141–151
- 30 Laio A, Parrinello M. *Proc Natl Acad Sci USA*, 2002, 99: 12562–12566
- 31 Gao YQ. *J Chem Phys*, 2008, 128: 064105
- 32 Wang L, Friesner RA, Berne BJ. *J Phys Chem B*, 2011, 115: 9431–9438
- 33 Krämer A, Pickard Iv FC, Huang J, Venable RM, Simonett AC, Reith D, Kirschner KN, Pastor RW, Brooks BR. *J Chem Theor Comput*, 2019, 15: 3854–3867
- 34 Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Miguez AN, Bickel J, Wang Y, Pincay J, Wu Q, Simmerling C. *J Chem Theor*

- Comput*, 2020, 16: 528–552
- 35 Damm W, Dajnowicz S, Ghoreishi D, Yu Y, Ganeshan K, Madin O, Rudshteyn B, Hu R, Wu M, Shang Y, Albanese S, Zou Y, Ye M, Johnson Z, Trzoss M, Rafi S, Kapilashrami K, Saleh N, Ghanakota P, Zhang Y, Sampson J, Chen W, Wang L, Dahlgren M, Russell E, Leffler A, Abel R, Friesner R, Harder E. OPLS5: Addition of Polarizability and Improved Treatment of Metals. Preprint at <https://chemrxiv.org/engage/chemrxiv/article-details/66ba761201103d79c5c60641>, 2024
 - 36 Shaw DE, Adams PJ, Azaria A, Bank JA, Batson B, Bell A, Bergdorf M, Bhatt J, Butts JA, Correia T, Dirks RM, Dror RO, Eastwood MP, Edwards B, Even A, Feldmann P, Fenn M, Fenton CH, Forte A, Gagliardo J, Gill G, Gorlatova M, Greskamp B, Grossman JP, Gullingsrud J, Harper A, Hasenplaugh W, Heily M, Heshmat BC, Hunt J, Ierardi DJ, Iserovich L, Jackson BL, Johnson NP, Kirk MM, Klepeis JL, Kuskin JS, Mackenzie KM, Mader RJ, McGowen R, McLaughlin A, Moraes MA, Nasr MH, Nociolo LJ, O'Donnell L, Parker A, Petricolas JL, Pocina G, Predescu C, Quan T, Salmon JK, Schwink C, Shim KS, Siddique N, Spengler J, Szalay T, Tabladillo R, Tartler R, Taube AG, Theobald M, Towles B, Vick W, Wang SC, Wazlowski M, Weingarten MJ, Williams JM, Yuh KA. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. New York: Association for Computing Machinery, 2021. 1–11
 - 37 Shim KS, Greskamp B, Towles B, Edwards B, Grossman JP, Shaw DE. The Specialized High-Performance Network on Anton 3. In: *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. Los Alamitos: IEEE Computer Society, 2022. 1211–1223
 - 38 Xie T, Saleh T, Rossi P, Kalodimos CG. *Science*, 2020, 370: eabc2754
 - 39 Shamir M, Bar-On Y, Phillips R, Milo R. *Cell*, 2016, 164: 1302–1302.e1
 - 40 Wang Y, Lamim Ribeiro JM, Tiwary P. *Curr Opin Struct Biol*, 2020, 61: 139–145
 - 41 Prašnikar E, Ljubić M, Perdih A, Borišek J. *Artif Intell Rev*, 2024, 57: 102
 - 42 Ding X, Vilseck JZ, Brooks III CL. *J Chem Theor Comput*, 2019, 15: 799–802
 - 43 Mironov V, Alexeev Y, Mulligan VK, Fedorov DG. *J Comput Chem*, 2019, 40: 297–309
 - 44 Song Z, Xu Y, Zhang H, Ding Y, Huang J. *J Chem Theor Comput*, 2024, 20: 9857–9870
 - 45 Sirignano J, Spiliopoulos K. *J Comput Phys*, 2018, 375: 1339–1364
 - 46 Mattheakis M, Sondak D, Dogra AS, Protopapas P. *Phys Rev E*, 2022, 105: 65305
 - 47 Zheng T, Gao W, Wang C. Learning Large-Time-Step Molecular Dynamics with Graph Neural Networks. Contributed Talk presented at the NeurIPS 2021 AI for Science Workshop - AI for Science: Mind the Gaps. 2021, arXiv: 2111.15176
 - 48 Schaaf LL, Batatia I, Brunken C, Barrett TD, Tilly J. BoostMD: Accelerating molecular sampling by leveraging ML force field features from previous time-steps. arXiv: 2412.18633
 - 49 Li Z, Meidani K, Yadav P, Barati Farimani A. *J Chem Phys*, 2022, 156: 144103
 - 50 Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. *J Chem Phys*, 2018, 148: 241722
 - 51 Tsai ST, Kuo EJ, Tiwary P. *Nat Commun*, 2020, 11: 5115
 - 52 Kadupitiya JCS, Fox GC, Jadhao V. Solving Newton's Equations of Motion with Large Timesteps using Recurrent Neural Networks based Operators. arXiv: 2004.06493
 - 53 Vlachas PR, Zavavlav J, Praprotnik M, Koumoutsakos P. *J Chem Theor Comput*, 2022, 18: 538–549
 - 54 Schreiner M, Winther O, Olsson S. Implicit Transfer Operator Learning: Multiple Time-Resolution Surrogates for Molecular Dynamics. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 2023. 36449–36462, arXiv: 2305.18046
 - 55 Wu F, Li SZ. DIFFMD: A Geometric Diffusion Model for Molecular Dynamics Simulations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Washington: AAAI Press, 2023. 5321–5329, arXiv: 2204.08672
 - 56 Tuckerman ME, Berne BJ, Martyna GJ. *J Chem Phys*, 1991, 94: 6811–6815
 - 57 Tuckerman M, Berne BJ, Martyna GJ. *J Chem Phys*, 1992, 97: 1990–2001
 - 58 Fu X, Musaelian A, Johansson A, Jaakkola T, Kozinsky B. Learning Interatomic Potentials at Multiple Scales. Poster presented at the NeurIPS 2023 AI for Science Workshop—AI for Scientific Discovery: From Theory to Practice. New Orleans, 2023, arXiv: 2310.13756
 - 59 Sun F, Huang Z, Wang H, Cao Y, Luo X, Wang W, Sun Y. Graph Fourier neural ODEs: bridging spatial and temporal multiscales in molecular dynamics. arxiv: 2411.01600
 - 60 Kochkov D, Smith JA, Alieva A, Wang Q, Brenner MP, Hoyer S. *Proc Natl Acad Sci USA*, 2021, 118: e2101784118

- 61 Nam J, Liu S, Winter G, Gomez-Bombarelli R. Generative acceleration of molecular dynamics simulations for solid-state electrolytes. Contributed Talk presented at the ICML'24 ML4LMS Workshop - ML for Life and Material Science: From Theory to Industry Applications. Vienna, 2024
- 62 Wang Q, Mi Y, Wang H, Zhang Y, Chengze R, Liu H, Wen J-R, Sun H. MultiPDENet: PDE-embedded learning with multi-time-stepping for accelerated flow simulation. arXiv: 2501.15987
- 63 Young S. Predicting molecular trajectories using machine learning methods. Preprint at <https://chemrxiv.org/engage/chemrxiv/article-details/657d2fce66c13817294bbc16>, 2023
- 64 Winkler L, Müller KR, Saucedo HE. *Mach Learn-Sci Technol*, 2022, 3: 025011
- 65 Jing B, Stärk H, Jaakkola T, Berger B. Generative Modeling of Molecular Dynamics Trajectories. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. New York: Curran Associates, Inc., 2024. 40534–40564, arXiv: 2409.17808
- 66 Maradt A, Pasquali L, Wu H, Noé F. *Nat Commun*, 2018, 9: 5
- 67 Chen W, Sidky H, Ferguson AL. *J Chem Phys*, 2019, 150: 214114
- 68 Hernández CX, Waymunt-Steele HK, Sultan MM, Husic BE, Pande VS. *Phys Rev E*, 2018, 97: 062412
- 69 Klein L, Foong A, Fjelde T, Mlodozeniec B, Brockschmidt M, Nowozin S, Noé F, Tomioka R. Timewarp: Transferable Acceleration of Molecular Dynamics by Learning Time-Coarsened Dynamics. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New York: Curran Associates, Inc., 2023. 52863–52883, arXiv: 2302.01170
- 70 Noé F, Olsson S, Köhler J, Wu H. *Science*, 2019, 365: eaaw1147
- 71 Hsu T, Sadigh B, Bulatov V, Zhou F. *J Chem Theor Comput*, 2024, 20: 2335–2348
- 72 Noé F, Wu H, Prinz JH, Plattner N. *J Chem Phys*, 2013, 139: 184114
- 73 McGibbon R, Ramsundar B, Sultan M, Kiss G, Pande V. Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models. In: *Proceedings of the 31st International Conference on Machine Learning*. PMLR, 2014. 1197–1205, arXiv: 1405.1444
- 74 Wu H, Noe F. Spectral Learning of Dynamic Systems from Nonequilibrium Data. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. New York: Curran Associates, Inc., 2016. 4186–4194, arXiv: 1609.00932
- 75 Dominic Iii AJ, Cao S, Montoya-Castillo A, Huang X. *J Am Chem Soc*, 2023, 145: 9916–9927
- 76 Cao S, Qiu Y, Kalin ML, Huang X. *J Chem Phys*, 2023, 159: 134106
- 77 Dominic Iii AJ, Sayer T, Cao S, Markland TE, Huang X, Montoya-Castillo A. *Proc Natl Acad Sci USA*, 2023, 120: e2221048120
- 78 Wu Y, Cao S, Qiu Y, Huang X. *J Chem Phys*, 2024, 160: 121501
- 79 Cao S, Montoya-Castillo A, Wang W, Markland TE, Huang X. *J Chem Phys*, 2020, 153: 14105
- 80 Hoffmann M, Noé F. Generating valid Euclidean distance matrices. arXiv: 1910.03131
- 81 Simm GNC, Hernández-Lobato JM. A Generative Model for Molecular Distance Geometry. In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020. 8949–8958, arXiv: 1909.11459
- 82 Shi C, Luo S, Xu M, Tang J. Learning Gradient Fields for Molecular Conformation Generation. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021. 9558–9568, arXiv: 2105.03902
- 83 Hoogetboom E, Satorras VG, Vignac C, Welling M. Equivariant Diffusion for Molecule Generation in 3D. In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022. 8867–8887, arXiv: 2203.17003
- 84 Hou Y, Xie T, He L, Tao L, Huang J. *Commun Biol*, 2023, 6: 1098
- 85 Meller A, Bhakat S, Solieva S, Bowman GR. *J Chem Theor Comput*, 2023, 19: 4355–4363
- 86 Gavalda-Garcia J, Dixit B, Díaz A, Ghysels A, Vranken W. *J Mol Biol*, 2025, 437: 168900
- 87 Waymunt-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, Ovchinnikov S, Colwell L, Kern D. *Nature*, 2024, 625: 832–839
- 88 Schafer JW, Lee M, Chakravarty D, Thole JF, Chen EA, Porter LL. *Nature*, 2025, 638: E8–E12
- 89 Xie T, Song Z, Huang J. *PRX Life*, 2024, 2: 43001
- 90 Vani BP, Aranganathan A, Wang D, Tiwary P. *J Chem Theor Comput*, 2023, 19: 4351–4354
- 91 Guan X, Tang QY, Ren W, Chen M, Wang W, Wolynes PG, Li W. *Proc Natl Acad Sci USA*, 2024, 121: e2410662121
- 92 Lu J, Zhong B, Zhang Z, Tang J. Str2Str: A Score-Based Framework for Zero-Shot Protein Conformation Sampling. In: *Proceedings of the 12th International Conference on Learning Representations*. New York: Curran Associates, Inc., 2024. 1–32, arXiv: 2306.03117

- 93 Jin Y, Huang Q, Song Z, Zheng M, Teng D, Shi Q. P2DFlow: A Protein Ensemble Generative Model with SE(3) Flow Matching. arXiv: 2411.17196
- 94 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. *Science*, 2023, 379: 1123–1130
- 95 Lewis S, Hempel T, Jiménez-Luna J, Gastegger M, Xie Y, Foong AYK, Satorras VG, Abdin O, Veeling BS, Zaporozhets I, Chen Y, Yang S, Schneuing A, Nigam J, Barbero F, Stimper V, Campbell A, Yim J, Lienen M, Shi Y, Zheng S, Schulz H, Munir U, Clementi C, Noé F. Scalable emulation of protein equilibrium ensembles with generative deep learning. Preprint at <https://www.biorxiv.org/content/10.1101/2024.12.05.626885v1>
- 96 Xiao S, Song Z, Tian H, Tao P. *J Comput Biophys Chem*, 2023, 22: 489–501
- 97 Mansoor S, Baek M, Park H, Lee GR, Baker D. *J Chem Theor Comput*, 2024, 20: 2689–2695
- 98 Ruzmetov T, Hung TI, Jonnalagedda SP, Chen S, Fasihianifard P, Guo Z, Bhanu B, Chang CA. *J Chem Inf Model*, 2025, 65: 2487–2502
- 99 Wang Y, Wang L, Shen Y, Wang Y, Yuan H, Wu Y, Gu Q. Protein Conformation Generation via Force-Guided SE(3) Diffusion Models. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024. 56835–56859, arXiv: 2403.14088
- 100 Hu Y, Yang H, Li M, Zhong Z, Zhou Y, Bai F, Wang Q. *Adv Sci*, 2024, 11: 2400884
- 101 Xu J, Wang Y. *J Chem Inf Model*, 2024, 64: 9227–9239
- 102 Janson G, Valdes-Garcia G, Heo L, Feig M. *Nat Commun*, 2023, 14: 774
- 103 Janson G, Feig M, MacKerell A. *PLoS Comput Biol*, 2024, 20: e1012144
- 104 Bos PH, Houang EM, Ranalli F, Leffler AE, Boyles NA, Eyrieh VA, Luria Y, Katz D, Tang H, Abel R, Bhat S. *J Chem Inf Model*, 2022, 62: 1905–1915
- 105 Carlino L, Astles PC, Ackroyd B, Ahmed A, Chan C, Collie GW, Dale IL, O'Donovan DH, Fawcett C, di Fruscia P, Gohlke A, Guo X, Hao-Ru Hsu J, Kaplan B, Milbradt AG, Northall S, Petrović D, Rivers EL, Underwood E, Webb A. *J Med Chem*, 2024, 67: 8962–8987
- 106 Latallo MJ, Cortina GA, Faham S, Nakamoto RK, Kasson PM. *Chem Sci*, 2017, 8: 6484–6492
- 107 Wang L, Xue Y, Wang J-N, Mo Y, Mei Y. Rational Enzyme Evolution by Facilitating Correlated Motion along the Reaction. Preprint at <https://chemrxiv.org/engage/chemrxiv/article-details/678f6a0ffa469535b9f59b92>, 2025

Accelerating molecular dynamics simulations with artificial intelligence

Runtong Qian^{1,2}, Shiyue Yang^{1,2}, Zilin Song^{1,2}, Jing Huang^{1,2*}

¹ Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310024, China

² School of Life Sciences, Westlake University, Hangzhou 310024, China

*Corresponding author (email: huangjing@westlake.edu.cn)

Abstract: Molecular dynamics (MD) simulation is a powerful computational tool for investigating the structural, thermodynamic and dynamic properties of complex molecular systems. However, its practical application is limited by computational inefficiencies, particularly the gap between the time scales of chemical processes and those accessible through simulations. With the advent of deep learning, artificial intelligence (AI) has emerged as a promising solution for accelerating MD simulations. This review examines AI-driven acceleration approaches at various levels: temporal acceleration to overcome step-size limitations, sequential acceleration to capture conformational transitions, and spatial sampling for optimized conformation exploration. We also discuss the challenges of maintaining physical consistency and multi-scale adaptability in current methods. Future advancements integrating multi-scale approaches with physical constraints are expected to enhance the accuracy, efficiency, and interpretability, thereby expanding the utility of MD simulations in mechanistic analysis and molecular design.

Keywords: molecular dynamics simulation, artificial intelligence, conformational ensemble, conformational dynamics, deep learning, AI for science

doi: [10.1360/SSC-2025-0055](https://doi.org/10.1360/SSC-2025-0055)