

专题: 生物分子模拟中的机器学习

生物分子模拟中的机器学习专题编者按

DOI: [10.7498/aps.72.240101](https://doi.org/10.7498/aps.72.240101)

分子模拟技术是人们从分子层次探究生命现象物理原理的重要手段, 被广泛应用于蛋白质等生物大分子的结构与动力学研究. 自从 20 世纪 70 年代 Karplus 等科学家首次将分子动力学模拟应用于蛋白质研究以来, 分子模拟技术在生物分子体系研究中的应用范围不断扩展, 深刻影响了生物物理学与分子生物学研究的基本范式. 生物大分子的结构动力学涉及皮秒到毫秒甚至更长时间尺度, 如何精确表征具有复杂能量面特征的生物大分子结构与动力学的多尺度特性是生物分子模拟领域的核心难题. 通过物理、化学以及计算机科学等多个领域科学家近 50 年的不懈努力, 人们在生物分子力场准确度提升、各种相互作用的准确描述和计算、增强采样与自由能计算、高维分子模拟数据信息挖掘以及多尺度理论模拟算法构建等方面取得了多个突破. 目前, 人们不仅能够实现对一些蛋白质分子体系毫秒时间尺度的折叠全过程进行分子模拟, 而且能够实现对病毒颗粒、细胞质、甚至染色质等超大分子体系进行分子模拟, 在推动生命科学研究向量化转变中发挥了重要作用. 近年来, 机器学习技术的突飞猛进为解决生物分子模拟中的挑战难题提供了新思路. 人们开始广泛利用深度学习技术构建高精度分子力场、增强分子模拟采样效率、分析高维复杂的分子模拟数据、提取结构及动力学特征等, 取得了一系列重要进展. 结合机器学习算法的分子模拟技术已经在生物物理机制探究、药物设计、结构与动力学预测等基础与应用研究中展现出其实用性与巨大发展潜力.

鉴于机器学习算法在推动生物分子模拟技术发展和生物物理研究中的关键作用, 《物理学报》特组织本专题, 邀请国内部分活跃在该领域前沿的学者撰稿, 深入探讨生物分子模拟与机器学习融合应用的最新研究成果, 并对该领域当前面临的重要挑战及未来研究中可能的突破方向进行综述和展望. 相关论文涵盖了基于机器学习算法的蛋白质分子模拟构象空间搜索、RNA 扭转角预测、蛋白质等生物大分子 pK_a 值预测、生物大分子构象过渡态搜索、蛋白质结构模型质量评估、靶标特异性药物筛选、蛋白质分子设计、高分子塌缩相变和临界吸附相变以及分子体系高维自由能地貌图构建等十余篇研究和综述论文, 分两期刊出. 这些研究论文和综述从不同的角度展示了国内外该领域的最新进展和研究现状. 希望本专题有助于读者了解该领域的前沿研究课题, 并能对促进国内生物分子模拟学术交流发挥作用. 本专题讨论的研究领域涉及多个学科的交叉融合, 且突破性的研究成果不断涌现, 因此本专题所涵盖的代表性成果和前沿进展介绍难免有所遗漏, 不足之处敬请谅解.

(客座编辑: 李文飞, 王伟 南京大学; 周昕 中国科学院大学)

Special Topic—Machine learning in biomolecular simulations

Preface to the special topic: Machine learning in biomolecular simulations

DOI: [10.7498/aps.72.240101](https://doi.org/10.7498/aps.72.240101)

专题: 生物分子模拟中的机器学习

生物大分子过渡态搜索算法及其中的机器学习*

杨建宇# 席昆# 竺立哲†

(香港中文大学(深圳)医学院, 瓦谢尔计算生物研究院, 深圳 518172)

(2023年8月13日收到; 2023年9月9日收到修改稿)

过渡态是物理化学家理解和调控生物大分子相关功能微观机制的关键. 因其存在时间极短, 难以被实验手段捕捉, 全面刻画其结构必须通过物理定律驱动模拟计算搜索予以实现. 然而, 与化学反应过程只涉及少量原子不同, 生物大分子的功能性构象变化所涉的原子和坐标数量巨大, 搜索其过渡态将不可避免地遭遇维数灾难, 即反应坐标问题, 因而催生了多种应对策略和算法. 同时, 随着近年来新型机器学习算法的大量涌现和日臻成熟, 融入机器学习范式的过渡态搜索算法也已出现. 本文首先回顾和梳理过渡态搜索代表性算法的设计思想, 包括依赖集合变量的温和爬升动力学 (gentlest ascent dynamics, GAD)、有限温度弦方法 (finite temperature string, FTS)、快速断层扫描法 (fast tomographic)、基于旅行商的自动路径搜索算法 TAPS, 以及过渡路径采样法 (transition path sampling, TPS). 然后, 重点介绍 TPS 与强化学习融合而成的新型路径采样算法, 解析强化学习在其中的作用, 并厘清其适用场景. 最后, 我们提出一种将降维算法与 GAD 深度融合的新构想, 讨论研发可保留过渡态信息的新型降维算法的必要性及可行性.

关键词: 过渡态搜索, 温和爬升动力学, 路径算法, 强化学习, 生成模型**PACS:** 87.10.Tf, 87.15.A-, 87.15.H-, 87.15.hp**DOI:** 10.7498/aps.72.20231319

1 引言

生物分子实现功能时, 常伴随着结构的巨大转变, 即生物分子的功能性构象变化^[1-3]. 利用实验方法, 往往只能获取上述转变过程前后重要的稳态结构, 如 X 射线 (X-ray macromolecular crystallography)^[4]、核磁共振 (nuclear magnetic resonance, NMR)^[5]、冷冻电子显微镜 (cryo-electron microscopy, cryo-EM)^[6] 等; 或者揭示分子结构变化中的部分特征, 如荧光共振能量转移 (fluorescence resonance energy transfer, FRET) 可给出少数目标残基间的距离变化^[7] 等. 因此, 仅依赖实验方法难以阐明生物分子转变过程的完整信息.

全原子 (all-atom) 分子动力学 (molecular dynamics, MD) 是从原子尺度全面描述生物分子动态行为的标准手段^[8]. 但和化学反应仅涉及反应活性中心内的数十个原子不同, 构象变化所涉及的原子数目巨大, 极端情况下可包括溶质的全部原子, 甚至环境中脂类和溶剂分子的原子^[9-36]. 众多的原子及其三维坐标带来了两个重要的瓶颈.

首先, 在计算效率方面, 复杂大分子百万级的原子数量意味着需要计算万亿级数量的原子间作用力, 即使在目前最优的通用硬件上, 人们所能完成的 MD 模拟时长也仅在微秒量级^[8,37], 距离生物分子的实际功能性动力学行为毫秒级的发生时间仍有巨大差距. 为缓解该效率瓶颈, 数十年来, 人们发展了各类增强采样算法, 其中较有代表性的算法

* 国家自然科学基金 (批准号: 31971179) 和深圳市科技创新委员会 (批准号: JCYJ20200109150003938, RCYX20200714114645019) 资助的课题.

同等贡献作者.

† 通信作者. E-mail: zhulizhe@cuhk.edu.cn

包括副本交换^[38-45], 选择性温度积分增强采样 (selective integrated tempering sampling)^[46-49]、局部抬升 (local elevation)^[50-53]、构象洪泛 (conformational flooding)^[54-56]、元动力学 (metadynamics)^[57-59]、高斯加速动力学^[60-62] 等。

更为重要的是, 在数据分析层面, 尤其是在提取过渡态信息这类理论化学家最关心的问题上, 巨大的原子数量导致了维数灾难. 搜寻过渡态的结构或特征信息是准确刻画和解释所采样本中动力学机制的重中之重. 然而, 即使是在采样数据充足的情况下, 使用不恰当的分析手段 (即机器学习语境下的降维算法), 过渡态区域都将被扭曲以致相关信息丢失.

在已有大量模拟数据的场景中, 可借助 tICA (time-lagged independent component analysis)^[63-65] 利用已有数据中蕴含的动力学信息进行降维, 或运用马尔可夫态模型 (Markov state models)^[66-78] 等分析算法提取动力学信息来应对维数灾难, 并间接推测过渡态信息. 但这类算法中并不直接含有过渡态的定义, 因而超出了本文范畴. 对此类算法感兴趣的读者可参看其他综述^[63,66,68,75-78].

在生物大分子模拟领域, 因其计算效率低下, 数据匮乏是常态, 因此人们对能高效搜寻过渡态的采样算法需求强烈. 但受限于维数灾难, 仅有以下两类采样策略可供选择.

1) 依赖 CV 的定向降维. 在不具备先验数据时, 依据直觉猜测少量有物理意义且可能重要的坐标, 即集合变量 (collective variable, CV), 强行定向降维到该预选的低维 CV 空间, 而后在 CV 空间内搜寻过渡态^[79-95]. 代表性方法: 温和爬升动力学 (gentlest ascent dynamics, GAD)^[79-81]、有限温度弦方法 (finite temperature string, FTS)^[82-87]、快速断层扫描法 (fast tomographic, FT)^[88-90]、基于旅行商的路径搜索 (travelling-salesman based automated path searching, TAPS)^[91-95].

2) 非 CV 依赖的高维搜索. 事先不降维, 坚持在高维空间内完成采样和过渡态搜索过程, 事后再进行降维分析^[96-101]. 代表性方法有过渡路径采样 (transition path sampling, TPS)^[98-101].

尽管上述算法已在一定范围内取得成功, 但在面对复杂生物分子时, 仍面临诸多限制. 其中, 对于依赖 CV 的搜索算法, 最直接的问题便是如何从较高维度空间中选取合适的 CV; 而对于非 CV 依

赖的路径采样算法, 则是计算资源消耗过大和有效采样率过低的问题.

近年来快速发展的机器学习及相关衍生算法 (如强化学习、生成式建模等), 已成功应用于解决诸多传统的复杂生物问题^[102-112], 如生物结构预测及生物分子相互作用的研究^[105], 或基于人工智能开发蛋白质从头设计算法^[106], 或借助于机器学习实现蛋白质结构准确预测的 trRosetta 线上服务^[107], 或实现生物分子冷冻电镜高分辨率结构重建的解析算法^[108] 和蛋白质间相互作用位点的快速预测^[109], 以及蛋白质与小分子、RNA 等复合物结构性质的预测^[110,111]. 因此, 将机器学习与现有过渡态搜索算法进行有效融合, 有望成为未来过渡态搜索研究实现进一步突破的可行方向.

本文将首先回顾依赖 CV 的过渡态搜索算法的发展历程, 厘清其基本原理及潜存问题. 随后, 聚焦于非 CV 依赖的 TPS 路径采样算法, 着重介绍其融合了强化学习的最新版本. 最后, 探讨一种新型的过渡态搜索策略, 即结合生成模型和 GAD, 在保留原高维空间过渡态信息的低维空间内实现过渡态搜索. 完整的算法总结已展示于表 1 中.

2 依赖 CV 的过渡态搜索算法

如前所述, 为了准确阐明生物分子功能性动力学的微观机制, 需要在传统采样算法的基础上, 发展可获取上述转变过程过渡态信息的过渡态搜索算法, 包括依赖 CV^[82-95] 和非 CV 依赖算法^[96-101] 两大类. 对于依赖 CV 的算法, 需在缺乏对体系的先验数据和认知的条件下, 将高维相空间 $\{x\}$ “定向降维”至少量的依据经验或直觉定义的 CV 上 (arbitrary guess). 而后续的计算采样和过渡态搜索则发生在由这些 CV 构成的低维空间 (CV1, CV2, ...) 内 (图 1(a)).

低维 CV 空间中的过渡态搜索, 依照采样开始时的已知信息可分为非路径算法和路径算法. 非路径算法以 GAD 算法为代表, 而路径算法以 finite temperature string^[82-87] 和快速断层扫描法^[88-90] 为代表. 前者可在仅有一个稳定态已知时开启过渡态搜索, 而后者需事先已知至少两个稳定态, 通过寻找两个稳定态之间的最小自由能路径 (minimum free energy path, MFEP), 而后获得沿路径的自由能分布确定过渡态位置. 此外, 两者的区别

表 1 主要过渡态搜索算法的总结分类
Table 1. Classification of the algorithms for transition state searching.

过渡态搜索算法分类	代表性算法	参考文献	备注
传统方法	依赖CV	Gentlest ascent dynamics (GAD)	[79—81] 非路径方法
		Finite temperature string	[82—87]
	预设低维	Fast tomographic	[88—90]
	空间搜索	基于旅行商的路径搜索 TAPS	[91—95] 路径方法
融合AI	不依赖CV 高维空间搜索	Transition path sampling	[98—101]
		Reinforcement path sampling	[113]
	保留过渡态信息的降维 低维空间搜索	融合生成模型及GAD的过渡态搜索(待研发)	无

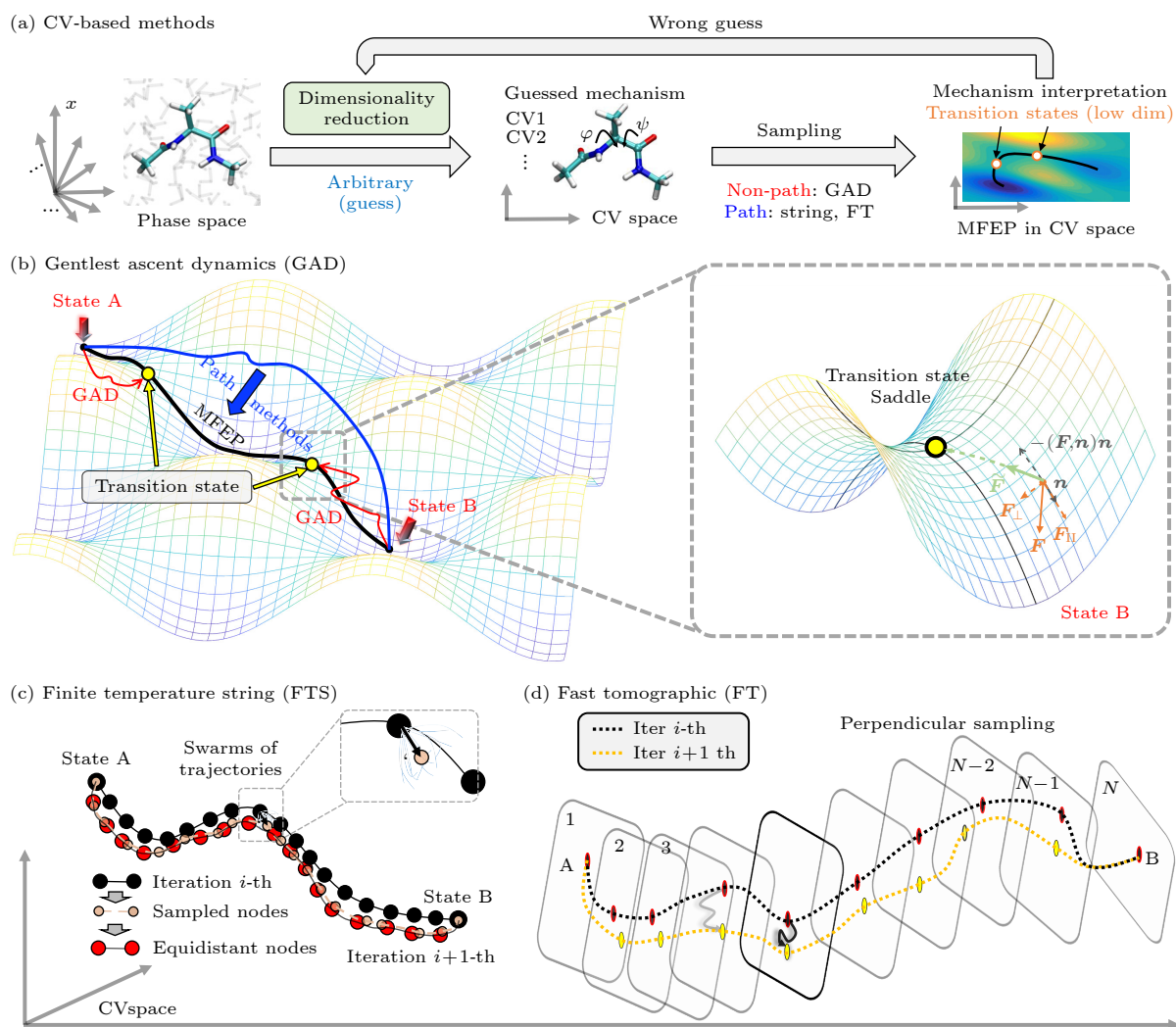


图 1 (a) 依赖集合变量的过渡态搜索示意图, 需由生物分子 (以丙戊酸二肽为例) 体系所在的高维相空间 (phase space) 选取少量集合变量 CV 强行“定向降维”, 后在此低维 CV 空间利用非路径类方法或路径方法, 找到过渡态 (Transition State), 并给出微观机制解释 (mechanism interpretation); (b) 非路径类的 GAD 算法原理示意图; (c), (d) 两类路径类搜索算法原理示意图

Fig. 1. (a) Illustration of the flow-chart of the collective variables (CVs) based transition state searching. A low dimensional space must be constructed with the CVs, which are arbitrary a priori guess about the mechanism. The transition state(s) is then determined by either the non-path or path methods. (b) The non-path method GAD. Path methods of (c) finite temperature string and (d) fast tomographic.

还有,前者采样过程是主动“爬山”(即向高能区域运动,图 1(b)左红),而后者是先通过施加外力促使分子强行翻山越岭得到能量过高的初始路径(图 1(b)左蓝),再设法使路径“整体下山”,落入附近的最优路径 MFEP (图 1(b)左黑).

2.1 非路径类过渡态搜索

GAD 是非路径类过渡态搜索的代表性算法,在预设的低维 CV 空间,从亚稳态或任意状态出发,可在低维势能面空间内,直接完成过渡态搜索^[79-81].如图 1(b)所示,此算法的原理为由低维势能面空间内的任意一点出发,根据以下规则:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) - 2\tilde{\mathbf{F}}, \quad (1a)$$

$$\gamma\dot{\mathbf{n}} = -H\mathbf{n} + (\mathbf{n}, H\mathbf{n})\mathbf{n}, \quad (1b)$$

来确定每轮迭代时移动至下一步的位移方向,即沿势能函数梯度变化率的最小方向进行小步长移动,最终收敛于鞍点位置(即过渡态).其中 $\tilde{\mathbf{F}} = (\mathbf{F}(\mathbf{x}), \mathbf{n})\mathbf{n}$, $\mathbf{F}(\mathbf{x})$ 为分子体系在根据当前低维 CV 空间内的势能梯度计算得到的作用力;而 \mathbf{n} 被设定为趋近于势能函数海森矩阵最小特征值对应的特征向量,即指向曲率最小方向,其需要基于(1b)式反复迭代达到收敛,在此期间, γ 则控制 H 对 \mathbf{n} 变化的影响能力,以此消除势能函数中的噪音.简单而言,(1)式的规则将引导分子不断沿势能坡度最缓的方向逆势攀登,直至收敛停滞于过渡态.

2.2 基于路径优化的过渡态搜索

对于基于路径优化进行过渡态搜索的算法,根据其输入不同,可主要分为两类:1)需要高质量预选集合变量 CV 的路径优化算法,包括 finite temperature string^[82-87]和快速断层扫描法^[88-90];2)基于路径集合变量(path collective variable, PCV)的路径优化算法,即基于 TAPS 算法^[91-95],此方法中避免了高质量预选集合变量的困境,可高效且快速找到最优转变路径.当构建完路径优化的低维空间后,需要从目标系统的两个稳定态结构出发,产生一条较为粗糙的转变路径^[114-116],而后对此路径进行迭代优化(路径整体下山),并最终收敛于最优路径(MFEP)^[82-95];继而便可通过计算 MFEP 的自由能图景,准确给出微观转变机制和过渡态信息^[57-59,117].

2.2.1 Finite Temperature String

当基于传统的增强采样算法(如 steered MD, climber MD, targeted MD 等^[114-116])快速得到描述目标生物分子过程的转变路径后,前人发现还需要通过选取合适的集合变量信息,来构建低维空间和完成对初始转变路径的进一步优化,从而得到最优路径,即最小自由能路径(minimum free energy path, MFEP).作为研究此类问题中的代表算法,finite temperature string 的优化策略^[82-87]较为简洁(以 swarms-of-trajectories 版本为例^[87]),见图 1(c).通过对连接转变路径(由 State A 到 State B)的所有节点,依次分别完成大量(swarms)非常短时的随机初始速率 MD 采样后,在预选的低维空间对采样结果聚类,找到出现概率最高的构象,作为代表性的采样节点(图 1(c)中 sampled node).这样做是为了在路径上各节点附近做非常局部的采样,从而估计各节点目前所在位置的自由能梯度,等效于让各节点沿着当前所在位置的自由能梯度最大方向稍作移动(下山),类似于势能最小化问题中的最速下降法;通过再优化节点分布来保证相邻节点间距离相近(equidistant nodes,图 1(c)),进而得到新一轮的转变路径.

通过不断重复上述迭代策略,路径将最终收敛到达最小自由能路径 MFEP.最终便可通过伞形采样等^[117]方法获取沿此 MFEP 的自由能景观(free energy landscape)^[82-87],进而给出微观机制解释和得到相应的过渡态信息.

2.2.2 快速断层扫描法

快速断层扫描法与前述的 finite temperature string 方法较为相似,亦需基于经验或随机预选集合变量来构建低维空间^[88-90],而后在此低维空间进行路径搜索,找到 MFEP,如图 1(d)所示:

首先,在选定的低维度空间内,均匀选取转变构象(每个构象称为节点,共 N 个节点)来代表初始转变路径(由 State A 到 State B);随后,对于每个节点,都在垂直于当前路径的超平面空间内进行相同时长的 MD 模拟采样,在采样过程中还需引入 SHAKE 算法^[118]以避免其离超平面空间过远,同时,结合自适应偏势 MD 方法(adaptively biased molecular dynamics, ABMD)^[119]来提高其采样效率;接着,针对每个节点的采样轨迹,直接将采样的终态结构进行连接,保存为新的转变路径

(如图 1(d) 中黑色虚线代表的第 i 轮结果和黄色虚线代表的第 $i+1$ 轮结果). 按照上述流程反复迭代, 将最终得到 MFEP, 及相应自由能景观分布, 从而阐明其微观转变机制并确定目标过渡态信息.

2.2.3 基于旅行商的自动化路径搜索算法

在基于集合变量的搜索算法中, 还存在一种基于路径集合变量 PCV 的新型算法^[120], 即基于旅行商问题的自动路径搜索算法 (TAPS). TAPS 巧妙地避开了其他路径优化算法中集合变量的选取问题, 同时基于并行化和 GPU 加速, 快速得到较高维度空间中的最优路径 (MFEP), 给出相应的微观转变机制和过渡态信息 (图 2)^[91-95].

具体来讲, 在使用 TAPS 方法时, 需提供目标生物系统的两个稳态结构和连接其转变过程的初始路径; 而后从初始路径中确定转变过程中变化较大的所有结构域, 并以这些结构域的重原子 (图 2(a) 中丙戊酸二肽结构中以球形显示的原子) 为参考, 通过计算构象间均方根位移偏差 (root mean square distance, RMSD) 来评估构象差异, 并从初始路径中在保证相邻构象间适度的差异基础上, 均匀选取构象 (即节点) 来代表整个转变过程; 接着, 基于此少量节点组成的转变路径, 便可利用 PCV 的计算公式得到二维的路径集合变量低维空间: 即 PCV- s 和 PCV- z . 其中, 对于任意构象 x , 参照目标路径计算得到的 PCV- s 代表其沿路径

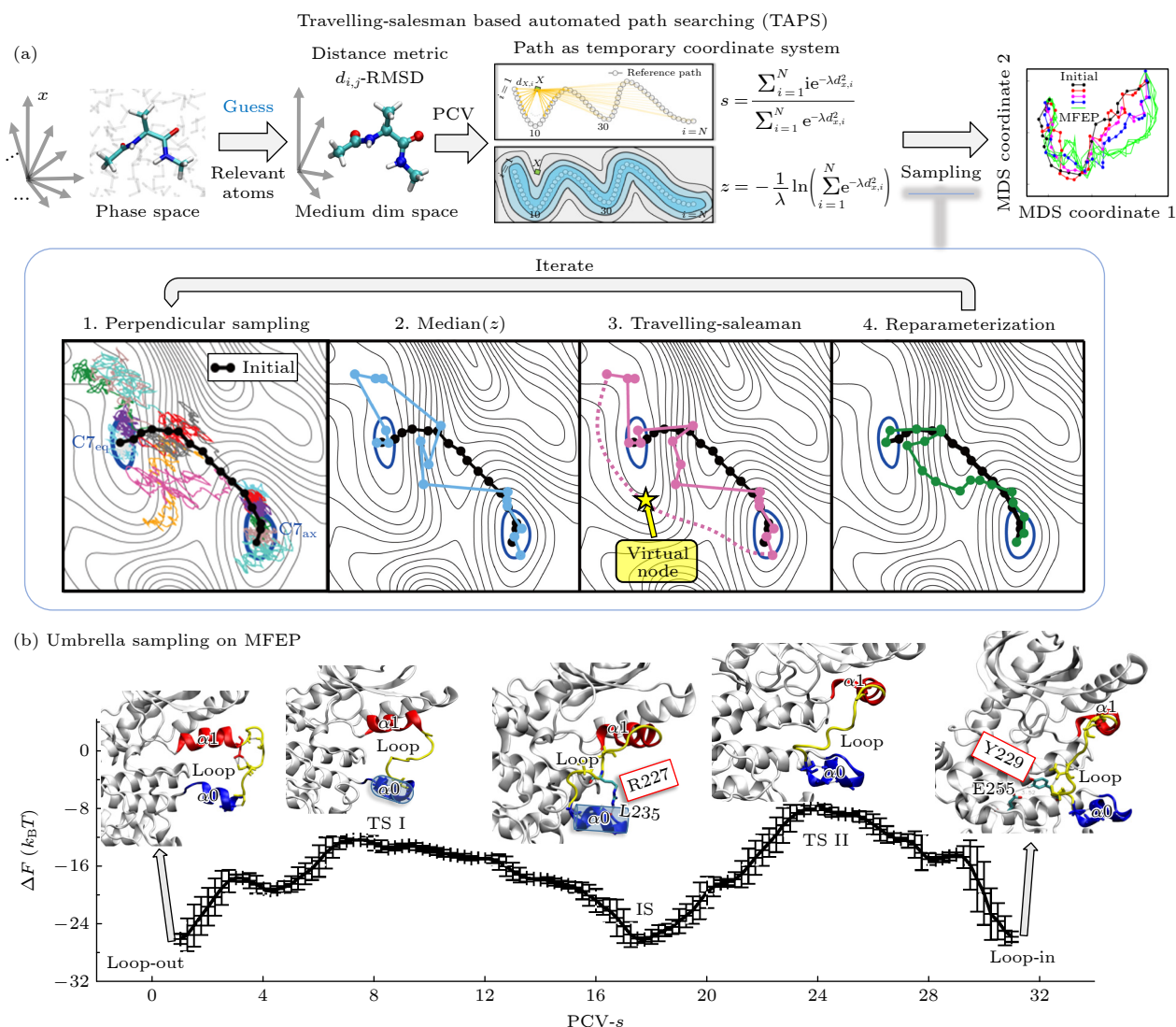


图 2 (a) PCV 构建^[120]和 TAPS Method^[91-95,121]算法原理示意图; (b) 基于伞形采样方法得到的 TAPS 算法确定的 MEK1 由 Loop-Out 到达 Loop-In 转变过程最小自由能路径 (MFEP) 的自由能图景及相应的微观转变机制^[92]

Fig. 2. (a) Illustration for the construction of PCV and the flow-chart of the TAPS method; (b) TAPS revealed the free energy landscape and the transition states for the transition from the Loop-Out state of MEK1 to its Loop-In state^[92].

方向的投影位置; 而 PCV- z 表示其距离参考路径的平均距离, 见图 2(a)^[120]. 通过在此路径集合变量空间内, 快速完成路径搜索, 将最终确定目标转变过程的最优路径 (MFEP), 如图 2(a) 中基于多维度标度方法 (multidimensional scaling method, MDS)^[122] 得到的二维路径搜索过程展示, 从黑色的初始路径快速搜索到达绿色的最优路径 (MFEP).

此处以丙戊酸二肽由 C7_{eq} 到 C7_{ax} 的转变为例, 完整展示 TAPS 进行路径优化的主要过程, 包括以下四步 (见图 2(a) 中下方白色框内的 TAPS 迭代流程).

步骤 1 基于转变路径节点间结构差异 ($d_{\mathbf{x}, i}$) 和节点编号 ($i = 1, 2, \dots, N$) 信息, 利用 PCV^[120] 构建路径优化的二维空间: 沿路径方向, PCV- s ((2a) 式) 和垂直于路径方向, PCV- z ((2b) 式), 而后从每个节点出发做采样, 采样时在 PCV- s 方向加入限制偏势, 阻止分子在平行于当前路径的方向运动, 但允许其在垂直于当前路径的超平面内任意运动; 同时, 为了后续步骤 4 补入节点时能有更多候选构象, 在 PCV- s 进行元动力学 (well-tempered metadynamics^[123]) 采样.

$$s = \frac{\sum_{i=1}^N i e^{-\lambda d_{\mathbf{x}, i}^2}}{\sum_{i=1}^N e^{-\lambda d_{\mathbf{x}, i}^2}}, \quad (2a)$$

$$z = -\frac{1}{\lambda} \ln \left(\sum_{i=1}^N e^{-\lambda d_{\mathbf{x}, i}^2} \right), \quad (2b)$$

步骤 2 对于每个节点的采样轨迹, 通过获取最接近轨迹 PCV- z 中位值的结构, 并按照上轮编号连接为新的转变路径 (蓝色实线).

步骤 3 经步骤 1 非局部的垂直空间采样后, 节点顺序很可能已发生改变需要重排. 本算法将节点重排转化为旅行商问题^[121], 并通过插入虚拟点 (即与其他任何节点间的距离为零) 来将旅行商问题的闭环解转化为节点顺序编号.

步骤 4 去除转变路径范围外节点, 并在距离较远的相邻节点间补入新节点.

最终, 通过不断重复迭代上述 1—4 步的路径优化过程, 将最终搜索到 MFEP 并结合伞形采样等算法^[117] 得到沿 MFEP 的自由能景观分布, 进而给出微观转变机制解释和确定相应的过渡态信息.

以 TAPS 对丝裂原激活蛋白激酶激酶 (MEK1) 由 Loop-Out 状态转变为 Loop-In 状态的研究为例 (图 2(b)), 实验发现其在传递生物信号中时需经历 Loop-Out 态到 Loop-In 态的转变, 即两个 α 螺旋 ($\alpha 0$ 和 $\alpha 1$) 的局部翻转以及连接螺旋的 Loop 进入激活口袋; 利用 TAPS 方法同时考察上述过程中涉及的所有重要残基, 在较短的采样总时间 (短于 32.6 ns) 内便得到了 MFEP (图 2(a) 最右侧的 MDS 结果内的绿色线)^[92]; 沿收敛的 MFEP 进一步得到了相应的自由能图景 (图 2(b)), 进而获得了主要转变机制和两个关键过渡态结构 (TS I 和 II). 此研究所新发现的 R227:L235 及 Y229:E255 极性接触作用, 也被成功用于解释实验关于 R227 或 Y229 的点突变造成 MEK1 无法激活的现象^[124,125].

尽管 TAPS 算法巧妙地规避了预选 CV 空间定向降维带来的试错成本, 但仍需选择计算 RMSD 所需的原子集作为输入信息. 这意味着在复杂大分子的过渡态搜索中, 即便 TAPS 的整体效率相比依赖 CV 的方法已有大幅提升, 它仍在事先对所研究构象变化的机制做出了一定假设.

3 基于路径采样的过渡态搜索

目前所有算法中, 只有以 TPS 为代表的路径采样方法在事先对构象变化机制未作任何假设, 因为 TPS 将构象转变路径直接定义在了高维相空间内. 传统 TPS 通过大量随机的不外加偏执势的无偏采样, 得到一个过渡路径系综 (transition path ensemble, TPE), 见图 3(a). 最终通过对 TPE 的后处理分析, 选取合适的集合变量以描述过渡态^[98-101] (图 3(b) 左); 最近, 通过引入强化学习范式 (reinforcement learning), 该方法实现了自适应无偏采样 (图 3(b) 右), 并采用符号回归 (symbolic regression) 完成机制解析^[113,126].

3.1 过渡路径采样

3.1.1 相空间中过渡态的定义 committor probability

由于 TPS 中的路径直接定义在相空间, 相应地过渡态也无法直接套用低维空间中的鞍点 (saddle) 来具象地表征. 假设我们能通过某些 CV

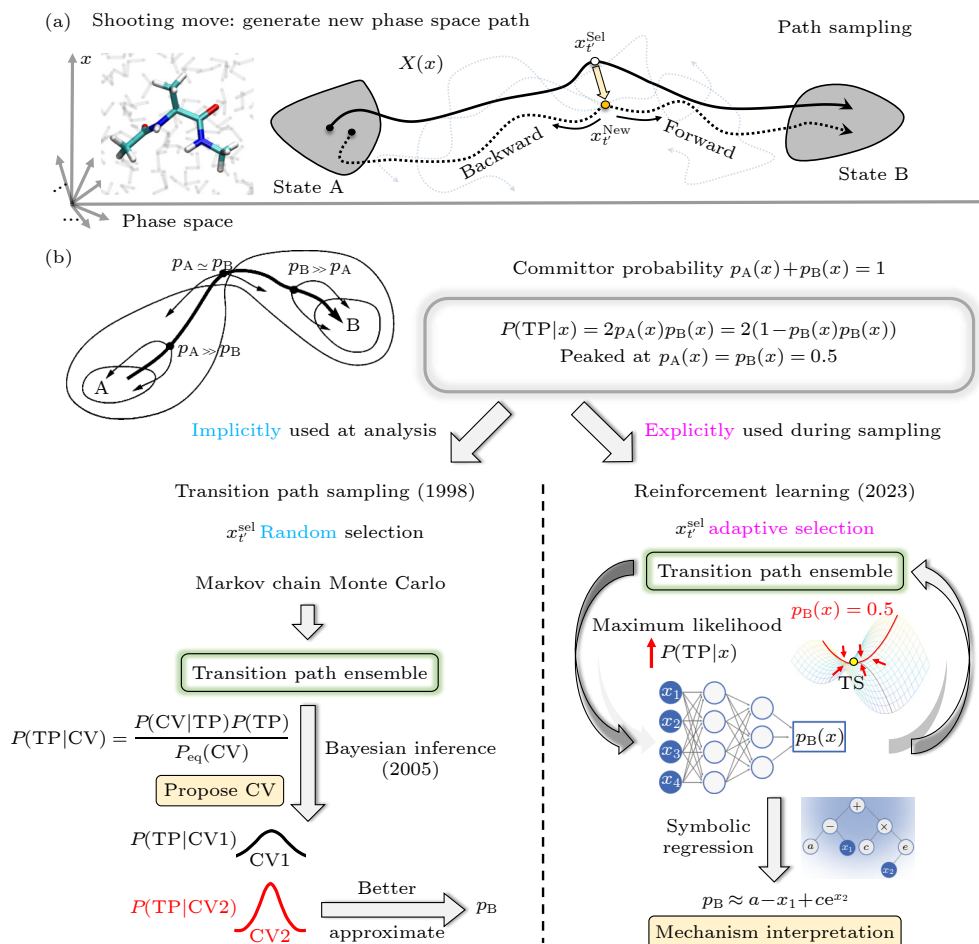


图 3 路径采样算法的基本原理示意图 (a) 路径采样中生成新相空间路径的 shooting move; (b) 传统过渡路径采样 (左侧) 的随机蒙特卡罗采样与过渡态分析原理^[98-101], 融合强化学习的路径采样 (右侧) 在学习过程中不断促进采样起始点选择向过渡态集中^[113]

Fig. 3. Schematics of path sampling methods. (a) Shooting move: select a phase space point on the current path, make a small perturbation to this point (redraw random initial velocities) and perform a set of simulations. (b) Path sampling is built upon the committor probability p_B . The traditional transition path sampling (left)^[98-101] selects shooting points randomly and uses Monte Carlo for sampling; the transition state is characterized through post-analysis: choosing the CVs with the highest and narrowest distribution of $P(TP|CV)$; the new reinforcement path sampling (right)^[113] chooses shooting points adaptively and directly learns the committor probability p_B with maximized $P(TP|x)$. Symbolic regression of p_B is used for mechanism interpretation.

定义出两个稳定态 A 和 B (并同时假设 A 和 B 中间不存在第 3 个稳定态 C), 那么 A 和 B 之间的过渡态就能通过 committor probability 来定义.

对相空间中的任一点, 都可以从其出发运行大量 MD 模拟并统计其中有多少比率分子是在抵达稳态 B 之前到达了 A, 另有多少比率相反在到达了 A 之前抵达了 B. 这两种比率 p_A 和 p_B 就是这一点对稳态 A 和 B 的 committor probability. 显然在不存在第 3 个稳态的前提下 $p_A + p_B = 1$. 相应地, 过渡态则可以定义为由相空间内所有 $p_A = p_B = 0.5$ 的点所组成的集合. 同时, 依据过渡路径理论 (transition path theory)^[96], 我们知道对相空间中的任一点 x 而言, 它是属于连接 A 和 B 反应

路径, 即过渡路径 (transition path, TP) 的其中一点的条件概率是

$$P(TP|x) = 2p_A(x)p_B(x) = 2(1 - p_B(x))p_B(x). \quad (3)$$

而此条件概率在过渡态上 $p_A = p_B = 0.5$ 时将达到其峰值, 即过渡态上的点是所有相空间中最有可能属于某条反应路径的. 这一点对路径采样算法至关重要.

3.1.2 Shooting move 新相空间路径的生成

假设已利用传统增强采样算法 (如 climber method/steered MD/targeted MD 等^[114-116]) 得到一条连接 A 到 B 的转变路径, 便可以在此转变路径中抽选一个点 x^{sel} ; 随后, 对 x^{sel} 做出微扰 Δx

(典型做法为根据给定温度的麦克斯韦-玻尔兹曼随机重置所有分子的初始速率),而后以 $\mathbf{x}^{\text{new}} = \mathbf{x}^{\text{sel}} + \Delta\mathbf{x}$ 为新的初始条件进行多次无偏 MD 模拟采样. 其中,每次 MD 模拟采样的终止条件为此采样路径到达了目标态 A 或 B 中的一个;当这些轨迹中既有到达过 A 也有到达过 B 态时,将到达过 A 态的任意路径和到达过 B 态的任意路径连接便成为由 A 态到达 B 态的转变路径. 该过程被称为 shooting move (图 3(a))^[127].

路径采样过程就是不断迭代选定 \mathbf{x}^{sel} ,而后进行 Shooting 的过程. 经过迭代最终会得到从 A 到 B 转变的路径系综 TPE^[128,129]. 但传统 TPS 和其强化学习新版本在 \mathbf{x}^{sel} 的选择策略上有所不同.

3.1.3 过渡路径采样的 shooting move 策略

在原版 TPS 中, \mathbf{x}^{sel} 的选择是完全随机的. 同时, shooting move 的迭代是马尔科夫链蒙特卡罗的串行过程 (图 3(b) 左). 因此, TPS 天然欠缺并行化能力.

3.1.4 从路径系综中提取过渡态信息

经 shooting move 迭代得到路径系综后,传统 TPS 需要用户自行定义 CV 来帮助解释其中蕴含的机制、提取过渡态信息. 根据 (3) 式,如果所选的 CV 能够较好地表征过渡态,即无限趋近 p_B ,那么 $P(\text{TP}|\text{CV})$ 应该呈现窄而高的分布. 但由于 $P(\text{TP}|\text{CV})$ 无法直接计算,需要通过贝叶斯推测间接计算:

$$P(\text{TP}|\text{CV}) = \frac{P(\text{CV}|\text{TP})P(\text{TP})}{P_{\text{eq}}(\text{CV})}, \quad (4)$$

其中 $P(\text{CV}|\text{TP})$ 可直接从 TPE 计算获得, $P(\text{TP})$ 需经额外长时间无偏采样算出,而 $P_{\text{eq}}(\text{CV})$ 是 CV 上的平衡态分布,也需通过额外的伞形采样获得. 在用户选择的 CV 中,以 $P(\text{TP}|\text{CV})$ 分布最窄最高者最能表征过渡态和 A 到 B 的转变机制^[98-101].

3.2 基于强化学习的路径采样

仔细分析原版 TPS 的后处理分析过程,不难看出其对蒙特卡罗迭代采样结果的要求较高,需确保所得 TPE 在过渡态附近有充足样本,但由于其 \mathbf{x}^{sel} 的选择是完全随机,这在面临较大的生物分子体系时是难以实现的.

因此, Jung 等^[113] 于近期开发了基于强化学习 (reinforcement learning) 的路径采样算法. 与原

版 TPS 仅在数据处理分析阶段隐性地使用 (4) 式不同,新框架直接将 $P(\text{TP}|\mathbf{x})$ 用作了强化学习中的目标函数 (通过最大似然估计将其最大化),用以训练以神经网络表达的 committor probability p_B (图 3(b) 右). 因此,在此强化学习过程中, $P(\text{TP}|\mathbf{x})$ 的最大化意味着算法会自适应地选择 \mathbf{x}^{sel} ,自发将其聚焦至过渡态附近 (即 $p_B = 0.5$, 图 3(b) 红线).

而后续对转变机制的解释,即神经网络 p_B 物理含义的挖掘则可通过符号回归 (symbolic regression) 达成,将 $p_B(\mathbf{x})$ 的神经网络表达为容易理解的简单解析式^[125,126].

3.3 路径采样算法的适用场景

值得强调的是,无论是传统 TPS 还是强化学习路径采样,二者的理论基础都是 $p_A + p_B = 1$,即不允许稳态 A 和 B 之间有第 3 个稳定态存在. 这意味着路径采样只能处理单个能垒,即只能表征单个过渡态. 然而,生物大分子的运动复杂,亚稳态数量众多,很难保证已知的两个稳定态之间只有一个能垒. 这也限制了路径采样在生物大分子模拟中的应用.

4 融合 GAD 与降维算法的可能方案

经过对上述算法的简单回顾,可以看出近年来依赖 CV 的路径搜索算法和非 CV 依赖的路径采样算法都已呈现与计算机科学和机器学习算法深度融合迈向自动化的发展趋势,但依赖 CV 的 GAD 方法尚无相似案例可循. 我们推测一个可能的发展方向是将 GAD 在低维空间搜索过渡态的能力与降维算法结合起来. 自然地,这对降维算法的性能提出了新的要求. 因此,有必要先对现有降维算法的设计思想进行简要梳理.

4.1 现有降维算法

降维是无监督机器学习的传统分支,其在生物分子模拟中的广泛应用已有综述阐明^[130],此处不再赘述. 但在目前众多的降维算法中,显式利用时间序列信息,即动力学信息,进行降维的仅有时间结构独立成分分析 (time-lagged independent components analysis, tICA) 方法^[63-65]. 但经 tICA 降维所得的低维 tIC 空间已被限定只能是原高维空间的线性组合,而能够表征跃迁过程和过渡态的

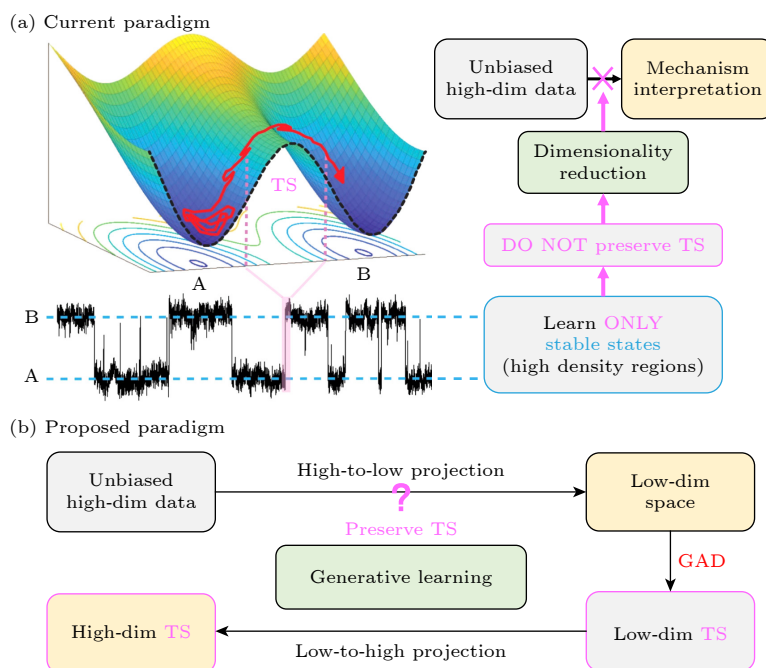


图4 物理化学家需要怎样的降维算法 (a) 现有降维算法范式不保留过渡态信息, 不利于机制解析; (b) 可能的替代范式, 基于生成模型研发可保留过渡态信息的可逆降维算法, 并与低维空间搜索过渡态的GAD联用

Fig. 4. Requirements on dimensionality reduction algorithms by physical chemists. (a) Current paradigm for dimensionality reduction and the main difficulties for the transition state searching. (b) Proposed alternative paradigm for transition state searching: combine dimensionality reduction that preserves transition state information with GAD.

坐标很可能是原高维坐标的非线性函数. 其他现存降维算法, 因在降维过程中, 只关注保留高密度区域信息 (即稳定态信息), 常会将高维空间过度扭曲以致过渡态信息丢失 (图4(a)). 因此, 现存降维算法都无法与GAD联用.

4.2 基于生成模型的可逆降维及过渡态搜索

近年来, 可逆神经网络和生成模型的发展, 为研发能够保留过渡态信息的新型降维算法提供了良好契机. 首先, 通过可逆神经网络, 我们可以期望利用深度学习训练出一个可以进行双向映射的生成模型, 即在将高维的全原子轨迹信息映射到某一低维空间的同时, 拥有把生成的低维空间样本逆投影回原空间的能力. 这样便可利用GAD在低维空间搜得鞍点结构, 再经逆投影自动得到完整的高维过渡态结构.

当然, 这一构想的实现难点是必须保证在降维过程中, 低维空间保有和原高维空间一致的动力学特征以及概率密度信息, 即保留过渡态信息. 这里我们建议参考tICA中直接使用动力学信息进行降维的做法. 此外, 为保障GAD在低维空间的顺

利运行, 该生成模型应能为低维空间自动拟合出连续可导的自由能面.

5 结论

生物分子功能机制的有效调控有赖于对其转变过程微观机制的全面考察, 其中以获取其主要转变路径中的过渡态信息最为关键. 当预设静态集合坐标较为容易、可强行定向降维时, 前人开发的GAD算法、finite temperature string和快速断层扫描法, 已成功阐明了诸多生物过程的微观转变机制, 但当面对复杂转变过程时, 仍易出现预设集合变量常不合理, 需要消耗大量资源试错. 近年出现的基于旅行商的自动路径搜索算法TAPS, 则有效避免了集合变量的预设问题, 还在并行化和GPU加速的基础上, 提升了自动化程度和过渡态搜索效率.

在完全无需事前降维、不依赖集合变量的路径采样类算法中, 也已出现了通过融入强化学习思想实现自适应的高效率采样及过渡态分析优秀变体. 但只能处理单个能垒和过渡态搜寻的特点限制了这类算法在生物分子模拟中的应用.

因此, 研发可保留过渡态信息的新型降维算法

或是将机器学习进一步融入过渡态搜索的可行方向。在此,我们建议基于生成模型研发此种高质量降维方法,并将之与 GAD 联用,从而做到从任意状态出发,快速捕捉其周围的过渡态信息。

参考文献

- [1] Edman L, Földes-Papp Z, Wennmalm S, Rigler R 1999 *Chem. Phys.* **247** 11
- [2] Evenäs J, Malmendal A, Thulin E, Carlström G, Forsén S 1998 *Biochemistry* **37** 13744
- [3] Hanson J A, Duderstadt K, Watkins L P, Bhattacharyya S, Brokaw J B, Chu J W, Yang H 2007 *Proc. Natl. Acad. Sci. USA* **104** 18055
- [4] Moffat K 1989 *Annu. Rev. Biophys. Chem.* **18** 309
- [5] Huang C, Kalodimos C G 2017 *Annu. Rev. Biophys. Chem.* **46** 317
- [6] Weissenberger G, Henderikx R J M, Peters P J 2021 *Nat. Methods* **18** 463
- [7] Clegg R M 1995 *Curr. Opin. Biotechnol.* **6** 103
- [8] Karplus M, McCammon J A 2002 *Nat. Struct. Biol.* **9** 646
- [9] Hollingsworth S A, Dror R O 2018 *Neuron* **99** 1129
- [10] Bernèche S, Roux B 2001 *Nature* **414** 73
- [11] Khafizov K, Perez C, Koshy C, Quick M, Fendler K, Ziegler C, Forrest L R 2012 *Proc. Natl. Acad. Sci. USA* **109** E3035
- [12] Li J, Shaikh S A, Enkavi G, Wen P C, Huang Z, Tajkhorshid E 2013 *Proc. Natl. Acad. Sci. USA* **110** 7696
- [13] Dror, R O, Green H F, Valant C, Borhani D W, Valcourt J R, Pan A C, Arlow D H, Canals M, Lane J R, Rahmani R, Baell J B, Sexton P M, Christopoulos A, Shaw D E 2013 *Nature* **503** 295
- [14] Wacker D, Stevens R C, Roth B L 2017 *a Cell* **170** 414
- [15] Wacker D, Wang S, McCorvy J D, Betz R M, Venkatakrishnan A J, Levit A, Lansu K, Schools Z L, Che T, Nichols D E, Dror R O, Roth B L 2017 *Cell* **168** 377
- [16] McCorvy J D, Butler K V, Kelly B, Rechsteiner K, Karpiak J, Betz R M, Kormos B L, Shoichet B K, Dror R O, Jin J, Roth B L 2018 *Nat. Chem. Biol.* **14** 126
- [17] Provasi D, Artacho M C, Negri A, Mobarec J C, Filizola M 2011 *PLoS Comput. Biol.* **7** e1002193
- [18] Cordero-Morales J F, Jogini V, Lewis A, Vásquez V, Cortes D M, Roux B, Perozo E 2007 *Nat. Struct. Mol. Biol.* **14** 1062
- [19] Fields J B, Németh-Cahalan K L, Freitas J A, Vorontsova I, Hall J E, Tobias D J 2017 *J. Biol. Chem.* **292** 185
- [20] Groban E S, Narayanan A, Jacobson M P 2006 *PLoS Comput. Biol.* **2** e32
- [21] Liu Y, Ke M, Gong H 2015 *Biophys. J.* **109** 542
- [22] Delemotte L, Tarek M, Klein M L, Amaral C, Treptow W 2011 *Proc. Natl. Acad. Sci. USA* **108** 6109
- [23] Lindorff-Larsen K, Piana S, Dror R O, Shaw D E 2011 *Science* **334** 517
- [24] Snow C D, Nguyen H, Pande V S, Gruebele M 2002 *Nature* **420** 102
- [25] Dror R O, Arlow D H, Maragakis P, Mildorf T J, Pan A C, Xu H, Borhani D W, Shaw D E 2011 *Proc. Natl. Acad. Sci. USA* **108** 18684
- [26] Dror R O, Pan A C, Arlow D H, Borhani D W, Maragakis P, Shan Y, Xu H, Shaw D E 2011 *Proc. Natl. Acad. Sci. USA* **108** 13118
- [27] Gu Y, Shrivastava I H, Amara S G, Bahar I 2009 *Proc. Natl. Acad. Sci. USA* **106** 2589
- [28] Latorraca N R, Fastman N M, Venkatakrishnan A J, Frommer W B, Dror R O, Feng L 2017 *Cell* **169** 96
- [29] Stelzl L S, Fowler P W, Sansom M S, Beckstein O 2014 *J. Mol. Biol.* **426** 735
- [30] Buch I, Giorgino T, De Fabritiis G 2011 *Proc. Natl. Acad. Sci. USA* **108** 10184
- [31] Liang R, Swanson J M J, Madsen J J, Hong M, DeGrado W F, Voth G A 2016 *Proc. Natl. Acad. Sci. USA* **113** E6955
- [32] Suomivuori C M, Gamiz-Hernandez A P, Sundholm D, Kaila V R I 2017 *Proc. Natl. Acad. Sci. USA* **114** 7043
- [33] Tajkhorshid E, Nollert P, Jensen M Ø, Miercke L J W, O'Connell J, Stroud R M, Schulten K 2002 *Science* **296** 525
- [34] Watanabe A, Choe S, Chaptal V, Rosenberg J M, Wright E M, Grabe M, Abramson J 2010 *Nature* **468** 988
- [35] Dedmon M M, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson C M 2005 *J. Am. Chem. Soc.* **127** 476
- [36] Nguyen H D, Hall C K 2004 *Proc. Natl. Acad. Sci. USA* **101** 16180
- [37] Levitt M 1983 *J. Mol. Biol.* **168** 595
- [38] Sugita Y, Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141
- [39] Rhee Y M, Pande V S 2003 *Biophys. J.* **84** 775
- [40] Zhang W, Wu C, Duan Y 2005 *J. Chem. Phys.* **123** 154105
- [41] Zhou R 2006 *Protein Folding Protocols* (Humana Totowa, NJ: Springer) pp205–223
- [42] Sindhikara D, Meng Y, Roitberg A E 2008 *J. Chem. Phys.* **128** 024103
- [43] Buchete N V, Hummer G 2008 *Phys. Rev. E* **77** 030902
- [44] Rosta E, Hummer G 2009 *J. Chem. Phys.* **131** 165102
- [45] Stelzl L S, Hummer G 2017 *J. Chem. Theory Comput.* **13** 3927
- [46] Yang L J, Gao Q Y 2009 *J. Chem. Phys.* **131** 214109
- [47] Yang L, Liu C W, Shao Q, Zhang J, Gao Y Q 2015 *Acc. Chem. Res.* **48** 947
- [48] Yang Y I, Zhang J, Che X, Yang L J, Gao Y Q 2016 *J. Chem. Phys.* **144** 094105
- [49] Yang Y I, Shao Q, Zhang J, Yang L J, Gao Y Q 2019 *J. Chem. Phys.* **151** 070902
- [50] Huber T, Torda A E, Van Gunsteren W F 1994 *J. Comput. -Aided Mol. Des.* **8** 695
- [51] Wada T, Kuroda K, Yoshida Y, Ogasawara K, Ogawa A, Endo S 2006 *Neurosurg. Rev.* **29** 242
- [52] Hansen H S, Hünenberger P H 2010 *J. Comput. Chem.* **31** 1
- [53] Perić-Hassler L, Hansen H S, Baron R, Hünenberger P H 2010 *Carbohydr. Res.* **345** 1781
- [54] Grubmüller H 1995 *Phys. Rev. E* **52** 2893
- [55] Schulze B G, Grubmüller H, Evanseck J D 2000 *J. Am. Chem. Soc.* **122** 8700
- [56] Bouvier B, Grubmüller H 2007 *Biophys. J.* **93** 770
- [57] Barducci A, Bonomi M, Parrinello M 2011 *WIREs Comput. Mol. Sci.* **1** 826
- [58] Tiwary P, Parrinello M 2013 *Phys. Rev. Lett.* **111** 230602
- [59] Bussi G, Laio A 2020 *Nat. Rev. Phys.* **2** 200
- [60] Miao Y, Feher V A, McCammon J A 2015 *J. Chem. Theory Comput.* **11** 3584
- [61] Miao Y, McCammon J A 2017 *Annu. Rep. Comput. Chem.* **13** 231
- [62] Wang J, Arantes P R, Bhattarai A, et al. 2021 *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **11** e1521
- [63] Naritomi Y, Sotaro F 2011 *J. Chem. Phys.* **134** 065101
- [64] Schwantes C R, Vijay S P 2013 *J. Chem. Theory Comput.* **9** 2000
- [65] Perez-Hernandez G, Paul F, Giorgino T, Fabritiis G D, Noé

- F 2013 *J. Chem. Phys.* **139** 015102
- [66] Bowman G R, Huang X H, Pande V S 2009 *Methods* **49** 197
- [67] Metzner P, Noé F, Schütte C 2009 *Phys. Rev. E* **80** 021106
- [68] Pande V S, Beauchamp K A, Bowman G R 2010 *Methods* **52** 99
- [69] Prinz J H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera J D, Schütte C, Noé F 2011 *J. Chem. Phys.* **134** 174105
- [70] Kellogg E H, Lange O F, Baker D 2012 *J. Phys. Chem. B* **116** 11405
- [71] Yao Y, Cui R Z, Bowman G R, Silva D A, Sun J, Huang X H 2013 *J. Chem. Phys.* **138** 174106
- [72] McGibbon R T, Schwantes C R, Pande V S 2014 *J. Phys. Chem. B* **118** 6475
- [73] Nuske F, Keller B G, Pérez-Hernández G, Mey A S J, Noé F 2014 *J. Chem. Theory Comput.* **10** 1739
- [74] Sheong F K, Silva D A, Meng L, Zhao Y, Huang X H 2015 *J. Chem. Theory Comput.* **11** 17
- [75] Zhu L Z, Sheong F K, Zeng X, Huang X H 2016 *Phys. Chem. Chem. Phys.* **18** 30228
- [76] Wang W, Cao S, Zhu L Z, Huang X H 2018 *WIREs Comput. Mol. Sci.* **8** e1343
- [77] Husic B E, Pande V S 2018 *J. Am. Chem. Soc.* **140** 2386
- [78] Konovalov K A, Unarta I C, Cao S, Goonetilleke E C, Huang X H 2021 *J. Am. Chem. Soc. Au.* **1** 1330
- [79] E W, Zhou X 2011 *Nonlinearity* **24** 1831
- [80] Samanta A, Chen M, Yu T Q, Tuckerman M E 2014 *J. Chem. Phys.* **140** 164109
- [81] Chen M, Yu T Q, Tuckerman M E 2015 *Proc. Natl. Acad. Sci. USA* **112** 3235
- [82] E W, Ren W, Vanden-Eijnden E 2002 *Phys. Rev. B* **66** 052301
- [83] E W, Ren W, Vanden-Eijnden E 2005 *J. Phys. Chem. B* **109** 6688
- [84] Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G 2006 *J. Chem. Phys.* **125** 024106
- [85] Ren W, Vanden-Eijnden E 2007 *J. Chem. Phys.* **126** 164103
- [86] Maragliano L, Vanden-Eijnden E 2007 *Chem. Phys. Lett.* **446** 182
- [87] Pan A C, Sezer D, Roux B 2008 *J. Phys. Chem. B* **112** 3432
- [88] Chen C, Huang Y, Xiao Y 2012 *Phys. Rev. E* **86** 031901
- [89] Chen C, Huang Y, Ji X, Xiao Y 2013 *J. Chem. Phys.* **138** 164122
- [90] Chen C J, Huang Y Z, Jiang X W, Xiao Y 2014 *J. Chem. Phys.* **141** 154109
- [91] Zhu L Z, Sheong F K, Cao S, Liu S, Unarta I C, Huang X H 2019 *J. Chem. Phys.* **150** 124105
- [92] Xi K, Hu Z, Wu Q, Wei M, Qian R, Zhu L Z 2021 *J. Chem. Theory Comput.* **17** 5301
- [93] Wang L, Xi K, Zhu L Z, Da L T 2022 *J. Chem. Inf. Model.* **62** 3213
- [94] Xi K, Zhu L Z 2022 *Int. J. Mol. Sci.* **23** 14628
- [95] Xi K, Zhu L Z 2023 *A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules* (AIP Publishing) pp9-1-9-24
- [96] Vanden-Eijnden E 2006 *Computer Simulations in Condensed Matter: From Materials to Chemical Biology* (Berlin: Springer) pp453-493
- [97] Vanden-Eijnden E 2010 *Annu. Rev. Phys. Chem.* **61** 391
- [98] Dellago C, Bolhuis P G, Csajka F S, Chandler D 1998 *J. Chem. Phys.* **108** 1964
- [99] Bolhuis P G, Dellago C, Chandler D 1998 *Faraday Discuss.* **110** 421
- [100] Dellago C, Bolhuis P G, Chandler D 1999 *J. Chem. Phys.* **110** 6617
- [101] Dellago C, Bolhuis P G, Geissler P L 2002 *Adv. Chem. Phys.* **123** 1
- [102] Noé F, Tkatchenko A, Müller K R, Clementi C 2020 *Annu. Rev. Phys. Chem.* **71** 361
- [103] AlQuraishi M, Sorger P K 2021 *Nat. Methods* **18** 1169
- [104] Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S, Yang L 2021 *Nat. Rev. Phys.* **3** 422
- [105] Ju F, Zhu J, Shao B, Kong L, Liu T Y, Zheng W M, Bu D 2021 *Nat. Commun.* **12** 2535
- [106] Huang B, Xu Y, Hu X H, Liu Y R, Liao S H, Zhang J H, Huang C D, Hong J J, Chen Q, Liu H Y 2022 *Nature* **602** 523
- [107] Du Z, Su H, Wang W, Ye L, Wei H, Peng Z, Anishchenko I, Baker D, Yang J 2021 *Nat. Prot.* **16** 5634
- [108] Dai M, Dong Z, Xu K, Zhang Q C 2023 *J. Mol. Biol.* **435** 168059
- [109] Yuan Q M, Chen J W, Zhao H Y, Zhou Y Q, Yang Y D 2022 *Bioinformatics* **38** 125
- [110] Su M Y, Feng G, Liu Z, Li Y, Wang R 2020 *J. Chem. Inf. Model.* **60** 1122
- [111] Zeng C, Jian Y, Vosoughi S, Zeng C, Zhao Y 2023 *Nat. Commun.* **14** 1060
- [112] Noé F, Olsson S, Köhler J, Wu H 2019 *Science* **365** 6457
- [113] Jung H, Covino R, Arjun A, Leitold C, Dellago C, Bolhuis P G, Hummer G 2023 *Nat. Comput. Sci.* **3** 334
- [114] Weiss D R, Levitt M 2009 *J. Mol. Biol.* **385** 665
- [115] Isralewitz B, Gao M, Schulten K 2001 *Curr. Opin. Struct. Biol.* **11** 224
- [116] Schlitter J, Engels M, Krüger P 1994 *J. Mol. Graph.* **12** 84
- [117] Torrie G M, Valleau J P 1977 *J. Comput. Phys.* **23** 187
- [118] Ryckaert J P, Ciccotti G, Berendsen H J 1977 *J. Comput. Phys.* **23** 327
- [119] Babin V, Roland C, Sagui C 2008 *J. Chem. Phys.* **128** 134101
- [120] Branduardi D, Gervasio F L, Parrinello M 2007 *J. Chem. Phys.* **126** 054103
- [121] Applegate D L, Bixby R E, Chvátal V, Cook W J 2011 *The Traveling Salesman Problem: A Computational Study* (Princeton University Press) pp1-58
- [122] Cox M A A, Cox T F 2008 *Handbook of Data Visualization* (Berlin: Springer) pp315-347
- [123] Barducci A, Bussi G, Parrinello M 2008 *Phys. Rev. Lett.* **100** 020603
- [124] Fischmann T O, Smith C K, Mayhood T W, Myers J E, Reichert J P, Mannarino A, Carr D, Zhu H, Wong J, Yang R S, Le H V, Madison V S 2009 *Biochemistry* **48** 2661
- [125] Hanrahan A J, Sylvester B E, Chang M T, et al. 2020 *Cancer Res.* **80** 4233
- [126] Schmidt M, Lipson H 2009 *Science* **324** 81
- [127] Jung H, Okazaki K, Hummer G 2017 *J. Chem. Phys.* **147** 152716
- [128] Swenson D W H, Prinz J H, Noe F, Chodera J D, Bolhuis P G 2019 *J. Chem. Theory Comput.* **15** 813
- [129] Swenson D W H, Prinz J H, Noe F, Chodera J D, Bolhuis P G 2019 *J. Chem. Theory Comput.* **15** 837
- [130] Glielmo A, Husic B E, Rodriguez A, Clementi C, Noé F, Laio A 2021 *Chem. Rev.* **121** 9722

SPECIAL TOPIC—Machine learning in biomolecular simulations

Transition state searching for complex biomolecules: Algorithms and machine learning*

Yang Jian-Yu # Xi Kun # Zhu Li-Zhe †

(Warshel Institute for Computational Biology, School of Medicine, The Chinese University of Hong Kong, Shenzhen 518172, China)

(Received 13 August 2023; revised manuscript received 9 September 2023)

Abstract

Transition state is a key concept for chemists to understand and fine-tune the conformational changes of large biomolecules. Due to its short residence time, it is difficult to capture a transition state via experimental techniques. Characterizing transition states for a conformational change therefore is only achievable via physics-driven molecular dynamics simulations. However, unlike chemical reactions which involve only a small number of atoms, conformational changes of biomolecules depend on numerous atoms and therefore the number of their coordinates in our 3D space. The searching for their transition states will inevitably encounter the curse of dimensionality, i.e. the reaction coordinate problem, which invokes the invention of various algorithms for solution. Recent years, new machine learning techniques and the incorporation of some of them into the transition state searching methods emerged. Here, we first review the design principle of representative transition state searching algorithms, including the collective-variable (CV)-dependent gentlest ascent dynamics, finite temperature string, fast tomographic, travelling-salesman based automated path searching, and the CV-independent transition path sampling. Then, we focus on the new version of TPS that incorporates reinforcement learning for efficient sampling, and we also clarify the suitable situation for its application. Finally, we propose a new paradigm for transition state searching, a new dimensionality reduction technique that preserves transition state information and combines gentlest ascent dynamics.

Keywords: transition state, gentlest ascent dynamics, path methods, reinforcement learning, generative models

PACS: 87.10.Tf, 87.15.A–, 87.15.H–, 87.15.hp

DOI: [10.7498/aps.72.20231319](https://doi.org/10.7498/aps.72.20231319)

* Project supported by the National Natural Science Foundation of China (Grant No. 31971179) and the Science Technology and Innovation Commission of Shenzhen Municipality, China (Grant Nos. JCYJ20200109150003938, RCYX2020071411 4645019).

These authors contributed equally.

† Corresponding author. E-mail: zhulizhe@cuhk.edu.cn

专题: 生物分子模拟中的机器学习

蛋白质结构模型质量评估方法综述*

刘栋 崔新月 王浩东 张贵军†

(浙江工业大学信息工程学院, 杭州 310014)

(2023年6月30日收到; 2023年8月1日收到修改稿)

蛋白质模型质量评估方法是蛋白质结构预测的关键技术, 自 CASP7 以来一直是结构生物信息学领域的研究热点. 模型质量评估方法不仅可以指导蛋白质结构模型的精修, 还能够从多个候选构象中筛选出最佳模型, 具有重要的生物学研究和实际应用价值. 本文首先回顾了国际蛋白质结构预测关键评估竞赛 (CASP)、全球持续蛋白质结构预测竞赛 (CAMEO) 以及单体蛋白和复合物的模型评估指标, 主要梳理了近 5 年来包括共识方法 (多模型方法)、准单模型方法和单模型方法在内的模型质量评估方法的发展历程, 并介绍 CASP15 中的复合物模型评估方法; 鉴于深度学习在蛋白质预测领域所取得的巨大进展, 重点分析了深度学习在单模型方法数据集生成、蛋白质特征提取以及网络架构构建方面的深入应用, 并进一步介绍了本课题组近年来在模型质量评估方面开展的工作; 最后, 总结分析了目前蛋白质模型质量评估技术的局限性及所面临的挑战, 并对未来发展趋势进行了展望.

关键词: 蛋白质模型质量评估, 深度学习, 单模型方法, 复合物模型评估**PACS:** 87.10.Vg, 87.14.E-, 87.16.A-, 87.55.de**DOI:** 10.7498/aps.72.20231071

1 引言

蛋白质参与生命活动的各个过程, 是生命体的重要组成部分. 了解蛋白质结构可以进一步揭示生命过程中生物分子复杂的相互作用机制^[1-3]. 经过实验科学家近 60 年来巨大的努力, 已经解析出了二十余万种蛋白质结构. 然而, 由于生物实验过程耗时长且成本较高, 致使实验解析结构仅占已知两亿多蛋白质序列数量的 0.1%^[4], 因此, 通过高效且准确的计算方法实现大规模蛋白质结构预测成为 50 多年来计算生物学家努力的方向^[5]. 广泛使用的 Rosetta^[6], I-TASSER^[7] 是蛋白质领域经典结构预测方法, 随着深度学习技术在该领域研究的广泛应用, 国内外学者陆续提出了 RaptorX^[8], trRosetta^[9], AlphaFold2^[5], PAtreader^[10], ESMFold^[11] 等方法.

尤其是 DeepMind 和 Meta 研究团队基于 AlphaFold2 和 ESMFold 的方法, 分别构建了约两亿预测结构的数据库 AlphaFold Protein Structure Database^[12] 和约七亿预测结构的数据库 ESM Metagenomic Atlas^[11]. 针对同一序列, 上述方法预测出的结构存在显著差异. 为解决此类问题, 模型精度估计或者模型质量评估方法 (estimation of model accuracy, EMA)^[13] 就成为蛋白质结构预测流程中一个关键的环节. EMA 方法主要目的是估计参考结构与预测模型在整体拓扑 (全局结构) 和残基级别 (局部结构) 相似的程度, 并能够进一步实现模型单残基、连续残基块的拓扑精修, 常用的指标包括 GDT-TS^[14], TM-score^[15], lDDT^[16], CAD^[17], SG^[18] 等.

Moult 等^[19]1994 年创立的蛋白质结构预测的关键评估 (CASP) 被誉为蛋白质结构预测领域的

* 科技创新 2030—“新一代人工智能”重大项目 (批准号: 2022ZD0115103)、国家自然科学基金 (批准号: 62173304) 和浙江省自然科学基金重点项目 (批准号: LZFO30002) 资助的课题.

† 通信作者. E-mail: zgj@zjut.edu.cn

奥林匹克竞赛. CASP 每两年举办一次, 目前开展了 15 届, 已经成为蛋白质结构预测技术发展的风向标^[20,21]. 在 2006 年 CASP7 中引入了模型质量评估方法的评测, 这足以说明 EMA 方法对结构预测的重要性. 此外, 另一个重要的国际赛事 CAMEO^[22] 自 CASP12 之后引入了每周在线的自动盲测评估服务器, 成为 CASP 两年间评测的重要补充平台. 值得一提的是, AlphaFold2 在 CASP14 中取得巨大的突破, 使得单体结构预测几乎到达了实验解析的精度^[23]. 因此, 在 CASP15 中接触预测、优化和单体模型质量评估被取消, 而新增 RNA 结构、蛋白质与配体复合物、复合物结构及其界面的质量评估类别^[24], 对于复合物评估, 除了全局结构与局部结构的精度估计之外, 还新增接触界面精度估计, 如 DockQ^[25] 和 QS-score^[26].

自 CASP7 至目前为止, 已经开发出许多蛋白质模型质量评估方法和在线服务器, 如图 1 所示. 本文梳理了最近 5 年主流模型质量评估方法, 主要分为共识方法 (多模型方法)、准单模型方法、单模型方法^[27]. 共识方法假设正确的结构包含在重复结构模式集合中, 通过聚类提取来自多个方法或不同模板生成的蛋白质结构模型的共识信息, 代表性方法有 Cheng 课题组开发的 MULTICOM 系列^[28-30], Xu 和 Shang 课题组开发的 MUFOLDQA 系列^[31,32] 等. 在 CASP7—15 评测中, 共识方法在大多数情况下都比单模型方法表现得更好. 准单模型方法将单个模型输入的便利性与共识方法预测能

力的优势相结合, 通过内部参考结构生成方法产生的一组蛋白质结构对预测模型进行评分, 代表性的方法有 McGuffin 课题组^[33-35] 开发的 ModFOLD 系列等. 单模型方法基于单一蛋白质模型特征提取 (序列信息、几何结构、理化信息), 通过神经网络来评估残基或者拓扑的质量. 随着机器学习和深度学习技术在蛋白质结构预测领域广泛、深入地应用, 单模型方法在性能逐渐与多模型方法持平甚至超越, 成为 EMA 方法中一个热点研究方向, 代表性的方法主要有 Baker 课题组^[27] 开发的 DeepAccNet 系列、Elofsson 课题组^[36,37] 开发的 ProQ 系列, Venclovas 课题组^[38-40] 开发的 Voro 系列, 杨建益课题组^[41] 开发的 Yang_TBM, 张贵军课题组^[42-44] 开发的 DeepUMQA 系列等.

本文将按顺序介绍 CASP 和 CAMEO, 其次详细讨论蛋白质模型质量评估的指标体系, 包括单体蛋白、复合物的评估指标以及综合性能分析指标. 然后, 对近 5 年来主流的共识方法、准单模型方法和单模型方法进行梳理, 并介绍 CASP15 的复合物模型质量评估方法. 考虑到深度学习对蛋白质领域的影响, 本文重点讨论单模型方法中的数据集、蛋白质特征和网络架构这三个方面, 并介绍了本课题组近年来在模型质量评估方面所开展的一些工作. 最后, 分析给出了蛋白质模型质量评估方法所面临的一些关键挑战, 并对未来可能的发展趋势进行了展望.

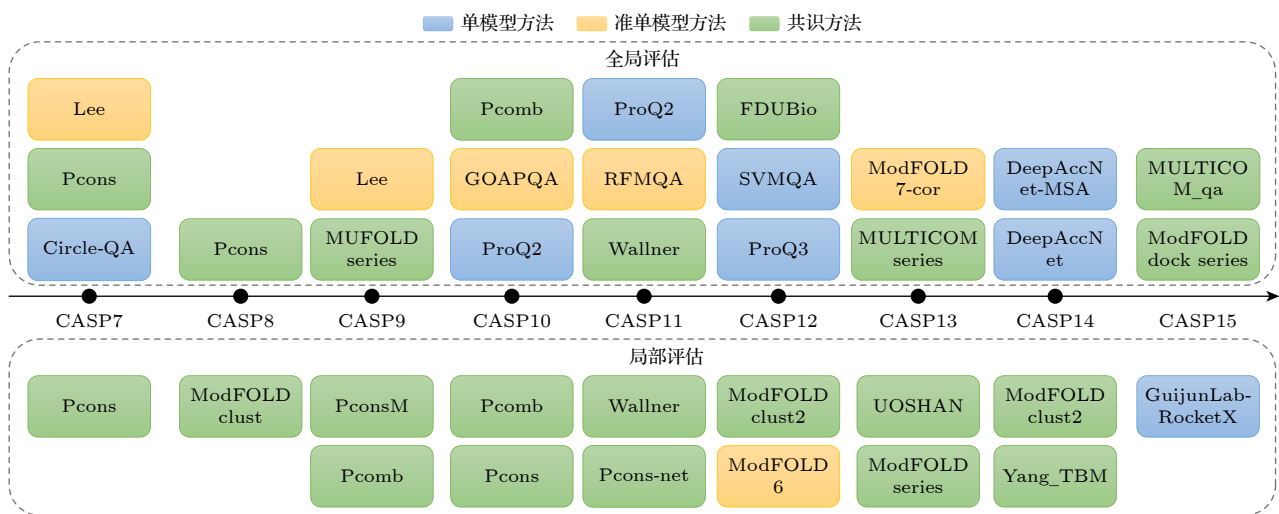


图 1 在 CASP 中主流的模型质量评估方法
Fig. 1. Mainstream model quality assessment methods in CASP.

2 国际蛋白质结构预测的关键评估竞赛 (CASP) 和全球连续自动模型评估竞赛 (CAMEO)

CASP^[19] 自 1994 年以来, 已成功举办了 15 届. CASP 为研究团队提供了一个客观测试蛋白质结构预测方法的平台, 并为研究团队和软件用户提供了对蛋白质结构建模最新技术水平的独立评估. 在 CASP7 中引入了蛋白质模型质量评估的评测, 其中蛋白质模型结构由三维结构预测组提交, 为评估模型质量方法提供了测试数据集. CASP 的评估过程分为两个阶段. 在第 1 阶段, 通过共识方法为每个蛋白质目标选择约 20 个蛋白质结构模型, 覆盖了整个模型质量范围进行评估; 在第 2 阶段, 选择前 150 个模型用于质量评估. 在这两个阶段中, EMA 方法需要评估每个模型的全局拓扑质量和残基级别的局部质量^[45,46]. 第 1 阶段的结果仅用于与第 2 阶段的结果比较, 以确定 EMA 方法是否是单模型方法^[47]. 在每届 CASP 比赛中, 表现最好的 EMA 方法通常代表了蛋白质质量评估领域的最新发展水平.

此外, 瑞士生物信息研究所和巴塞尔大学联合举办 CAMEO^[48] 是一个全球持续进行的蛋白质结构预测平台, 被认为是蛋白质结构预测领域最重要的比赛之一. CAMEO 中每位参赛者每周对由世界范围内的结构生物学家最新破解出的 20 个蛋白质结构进行预测. 在 CAMEO-QE 中, 预测出的结构由模型质量评估参赛者进行评估并在线提交. 多年来, CASP 和 CAMEO 不断进步和相互促进, 为 EMA 研究带来了新的思路和方法, 并推动了这一领域的不断突破和发展.

3 蛋白质模型质量的评估指标

蛋白质结构的准确性和可靠性对于理解生命活动过程至关重要. 为了评估计算方法的性能, 必须使用有效的评估指标来衡量蛋白质模型的质量. 这些评估指标能够判断蛋白质模型与实验解析结构之间的相似程度, 并识别模型中可能存在的结构缺陷或误差, 从而进一步改进和优化模型. 此外, 蛋白质评估指标对于蛋白质设计和药物设计等领域也具有重要意义. 随着多年来蛋白质结构领域的发展, 衍生出了多种评估指标, 特别是在最近 CASP

或 CAMEO 比赛中采用的指标. 总体上来讲, 这些指标大致分为“单体结构质量评估指标”和“复合物结构质量评估指标”, 其中单体结构质量评估指标主要侧重于局部评估指标和全局评估指标, 下面将分别介绍一些常用的评估指标及其应用场景.

3.1 单体结构质量评估指标

对于 CASP 评估者而言, 其中一个主要挑战是定义合适的数值指标, 以量化预测与实验结构之间的准确度. 在 CASP 评估过程中, 研究者通过评估预测模型质量来反映结构预测技术的最新水平^[16]. 均方根误差 (root mean square deviation, RMSD) 在 CASP 早期作为主要评估标准^[49,50], 然而 RMSD 存在极易受到预测不准确区域的异常值影响、对模型中的缺失部分不敏感、对参考结构的叠加具有较高依赖性的问题^[17]. 为了更为客观地评估蛋白质结构模型的质量, 研究者相应提出了多种评估指标来综合描述蛋白质结构的质量.

GDT-score (global distance test score)^[14] 从 CASP4 引入以来一直被广泛使用. GDT-score 通过将预测与实验参考结构进行叠合后, 计算模型结构中某种原子 (如 C_α) 落在实验结构对应位置的某个阈值范围内所得到最大的原子数目. 通常 GDT-HA 使用的阈值为 0.5, 1, 2 和 4 Å, GDT-TS 使用的阈值为 1, 2, 4 和 8 Å, 计算公式^[14] 如下:

$$\text{GDT-TS}_{(M_p, M_r)} = \frac{(P_1 + P_2 + P_4 + P_8)}{4}, \quad (1)$$

其中 M_p 是预测模型; M_r 是参照模型; P_1, P_2, P_4 和 P_8 是 M_p 中的 C_α 原子与 M_r 的 C_α 原子距离小于 1, 2, 4 和 8 Å 的概率. 此外, 根据所比较的原子类型, 分为使用侧链的原子 GDC_SC^[51] 和全原子 GDC_ALL. 与 RMSD 相比, 局部低精度的原子不会对质量分数产生显著影响. 然而, GDT-score 对于蛋白质的大小具有依赖性. 当蛋白质序列的长度较短时, 它可能接近于随机选择结构模型. 这种显著依赖于序列长度的现象使得评分绝对值大小可能变得毫无意义^[15]. 此外, GDT-score 评估中的缺失片段会导致较低的质量得分, 而类似于 GDT-score 这种基于全局叠加比对的度量方法, 其主要局限性在具有多个结构域的柔性蛋白质时更为突出. 全局刚体叠合会由最大的结构域主导, 因此较小的结构域无法正确匹配, 导致不合适的质量分数. 而且结构域相对位置轻微变化 (在生物学上可

能是可以忽略的)可能会强烈影响 GDT-score. 这导致在 CASP 中需要将蛋白质模型分割成评估单元 (AU) 来减少结构域的影响, 并对其进行单独评估.

TM-score^[15] 利用蛋白质长度相关的数值来消除之前评估指标中对于蛋白质长度的依赖性. 其次, 与设置特定距离阈值并仅计算低于阈值误差的部分不同, TM-score 会对齐预测模型与参考结构之间所有残基对进行评估, 计算公式^[15]如下:

$$\text{TM-score} = \max \left[\frac{1}{L_{\text{ref}}} \sum_i^{L_{\text{aligned}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{ref}})} \right)^2} \right], \quad (2a)$$

$$d_0(L_{\text{ref}}) = 1.24 \sqrt[3]{L_{\text{ref}} - 1.5} - 1.8, \quad (2b)$$

其中 L_{aligned} 和 L_{ref} 分别是对齐的预测和参考结构的序列长度, d_i 是指预测蛋白中的残基与参考蛋白中相应残基之间的距离, $d_0(L_{\text{ref}})$ 是用来归一化 d_i 的距离. 由于 TM-score 是基于两个结构之间单个叠加比对计算得出的分数, 当蛋白质长度依赖性对模型评估没有影响时, GDT-score 可以在多个阈值距离下进行评估, 综合考虑了更多的结构信息, 从而提供了更全面的相似性度量^[17].

一般来讲, 单体蛋白全局结构模型质量的评估指标是从整体拓扑上比较预测结构与参考结构的相似度, 而局部结构质量评估指标能够细致地分析蛋白质中局部区域的结构特征和稳定性, 帮助研究者们识别和定位潜在的结构问题和缺陷.

为了更好地理解单体蛋白质主链中局部原子的相互作用, 验证其立体化学的合理性. IDDT (local distance difference test)^[16] 通过比较参考结构中一定范围内较近的、不属于同一残基的原子对之间的距离进行计算. 如果模型中的距离与参考结构中的距离在一定的阈值范围内 (如 0.5, 1, 2 和 4 Å), 则被认为是符合要求的距离. 通过计算保留距离的比例, 可以得到预测模型的 IDDT. 其能够捕获结合位点中的局部几何结构, 并且对结构域的方位变化不敏感, 使得绝对值分数具有指导性的意义. 并且, 该指标可用于进一步指导结构模型的精细修正和拓扑微调.

由于蛋白质的空间结构是通过残基的相互作用形成, 而这种互作模式可以用空间结构上的接触表示. 因此, 通过量化蛋白质模型结构的接触预测相对于参考结构偏差, 并且不需要两个结构之间的

对齐, 从而避免一些叠合对齐的问题. 基于接触面积差异的评估指标接触区域差异 CAD (contact area difference)^[17], 它通过计算残基之间的接触面积差异来量化模型与参考结构之间的接触, 计算公式^[17]如下:

$$\text{CAD}_{(i,j)} = |T_{(i,j)} - M_{(i,j)}|, \quad (3a)$$

$$\text{CAD}_{(i,j)}^{\text{bounded}} = \min(\text{CAD}_{(i,j)}, T_{(i,j)}), \quad (3b)$$

$$\text{CAD-score} = 1 - \frac{\sum_{(i,j) \in G} \text{CAD}_{(i,j)}^{\text{bounded}}}{\sum_{(i,j) \in G} T_{(i,j)}}, \quad (3c)$$

其中 i 和 j 代表预测模型和参考结构中的残基, G 是参考结构中的接触残基对的集合, $T_{(i,j)}$ 和 $M_{(i,j)}$ 分别表示参考结构和预测模型中的接触面积. CAD-score 可以单独考虑残基主链和侧链, 具有处理模型中缺失残基的能力, 并且类似于 GDT-score, 能够对完整和不完整的模型进行排名. 此外, 另一个指标是 Sphere Grinder (SG)^[18], 通过简单直观的方式识别预测模型中不正确的区域.

对于单体蛋白质模型的质量评估, 局部指标和全局指标相互弥补, 有效地揭示蛋白质模型的局部和整体结构质量, 并为蛋白质结构预测提供更可靠的指导.

3.2 复合物结构质量评估指标

随着人工智能技术在单体结构预测领域的突破, 之前的评估指标更适用于描述单体结构的质量, 而研究的重点逐步向复合物转移. 为了探究蛋白质与蛋白质之间的相互作用, 研究者们设计了专门用于复合物 (多聚体) 的评估指标, 这对于预测复合物的结构发展至关重要.

蛋白质相互作用的关键评估竞赛 (CAPRI) 旨在评估蛋白质对接方法和预测蛋白质与蛋白质相互作用关系^[52]. CAPRI 引入 F_{nat} , LRMS 和 iRMS 指标用于评估模型^[25]. F_{nat} 衡量了预测复合物界面中在实验参考结构中界面接触残基所占的比例, 界面接触被定义为两个相互作用的蛋白质 (受体和配体) 之间任意一对重原子之间的距离在 5 Å 以内. LRMS 是在将预测和参考复合物的受体 (两个蛋白质中较大的一个) 进行叠合比对后, 计算配体 (较小的蛋白质) 预测和参考复合物的 RMSD. LRMS 是一个全局指标, 取决于配体的大小. 因此, 在接

触界面区域的匹配情况中, 它可能不是一个较好的评估指标. iRMS 仅针对接触界面残基的 RMSD, 其接触界面的残基距离范围重新定义为 10 Å 以内, 即 F_{nat} 定义界面阈值的两倍. 虽然这些评估指标可以量化蛋白质对接模型质量的不同方面, 但在对模型排序、模型质量与评分函数的相关性分析以及在机器学习算法中作为目标函数时存在一定限制. 因此, 需要综合考虑多个指标, 以更准确地评估模型的质量. DockQ^[25] 将 F_{nat} , LRMS 和 iRMS 综合到一个介于 0 到 1 之间的单一评估指标中, 可以更加定量地评估蛋白质对接模型的质量, 计算公式^[25] 如下所示:

$$\text{RMS}_{\text{scaled}}(\text{RMS}, d_i) = 1/[1 + (\text{RMS}/d_i)^2], \quad (4a)$$

$$\text{DockQ} = \frac{(F_{\text{nat}} + \text{RMS}_{\text{scaled}}(\text{LRMS}, d_1) + \text{RMS}_{\text{scaled}}(\text{iRMS}, d_2))}{3}, \quad (4b)$$

其中 $\text{RMS}_{\text{scaled}}$ 表示与 LRMS 或 iRMS (RMS) 中的任何一项相对应的缩放后的 RMS 偏差, d_i 是一个缩放因子, d_1 用于 LRMS, d_2 用于 iRMS. F_{nat} 被定义为预测的复合物界面中保留的原生界面接触的比例. 在评估 CAPRI 中的蛋白模型时, DockQ 几乎可以重现原始的 CAPRI 分类, 这意味着不需要使用阈值对预测模型进行分类, 并且可以使用 Z-score 来评估模型质量, 类似于 CASP 中使用的方法.

在蛋白质与蛋白质对接模型评估指标的发展历程中, 主要集中在二聚体的相互作用. 然而, 对于多聚体 (链数大于两条) 需要将其分解为二聚体可能需要大量的比较工作, 并且可能会缺失一些整体结构的接触界面残基. 因此, 研究者设计了 QS-score^[26], 用于量化界面之间的相似性, 该相似性取决于共同的界面接触. 其能够区分不同的多聚体结构和结合模式, 计算公式^[26] 如下所示:

$$Q = \sum_{s(i,j)} w(\min(d_i, d_j)) + \sum_{n-s(i)} w(d_i) + \sum_{n-s(j)} w(d_j), \quad (5a)$$

$$\text{QS-score} = (1/Q) \times \sum_{\text{shared}(i,j)} w(\min(d_i, d_j)) \left(1 - \frac{|d_i - d_j|}{12}\right), \quad (5b)$$

$$w(d) = \begin{cases} 1, & d \leq 5, \\ e^{-2\left(\frac{d-5}{4.28}\right)^2}, & d > 5, \end{cases} \quad (5c)$$

其中 d 代表残基之间的欧式空间 C_β 距离, $|d_i - d_j|$ 代表相对误差 (将 12 Å 作为最大误差), w 是加权函数. 当涉及的所有残基都被“映射”时, 形成的接触被定义为 s . 而那些接触但未被“映射”的残基对, 或者只在其中一个寡聚体中形成接触被定义为 $n - s$. 这里所提及的“映射”是指一个复合物中的蛋白质链与另一个复合物中蛋白质链之间的对应关系. QS-score 能够评估组装界面的质量, 适用于比较链的相对方位. 在最近的 CASP15 中, 评估者还使用界面接触分数 (ICS) 和接触区域分数 (IPS) 来评估模型. ICS 以 F1-score^[53] 的形式计算, 用于衡量预测的链间接触的精准率和召回率之间的关系. IPS 则通过计算模型预测的接触残基与参考结构接触残基之间的部分, 得出 Jaccard^[54] 系数.

伴随着结构预测领域的发展, 复合物结构的评估逐渐变得尤为关键. 复合物的评估指标可以从多个独立计算却相关的指标合成一个评估指标, 并且可以从二聚体扩展到多聚体的评估指标.

3.3 评估结构精度估计的指标

模型质量评估 (EMA) 是 CASP 重要的组成部分, 理想情况下, EMA 方法可以提供与计算的评估指标分数相关的模型质量估计. 在 CASP14 之前的比赛中约有 70 多种参赛方法^[55], 这凸显了模型质量评估对蛋白质结构预测的重要性, 并且研究人员通常将模型质量估计整合到建模流程. 蛋白质模型的精度估计包括了每个模型的全局精度评估和每个残基的局部精度估计. 此外, CASP 对参赛组进行分别排名, 这些排名通常使用多个评估指标综合计算得出.

评估全局结构精度估计包含 Top1 loss^[47], AUC (area under the curve)^[56], 相关性和绝对误差分析. Top1 loss 用于对比蛋白质结构预测模型的精度估计, 并选择排名第一的模型作为最佳模型. 在不同指标下, 计算选定的最佳模型与实际最佳模型质量的绝对误差. 相关性分析使用 Pearson 和 Spearman^[57] 来评估预测全局模型与真实模型质量之间的相关性. 通过绝对误差分析 (MAE 或 MSE), 分析不同指标下模型质量预测值与真实值之间的差

异. AUC^[56] 用于判断预测模型质量是否可以接受, 它通过计算 ROC 曲线下的面积衡量模型的性能, 而 ROC 曲线则反映了在不同质量阈值下, 准确和 inaccurate 模型的真阳性率和假阳性率之间的关系.

局部结构精度评估是在评估单元 (EUs)^[47] 级别进行. ASE (average S-score error)^[47] 是通过计算每个残基的 S-score 误差的平均值来评估:

$$ASE = \left(1 - \frac{1}{N} \sum_{i=1}^N |S(e_i) - S(d_i)| \right) \times 100, \quad (6)$$

其中第 i 个残基的 S-score 误差是对预测模型中评估单元 (EU) 的第 i 个 C_α 原子的预测距离误差 (e_i) 和实际距离误差 (d_i) 之间的差值. 通过 LGA^[14] 在评估单元的叠合后, 使用 S-function 函数来计算, N 是评估单元中的残基数目. ULR (unreliable local region)^[47] 是由预测模型中 3 个或更多连续残基组成的区域, 其在最佳叠合下与相应参考结构的残基之间的距离偏差超过 3.8 Å. 相隔一个残基的两个 ULR 将合并为一个 ULR. 确定 ULR 后, 计算它们的准确度和覆盖率, 并在实际 ULR 边界上以及在两个残基以内的预测被认为是准确预测. 对于每个 CASP 评估组, 通过调整阈值计算以最大化平均 F1-score^[53]. 在 CASP 中, 组的排名往往是根据蛋白质目标的评估指标对应平均 Z-score 统计, 其中每个组的 Z-score 是对每个目标的结果计算的均值和标准差, 将 Z-score 设置为 -2—2.

随着 AlphaFold2 在单体结构预测方面的巨大

进展, 几乎解决了单体结构预测问题, 促使 CASP15 将重点转向复合物的预测和模型质量评估. 其中, 整体模型拓扑质量评估采用 GTD-Score 和 TM-Score 指标; 链间相互作用质量评估采用 DockQ 和 QS-Score 进行衡量; 界面接触残基质量评估采用 CAD-Score, lDDT, PatchQS 和 PatchDockQ^[24] 指标衡量. CASP 参赛组的性能往往是通过这些指标对应的 Pearson, Spearman, AUC 和 Loss 进行综合加权给出最终排名.

在蛋白质结构预测领域, 质量评估对于建模过程具有重要意义. 质量评估指标提供了一种客观、量化的方法来评估模型的准确性和质量, 同时为改进和优化建模过程提供了指导和依据.

4 蛋白质模型质量方法

在最近的 CASP 中, 研究者已经开发了许多方法, 包括共识、准单模型和单模型的质量评估方法, 主要步骤如图 2 所示. 此外, 鉴于复合物模型评估的重要性, 我们回顾了 CASP15 中的复合物质量评估方法. 最后, 介绍了本课题组近年来在模型质量评估方面开展的工作.

4.1 数据集

训练数据集在神经网络中起着至关重要的作用, 它是神经网络学习和理解模式的基础^[58]. 通过训练数据, 神经网络可以从中学到输入与输出之

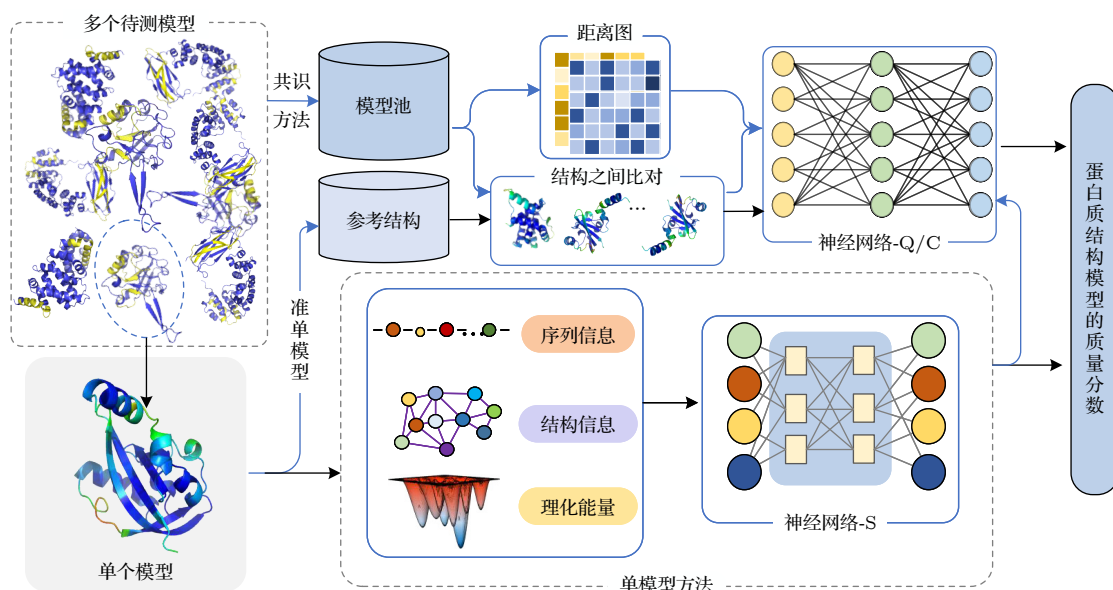


图 2 模型质量评估三类方法示意图

Fig. 2. Schematic diagram of three methods of model quality assessment.

间的关联性,使其能够对新数据进行准确的预测和推断.丰富、多样且代表性的训练数据可以帮助神经网络克服过拟合和欠拟合等问题,提高模型的泛化能力和稳定性.因此,对基于神经网络的蛋白质模型质量评估而言,高质量数据集需要包含不同精度的结构并且达到一定程度的数量,这可以使网络学习到蛋白质的结构与质量的潜在映射关系.

CASP1-CASP15数据集由每届参加CASP结构预测组提交的模型构成.每个蛋白质目标至少包含150个预测结构,这些结构的精度各不相同,往往被用于训练和测试模型.截止至2023年6月28日,CAMEO-QE数据已经持续评估了74704个蛋白质预测模型,针对每个蛋白质目标的模型数大约为10个,相比于CASP,模型的相似度较高且预测难度较低.AlphaFoldDB和ESM Metagenomic Atlas分别是AlphaFold2与ESMfold预测的高精度蛋白质模型数据库.虽然大部分结构还未通过实验解析出来,但是这两个数据集对于蛋白质结构领域的研究具有重要的意义.Zhanglab服务器中非冗余的蛋白质目标所生成的诱饵结构包含3DRobot数据集、I-TASSER数据集、QUARK数据集等.而DeepAccNet, GNNRefine, DeepUMQA, DeepUMQA3, GraphCPLMQA和GraphGPSM这些方法都采用大致相同的数据集制作思路:从PDB库中筛选出一批非冗余的蛋白质目标,通过不同的方法生成预测模型结构(Decoys)用于训练神经网络.在开发基于深度学习模型质量评估的方法,往往可以组合这些数据进行训练,如表1所列.

4.2 共识方法

共识方法在CASP蛋白质模型精度评估上具有显著优势.Cheng课题组^[28-30]开发的MULTICOM系列结合了各种质量评估技术,包括半聚类方法、单模型机器学习方法以及组方法.其中,MULTICOM-cluster和MULTICOM-construct^[29]在CASP质量评估测试中表现优异.MULTICOM系列评估方法通过结合来自12种不同EMA方法(9种单模型方法和3种多模型方法)以及1种蛋白质接触预测方法(DNCON2^[47])的预测结果,生成10个质量分数作为预训练深度神经网络的输入特征.对于MULTICOM-construct,这10个质量分数取平均值.而MULTICOM-cluster则将13个初步预测结果和10个DNNs预测结果的组合输入另一个DNN,进一步预测最终的质量分数.该研究方法表明,使用残基与残基接触特征可以显著提高该方法的性能.在MULTICOM-AI^[16]中,基于深度学习技术和共进化分析,新增了残基间距离特征,其计算一组结构模型中的残基距离与DeepDist^[30]预测的距离之间的相关性.此外,MULTICOM-AI还使用了基于DNCON4生成残基间接触特征.

Xu和Shang课题组开发的MUfoldQA^[31,32]系列方法,在CASP13中涵盖了MUfoldQA_M和MUfoldQA_T两种方法,其核心思想是利用一组参考模型对每个候选模型进行评分.它们之间的区别在于选择参考模型和计算给定一组参考模型的候选模型评分方式.MUfoldQA结合了准单模型的质量评估方法,首先通过在PDB数据库中搜索蛋白质序列来获得一组模板.然后,从候选模型中选

表1 模型质量评估的蛋白质结构数据集(诱饵)
Table 1. Protein structure dataset (Decoys) for model quality assessment.

Data sets	URLs
CASP	https://predictioncenter.org/download_area/
CAMEO	https://www.cameo3d.org/
Zhanglab	https://zhanglab.cmb.med.umich.edu/decoys/
AlphaFoldDB	https://alphafold.ebi.ac.uk/
ESM Metagenomic Atlas	https://esmatlas.com/resources?action=search_structure
DeepAccNet	https://github.com/hiranumn/DeepAccNet
GNNRefine	http://raptorx.uchicago.edu/download/
DeepUMQA	https://academic.oup.com/bioinformatics/article/38/7/1895/6520805?login=true
DeepUMQA3	https://www.biorxiv.org/content/10.1101/2023.04.24.538194v1.full.pdf+html
GraphCPLMQA	https://www.biorxiv.org/content/10.1101/2023.05.16.540981v1.full.pdf+html
GraphGPSM	https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbad219/7197734?searchresult=1#supplementary-data

择一个子集作为参考模型,并根据与模板的相似性对每个参考模型进行评分.最后,每个候选模型根据其参考模型的相似性进行评分,并考虑到参考模型的评分进行加权.此外, MUfoldQA_G^[59]结合了蛋白质模板和参考模型的信息,以优化最大化皮尔逊相关系数的 QA 指标. MUfoldQA_Gr 通过重采样训练数据并训练模型,学习到更好的共识模式,同时最小化了平均 GDT-TS 误差. MUfoldQA_G 将 MUfoldQA_Gr 和 MUfoldQA_Gp 的结果相结合,使最终的预测结果接近 MUfoldQA_Gr 的低平均 GDT-TS 误差,并保持与 MUfoldQA_Gp 结果相同皮尔逊相关系数.

McGuffin 开发的 ModFOLDclust2^[60]是一种基于自动聚类的领先方法,用于对局部和全局模型的质量评估. ModFOLDclust2 服务器在 CASP9-CASP14 中测试的方法基本相同. ModFOLDclust2 最初的开发目标是减少计算代价,并提供比 ModFOLDclust^[61]更高的预测精度. ModFOLDclust2 的全局质量分数为 ModFOLDclustQ 和 ModFOLDclust 全局质量评估分数的平均值.为了进行全面的比较模型,使用了一种修改后的无结构比对的 Q-measure^[62]. ModFOLDclust2 的残基的质量评估分数是直接从 ModFOLDclust 中获取.

杨建益课题组^[41]开发 QDistance(Yang_TBM)是基于 trRosetta 预测的残基间距离估计全局和局部质量. QDistance 使用 trRosetta 预测查询蛋白的残基间距离和结构模型.为了预测每个模型的全局质量评估分数,设计了三组特征,包括基于 2D 距离矩阵比对、势能分数和其他单一 QA 方法以及 1D 结构特征比较的特征.这些特征被输入到线性回归模型中,以预测 GDT_TS.为了进行局部 QA 预测,首先选择排名靠前的模型(根据预测的 GDT_TS 分数),然后使用共识分析来推断每个模型的局部质量分数.

clustQ 是 Bhattacharya 课题组^[63]基于加权距离比较的无超聚 (superposition-free) 方法评估质量. clustQ 对在序列中相隔较远的残基,分配了较高的权重.这类残基之间相互作用相对于局部短程相互作用提供了更多的信息,并且使用基于 Q-score^[62]扩展的 WQ-score 对模型之间进行了配对比较,以估计预测模型质量精度.

此外, UOSHAN^[64]是基于聚类 SARTclust_G 和 SARTclust_L 的评估方法.在全局和局部评分

中,根据 SART_G 分数对预测模型进行排名,形成一个包含前 N 个模型的参考集合.然后,将待评估模型与参考集合中的所有模型进行 TM-score 比对.对于全局评分,计算 N 个比较得到的 GDT_TS 分数,并使用 SARTclust_G 对这些分数进行加权平均.对于局部评分,计算相应残基之间的 N 个距离值,然后使用 SARTclust_G 对这些 S-score 进行加权平均. MESHConsensus^[65]是基于 LightGBM^[66] 随机森林回归器,利用结构、序列和共识特征来估计蛋白质模型的质量.

4.3 准单模型方法

共识方法在 CASP 测试中表现出色,因为它们能够利用多个模型之间的信息来生成更准确的预测.然而,共识方法的性能很大程度上受候选模型池质量和全面性的影响.如果候选模型池质量较低或缺乏全面性,那么共识方法的性能可能会受到影响.鉴于共识方法的局限性,准单模型方法通过参考其内部方法生成的一组蛋白质结构来评估预测模型,从而避免了依赖于候选模型池的问题.

McGuffin^[35]开发 ModFOLD 系列方法作为准单模型方法在 CASP 测试中表现出色,其中 ModFOLD6^[67], ModFOLD7^[68] 和 ModFOLD8^[33] 在 CASP 评测中表现突出.它们具有类似的工作流程,通过使用不同的单模型和准单模型方法对蛋白质模型进行独立评估,并生成局部质量评分.这些局部质量评分被视为特征,并输入到神经网络中,以推导出最终的预测的全局评分. ModFOLD6 采用了多个评估方法,如 ProQ2^[36]、接触距离一致性 (CDA)、二级结构一致性 (SSA)、无序 B-factor 一致性 (DBA)、ModFOLD5(MF5s) 和 ModFOLDclustQ (MFcQs).在 ModFOLD6^[69] 中,为了提高局部质量预测的准确性和单模型排名的一致性,它采用了与之前类似的十种单模型和准单模型方法. ModFOLD7 还提供了两个版本,分别是在排序 Top 1 模型方面表现最好的 ModFOLD7-rank 和在反映估计绝对误差方面表现良好的 ModFOLD7-cor. ModFOLD8^[35] 结合了来自 13 种评估方法 (包括 9 个单模型和 4 个准单模型) 进一步发挥多个单模型和准单模型方法的各自优势提高预测准确性.

此外, QMEANDisco^[70] 利用与同源模型结构的距离分布,使用训练神经网络将多模板 DisCo 分数和单模型 QMEAN^[71] 分数加权组合,得到 QMEANDisCo 复合分数.

4.4 单模型方法

随着机器学习和深度学习的发展,在蛋白质领域单模型评估方法得到越来越多关注与研究. 这些方法只需要一个模型作为输入,并能够表现出与共识方法相似或更好的性能. 单模型方法可以分为基于传统机器学习和基于深度学习的评估方法,并鉴于深度学习对蛋白质领域的影响,将对基于深度学习模型评估方法从特征、网络以及架构展开描述.

基于传统机器学习的单模型质量评估方法通常使用多种特征作为输入,包括基于能量的特征、基本的物理化学特征和统计特征. 例如 SVMQA^[72]方法则将基于势能的特征和基于一致性的特征作为输入,使用随机森林算法预测全局质量. 此外,还通过改变特征组合改善质量得分. MESH1-enrich-server, MESH1-corr-server 和 MESH1-server 使用机器学习训练的 3 种不同损失函数分析对该方法性能的影响.

对基于深度学习的单模型质量评估而言,蛋白质模型特征和网络架构对于方法的性能有关键影响. 特征可以显性刻画蛋白质的属性,其中包括蛋白质的结构特征和非结构特征. 对于结构的特征,3DCNN^[73]仅利用 3D 结构的原始原子密度作为特征,没有进行任何特征调整. Ornate^[74]表示基于体系化特征的蛋白质拓扑结构,这些体系化特征根据骨架中原子的方向构建立方图,描绘了残基及其邻域. Atom-ProteinQA 设计了两个提取几何和拓扑原子级关系模块. 几何感知模块捕捉输入蛋白质的几何特征,生成细粒度的原子级预测,基于化学键构建原子级图通过拓扑感知模块的消息传递并行输出残基级别的预测. 这些方法通过低维空间关系来表示蛋白质几何模型结构.

对于非结构特征,ProQ3D^[75]采用了基于 Rosetta 能量项的两个特征,即全原子 Rosetta 能量项和粗粒化中心点 Rosetta 能量项. Venclovas 课题组^[38]开发的 VoromQA,将统计势的概念与原子球的 Voronoi^[76]分割相结合评估模型质量. 其将蛋白质结构表示为一组原子球,每个球具有对应于原子类型的范德瓦耳斯半径分配的空间区域,并使用 Voronoi 面和球面的三角表示,接触面积被计算为对应三角的面积. 其中, VoromQA-A 通过使用 SCWRL4^[77]重构其侧链对输入模型进行预处理,而 VoromQA-B 在评估之前不会修改输入模型. 此外,特别是,序列信息中在包含潜在的蛋白质进

化关系,可以提高模型评估的准确性. ProQ4^[78]使用多序列比对的统计信息熵提升原有评估的精度. Bhattacharya-QDeepU(QDeep^[79]的变体方法)使用从全基因组序列数据库与宏基因组数据库合并生成的多序列比对信息(MSA)进行训练. VoroCNN-GEMME 使用 GEMME^[80]计算了每个残基的共进化描述符,其预测了在该序列位置发生突变对其他每个氨基酸的影响程度, GEMME 的输入也是 MSA 信息. DeepAccNet-MSA^[27]通过 trRosetta^[9]网络将 MSA 信息转换为几何约束特征输入神经网络预测质量分数.

深度学习网络可以捕获蛋白质内部的潜在联系. Venclovas 课题组^[81]开发 VoromQA-dark 是基于部分 VoromQA,通过神经网络(NN)来预测局部(每残基)CAD-score 值. 其针对每个氨基酸残基输出包括 3 个 CAD-score: CAD-score-level0 是基于涉及中心残基的所有氨基酸残基间接触; CAD-score-level1 是基于涉及至少一个来自中心残基的第一层邻居(直接邻居)的所有氨基酸残基间接触; CAD-score-level2 是基于中心残基的直接邻居和直接邻居的邻居与所有氨基酸残基之间的间接接触来计算的. 输入向量已经进行了预卷积操作,最终只使用了一个全连接隐藏层. VoroCNN^[40]是一种基于深度卷积神经网络的模型质量评估方法,它处理无向加权图表示的蛋白质模型. 为了处理这些图, VoroCNN 由一个基于消息传递图卷积层和一个池化层组成. 此外, VoroCNN-GDT 网络输出层之前增加了一个 1D 卷积层,以实现在蛋白质序列上有更好的局部质量预测的平滑性. Bhattacharya 课题组^[79]提出的 QDeep (Bhattacharya-QDeep)采用堆叠式深度 ResNet 估计模型在四个不同距离阈值 1, 2, 4 和 8 Å 下每残基的误差. 其中, 4 个 ResNet 网络独立训练. DeepQA^[82]使用多个特征(包括能量、物理化学性质和结构信息)输入到深度置信网络中预测质量,该网络由受限玻尔兹曼机(RBM)^[83]隐藏层和逻辑回归层构成的网络结构. AngularQA^[84]将原子结构信息转化为二面角和键长,并将序列信息通过 LSTM^[85]神经网络输入. 它使用每个残基作为时间步,预测模型的质量,并考虑 LSTM 单元的回值. GraphQA^[86]使用图卷积网络并使用与 ProQ4 相同的特征,将蛋白质分子转化为具有旋转不变性的图形来评估质量. tFold^[87]通过更改消息传递网络(MPNN)^[88]

的图形通用架构,学习了残基之间的相互作用对模型进行评分.

通过构建编解码可以更好地利用神经网络的模块,以实现更准确的预测. Baker 课题组^[27]开发的 DeepAccNet 是基于一维、二维和三维特征的模型,在不同层次上反映蛋白质模型. 它通过对三维原子网格在旋转不变的局部框架中对每个残基周围执行三维卷积操作来捕捉高分辨率原子空间结构. 二维特征提取了模型结构中所有残基对的信息,包括 Rosetta 残基间的相互作用项,进一步描述原子间相互作用的细节,而残基与残基的距离和角度特征提供了较低分辨率的结构信息. 在每个残基水平上的一维特征包括氨基酸序列、主链扭转角和 Rosetta 残基能量项. 该网络使用三维卷积评估局部原子环境,然后通过二维卷积提供全局环境来预测蛋白质的局部质量,并预测每个残基的质量精度和蛋白质模型中残基间的距离误差,并利用这些预测来指导蛋白质结构的精修和优化. 此外, AlphaFold2 通过 Evoformer 编码序列信息,并在 Structure 模块解码中预测原子坐标和结构的质量.

4.5 复合物结构模型评估方法

在 CASP15 中,模型质量评估从单体质量评估转移到复合物的质量评估. MULTICOM_qa 是结合了基于深度学习链间接触预测和界面接触概率评分的方法,使用一个蛋白质目标的多聚体模型池作为输入,预测它们的全局质量得分. 并使用 MMalign^[89] 将多聚体模型相互比对,并计算模型与池中其他模型之间的平均 TM-score 作为模型质量的度量. 此外,对于每个多聚体目标蛋白质,使用基于深度学习方法^[18] 预测的多聚体残基间接触或距离,计算链间残基接触的概率,并将其平均值作为模型全局质量的另一个度量. 最后,通过加权计算得到池中每个多聚物模型的最终预测质量得分. MULTICOM_egnn 基于 DProQA^[90] 将多聚体模型作为输入并将其表示为三维图,使用门控图 Transformer 架构预测 DockQ 质量分数. 此外, MULTICOM_deep 采用类似的方式.

McGuffin 课题组^[91] 开发了 ModFOLDdock 的三种变体: ModFOLDdock, ModFOLDdockR 和 ModFOLDdockS. 这些变体结合了一系列单模型、聚类和深度学习方法形成共识来计算评估复合物质量. ModFOLDdock 优化了预测分数与参考分数

的相关性, ModFOLDdockR 优化了挑选 Top 1 模型的能力,而 ModFOLDdockS 使用 MultiFOLD 方法从输入序列生成参考模型集,并使用多个评分方法将每个模型与参考集进行比较.

MUFold 和 MUFold2^[32] 结合 AlphaFold-Multimer^[92] 作为蛋白质复合物质量评估的方法. MUFold 采用了基于 AlphaFold-Multimer 预测结果的单阶段机器学习方法,而 MUFold2 则采用了两阶段机器学习方法. 在 MUFold2 中,首先使用 AlphaFold-Multimer 的输出结果训练一个模型进行初始预测,然后使用第二个预训练的模型生成更准确的预测结果.

VoroIF-jury^[93] 包含了两种界面评分方法:一种是通用的基于原子间接触面积的能量势函数,该势函数是从蛋白质界面的 VoroMQA 势能函数推导出来的;另一种 VoroIF-GNN^[93] 方法是基于接受由 Voronoi 镶嵌派生的蛋白质链间界面接触图的图注意力网络 (GAT) 预测复合物模型中的残基级别界面精度. 此外, APOLLO^[94] 使用基于能量模型 (EBM) 来评估整体折叠、界面准确性以及界面残基的置信度得分.

4.6 DeepUMQA 系列

张贵军课题组在最近几年开发了 DeepUMQA 系列、GraphGPSM 等模型质量局部及全局评估方法. 基于 DeepUMQA^[42-44] 系列算法开发的 Guijun-Lab-RocketX 服务器与基于 GraphGPSM^[95] 算法开发的 GuijunLab-Threader 服务器首次参加了 2022 年举行 CASP15, 并表现出了不错的性能.

DeepUMQA^[42] 基于超快速形状识别 (USR)^[96] 来补充对于描述残基级别的拓扑信息可能不足的情况,其能够与深度学习方法相结合进一步反映残基级别拓扑的特征来提高模型质量评估的性能. 体素化方法有效地描述了残基的局部结构信息,但它并未完全反映残基与整体结构之间的拓扑关系. 此外,体素化特征向量的计算和三维卷积非常复杂且耗时. 因此,通过选择适当的一组原子间距离,可以几乎不增加额外的计算成本快速捕捉蛋白质结构的拓扑信息. 具体而言,考虑了四个参考位置有效代表蛋白质结构中心和边界关系,并利用它们之间的距离子集构建蛋白质整体结构的拓扑关系.

DeepUMQA2^[44] 是基于 DeepUMQA 的显著改进版本. 在基于之前特征基础上,结合了来自多

序列比对的序列信息和同源模板的结构特征, 对模型的潜在属性进行表征. DeepUMQA2 首先根据输入模型的序列进行多序列比对 (MSA) 和同源模板搜索, 然后提取序列特征和模板结构特征, 并与输入模型相关特征结合, 形成初始残基对信息. 通过基于三角乘法更新和轴向注意机制的网络迭代更新残基对信息. 然后, 使用两个分支网络分别预测残基间距离偏差和接触图 (阈值为 15 Å), 进一步计算模型的每个残基的准确性.

DeepUMQA3^[97] 适用于评估蛋白质复合物模型质量的方法. 在 DeepUMQA 和 DeepUMQA2 的基础上, 为复合物结构设计了新的特征, 并使用改进的深度神经网络预测了每个残基的 IDDT 和界面残基的准确性. DeepUMQA3 在 CASP15 的蛋白质复合物界面残基准确性估计中名列第一, 参见图 3. 其 Web 服务器为蛋白质复合物提供了快速准确的界面残基准确性预测和每个残基的 IDDT 预测服务. 对于待评估的复合物结构, DeepUMQA3 从三个层次描述它: 整体复合物特征、单体内特征和单体间特征. 在整体复合物层次上, 将整个复合物视为一个大的单体结构. 考虑到蛋白质复合物在序列上是不连续的, 提取了与残基顺序无关的特征, 包括整体 USR、残基体素化、残基间距离和方向以及氨基酸性质. 在单体内层次上, 分别提取了

每个单体的特征, 包括由 ESM-1b^[98] 生成的序列嵌入、二级结构和 Rosetta 能量项. 在单体间层次上, 使用单体间成对序列的注意力图描述了单体之间的序列关系. 此外, 设计了单体间 USR 来描述一个单体中残基与其他单体的拓扑关系. 这三个层次的特征被输入带有三角形更新和轴向注意力的深度卷积神经网络, 以预测残基间距离偏差和阈值为 15 Å 的残基间接触图, 从而计算每个残基的 IDDT 和界面残基准确性.

在 DeepUMQA 系列算法基础上, 张贵军课题组^[99] 进一步结合图耦合网络开发了 GraphCP LMQA 算法. 算法利用蛋白质语言模型的嵌入来评估残基级别的蛋白质模型质量. GraphCPLMQA 由图编码模块和基于变换的卷积解码模块组成. 在编码模块中, 利用具有 ESM 蛋白质语言模型提取序列和高维几何结构的潜在关系表示, 能够捕捉蛋白质模型的序列和结构特征的重要信息. 在解码模块中, 利用提取的嵌入表示和低维特征推断蛋白质结构与质量之间的映射关系. 为了增强局部结构和整体拓扑之间的关联性, 设计了三角定位和残基级别接触顺序特征. 其中, 三角定位基于 DeepUMQA 中的 USR 引入了残基之间方向的信息, 可以更为充分地描述蛋白质局部空间的结构. 接触序 (contact order)^[100] 用于描述整体拓扑的复杂性, 并扩

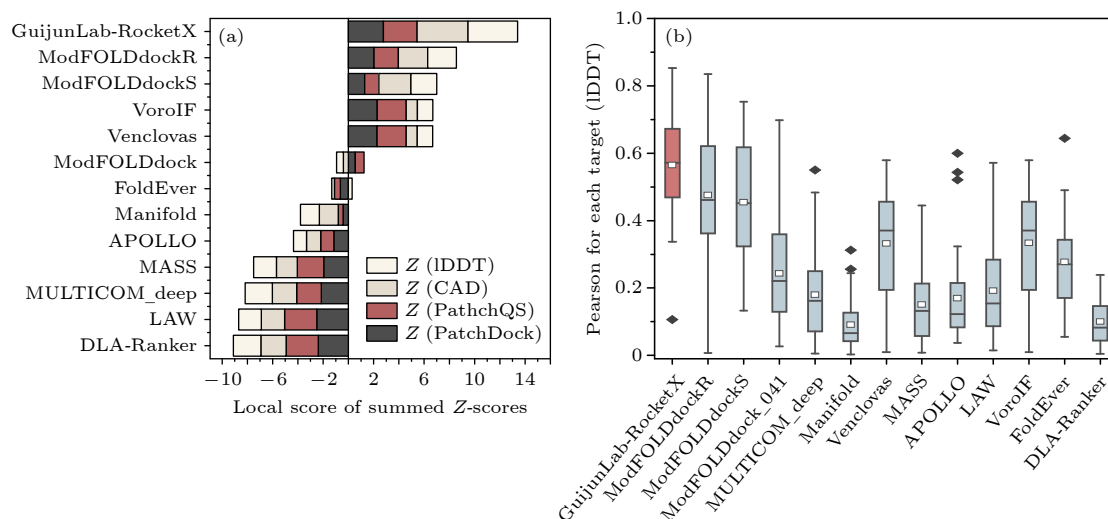


图 3 (a) IDDT, CAD, PatchDockQ 和 PatchQS 的平均 Z 分数之和, CASP15 官方公布各个小组在界面残基精确度估计排名 (数据来自 <https://predictioncenter.org/casp15>). CASP15 中 DeepUMQA3 的组名称为 “GuijunLab-RocketX”; (b) 针对 CASP15, 每个蛋白质目标上的预测的 IDDT 质量与真实 IDDT 质量的 Pearson 相关性, 其中, 白色方框是均值, 中间横线是中位数

Fig. 3. (a) The sum of average Z-scores of IDDT, CAD, PatchDockQ and PatchQS, CASP15 officially announces the ranking of each group in the interface residue accuracy estimation (data from <https://predictioncenter.org/casp15>). The group name of DeepUMQA3 in CASP15 is “GuijunLab-RocketX”. (b) Pearson correlation of predicted and true IDDT quality on each protein target. The white box is the mean and the middle horizontal line is the median.

表 2 CAMEO-QE: 模型质量评估性能 (数据来自官网 2022-6-24—2023-6-17)
Table 2. CAMEO-QE: Model Quality Evaluation Performance (Data from official website 2022-6-24—2023-6-17).

Predictor Name	ROC ^{normalized}		PR ^{normalized}		Models
	AUC _{0,1}	AUC* _{0,0.2}	AUC _{0,1}	AUC* _{0.8,1}	Received
ZJUT-GraphCPLMQA	0.82	0.73	0.79	0.54	5143
DeepUMQA2	0.72	0.62	0.68	0.47	4468
DeepUMQA	0.73	0.60	0.67	0.45	4611
ModFOLD9	0.63	0.52	0.59	0.36	4309
QMEANDisCo3	0.9	0.66	0.79	0.49	6348
ProQ3D_LDDT	0.74	0.55	0.67	0.43	5171
QMEAN3	0.88	0.65	0.77	0.43	6348
ProQ3	0.72	0.53	0.66	0.39	5126
VoroMQA_v2	0.89	0.64	0.77	0.45	6350
ProQ2	0.86	0.59	0.74	0.39	6337
ProQ3D	0.70	0.47	0.61	0.35	5119
ModFOLD7_IDDT	0.84	0.53	0.69	0.41	6191
ModFOLD8	0.79	0.50	0.65	0.38	5802
Baseline Potential	0.80	0.51	0.66	0.32	6350
VoroMQA_sw5	0.82	0.50	0.65	0.36	6349
ModFOLD6	0.73	0.42	0.57	0.35	5380

展到残基级别特征以描述局部结构之间的复杂性. 这些特征有助于捕捉蛋白质模型的局部结构元素与全局折叠模式之间的关系. 通过结合图编码模块和基于变换的卷积解码模块, 能够评估蛋白质模型的残基级别的质量. GraphCPLMQA 持续参加了一年的 CAEMO (<https://www.cameo3d.org>), 结果如下表 2 所列.

此外, 本课题组^[95]还开发了全局质量评估模型 GraphGPSM, 该模型利用高斯径向基函数对原子级别的主链特征进行编码, 基于 DeepUMQA 的 USR, Rosetta 能量项、距离和方向、序列的独热编码以及残基的位置嵌入来描述蛋白质结构. 这些特征被配置到初始图的节点和边上, 并与坐标嵌入相结合, 构建了 EGNN^[101]的初始架构. 通过堆叠 EGNN 架构形成了一个密集的消息传递网络. 最后, 通过多层感知器 (由 Dropout 层、激活函数和线性层组成) 生成结构模型的全局评分. 特别地, GraphGPSM (GuijunLab-Threader) 在 CASP15 性能如表 3 所列.

深度学习在蛋白质模型质量评估领域得到广泛应用, 并成为主流技术, 评估质量的效果也显著提升. 回顾模型质量评估方法, 可以得出以下几点结论:

1) 近三年来开发出的单模型方法大多都是基于深度学习. 尤其, 与之前 CASP 中最佳的单模型方

法以及 CASP 中最佳的多模型方法相比, CASP14 上最佳单模型方法 (DeepAccNet 和 DeepAccNet-MSA) 在全局结构准确性评估方面取得显著的提升. 虽然, 在 CASP15 全局质量评估和接口界面评估中最好的两种方法分别是 MULTICOM_qa 和 ModFOLDdock 这两种共识方法. 但是, 在局部接触界面的质量评估方法基于深度学习的 DeepUMQA3 相比于排名第二的共识方法具有显著的优势, 单模型方法依然是未来的发展趋势.

表 3 在所有蛋白质目标与 CASP15 服务器的性能比较 (数据来自 GraphGPSM)

Table 3. Performance comparison with CASP15 server on all protein targets (data from GraphGPSM).

Method	Average TM-score	Average Pearson	Average bias
GraphGPSM	0.730	0.633	0.126
MULTICOM_qa	0.485	0.715	0.258
ModFOLDdock	0.515	0.636	0.241
ModFOLDdockR	0.666	0.635	0.165
Venclovas	0.449	0.494	0.339
Manifold	0.582	0.541	0.179
Bhattacharya	0.387	0.474	0.361
*Real value	0.716	None	None

注: *Real value 代表 CASP15 中所有蛋白质目标所有模型的真实平均 TM-score 分数.

Note: *Real value represents the real average T-score of all targets in CASP15.

2) 从 CASP13—CASP15 模型质量评估的参赛组可以看出: 在 CASP13 中分别有 51 个和 29 个参赛组提交了全局和局部精度估计; 在 CASP14 中分别有 72 个和 38 个参赛组提交了对全局和局部精度估计; 在 CASP15 中分别有 22 个, 13 个和 17 个参赛组提交了全局, 局部和接触界面精度估计. 从 CASP13 至 CASP14 对于评估质量的参赛组的数量呈现上升的趋势, 但是从 CASP14 至 CASP15 的参赛数量非常明显的减少. 这可能的原因是: ①对于复合物的模型质量评估, 很多之前的参赛组并没有开发出相应的方法. ②现阶段复合物的结构模型质量评估依旧存在挑战.

3) 通过深度学习的发展历程可以看出, 在网络层面, 从 ProQ3D 简单的几层神经网络逐步引入了更加复杂的模型, 即 3DCNN 的 3 维卷积网络、AngularQA 的 LSTM 网络、GraphQA 的图神经网络、GraphPSM 的等变图网络, DeepUMQA2 的注意力机制网络以及编解码模块 AlphaFold2 或者 GraphCPLMQA. 在特征层面, 距离图的特征和序列编码向表征局部空间结构, 全局拓扑结构和进化信息设计特征描述蛋白质模型, 如 USR, 体素化, MSA 多序列比对信息等. 这表明深度网络的架构和蛋白质特征对网络模型性能的提升产生关键作用.

5 模型质量评估方法的挑战与发展趋势

模型质量评估方法在蛋白质结构预测中扮演着关键角色, 并持续成为该领域的研究热点. 然而, 这一领域依然面临许多挑战, 以下从单体模型评估、复合物模型评估和模型评估的共性问题三个方面进行讨论.

在单体模型评估方面, 尽管 AlphaFold2 已经取得了卓越的精度, 但对于缺乏多序列比对 (MSA) 数据或模板质量较低的情况, 建模精度仍存在局限性. 目前关键问题在于如何区分高质量模型 (如 AlphaFold2 生成的模型) 和低质量模型, 并评估高质量模型中需要改进的相对不正确区域. 此外, 目前蛋白质预测的结构数据库规模庞大, 如 AlphaFold Protein Structure Database (~2 亿) 和 ESM Metagenomic Atlas (~7 亿). 虽然这些预测结构有自评估的质量分数, 但是这些分数与预测的结构相

关性依然需要提升, 特别是在局部区域. 如何通过模型质量评估合理利用这些预测数据促进生物学研究值得深思.

在复合物评估方面, 研究者们面临着许多需要进一步探索的问题, 这些问题源于复合物结构的复杂性和多样性. 首先, 复合物的质量评估需要解决基于深度学习的方法如何构建适当的训练数据集的问题. 由于复合物模型可能包含多个链, 而蛋白质结构数据库中主要以双链结构为主, 如何有效地收集和组织复合物结构数据, 以便用于训练深度学习模型. 其次, 复合物的结构通常比单体结构更加复杂和庞大, 其复杂性意味着在网络训练过程中需要更大的计算和内存资源, 并且训练时间可能会显著增加. 最后, 复合物评估指标体系的建立和应用也需要进一步发展. 目前, 许多复合物的评估指标仍在沿用单体结构的评估方法, 然而复合物具有独特的结构和功能特征, 需要开发适用于复合物质量评估的专用指标, 以更好地反映复合物的质量和功能特性, 并促进复合物结构预测领域的进一步发展.

除了在单体和复合物评估中面临的挑战之外, 模型评估中还存在一些共性问题需要解决. 首先, 对于模型的质量评估, 传统上常常依赖于多序列比对 (MSA) 和模板的信息来提高评估的准确性. 然而, 在某些情况下, 蛋白质的序列可能缺乏足够的相关信息或者没有相关的模板结构可供参考. 因此, 如何仅仅利用蛋白质的单序列和结构本身的信息来评估模型的质量成为一个重要的问题. 其次, 在模型评估中, 有时会发现模型的结构在局部区域被认为是较低质量的, 然而却缺乏对这些局部结构进一步处理的方法. 如何在模型评估的基础上进行结构的精修成为一个需要关注的问题.

综上所述, 未来模型质量评估的趋势将聚焦于复合物模型结构的评估. 借助深度学习网络和最新技术的融合, 以及对复合物模型的结构和序列特征进行工程化的探索, 以揭示不同类型复合物的互作方式. 同时, 引入更加全面和合理的评估指标体系, 将进一步推动复合物结构预测的发展, 并为模型评估提供更加可靠和准确的基础. 这一努力的成果将为蛋白质领域带来更为深入的认知和应用前景, 为研究者揭示复合物结构的复杂性和功能特征提供更精准的工具和方法.

参考文献

- [1] Thompson M C, Yeates T O, Rodriguez J A 2020 *F1000 Research* **9** 667
- [2] Bai X C, McMullan G, Scheres S H 2015 *Trends Biochem. Sci.* **40** 49
- [3] Wüthrich K 2001 *Nat. Struct. Biol.* **8** 923
- [4] Steinegger M, Mirdita M, Söding J 2019 *Nat. Methods* **16** 603
- [5] Junger J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl S A, Ballard A J, 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, Bodenstern S, Silver D, Vinyals O, Senior A W, Kavukcuoglu K, Kohli P, Hassabis D 2021 *Nature* **596** 583
- [6] Rohl C A, Strauss C E, Misura K M, Baker D 2004 *Methods in Enzymology* (Amsterdam: Elsevier) pp66–93
- [7] Zhang Y 2008 *BMC Bioinf.* **9** 40
- [8] Källberg M, Wang H P, Wang S, Peng J, Wang Z Y, Lu H, Xu J B 2012 *Nat. Protoc* **7** 1511
- [9] Yang J Y, Anishchenko I, Park H, Peng Z L, Ovchinnikov S, Baker D 2020 *PNAS* **117** 1496
- [10] Zhao K L, Xia Y H, Zhang F J, Zhou X G, Li S Z, Zhang G J 2023 *Commun. Biol.* **6** 243
- [11] Lin Z M, Akin H, Rao R, Hie B, Zhu Z K, Lu W T, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Costa S D A, Zarandi F M, Sercu T, Candido S, Rives S 2023 *Science* **379** 1123
- [12] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A 2022 *Nucleic Acids Res.* **50** D439
- [13] Chen J R, Siu S W 2020 *Biomolecules* **10** 626
- [14] Zemla A J 2003 *Nucleic Acids Res.* **31** 3370
- [15] Zhang Y, Skolnick J 2004 *Proteins Struct. Funct. Bioinf.* **57** 702
- [16] Mariani V, Biasini M, Barbato A, Schwede T J 2013 *Bioinformatics* **29** 2722
- [17] Olechnovič K, Kulberkytė E, Venclovas Č 2013 *Proteins Struct. Funct. Bioinf.* **81** 149
- [18] Antczak P L M, Ratajczak T, Lukasiak P, Blazewicz J 2015 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* Washington D. C., November 9–12, 2015 p665
- [19] Moulton J, Fidelis K, Kryshchafovich A, Schwede T, Tramontano A 2016 *Proteins Struct. Funct. Bioinf.* **84** 4
- [20] Kryshchafovich A, Schwede T, Topf M, Fidelis K, Moulton J 2019 *Proteins Struct. Funct. Bioinf.* **87** 1011
- [21] Moulton J, Pedersen J T, Judson R, Fidelis K 1995 *Proteins Struct. Funct. Bioinf.* **23** R2
- [22] Robin X, Haas J, Gumienny R, Smolinski A, Tauriello G, Schwede T 2021 *Proteins Struct. Funct. Bioinf.* **89** 1977
- [23] Fowler N J, Williamson M P 2022 *Structure* **30** 925
- [24] Kryshchafovich A, Antczak M, Szachniuk M, Zok T, Kretsch R C, Rangan R, Pham P, Das R, Robin X, Studer G, Durairaj J, Eberhardt J, Sweeney A, Topf M, Schwede T, Fidelis K, Moulton J 2023 *Proteins Struct. Funct. Bioinf.* **91** 1550
- [25] Basu S, Wallner B 2016 *PLoS One* **11** e0161879
- [26] Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T 2017 *Sci. Rep.* **7** 10480
- [27] Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J Baker D 2021 *Nat. Commun.* **12** 1340
- [28] Wang Z, Eickholt J, Cheng J L 2010 *Bioinformatics* **26** 882
- [29] Cheng J L, Wang Z, Tegge A N, Eickholt J 2009 *Proteins Struct. Funct. Bioinf.* **77** 181
- [30] Wu T Q, Guo Z Y, Hou J, Cheng J L 2021 *BMC Bioinf.* **22** 1
- [31] Wang J L, Wang W B, Shang Y, Xu D 2022 *IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)* Las Vegas, NV, USA & Changsha, China, December 6–8, 2022 p84
- [32] Wang W B, Li Z Y, Wang J L, Xu D, Shang Y 2019 *Nucleic Acids Res.* **47** W443
- [33] McGuffin L J, Aldowsari F M, Alharbi S M, Adiyaman R 2021 *Nucleic Acids Res.* **49** W425
- [34] McGuffin L J, Buenavista M T, Roche D B 2013 *Nucleic Acids Res.* **41** W368
- [35] McGuffin L J 2008 *Bioinformatics* **24** 586
- [36] Uziela K, Wallner B 2016 *Bioinformatics* **32** 1411
- [37] Uziela K, Shu N, Wallner B, Elofsson A 2016 *Sci. Rep.* **6** 33509
- [38] Olechnovič K, Venclovas Č 2017 *Proteins Struct. Funct. Bioinf.* **85** 1131
- [39] Olechnovič K, Venclovas Č 2019 *Nucleic Acids Res.* **47** W437
- [40] Igashov I, Olechnovič K, Kadukova M, Venclovas Č, Grudinina S 2021 *Bioinformatics* **37** 2332
- [41] Ye L S, Wu P K, Peng Z L, Gao J Z, Liu J, Yang J Y 2021 *Bioinformatics* **37** 3752
- [42] Guo S S, Liu J, Zhou X G, Zhang G J 2022 *Bioinformatics* **38** 1895
- [43] Liu J, Liu D, He G X, Zhang G J 2023 *Proteins Struct. Funct. Bioinf.* **91** 1861
- [44] Liu J, Zhao K L, Zhang G J 2023 *Brief. Bioinform.* **24** bbac507
- [45] Kryshchafovich A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A 2014 *Proteins Struct. Funct. Bioinf.* **82** 112
- [46] Kryshchafovich A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A 2018 *Proteins Struct. Funct. Bioinf.* **86** 345
- [47] Won J, Baek M, Monastyrskyy B, Kryshchafovich A, Seok C 2019 *Proteins Struct. Funct. Bioinf.* **87** 1351
- [48] Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, Mostaguir K, Gumienny R, Schwede T 2018 *Proteins Struct. Funct. Bioinf.* **86** 387
- [49] Jones T A, Kleywegt G J 1999 *Proteins Struct. Funct. Bioinf.* **37** 30
- [50] Martin A C, MacArthur M W, Thornton J M 1997 *Proteins Struct. Funct. Bioinf.* **29** 14
- [51] Keedy D A, Williams C J, Headd J J, Arendall III W B, Chen V B, Kapral G J, Gillespie R A, Block J N, Zemla A, Richardson D C, Richardson 2009 *Proteins Struct. Funct. Bioinf.* **77** 29
- [52] Janin J, Henrick K, Moulton J, Eyck T L, Sternberg G E, Vajda S, Vakser L, Wodak S J 2003 *Proteins Struct. Funct. Bioinf.* **52** 2
- [53] Lipton Z C, Elkan C, Narayanaswamy B 2014 *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014*, Nancy, France, September 15–19, 2014 p225
- [54] Ozden B, Kryshchafovich A, Karaca E 2021 *Proteins Struct. Funct. Bioinf.* **89** 1787
- [55] Kwon S, Won J, Kryshchafovich A, Seok C 2021 *Proteins Struct. Funct. Bioinf.* **89** 1940

- [56] Lobo J M, Jiménez-Valverde A, Real R 2008 *Global Ecol. Biogeogr.* **17** 145
- [57] Spearman correlation coefficients, differences between, Myers L, Sirois M J <https://doi.org/10.1002/0471667196.ess5050.pub2> [2023-11-21]
- [58] Ron K, Foster P 1998 *J. Mach. Learn.* **30** 271
- [59] Wang W B, Wang J L, Li Z Y, Xu D, Shang Y 2021 *Comput. Struct. Biotechnol. J.* **19** 6282
- [60] McGuffin L J, Roche D B 2010 *Bioinformatics* **26** 182
- [61] McGuffin L J 2009 *Proteins Struct. Funct. Bioinf.* **77** 185
- [62] Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman J L, Levy Y 2009 *Proteins Struct. Funct. Bioinf.* **77** 50
- [63] Alapati R, Bhattacharya D 2018 *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* Washington DC, USA, August 29–September 1, 2018 p307
- [64] Cheng J L, Choe M H, Elofsson A, Han K S, Hou J, Maghrabi A H, McGuffin L J, Menéndez-Hurtado D, Olechnovič K, Schwede T, Studer G, Uziela K, Venclovas Č, Wallner B 2019 *Proteins Struct. Funct. Bioinf.* **87** 1361
- [65] Bitton M, Keasar C 2022 *Sci. Rep.* **12** 14074.
- [66] Ke G L, Meng Q, Finley T, Wang T F, Chen W, Ma W D, Ye Q W, Liu T Y 2017 *Adv. Neural Inf. Process. Syst.* **30** 3149
- [67] Maghrabi A H, McGuffin L J 2017 *Nucleic Acids Res.* **45** W416
- [68] Maghrabi A H, McGuffin L J 2020 *Protein Struct. Prediction* **2165** 69
- [69] McGuffin L J, Shuid A N, Kempster R, Maghrabi A H, Nealon J O, Salehe B R, Atkins J D, Roche D B 2018 *Proteins Struct. Funct. Bioinf.* **86** 335
- [70] Studer G, Rempfer C, Waterhouse A M, Gummienny R, Haas J, Schwede T 2020 *Bioinformatics* **36** 1765
- [71] Benkert P, Tosatto S C, Schomburg D 2008 *Proteins Struct. Funct. Bioinf.* **71** 261
- [72] Manavalan B, Lee J 2017 *Bioinformatics* **33** 2496
- [73] Derevyanko G, Grudinin S, Bengio Y, Lamoureux G 2018 *Bioinformatics* **34** 4046
- [74] Pagès G, Charmettant B, Grudinin S 2019 *Bioinformatics* **35** 3313
- [75] Uziela K, Menéndez Hurtado D, Shu N, Wallner B, Elofsson A 2017 *Bioinformatics* **33** 1578
- [76] Rother K, Hildebrand PW, Goede A, Gruening B, Preissner R 2009 *Nucleic Acids Res.* **37** D393
- [77] Krivov G G, Shapovalov M V, Dunbrack Jr R L 2009 *Proteins Struct. Funct. Bioinf.* **77** 778
- [78] Hurtado D M, Uziela K, Elofsson A 2018 [arXiv:1804.06281](https://arxiv.org/abs/1804.06281) [q-bio.BM]
- [79] Shuvo M H, Bhattacharya S, Bhattacharya D 2020 *Bioinformatics* **36** i285
- [80] Laine E, Karami Y, Carbone A 2019 *Mol. Biol. Evol.* **36** 2604
- [81] Dapkūnas J, Olechnovič K, Venclovas Č 2021 *Proteins Struct. Funct. Bioinf.* **89** 1834
- [82] Cao R Z, Bhattacharya D, Hou J, Cheng J L 2016 *BMC Bioinf.* **17** 495
- [83] Fischer A, Igel C 2012 *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012*, Buenos Aires, Argentina, September 3–6, 2012 p14
- [84] Conover M, Staples M, Si D, Sun M, Cao R Z 2019 *Comput. Math. Biophys.* **7** 1
- [85] Yu Y, Si X S, Hu C H, Zhang J X 2019 *Neural Comput.* **31** 1235
- [86] Baldassarre F, Menéndez Hurtado D, Elofsson A, Azizpour H 2021 *Bioinformatics* **37** 360
- [87] Shen T, Wu J X, Lan H D, Zheng L Z, Pei J G, Wang S, Liu W, Huang J Z 2021 *Proteins Struct. Funct. Bioinf.* **89** 1901
- [88] Gilmer J, Schoenholz S S, Riley P F, Vinyals O, Dahl G 2017 *International Conference on Machine Learning* Sydney, Australia, August 6–11, 2017 p1263
- [89] Mukherjee S, Zhang Y 2009 *Nucleic Acids Res.* **37** e83
- [90] Chen X, Morehead A, Liu J, Cheng J L 2023 *Bioinformatics* **39** i308
- [91] McGuffin L J, Edmunds N S, Genc A G, Alharbi S, Salehe B R, Adiyaman R 2023 *Nucleic Acids Res.* **51** W274
- [92] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Židek A, Bates R, Blackwell S, Yim J, Ronneberger O, Bodenstein S, Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K, Jain R, Clancy E, Kohli P, Jumper J, Hassabis D 2022 [bioRxiv 2021.10.04.463034](https://arxiv.org/abs/2021.10.04.463034)
- [93] Olechnovic K, Venclovas Č 2023 *Proteins Struct. Funct. Bioinf.* **91** 1879
- [94] Wang Z, Eickholt J, Cheng J L 2011 *Bioinformatics* **27** 1715
- [95] He G, Liu J, Liu D, Zhang G 2023 *Brief. Bioinform.* **24** 4
- [96] Ballester P J, Richards W G 2007 *J. Comput. Chem.* **28** 1711
- [97] Liu J, Liu D, Zhang G 2023 [bioRxiv 2023.04.24.538194](https://arxiv.org/abs/2023.04.24.538194)
- [98] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A 2021 *Adv. Neural Inf. Process. Syst.* **34** 29287
- [99] Ivankov D N, Garbuzynskiy S O, Alm E, Plaxco K W, Baker D, Finkelstein A V 2003 *Protein Sci.* **12** 2057
- [100] Liu D, Zhang B, Liu J, Li H, Song L, Zhang G 2023 [bioRxiv 2023.05.16.540981](https://arxiv.org/abs/2023.05.16.540981)
- [101] Satorras V G, Hoogeboom E, Welling M 2021 *International Conference on Machine Learning* Vienna, Austria, July 18–24, 2021 p9323

SPECIAL TOPIC—Machine learning in biomolecular simulations

Recent advances in estimating protein structure model accuracy^{*}

Liu Dong Cui Xin-Yue Wang Hao-Dong Zhang Gui-Jun[†]*(School of Information Engineering, Zhejiang University of Technology, Hangzhou 310014, China)*

(Received 30 June 2023; revised manuscript received 1 August 2023)

Abstract

The quality assessment of protein models is a key technology in protein structure prediction and has become a prominent research focus in the field of structural bioinformatics since advent of CASP7. Model quality assessment method not only guides the refinement of protein structure model but also plays a crucial role in selecting the best model from multiple candidate conformations, offering significant value in biological research and practical applications. This study begins with reviewing the critical assessment of protein structure prediction (CASP) and continuous automated model evaluation (CAMEO), and model evaluation metrics for monomeric and complex proteins. It primarily summarizes the development of model quality assessment methods in the last five years, including consensus methods (multi-model methods), single-model methods, and quasi-single-model methods, and also introduces the evaluation methods for protein complex models in CASP15. Given the remarkable progress of deep learning in protein prediction, the article focuses on the in-depth application of deep learning in single-model methods, including data set generation, protein feature extraction, and network architecture construction. Additionally, it presents the recent efforts of our research group in the field of model quality assessment. Finally, the article analyzes the limitations and challenges of current protein model quality assessment technology, and also looks forward to future development trends.

Keywords: protein model quality assessment, deep learning, single-model methods, complex model evaluation**PACS:** 87.10.Vg, 87.14.E-, 87.16.A-, 87.55.de**DOI:** [10.7498/aps.72.20231071](https://doi.org/10.7498/aps.72.20231071)

^{*} Project supported by the Scientific and Technological Innovation 2030—“New Generation Artificial Intelligence”, China (Grant No. 2022ZD0115103), the National Nature Science Foundation of China (Grant No. 62173304), and the Key Project of Zhejiang Provincial Natural Science Foundation of China (Grant No. LZ20F030002).

[†] Corresponding author. E-mail: zgj@zjut.edu.cn

专题: 生物分子模拟中的机器学习

蛋白质 pK_a 预测模型研究进展*

罗方芳 蔡志涛 黄艳东†

(集美大学计算机工程学院, 厦门 361021)

(2023年8月20日收到; 2023年9月1日收到修改稿)

pH 表征溶液的酸碱性, 是许多与人类重大疾病密切相关的生命活动的调控因子. pK_a 决定可滴定基团在一定 pH 条件下的去质子化平衡, 是研究 pH 调控的生物化学过程的重要参量. 然而, 由于蛋白质结构的复杂性以及实验条件的限制, 蛋白质 pK_a 通常需要借理论预测. 近 30 年, 研究者们开发了各种基于先验知识的 pK_a 预测模型. 随着近几年人工智能技术的快速发展, 人们开始尝试将人工智能算法应用于蛋白质 pK_a 预测工具的开发. 本文介绍 pK_a 理论预测近年来的一些重要研究进展, 主要包括恒定 pH 分子动力学以及基于泊松-玻尔兹曼方程、经验函数和机器学习的 pK_a 预测模型. 在此基础上, 讨论蛋白质 pK_a 预测模型的未来发展方向和应用前景.

关键词: 分子动力学, 泊松-玻尔兹曼方程, 机器学习, pK_a 预测

PACS: 87.15.ap, 87.14.E-, 87.10.Vg, 87.15.A-

DOI: 10.7498/aps.72.20231356

1 引言

为保证正常的生命活动, 人体细胞的细胞质、细胞核以及各个细胞器需维持在特定的 pH 水平. 例如, 线粒体和溶酶体的 pH 分别是 8.0 和 4.7, 偏离细胞质的 7.2^[1]. 其中, 用于表征溶液的酸碱度的 pH 为氢离子浓度的对数取负 ($pH = -\log[H^+]$), 其是人体中许多重要生物过程的调控因子, 例如物质跨膜转运^[2]、酶催化^[3]、蛋白质折叠^[4]、多肽聚集^[5]、脂质分子自组装^[6]、病毒入侵细胞^[7]和细胞能量代谢^[8]. 从微观的角度, 以上生物过程均与关键可离子化基团的质子化 (protonation) 或去质子化反应 (deprotonation) 相关联. 可离子化基团的去质子化 (正反应) 和质子化反应 (逆反应): $AH \rightleftharpoons A^- + H^+$, 其中, AH 是一种可离子化基团的质子化态, A^- 是去质子化态.

以 β 分泌酶 BACE1 为例阐述蛋白质功能和

可离子化基团质子化/去质子化的关系. BACE1 的生物功能是裂解 β 淀粉样前体蛋白 APP. 它与神经退行性疾病阿尔茨海默症密切相关, 是典型的结构和功能依赖于 pH 的蛋白质. 该蛋白的催化中心含两个天冬氨酸 Asp32 和 Asp228 (图 1(a)). 实验指出, BACE1 仅在一个狭小的 pH 范围内具有活性^[9]. 如图 1(b) 所示, 在最适 pH 条件下 (约等于 4.5), Asp32 处于质子化态, 扮演质子供体 (proton donor); Asp228 处于去质子化态, 扮演亲核试剂 (nucleophile). 然而, 当溶液 pH 偏离 4.5, 两个天冬氨酸同时质子化或去质子化, BACE1 无法行使其生物功能^[10].

当一个可离子化基团的质子化和去质子化达到平衡, 可由以下公式计算解离常数 K_a :

$$K_a = \frac{[H^+][A^-]}{[AH]}, \quad (1)$$

其中, $[H^+]$, $[A^-]$ 和 $[AH]$ 分别代表溶液中氢离子

* 国家自然科学基金 (批准号: 11804114, 62006096)、福建省自然科学基金 (批准号: 2023J01329, 2020J05146)、厦门市自然科学基金 (批准号: 3502Z20227205) 和集美大学校启动金 (批准号: ZQ2020027) 资助的课题.

† 通信作者. E-mail: yandonghuang@jmu.edu.cn

以及该基团去质子化和质子化态下的浓度. K_a 代表一种酸 (如 AH) 离解氢离子的能力. 将方程 (1) 的两边对数取负, 可得到著名的 Henderson-Hasselbalch 方程:

$$\text{pH} = \text{p}K_a + \log \left(\frac{A^-}{AH} \right) \quad (2)$$

其中, $\text{p}K_a$ 为解离常数 K_a 的对数取负, 代表一种酸 (如 AH) 去质子化的难易程度. 例如, 溶液中天冬氨酸的 $\text{p}K_a$ 测量值是 3.7^[11]. 根据 (2) 式, 天冬氨酸在中性 ($\text{pH} = 7.0$) 水溶液中处于去质子化态 (A^-); 在 pH 小于 3.7 的酸性溶液中, 天冬氨酸质子化 (AH); 当 pH 位于 $\text{p}K_a$ 附近, 质子化和去质子化态共存. 如上所述, $\text{p}K_a$ 决定了可离子化基团在任意 pH 条件下的质子化和去质子化反应平衡. 根据 $\text{p}K_a$ 值, 可以推断不同 pH 条件下生物大分子质子化态的分布, 进而讨论结构和功能的关系. 因此, $\text{p}K_a$ 是研究 pH 相关的生物化学过程的一个核心问题. 不仅如此, $\text{p}K_a$ 与药物研发中长期存在的靶向性和抗药性问题以及蛋白质设计密切相关. 然而, 由于蛋白质结构的复杂性以及实验条件的限制, 人们难于通过实验获取蛋白质中可离子化氨基酸残基的 $\text{p}K_a$, 需借助理论预测.

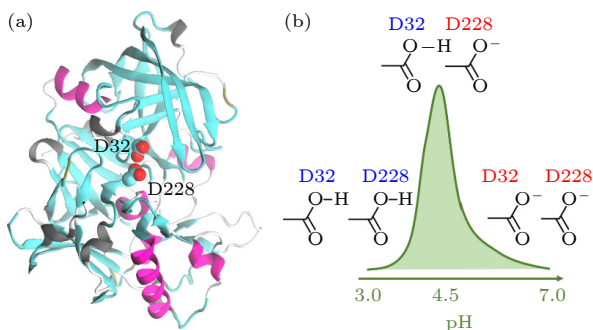


图 1 BACE1 催化中心质子化态和功能的关系 (a) BACE1 三维结构及其催化中心酸性二分体 D32 和 D228; (b) D32 和 D228 质子化态和蛋白质活性随 pH 的变化规律 (D 是 Asp 的缩写)

Fig. 1. Relationship between protonation state of BACE1 catalytic center and the function: (a) Crystal structure of BACE1 and the acidic dyad in the catalytic center; (b) protonation states of D32 and D228 and the activity as a function of pH (D is the abbreviation of Asp).

为此, 将以上 Henderson-Hasselbalch 方程转换为能量形式, 得到游离氨基酸关于 pH 和 $\text{p}K_a$ 的去质子化自由能 ΔG^{mod} 的表达式:

$$\Delta G^{\text{mod}} = \ln 10 \times k_B T \left(\text{p}K_a^{\text{mod}} - \text{pH} \right), \quad (3)$$

其中, k_B 和 T 分别是玻尔兹曼常数和温度; $\text{p}K_a^{\text{mod}}$ 为游离氨基酸的 $\text{p}K_a$, 是可测量值. 去质子化自由能可分解为成键作用部分 ΔG_{Bond} 和非键作用部分 ΔG_{NBond} . 其中, 成键作用部分描述共价键断裂的能量变化, 计算复杂度高, 不适用于生物大分子体系^[12]. 值得一提的是, 当溶剂中的可离子化氨基酸参与蛋白质的合成, 蛋白质环境对成键作用部分的影响可忽略不计. 基于该假设, 我们只需考虑非键作用部分. 因此, 可离子化氨基酸从溶剂到蛋白质的去质子化自由能改变量 $\Delta G - \Delta G^{\text{mod}}$ 可表示为

$$\Delta G - \Delta G^{\text{mod}} = \Delta G_{\text{NBond}} - \Delta G_{\text{NBond}}^{\text{mod}}. \quad (4)$$

根据 (3) 式, ΔG^{mod} 为已知量. 因此, 求解蛋白质中氨基酸残基的去质子化自由能 ΔG 的问题简化为计算蛋白质环境对非键作用部分的自由能微扰 $\Delta G_{\text{NBond}} - \Delta G_{\text{NBond}}^{\text{mod}}$.

基于以上框架, 人们发展了基于自由能计算的蛋白质 $\text{p}K_a$ 预测模型, 例如恒定 pH 分子动力学 (constant pH molecular dynamics, CpHMD)^[13]. 许多生物大分子含有不止一个功能构象, 并且构象的转变与质子化/去质子化反应相关联: 当活性位点质子化 ($\text{pH} < \text{p}K_a$), 蛋白处于构象 C_1 ; 去质子化 ($\text{pH} > \text{p}K_a$), 构象由 C_1 转变到 C_2 ; 当 pH 取 $\text{p}K_a$ 附近, 质子化和去质子化态共存, 构象 C_1 与 C_2 相互转变. 因此, 只有考虑了构象与质子化态耦合的理论模型, 才能得到和实验相一致的宏观 $\text{p}K_a$ (macroscopic $\text{p}K_a$)^[14]. CpHMD 通过分子动力学模拟实现在不同构象下对质子化态空间进行采样. 在蛋白质 $\text{p}K_a$ 预测精度方面, CpHMD 相对其他现有模型具有明显的优势^[15]. CpHMD 的缺点是 $\text{p}K_a$ 计算效率低. 例如, 完成一个蛋白质 $\text{p}K_a$ 的计算通常需要进行几个小时甚至几天的分子动力学模拟, 因此难以满足工业界大批量计算的需求. 目前, CpHMD 多被应用于结构和功能依赖于 pH 的药物靶向蛋白的分子机制研究^[16].

为了实现高通量的 $\text{p}K_a$ 计算, 人们发展了基于泊松-玻尔兹曼 (Poisson-Boltzmann, PB) 方程的模型, 主要包括 MCCE^[17], H++^[18], APBS^[19], DelPhi-PKa^[20] 和 PypKa^[21]. 基于 PB 的模型能够在几分钟内完成一个蛋白质的 $\text{p}K_a$ 计算, 极大地提高了计算效率. 然而, 基于 PB 的模型具有其理论局限性. 例如, 由于连续介质假设, PB 方程不适用于非水溶性的膜蛋白. 其次, 蛋白质结构的复杂性增加了

介电常数的不确定性, 因此即便是水溶性蛋白, 分子内部 (例如酶的催化反应中心) 的 pK_a 计算对介电常数敏感^[22].

除了以上基于能量的模型, 人们也可以用一个经验函数描述某可离子化氨基酸残基的蛋白质环境 (如疏水环境和氢键) 与其 pK_a 偏移量的映射关系. 蛋白质某氨基酸残基 pK_a 可表示为其游离状态下参考值 pK_a^{mod} 和偏移量 ΔpK_a 的和:

$$pK_a = pK_a^{\text{mod}} + \Delta pK_a. \quad (5)$$

2005 年, 基于前期的第一性原理计算工作^[12], 哥本哈根大学 Jensen 课题组^[23] 提出了一个计算蛋白质 pK_a 的经验函数 PropKa. 该模型提出一组经验公式分别计算库仑力、去溶剂化效应和氢键等关键因素对 pK_a 偏离参考值的贡献. PropKa 可在几秒内完成一个蛋白质的 pK_a 计算, 计算效率明显比基于 PB 的模型高, 近 20 年得到了广泛的应用, 其最新版本 PropKa 3.0 发表于 2011 年^[24].

直到 2021 年 12 月, 本课题组^[25] 发表了首个人工智能 (artificial intelligence, AI) 驱动的蛋白质 pK_a 预测模型 DeepKa. 随后, 美国卡内基·梅隆大学 Olexandr Lsayev、美国约翰斯·霍普金斯大学 Ana Damjanovic 和德国拜耳公司 Pedro Reis 研究小组陆续提出了基于机器学习的 pK_a 预测模型 pKa-ANI^[26], XGB-WMa^[27] 和 PKAI/PKAI+^[28]. 其中, DeepKa 和 PKAI/PKAI+ 主要依赖于数据集, 而为了在少样本情况下建立有效模型, pKa-ANI 和 XGB-WMa 需要一定程度的预训练或先验知识. 值得一提的是, 机器学习模型也能够在这几秒内完成一个蛋白质的 pK_a 计算.

上述的 CpHMD 以及基于 PB 方程、经验函数和机器学习的模型是目前 4 种主流的 pK_a 预测方法. 最近, 本课题组^[29] 采用 CpHMD 扩增了 pK_a 数据集, 进一步提高了 DeepKa 的预测精度. 值得一提的是, DeepKa 已展现出类似物理模型 (如 CpHMD) 的高鲁棒性, 进一步证明了人工智能算法在蛋白质 pK_a 预测领域的有效性. 下面将介绍这 4 种主流方法的理论基础及研究进展.

2 蛋白质 pK_a 预测方法

2.1 CpHMD

根据质子化态采样方法的不同, 恒定 pH 分子

动力学 CpHMD 分为随机采样 (discrete CpHMD, D-CpHMD)^[30] 和 λ 动力学 (continuous CpHMD, C-CpHMD)^[31]. 随机采样利用蒙特卡罗 (Monte Carlo, MC) 模拟在离散的质子化态空间 (反应坐标取 0 或 1) 进行采样^[30]. λ 动力学则采用取值范围 0 (质子化态) 到 1 (去质子化态) 的连续变量 λ 作为反应坐标对可离子化基团的电荷或体系哈密顿量进行标度^[31]. 如图 2 所示, 先使用以上基于 MC 或 λ 动力学的采样算法更新质子化态或者电荷. 基于更新后的电荷分布, 通过分子动力学模拟对构象进行采样. 更新位置坐标后, 进入下一轮质子化态的采样. 模拟结束后, 采用广义 Henderson-Hasselbalch 方程拟合 CpHMD 模拟产生的不同 pH 条件下某可离子化基团的去质子化概率 S , 进而获得其 pK_a 值, 即 $S = 0.5$ 所对应的 pH^[31].

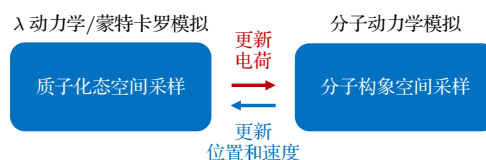


图 2 CpHMD 模拟框架

Fig. 2. Framework of a CpHMD simulation.

由于滴定动力学与构象动力学相关联, 提高质子化态和构象空间的采样是近 30 年 CpHMD 模型发展的主线. 下面将分别介绍 D-CpHMD 和 C-CpHMD.

2.1.1 D-CpHMD

D-CpHMD 用一个反应坐标 λ 表示某可离子化位点的质子化态. λ 只能取 0 或 1. 其中, 0 和 1 分别表示质子化态和去质子化态. 经过一定长度的分子动力学 (molecular dynamics, MD) 模拟, 随机选取一个可离子化基团, 尝试改变其质子化态. 例如, 将其 λ 值从 0 改为 1. 然后, 计算 λ 值改变引起的能量变化 ΔE . 将该能量变化代入 Metropolis 准则:

$$p = \begin{cases} 1, & \Delta E \leq 0, \\ \exp(-\Delta E/k_B T), & \Delta E > 0. \end{cases} \quad (6)$$

如果能量差小于或等于 0, 接受 λ 值改变的概率为 1. 如果能量差大于 0, 则接受改变的概率 p 小于 1. 在数值模拟中, 通常是随机生成一个取值范围为 $[0, 1]$ 的数 s . 只有 s 小于等于 p , 才接受 λ 值改变, 否则保留原值. 以上为一步的 MC, 和开始

的 MD 构成一个模拟周期. 因此, 在 MC 之后, 便是下一个周期的 MD 模拟. 显性溶剂下质子化或去质子化的能量变化较大, 导致较小的接受概率. 起初, 为了提高接受概率或质子化态的采样效率, MC 的能量计算使用隐性溶剂 (implicit solvent) 模型, 如广义玻恩 (generalized Born, GB)^[32-34] 和引言提到的 PB 模型^[31,35,36]. 当 MC 和 MD 均采用隐性溶剂, 计算效率最高, 但是牺牲了精度^[32,33]. 为了提高构象方面的采样精度, MD 可替换成显性溶剂, 即杂化溶剂^[31,34,35]. 其中, 基于 GB 和 PB 的模型分别在分子模拟软件 Amber 和 GROMACS 中已被实现. 需要指出的是, 隐性溶剂难以描述活性位点附近与功能相关的水分子或盐离子对去质子化平衡的影响^[37].

为提高显性水溶剂下 MC 的接受概率, 2007 年 Stern^[38] 提出了尝试改变 λ 值之后, 先进行一定长度的尝试性的分子动力学模拟, 再计算能量差. 该尝试性的 MD 使周围水溶剂构型得到调整, 可降低 λ 值改变前后的能量差. 然而, 以上尝试性 MD 的长度依赖于经验或不确定, 其应用可能受到限制. 尽管如此, 该模型为解决显性溶剂下质子化态空间的采样问题提供了一条新思路. 随着高性能计算的发展, 人们开始考虑将显性溶剂应用到蛋白质 D-CpHMD 的 MC 部分. 如无特别说明, 以下提到的显性溶剂均是分子动力学模拟中计算静电相互作用的标准算法 PME (particle mesh Ewald, PME)^[39]. 2015 年芝加哥大学的 Roux 课题组^[40] 提出了显性溶剂下的非平衡 MD/MC 模拟. 例如, 对于某可滴定位点在 MC 阶段的去质子化 (λ 由 0 变为 1) 尝试, 该模型在 0 和 1 之间添加了 m 个中间值. 对于每个 λ 值 (m 个中间值和两个边界值 0 和 1), 执行一定长度的非平衡 MD, 令可离子化基团周围的环境根据 λ 值在构型上作出调整, 减缓了因 λ 值改变而导致的能量涨落. 结束 $\lambda = 1$ 的非平衡 MD 后, 计算当 $\lambda = 1$ 和 $\lambda = 0$ 的能量差. 同样, 根据 Metropolis 准则, 如果接受该可滴定位点去质子化, 继续 $\lambda = 1$ 的 MD. 否则, 退回到非平衡 MD 前的时刻, 继续 $\lambda = 0$ 的 MD. 通过以上的非平衡模拟, 该模型提供了较合理的能量差的计算, 提高了总体接受概率. Roux 课题组^[40,41] 利用著名的 Jarzynski 方程将自由能变化与非平衡 MD 所做的功相关联, 使得以上非平衡 MD 的模拟时间可被量化. 值得一提的是, 该方法可被应

用于生物大分子, 目前在分子模拟软件 NAMD 中已有实现. 然而, 可滴定氨基酸的固有 pK_a (inherent pK_a) 是该模型的一个主要参量. 为了提高预测性能, 该模型要求固有 pK_a 尽可能接近真实值^[41]. 因此, D-CpHMD 一个潜在的研究方向是消除上述模型对固有 pK_a 的依赖.

2.1.2 C-CpHMD

本课题组统计了 4057 个蛋白质中可滴定氨基酸的个数^[29]. 这些蛋白质来自复旦大学王任小实验室^[42] 创建的蛋白质抑制剂复合物数据库 PDBbind 的精细集 v2016. 除了半胱氨酸 Cys, 蛋白质中其他可滴定氨基酸类型 (谷氨酸 Glu、天冬氨酸 Asp、赖氨酸 Lys、精氨酸 Arg、酪氨酸 Tyr、组氨酸 His) 的平均个数不低于 10^[29]. 理论上, 一个含有 N 个可滴定氨基酸残基的蛋白质包含 2^N 个质子化态. 然而, D-CpHMD 的 MC 每次只取一个可滴定位点来判断是否改变其质子化态, 采样效率较低^[34,43,44].

2004 年, 为了研究生物大分子体系 (如蛋白质, DNA 和 RNA) 的质子化和去质子化, 密西根大学 Brooks 课题组开发了首个 λ 动力学框架下^[45] 的恒定 pH 分子动力学 C-CpHMD^[31]. 每个可滴定位点对应一个反应坐标 λ , 取值范围同样是 0—1. 和 D-CpHMD 不同的是, C-CpHMD 的反应坐标是连续的变量. 值得一提的是, C-CpHMD 同时更新所有可滴定位点的质子化态. 哈密顿量 H 代表体系的总能量, 包括动能和势能. 除了真实的粒子, 如模拟体系中溶剂和溶质的原子, C-CpHMD 添加了虚粒子. 每个可滴定基团对应一个虚粒子. 这里用范围在 $[0, 1]$ 的连续变量 λ 作为虚粒子的坐标. 为了模拟虚粒子的滴定动力学, 可将其质量设为 10 (单位是原子质量). 以下是修正后的总哈密顿量:

$$\begin{aligned}
 H(\{\mathbf{r}_a\}, \{\lambda_j\}) &= \sum_a^{N_{\text{atom}}} \frac{1}{2} m_a \dot{\mathbf{r}}_a^2 + U^{\text{bond}}(\{\mathbf{r}_a\}) + U^{\text{nbond}}(\{\mathbf{r}_a\}, \{\lambda_j\}) \\
 &+ \sum_j^{N_{\text{virt}}} \frac{1}{2} m_j \dot{\lambda}_j^2 + U^*(\{\lambda_j\}), \quad (7)
 \end{aligned}$$

其中, N_{atom} 是总粒子数, \mathbf{r} 是原子的位置矢量, λ 是虚粒子的滴定坐标, m_a 和 m_j 是原子和虚粒子的质量. 第 1 和第 4 项的求和分别是原子和虚粒子的总动能. 第 2 项 U^{bond} 是键相互作用能, 包括键伸缩能、键角弯折能和二面角扭转能. 这里假设键

相互作用与 λ 无关. 第 3 项 U^{nbond} 是非键相互作用能, 包括静电 U^{elec} 和范德瓦耳斯 U^{vdw} 相互作用, 与 λ 相关. 最后一项 U^* 是偏置势, 利用经验势描述去质子化键断裂的能量变化, 只和 λ 相关.

以下介绍如何利用 λ 标度非键相互作用能和偏置势. 对于可滴定的氢原子和周围原子的范德瓦耳斯相互作用, 直接用 $1 - \lambda_i$ 标度势能函数 (这里采用 6-12 勒让德琼斯势 U^{LJ}):

$$U_{ij}^{\text{vdw}} = (1 - \lambda_i) U_{ij}^{\text{LJ}}. \quad (8)$$

可见, 当 $\lambda = 1$ 时, 残基 i 去质子化, 残基 i 的可滴定氢与 j 无相互作用.

对于两个可滴定氢之间的范德瓦耳斯相互作用, 采用 $1 - \lambda_i$ 和 $1 - \lambda_j$ 进行标度:

$$U_{ij}^{\text{vdw}} = (1 - \lambda_i)(1 - \lambda_j) U_{ij}^{\text{LJ}}. \quad (9)$$

范德瓦耳斯力是近程非键相互作用力, 主导疏水基团间的相互作用. 然而, 由于原子半径的差异, 氢 (半径约 1 Å) 几乎被与之成键 (键长约 1 Å) 的重原子 (半径约 2 Å) 包围, 使其难以接触到其他原子. 因此, 质子化和去质子化对范德瓦耳斯相互作用影响不大, 相对长程静电相互作用可以忽略不计. 对于静电相互作用, λ 标度的是原子电荷 [31]:

$$q_{a,j} = \lambda_j q_{a,j}^{\text{dep}} + (1 - \lambda_j) q_{a,j}^{\text{prot}}, \quad (10)$$

其中, $q_{a,j}^{\text{dep}}$ 和 $q_{a,j}^{\text{prot}}$ 是氨基酸残基 j 处于去质子化态和质子化态时原子 a 所带电荷. 静电相互作用 U^{elec} 计算复杂度高, 在分子动力学模拟中占据大多数计算资源, 特别是和溶剂相关的部分. 因为静电势是 pK_a 计算的关键因子, 我们将详细介绍不同溶剂条件下的 C-CpHMD.

早期为了提高计算效率, Brooks 课题组 [31] 采用隐性溶剂模型计算溶剂对溶质的平均效应. 如此一来, 总静电能 U^{elec} 的溶质内静电相互作用仍采用库仑势 ((11) 式第 1 项), 而溶质与溶剂的静电相互作用 U^{solv} 采用 GB 势能函数 ((12) 式):

$$U^{\text{elec}} = \sum_{a < b}^{N_{\text{atom}}^*} \frac{q_a q_b}{r_{ab}} + U^{\text{solv}}, \quad (11)$$

$$U^{\text{solv}} = -\frac{1}{2} \sum_{a,b}^{N_{\text{atom}}} \left(\frac{1}{\epsilon_p} - \frac{e^{-\kappa r_{ab}}}{\epsilon_w} \right) \times \frac{q_a q_b}{\sqrt{r_{ab}^2 + \alpha_a \alpha_b e^{-r_{ab}^2/4\alpha_a \alpha_b}}}, \quad (12)$$

$$\kappa^2 = \frac{8\pi q^2 I}{e k_B T}, \quad (13)$$

其中, 星号代表排除存在键相互作用的原子对; r_{ab} 是电荷 q_a 和 q_b 的距离; ϵ_p 和 ϵ_w 是蛋白质和水的介电常数; κ 是德拜长度取反 ((13) 式); I 是盐离子强度; q 是盐离子电荷; e 是基本电荷; k_B 是玻尔兹曼常数; T 是温度; α 是有效玻恩半径, 表征某原子埋在蛋白内部的程度, 为衡量 GB 模型精度的关键参数. 相对 PB 模型, GB 的计算复杂度较低, 并且是解析的, 适合需要对位置坐标求一阶导 (计算粒子所受合外力) 的分子动力学模拟. GB 模型的计算复杂度主要体现在有效玻恩半径的求解.

2004 和 2005 年 Brooks 课题组接连开发了 CH ARMM 软件中基于隐性溶剂 GBMV [31] 和 GBSW [46] 的 C-CpHMD, 证明了基于 GB 的 C-CpHMD 在 pK_a 预测方面的有效性. 相对 GBSW/GBMV 溶剂模型, GBNeck2 可提供更优的构象采样 [47]. 于是, 马里兰大学 Shen 课题组 [48] 在 2018 年开发了 Amber 软件中基于隐性溶剂 GBNeck2 的 C-CpHMD. 值得一提的是, 对于实验科学家关心的酶催化中心 (如图 1 活性位点 Asp32 和 Asp228), 该方法也表现较好, 目前已被应用于共价抑制剂靶点的预测 [49-51], 蛋白质 pK_a 数据集的建立 [25,29], 以及依赖于 pH 的蛋白质分子机制研究 [52,53]. 目前, 基于 GBSW 和 GBNeck2 的 C-CpHMD 均已实现 GPU 加速, 这进一步扩展了模型的应用范畴 [54,55].

为了提高构象采样精度以及扩展 C-CpHMD 的应用范围, Shen 课题组 [56] 提出了杂化溶剂 C-CpHMD: 构象动力学使用显性溶剂; 而滴定动力学保留隐性溶剂. 为此, 构象动力学和滴定动力学采用不同的哈密顿量. 前者去掉方程 (7) 的最后两项, 第 3 项不再包含反应坐标 λ , 令方程 (7) 回归到常规分子动力学. 该方法不仅维持了质子化态空间采样效率, 而且提高了构象采样精度. 起初人们会担心隐性溶剂 GB 的理论局限性 (例如偏弱的疏水效应) 会影响 pK_a 预测精度. 然而, Shen 课题组 [56] 发现, 显性溶剂 PME 可导致偏高的疏水效应, 一定程度上抵消了隐性溶剂导致的偏弱的疏水效应. 相对隐性溶剂, 该杂化溶剂 C-CpHMD 获得了广泛的应用, 如钠离子质子交换蛋白 [37,57], 质子通道 [58], 类药物分子的膜渗透 [59], 芬太尼激活 G 耦联受体 [60], 糖苷水解酶 [61], 络氨酸激酶药物发现 [62], 以及上文提及的 β 分泌酶 [10].

为了描述和功能相关的水分子或其他辅助因子(如金属离子和小分子)对去质子化平衡的影响, 滴定动力学部分也需采用显性溶剂. 起初, Brooks 课题组和 Shen 课题组分别选择了较简单的基于截断的显性溶剂 FSh (force shifting, FSh)^[63] 和 GRF (generalized reaction field, GRF)^[64]. 然而, 由于截断, 这两个模型均低估了长程静电力对可滴定位点的影响^[65]. 为此, Shen 课题组^[66] 开发了基于显性溶剂 PME 的 C-CpHMD. 最近, 该模型在分子模拟软件 Amber 中实现了 GPU 加速^[67]. 众所周知, PME 是满足周期性边界条件 (periodic boundary condition, PBC) 的分子模拟中计算静电相互作用的标准算法, 因此基于 PME 的 C-CpHMD 是 λ 动力学框架下所能达到的最优版本. 理论上, 如果不考虑取样问题, 该模型的 pK_a 预测应该最接近实验. 对于一个满足 PBC 的分子动力学模拟体系, PME 的总静电能是 3 个能量项的加和:

$$U^{\text{elec}} = U^{\text{dir}} + U^{\text{rec}} + U^{\text{corr}}, \quad (14)$$

其中, U^{dir} 是实空间静电相互作用, 在库仑势基础上增加一个补偿函数, 负责截断距离以内的短程静电相互作用 ((15) 式). U^{rec} 最为耗时, 为倒格空间 (reciprocal space) 下求解的长程静电能, 负责截断以外的长程静电相互作用 ((16) 式). U^{corr} 是修正项 ((20) 式)^[39].

$$U^{\text{dir}} = \frac{1}{2} \sum_{\mathbf{n}}^* \sum_{a,b=1}^{N_{\text{atom}}} \frac{q_a q_b \text{erf}(\beta |\mathbf{r}_b - \mathbf{r}_a + \mathbf{n}|)}{|\mathbf{r}_b - \mathbf{r}_a + \mathbf{n}|}, \quad (15)$$

其中, \mathbf{r}_a 和 \mathbf{r}_b 是中心元胞的位置矢量; \mathbf{n} 是元胞的位置矢量, 其表达式为 $\mathbf{n} = n_1 \mathbf{c}_1 + n_2 \mathbf{c}_2 + n_3 \mathbf{c}_3$, 其中 \mathbf{c}_1 , \mathbf{c}_2 和 \mathbf{c}_3 代表元胞的 3 个正交方向矢量; 星号代表被排除的原子对, 包括原子自身 ($a = b$), 形成化学键的原子对, 以及最近邻 (n 的大小为 1) 以外的镜像; erf 是补偿误差函数; 参数 β 决定 U^{dir} 和 U^{rec} 的相对收敛速度. 例如, β 越大, U^{dir} 计算收敛越快, 而 U^{rec} 计算收敛会越慢.

$$U^{\text{rec}} = \frac{1}{2\pi V} \sum_{\mathbf{m} \neq 0} \frac{\exp(-\pi^2 \mathbf{m}^2 / \beta^2)}{\mathbf{m}^2} S(\mathbf{m}) S(-\mathbf{m}), \quad (16)$$

式中 \mathbf{m} 是倒格矢, 其表达式为 $\mathbf{m} = m_1 \mathbf{c}_1^* + m_2 \mathbf{c}_2^* + m_3 \mathbf{c}_3^*$, 其中, m_1 , m_2 , m_3 是非零整数; \mathbf{c}_i^* 是以上 \mathbf{c}_i ($i = 1, 2, 3$) 的共轭倒格矢, 二者满足关系式 $\mathbf{c}_i^* \cdot \mathbf{c}_j = \delta_{ij}$, 这里 i 和 j 取 1, 2 和 3. 另外, $V = \mathbf{c}_1 \cdot$

$\mathbf{c}_2 \times \mathbf{c}_3$, 是元胞的体积. $S(\mathbf{m})$ 是结构因子:

$$S(\mathbf{m}) = \sum_{a=1}^{N_{\text{atom}}} q_a \exp(2\pi i \mathbf{m} \cdot \mathbf{r}_a). \quad (17)$$

该结构因子可近似表示为

$$\begin{aligned} S(\mathbf{m}) &\approx \sum_{k_1, k_2, k_3} Q(k_1, k_2, k_3) \\ &\times \exp \left[2\pi i \left(\frac{m_1 k_1}{K_1} + \frac{m_2 k_2}{K_2} + \frac{m_3 k_3}{K_3} \right) \right] \\ &= F(Q)(m_1, m_2, m_3), \end{aligned} \quad (18)$$

式中通过将元胞中的电荷分布 (B 样条) 插值到具有相同的 3 个维度 k_1 , k_2 , k_3 的网格来构造三维矩阵 Q ; k_i/K_i 是分数坐标, 其中, k_i ($i = 1, 2, 3$) 取值范围是 $(1, 2, 3, \dots, K_i)$, 正整数常数 K_i 代表元胞的尺寸; $F(Q)$ 是矩阵 Q 的三维快速傅里叶变换. 经过以上变换, U^{rec} 的表达式为

$$\begin{aligned} U^{\text{rec}} &= \frac{1}{2\pi V} \sum_{\mathbf{m} \neq 0} \frac{\exp \left(\left[-(\pi \mathbf{m} / \beta)^2 \right] \right)}{\mathbf{m}^2} \\ &\times F(Q)(\mathbf{m}) F(Q)(-\mathbf{m}). \end{aligned} \quad (19)$$

值得一提的是, U^{rec} 线性依赖于格点电荷, 因此对 λ 求一阶导和库仑势的一样简单.

$$\begin{aligned} U^{\text{corr}} &= -\frac{1}{2} \sum_{(a,b) \in M} \frac{q_a q_b \text{erf}(\beta |\mathbf{r}_b - \mathbf{r}_a|)}{|\mathbf{r}_b - \mathbf{r}_a|} \\ &- \frac{\beta}{\sqrt{\pi}} \sum_{a=1}^N q_a^2 - \frac{\pi}{2\beta^2 V} \left(\sum_a q_a \right)^2. \end{aligned} \quad (20)$$

U^{rec} 考虑整体的电荷分布, 并未排除存在键相互作用的原子对, 因此需采用和 U^{dir} 相同的函数形式进行修正 ((20) 式第 1 项). 此外, U^{corr} 第 2 项的作用是排除点电荷自相互作用, 第 3 项则是中和体系净电荷的背景电荷 (background plasma). 其中, 后面两个修正只依赖于原子电荷.

为了避免元胞之间不真实的静电相互作用, 常规 MD 通过添加补偿盐离子使体系呈电中性. 然而, CpHMD 模拟中电荷是动态变化的. 为了解决该问题, Shen 课题组^[64] 提出了将盐离子作为质子缓存器. 然而, 盐离子如果不带电会导致聚集, 于是改使用可滴定水分子^[68]. 酸性氨基酸 (例如 Asp 和 Glu) 与水阴离子 (hydroxide, TIPU) 耦合 ($\text{AH} + \text{OH}^- \rightleftharpoons \text{A}^- + \text{H}_2\text{O}$); 碱性氨基酸 (例如 Lys, Arg 和 His) 与水阳离子 (hydronium, TIPP) 耦合 ($\text{BH}^+ + \text{H}_2\text{O} \rightleftharpoons \text{H}_3\text{O}^+ + \text{B}$). 该耦合令反应式两端的电荷守

恒. 电中性的另一个好处是消除 U^{corr} 中会导致反常 $\text{p}K_a$ 偏移的背景电荷.

以上介绍了不同溶剂下静电能的具体求解. 下面介绍哈密顿量中只依赖于反应坐标 λ 的偏置势^[31]:

$$U^* (\{\lambda_j\}) = \sum_j^{N_{\text{tit}}} [-U^{\text{mod}} (\lambda_j) + U^{\text{pH}} (\lambda_j) + U^{\text{barr}} (\lambda_j)], \quad (21)$$

其中, 第 1 项 ((22) 式) 和第 2 项 ((23) 式) 分别是游离可滴定氨基酸去质子化的非键相互作用能和总自由能. 对于单个可滴定位点的氨基酸 (如赖氨酸), U^{mod} 是一个关于 λ 的一元二次函数. U^{pH} 由 λ 线性标度 ((23) 式). $U^{\text{pH}} - U^{\text{mod}}$ 是化学能改变量的近似解. 为了减少 λ 处于不真实的中间态 (如 $\lambda = 0.5$) 的概率, 另外添加了一个二次函数势垒 U^{barr} ((24) 式). U^{barr} 降低了 λ 的动力学, 对热力学统计没有影响. (23) 式和 (24) 式的参数为已知, 因此, C-CpHMD 的主要工作是确定 U^{mod} 的参数 (如 (22) 式中的 A_j 和 B_j):

$$U^{\text{mod}} (\lambda_j) = A_j (\lambda_j - B_j)^2, \quad (22)$$

$$U^{\text{pH}} (\lambda_j) = \ln (10) k_B T (\text{p}K_a^{\text{mod}} - \text{pH}) \lambda_j, \quad (23)$$

$$U^{\text{barr}} (\lambda_j) = 4\eta (\lambda_j - 0.5)^2, \quad (24)$$

其中, $\text{p}K_a^{\text{mod}}$ 是游离可滴定氨基酸的 $\text{p}K_a$ 测量值, η 决定势垒高度. 对于一个 C-CpHMD 模型, 需要通过平均力势 (potential of mean force, PMF) 模拟求 U^{mod} 函数中的系数. 这里可用单个可滴定位点的游离赖氨酸 (Lys) 为例. 固定 λ 值, 经过一定时间 (如 1 ns) 的 MD, 对作用在虚粒子上的力求时间平均, 即 $\langle dU/d\lambda \rangle$, 其中 λ 在 0—1 之间取离散的值. 基于线性响应理论, 用线性函数 $2A(\lambda - B)$ 拟合平均力, 确定模型参数 A 和 B . 同时, 可利用以下热力学积分求 PMF, 计算去质子化自由能改变量:

$$U^{\text{mod}} (\lambda) = \int_0^\lambda \left\langle \frac{\partial U (\lambda')}{\partial \lambda'} \right\rangle_{\lambda'} d\lambda'. \quad (25)$$

需要注意的是, 为了将 λ 约束在 $[0, 1]$, 需定义另一个变量 θ . λ 和 θ 的关系式为 $\lambda = \sin^2 \theta$. 于是, 数值模拟中进行迭代的是 θ , 而非反应坐标 λ .

对于含有两个可滴定位点的氨基酸, 需要定义反应坐标 x 来描述处于去质子化 (His) 或质子化 (Glu 和 Asp) 态时质子所处的可滴定位点^[46]. x 同

样是在 0 到 1 范围内的连续变量. 图 3 展示了 Asp 和 His 侧链 3 个质子化态对应的反应坐标值以及状态间的转化. 类似变量 λ , 可利用插值将 x 加入哈密顿量的各个能量项. 例如, 以下分别是 Asp 和 His 电荷关于 λ 和 x 的表达式:

$$q_{a,j}^{\text{D}} = \lambda_j q_{a,j}^{\text{ASP}} + (1 - \lambda_j) [x_j q_{a,j}^{\text{ASP2}} + (1 - x_j) q_{a,j}^{\text{ASP1}}], \quad (26)$$

$$q_{a,j}^{\text{H}} = \lambda_j [x_j q_{a,j}^{\text{HSE}} + (1 - x_j) q_{a,j}^{\text{HSD}}] + (1 - \lambda_j) q_{a,j}^{\text{HSP}}, \quad (27)$$

其中 $q_{a,j}^{\text{ASP2}}$ 和 $q_{a,j}^{\text{ASP1}}$ 分别是 Asp 侧链 j 上原子 a 在 $\text{O}_{\delta 2}$ 和 $\text{O}_{\delta 1}$ 质子化时所带的电荷, $q_{a,j}^{\text{ASP}}$ 是该侧链去质子化时原子 a 所带电荷; $q_{a,j}^{\text{HSE}}$ 和 $q_{a,j}^{\text{HSD}}$ 分别是 His 侧链 j 上原子 a 在 N_{δ} 和 N_{ϵ} 去质子化时所带的电荷, $q_{a,j}^{\text{HSP}}$ 是该侧链质子化时原子 a 所带电荷. 具有双可滴定位点的 Glu/Asp 和 His 的 U^{mod} 是关于 λ 和 x 的多项式, 需要取 λ 和 x 值的不同组合计算平均力, 然后通过 Brooks 课题组提出的方法计算多项式系数^[46].

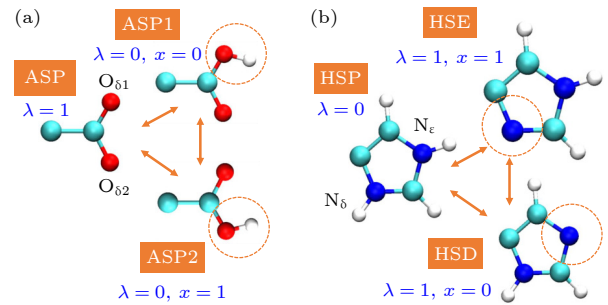


图 3 互变异构滴定模型的 3 个质子化态以及状态间的转化 (a) 天冬氨酸 Asp; (b) 组氨酸 His

Fig. 3. Three protonation states and their interconversion in the tautomeric titration model: (a) Aspartic acid; (b) histidine.

CpHMD 模拟同时对构象和质子化态采样. 根据设置的输出频率保存每个可离子化基团的滴定坐标 λ ($\lambda \in [0, 1]$) (图 4(a)). 统计处于质子化态 ($0 \leq \lambda \leq 0.1$) 的次数 N^{prot} 以及去质子化态 ($0.9 \leq \lambda \leq 1$) 的次数 N^{dep} , 计算不同 pH 条件下的去质子化概率 S (图 4(a))^[31]:

$$S = \frac{N^{\text{dep}}}{N^{\text{dep}} + N^{\text{prot}}}. \quad (28)$$

最后, 采用如下 Hill 函数 (广义 Henderson-Hasselbalch 函数) 拟合 S . $\text{p}K_a$ 便是 $S = 0.5$ 时对应的 pH (图 4(b)):

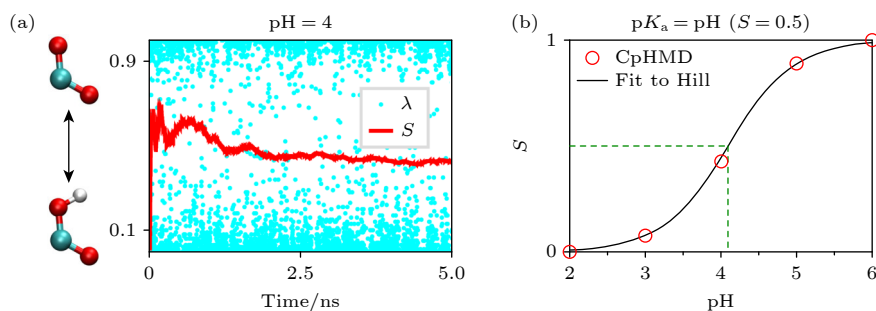

 图 4 基于 C-CpHMD 的 pK_a 计算 (a) 滴定坐标 λ 和去质子化概率 S 的轨迹; (b) 采用 Hill 函数拟合 S

Fig. 4. The pK_a calculation based on C-CpHMD: (a) Trajectories of titration coordinate λ and deprotonation fraction S ; (b) fitting S to Hill function.

$$S = \frac{1}{1 + 10^{h(pK_a - pH)}}, \quad (29)$$

其中 h 是 Hill 系数, 表征一个可离子化基团与周围可滴定基团的滴定动力学是否存在耦合. $h = 1$ 表示无耦合, 如位于分子表面的残基或游离氨基酸. $h < 1$ 表示负耦合, 如形成盐桥键的去质子化的 Asp 和质子化的 Lys. $h > 1$ 表示正耦合, 如酶活性位点距离相近的两个酸性氨基酸 (质子化的 Asp 或 Glu). h 偏离 1 越多, 耦合越强^[69].

当两个氨基酸的滴定动力学存在耦合, 可将二者看作一个整体, 利用以下公式计算宏观 pK_1 和 pK_2 (macroscopic sequential pK_a)^[64,70]:

$$N = \frac{10^{(pK_2 - pH)} + 2 \times 10^{(pK_1 + pK_2 - 2pH)}}{1 + 10^{(pK_2 - pH)} + 10^{(pK_1 + pK_2 - 2pH)}}, \quad (30)$$

其中 N 是一定 pH 条件下的平均质子数. 为获得 pK_1 和 pK_2 , 也可以采用以下非耦合模型 (31) 式^[71,72]:

$$S_1 + S_2 = \frac{1}{1 + 10^{(pK_1 - pH)}} + \frac{1}{1 + 10^{(pK_2 - pH)}}, \quad (31)$$

其中 S_1 和 S_2 分别是两个耦合的可滴定位点的去质子化概率.

当滴定动力学采用满足周期性边界条件的显性溶剂时, 需要考虑有限尺度效应^[73]. 由于采用耦合水离子实现了电中性, 有限尺度效应只剩下和水分子模型相关的离散溶剂效应 (discrete solvent effect)^[66]. 当某个可滴定氨基酸去质子化, 因离散溶剂效应引起的能量变化是

$$\Delta G^{\text{offset}} = \frac{2\pi}{3} \kappa \gamma q \rho, \quad (32)$$

其中, κ 是介电常数; ρ 是水数量密度, 等于水分子数 N 除以体积 V , 这里 N 指的是和蛋白有相互作用的水分子数, V 也是这些水包络范围内的体积; q 是可滴定氨基酸的电荷, Asp/Glu 是 $-1e$, His/Lys

为 $+1e$; γ 是显性溶剂模型范德瓦耳斯相互作用中心的电四极矩. 对于溶剂模型 TIP3P, γ 的值为 $0.764 e \cdot \text{\AA}^2$. 为了估算该有限尺度效应导致的 pK_a 偏移, 需要计算相对模型分子的能量变化^[66]:

$$\Delta \Delta G^{\text{offset}} = \frac{2\pi}{3} \kappa \gamma q \left(\frac{N}{V} - \frac{N^{\text{mod}}}{V^{\text{mod}}} \right), \quad (33)$$

其中, N 和 N^{mod} 分别是蛋白质和游离氨基酸模拟体系中与溶质有相互作用的水分子数; V 和 V^{mod} 是相应的周期性元胞体积. 将以上表达式转化为 pK_a 偏移量, 可得到^[66]

$$\Delta pK_a^{\text{corr}} = \pm \frac{\Delta \Delta G^{\text{offset}}}{\ln(10) RT}. \quad (34)$$

根据 N 和 V 的定义, 可以推断有限尺寸效应对 PME 影响较大. PME 考虑了周期性元胞内所有水分子, 蛋白质体积所占比例较小, 水数量密度 ρ 较大; 另一方面, GRF 和 FSh 仅考虑截断以内的水, 蛋白质体积所占比例较大, 水数量密度可忽略不计. 对于膜蛋白体系, 可参考 Roux 课题组^[74] 提出的方法做相应的修正.

以上介绍的 C-CpHMD 属于对电荷插值, 实现电荷对反应坐标的线性响应. 实际上, 由于库仑势对电荷线性依赖, 库仑势和电荷两者的线性插值是等效的. 因为两种情况下, 关于插值变量 (反应坐标 λ) 负的一阶导数 (作用在虚粒子上的合外力) 是相等的. 然而, 并不是所有和静电势相关的能量项和电荷线性相关, 如 PME 算法中对点电荷自相互作用和净电荷的修正项 ((20) 式)^[66]. 所以, 为了更好描述电荷变化对滴定动力学的影响, 基于截断的 GRF 和 FSh 较适合对静电势进行插值的 C-CpHMD, 因为它们静电势保留了对电荷的线性依赖. 德国马克斯普朗克研究所的 Grubmüller 课题组^[75] 在分子模拟软件 GROMACS 中开发的

C-CpHMD 便是对势函数进行插值. 最近, 芬兰的 Groenhof 课题组 [76,77] 基于该模型进行代码优化, 并实现基于 CHARMM 力场的 CpHMD 模拟. 然而, 该模型采用了显性溶剂 PME, 而不是基于截断的 GRF 或 FSh. 其次, 该模型没有像 Shen 课题组 [64] 一样考虑有限尺寸效应. 另一方面, 同样是对势能进行插值, Brooks 课题组 [63,71] 基于显性溶剂的 C-CpHMD 模型合理地采用了基于截断的 FSh. 除了以上正弦函数形式, Grubmüller 课题组和 Brooks 课题组提出了其他将 λ 约束在区间 $[0, 1]$ 的方法. 例如, Grubmüller 课题组 [75] 提出了余弦形式. Brooks 课题组 [78] 提出一个较复杂的指数形式. 对于显性溶剂 C-CpHMD, 体系电中性是一项重要的约束条件, Shen 课题组 [66] 和 Grubmüller 课题组 [79] 均采用了可滴定水分子实现体系净电荷恒等于 0. 然而, Brooks 课题组 [71] 的显性溶剂 C-CpHMD 还未考虑该约束. 因此, 为了避免溶质与其镜像的静电相互作用, 需对 FSh 静电势设置较小截断值.

从理论上讲, Shen 课题组 [66] 开发的基于 PME 的 C-CpHMD 可应用于分子力场能描述的任何体系, 似乎没有改进的空间. 实际上, 一个酸性氨基酸残基的去质子化或一个碱性氨基酸残基的质子化可诱导周围可极化原子 (原子核外电子云的中心偏离原子核) 或基团 (组氨酸咪唑环上的电子离域) 形成偶极子 [80]. 偶极子与电荷相互吸引, 一定程度上加强了该氨基酸残基带电状态的稳定性. 然而, 传统力场下电荷分布是固定的, 不会因为滴定引起周围电场的变化而做出调整, 这可能导致可滴定氨基酸残基偏爱电中性, 特别是位于蛋白质内部的氨基酸残基 [66]. 基于以上考虑, 如果采用极化力场 (如 CHARMM 的 Drude^[81]), C-CpHMD 的精度将得到进一步的提升. 其次, 大部分 CpHMD (包括该模型) 没有考虑质子化和去质子化对键相互作用的影响 [44].

随着显性溶剂 CpHMD 的快速发展, 急需解决质子化态和构象的采样问题. 2006 年 Brooks 课题组 [82] 率先将基于温度的副本交换 (replica exchange) 算法应用到 C-CpHMD, 即将副本以一定的概率交换到较高温度, 借助热涨落提高 CpHMD 模拟的采样. 受到哈密顿量副本交换算法的启发, 2011 年 Shen 课题组 [56] 提出了基于 pH 的副本交换算法: 将副本以一定的概率 p 交换到较高的 pH, 提高去质子化态的采样; 或交换到较低的 pH, 提

高质子化态的采样 ((35) 式). 因为实际进行交换的 pH 只存在于 U^{pH} ((23) 式), 交换前后总能量的变化 Δ/β 可简化为仅含 U^{pH} 的表达式 ((36) 式). 交换 pH 后, 两个副本将在新的 pH 条件下 (或新的 U^{pH}) 进行采样. 该算法效率极高, 同时操作简单, 已被应用到其他 CpHMD 模型 [83-86]. 为了增强质子化态空间采样, 美国国立卫生研究院 NIH 的 Brooks 课题组 [87] 提出结合包络分布采样 (enveloping distribution sampling, EDS) 和哈密顿量副本交换 (Hamiltonian replica exchange, HREX). EDS 通过定义一个参数 s 标度状态间的能垒. 较小的 s 对应较平滑的能垒, 方便了状态间的转化. 然而, 能垒的消除促进了虚拟中间态的采样, 这将影响物理态的采样. 为了避免中间态的采样, 在 EDS 基础上利用 HREX 提高离散的质子化态空间的采样效率. 接着, 该课题组 [86] 加入以上基于 pH 的副本交换, 构成二维的副本交换. 从算法的角度, 该方法确实提高了采样效率, 但代价是产生大量的副本以及模拟过程中副本的频繁通讯, 对计算能力要求较高. 近期, 为了在有限 GPU 显卡数量的条件下实现基于 pH 的副本交换, Shen 课题组 [88] 提出了副本同步交换.

$$p = \begin{cases} 1, & \Delta \leq 0, \\ \exp(-\Delta), & \Delta > 0, \end{cases} \quad (35)$$

$$\Delta = \beta(U^{\text{pH}}(\{\lambda_j\}; \text{pH}') + U^{\text{pH}}(\{\lambda'_j\}; \text{pH}) - U^{\text{pH}}(\{\lambda_j\}; \text{pH}) - U^{\text{pH}}(\{\lambda'_j\}; \text{pH}')), \quad (36)$$

其中, p 是副本交换的概率; $U^{\text{pH}}(\{\lambda_j\}; \text{pH})$ 和 $U^{\text{pH}}(\{\lambda'_j\}; \text{pH}')$ 是两个副本交换前的 U^{pH} . 将以上两项的 pH 和 pH' 进行互换, 得到 $U^{\text{pH}}(\{\lambda_j\}; \text{pH}')$ 和 $U^{\text{pH}}(\{\lambda'_j\}; \text{pH})$.

除了副本交换, 另一种增强采样的方法是对生物大分子进行粗粒化 (coarse graining, CG), 减少模拟体系中粒子的数量, 从而降低了构象空间的自由度. 该方法通常被应用于具有较大空间和时间尺度的生物过程, 如蛋白质折叠、多肽聚集和物质跨膜转运等 [89]. 近几年, 研究者们开始将 CG 与 CpHMD 结合, 发展 CpHMD 的粗粒化模型 [90-93]. 值得一提的是, 提出 Martini 粗粒化力场的 Marrink 课题组 [92] 已在分子模拟软件 GROMACS 中实现了 CpHMD 的粗粒化模拟.

2.2 基于 PB 的 pK_a 预测模型

实际上, 如果只考虑单个结构, 可以用 PB 方程计算相对去质子化自由能 $\Delta\Delta G = \Delta G - \Delta G^{\text{mod}}$. 其中, ΔG^{mod} 是某可离子化氨基酸 A 在游离状态下去质子化自由能改变量:

$$\Delta G^{\text{mod}} = G^{\text{mod}}(A^-) - G^{\text{mod}}(AH), \quad (37)$$

式中 $G^{\text{mod}}(A^-)$ 和 $G^{\text{mod}}(AH)$ 分别是去质子化 (A^-) 和质子化 (AH) 状态的自由能. 同理, 当该氨基酸参与蛋白质的合成, 它在蛋白质中的去质子化自由能改变量 ΔG 表示为

$$\Delta G = G(A^-) - G(AH). \quad (38)$$

基于蛋白质环境不影响成键作用部分 ΔG_{Bond} (见 (4) 式) 的假设, 以上两个自由能改变量的差可表示为

$$\Delta G - \Delta G^{\text{mod}} = (G_{\text{PB}}(A^-) - G_{\text{PB}}(AH)) - (G_{\text{PB}}^{\text{mod}}(A^-) - G_{\text{PB}}^{\text{mod}}(AH)), \quad (39)$$

其中, 下标 PB 表示用 PB 方程分别计算等式右边 4 个状态下的静电能. 令 $\Delta G(AH) = G_{\text{PB}}(AH) - G_{\text{PB}}^{\text{mod}}(AH)$ 和 $\Delta G(A^-) = G_{\text{PB}}(A^-) - G_{\text{PB}}^{\text{mod}}(A^-)$, 可得到

$$\Delta G + \Delta G_{\text{PB}}(AH) = \Delta G^{\text{mod}} + \Delta G_{\text{PB}}(A^-), \quad (40)$$

其中, $\Delta G_{\text{PB}}(AH)$ 和 $\Delta G_{\text{PB}}(A^-)$ 分别表示在水溶液中将质子化 (AH) 和去质子化 (A^-) 的氨基酸放入蛋白质的静电能改变量. 基于该等式, 可以得到如图 5 所示的热力学循环 (thermodynamic cycle). 相对去质子化自由能 $\Delta\Delta G$ 可表示为

$$\Delta\Delta G = \Delta G_{\text{PB}}(A^-) - \Delta G_{\text{PB}}(AH), \quad (41)$$

接着, 将 $\Delta\Delta G$ 代入关系式 $\Delta pK_a = \Delta\Delta G / (k_B T \ln 10)$ 计算 pK_a 偏移量 ΔpK_a . 最后, 利用 (5) 式计算 pK_a . 可见, 热力学循环 4 个状态的静电能计算决定了 pK_a 的预测精度. 目前, 基于 PB 计算静电能并预测蛋白质 pK_a 的方法包括 MCCE^[17,94], H++^[18], APBS^[19], DelPhiPKa^[20,95,96] 以及 PypKa^[21]. 其中, MCCE 和 PypKa 利用 MC 对侧链二面角进行采样, 一定程度上提高了预测精度, 但总体精度仍低于 CpHMD, 说明了空间构象充分采样的重要性^[15]. PB 方程的参数主要是介电常数, 原子的电荷和半径, 因此容易拓展到其他类型的体系. 例如, 除了蛋白质, DelPhiPKa 也适用于 DNA 和 RNA. 除了蛋白质单体, H++ 也考虑了含有配体的复合物.

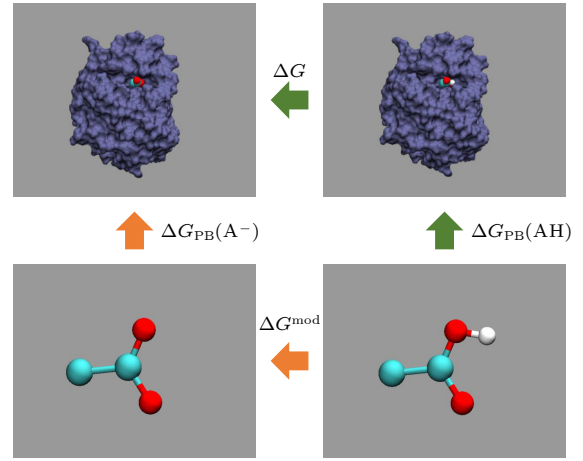


图 5 相对去质子化自由能计算的热力学循环

Fig. 5. Thermodynamic cycle of relative deprotonation free energy calculation.

2.3 基于经验函数的 pK_a 预测模型

以上物理模型 (CpHMD 和基于 PB 的模型) 需要计算体系的静电能, 计算复杂度较高. 为了进一步提高 pK_a 计算的效率 (例如将单个蛋白的 pK_a 计算时长缩短到秒量级), 2005 年哥本哈根大学的 Jensen 课题组^[23] 提出了一组经验函数 PropKa 分别描述点电荷相互作用 (Coulomb force)、去溶剂化效应 (desolvation) 和氢键相互作用 (hydrogen bonding) 对 pK_a 偏移量的贡献:

$$\Delta pK_a = \Delta pK_a^{\text{Columb}} + \Delta pK_a^{\text{Desolv}} + \Delta pK_a^{\text{HBond}}. \quad (42)$$

以上 3 项的函数均采用分段的一次函数, 计算复杂度低, 已被应用到蛋白质单体^[23], 蛋白质和小分子配体的复合物^[97]. 然而, 该版本的 PropKa 没区分可滴定氨基酸残基是处于蛋白质的表面还是内部.

为此, 2011 年 Jensen 课题组^[24] 提出了改进的 PropKa 3.0. 新版本考虑了相同的 ΔpK_a 决定因子, 将 (42) 式的氢键相互作用导致的 $\Delta pK_a^{\text{HBond}}$ 和去溶剂化效应导致的 $\Delta pK_a^{\text{Desolv}}$ 归为自能 $\Delta pK_a^{\text{Self}}$. 不同的是, PropKa 3.0 采取了一个折中的方案, 即部分使用能量公式. 例如, 点电荷相互作用采用经典的库仑势. 去溶剂化效应采用了和 GB 模型中求解有效波恩半径的倒数 ($1/\alpha$) 类似的原子体积 (V) 除以原子间距离的四次方 (r^4). 此外, 蛋白质表面和内部被赋予不同的介电常数. 对于氢键相互作用, 则保留了一次函数形式. 该模型参数化基于谷氨酸和天冬氨酸的 pK_a 实验值, 对酸性氨基酸的预测能力接近 CpHMD^[98]. 然而, 该模型对碱性

氨基酸 (如 Lys 和 His) 的预测效果较差^[25].

2.4 基于机器学习的 pK_a 预测模型

上述 PropKa 经验函数的提出较大程度依赖于科学家的先验知识. 理论上, 如果有足够多的 pK_a 实验测量值, 可以结合数据和机器学习算法训练出一个经验函数, 而不需要依靠已有的知识. 2018 年 波兰华沙大学 Siedlecki 课题组^[99] 提出首个基于深度学习的蛋白质配体结合亲和力 (binding affinity) 预测模型. 这里的配体通常指具有几何结构的小分子. 我们知道, pK_a 表征某可滴定基团去质子的难易程度. 换一种表达, pK_a 代表蛋白质和质子的结合亲和力. 可见, 蛋白质配体结合亲和力预测方法对 pK_a 预测具有参考价值^[25].

由于实验条件的限制, 迄今为止蛋白质可滴定氨基酸残基的 pK_a 实验测量值不到两千个^[100,101]. 于是, 本课题组采用基于隐性溶剂 GBNeck2 的 C-CpHMD^[48] 建立了一个蛋白质 pK_a 数据集 (包含 12809 个 pK_a)^[25]. 2021 年 12 月, 本课题组提出了国际上首个基于机器学习的蛋白质 pK_a 预测模型 DeepKa, 证明了引入人工智能方法解决蛋白质 pK_a 预测问题的可行性^[25]. 本课题组对现有的 pK_a 数据库 PKAD^[100] (包含 1350 个蛋白质 pK_a 实验测量值) 进行数据清洗, 得到了测试集 EXP67S. 首先, 根据氨基酸序列相似性比对排除了冗余数据. 剩下的 67 个蛋白质的 470 个 Asp, Glu, Lys 或 His 的 pK_a 构成数据集 EXP67. 接着, 对 EXP67 进行欠采样, 使得不同 ΔpK_a 区域分布均匀. 最后剩下的 167 个 pK_a 为该模型的测试集 EXP67S. 该测试集的优势将在下文的多模型对比体现出来 (图 6). 模型的大部分输入特征以及三维卷积神经网络 (convolutional neural network, CNN) 框架均借鉴 Siedlecki 课题组^[99] 提出的 Pafnucy 模型. 值得一提的是, 为了解决截断导致的边界问题, DeepKa 采用格点电荷 (Siedlecki 课题组^[99] 采用原子电荷) 描述对 pK_a 预测精度起决定性作用的静电环境^[25]. 虽然 DeepKa 第一版本的预测精度高于 PropKa 3.0, 但是和 CpHMD 还存在一定差距^[25]. 此外, 该工作只测试了 DeepKa 的总体性能, 并未对特定的问题 (如酶催化中心或无序蛋白) 进行讨论.

2022 年 1 月, 美国卡内基-梅隆大学 Lsayev 课题组^[26] 开发了基于神经网络势 ANI-2X 和原子环境矢量 AVE 的深度学习模型 pKa-ANI. 然而, 该

模型将所有的实验数据用于模型的训练, 不利于对其性能进行客观的评价. 另外, 该模型对结构敏感, 需要在预处理阶段对初始结构进行能量最小化, 否则将得到不合理的预测结果^[26]. 2022 年 3 月, 美国约翰斯-霍普金斯大学 Damjanovic 课题组^[27] 测试了 4 种基于树的机器学习算法. 其中, XGB-WMa 表现最好. 该小组同样采用有限的实验数据来训练和测试模型. 为了建立有效的模型, 他们在特征描述上加入了较多的经验知识: 首先, 统计可滴定基团参与的氢键数量; 其次, 计算可滴定基团的溶剂可及表面积 (solvent accessible surface area, SASA); 最后, 根据是否带电或亲水对可滴定基团附近氨基酸残基进行分类. 显然, 以上特征基本上覆盖了 PropKa 模型中影响 pK_a 偏移量的 3 个关键因素: 氢键相互作用、去溶剂化效应和点电荷相互作用. 2022 年 7 月, Reis 课题组^[102] 利用基于 PB 的 PypKa 建立了包含 1200 万个 pK_a 值的数据集, 并基于该数据集开发了深度学习模型 PKAI^[28]. 为了提高精度, 在 PKAI 基础上对损失函数进行正则化处理, 从而得到 PKAI+. 然而, PKAI+ 在其他测试集 (如 EXP67S) 的表现与 PKAI 相似, 说明上述的正则化处理缺乏普适性^[29]. 因此, 如果没有特别说明, 下文只讨论 PKAI.

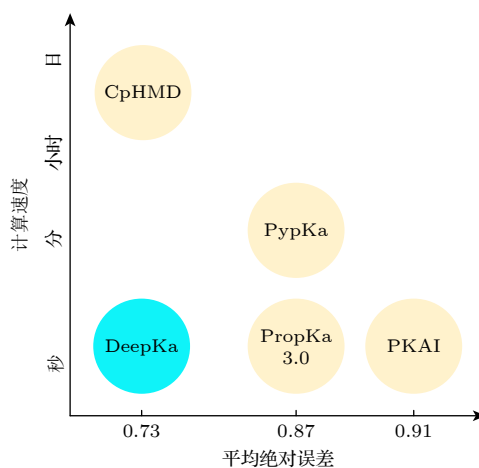


图 6 pK_a 预测模型性能对比

Fig. 6. Comparison of existing pK_a predictors.

2023 年 5 月, 本课题组发布了 DeepKa 的最新版本^[29]. 该版本的输入特征和模型框架与旧版本相同, 仅仅是增加了训练和验证集的 pK_a 样本量. 这些样本出自 549 个蛋白质的 26552 个 Asp, Glu, Lys 和 His. 相对旧版本, 该版本预测性能更接近 CpHMD. 此外, 在这个工作中特定的蛋白质体系

被用于进一步评估 DeepKa 的可靠性. 例如, 酶催化中心具有复杂的静电环境, 是 pK_a 预测的一个重要挑战. 新版本通过 pK_a 计算准确预测了 5 个酶催化中心的质子供体. 除了具有稳定三维结构的蛋白, 该模型也可被应用于无序蛋白. 理论预测 pK_a 偏移量较小的滴定位点往往容易做到预测精确, 但难以做到预测相关, 而即使在 pK_a 偏移量小于 1.0 的情况下, 理论和实验仍然表现出较高的相关性, 证明了该模型的高鲁棒性^[29]. 如无特别说明, 下文的 DeepKa 代表该新版本.

上述基于 AI 的模型均采用 PKAD 中的实验数据来训练或测试模型. 然而, pK_a -ANI, XGB-WMa 和 PKAI 忽略了存在于 PKAD 的冗余数据 (例如一个蛋白质有两组相同的 pK_a 值), 这可能导致过拟合. 其次, PKAD 中大多数 pK_a 处于参考值 pK_a^{mod} 附近, 因此测试结果并不能反应模型真实的预测能力^[25]. 值得一提的是, 本课题组创建的测试集 EXP67S 不存在以上两个问题, 可较为客观地对模型进行评价^[25]. 研究发现, 除了在实验和理论相关性方面仍旧低于 CpHMD, DeepKa 的预测精度明显高于其他主流 pK_a 预测模型, 包括 PypKa, PropKa, PKAI 和 pK_a -ANI^[29]. 其中, PypKa 代表基于 PB 的模型, PropKa 代表基于经验函数的模型, PKAI 和 pK_a -ANI 代表其他 AI 模型. 基于树的 XGB-WMa 没有开放源代码, 所以无法利用 EXP67S 对其进行测试. 因此, XGB-WMa 不参与下面的模型讨论. 同时考查精度和速度, 图 6 展示了 5 个模型的预测性能. 其中, 平均绝对误差用于表征模型的精度. 显而易见, 如果以 PropKa 的速度和 CpHMD 的精度作为参照, 目前只有 DeepKa 能提供准确的高通量 pK_a 计算^[29]. 最近, 加拿大国家研究委员会 Sulea 课题组^[103] 比较了现有的 7 种高通量 pK_a 预测模型, 包括基于经验函数的 PropKa 3.0^[24], 基于深度学习的 DeepKa^[29]、PKAI 和 PKAI+^[28] 以及基于 PB 方程的 DelPhiPKa^[95]、MCCE2^[94] 和 H++^[18]. 该研究指出在以上高通量模型中 DeepKa 的精度最高, 与图 6 的结论一致.

3 结论

pH 与温度、压强一样是基本的环境参量. 传统的分子动力学假设溶剂是中性水 (pH=7.0), 不考虑其他 pH 条件; 此外, 传统分子动力学假设电

荷是固定的, 不受溶质静电场的影响. 以上两个假设限制了传统分子动力学进一步探究细胞中许多与 pH 相关的生物过程, 而可靠的 pK_a 计算将有助于解决该难题. 本综述主要介绍了 4 类主流的 pK_a 预测方法. 显然, 对于不同理论的 pK_a 预测模型, 其适用范围也存在差异. 首先, 不论何种特定的问题, 如果不要求高通量计算, 可采用预测精度较高但计算效率较低的 CpHMD. 当涉及非水溶性蛋白 (如膜蛋白) 的 pK_a 计算, 目前理论上可行的模型为基于杂化溶剂^[37,56] 或显性溶剂^[66,67] 的 CpHMD. 另一方面, 需要开发高通量的 pK_a 预测模型, 从而满足工业界批量的 pK_a 计算需求. 由于隐性溶剂的理论局限性和实验条件的限制, 上述的高通量模型仅适用于水溶性蛋白. 对于水溶性蛋白质单体的 pK_a 计算, 在所有高通量模型中 DeepKa 无疑是最优的选择^[29,103]. 若只关心酸性氨基酸残基 (如 Asp 和 Glu) 的质子化态, 也可考虑 PropKa 3.0^[24]. 而对于主要的 4 种可离子化氨基酸残基 (Asp, Glu, Lys 和 His) 以外的可滴定基团 (如 Cys 和 Tyr), 可考虑基于 PB 的模型 (如 H++^[18] 和 PypKa^[21]).

随着计算机软件和硬件的快速发展, 国际著名的美国药物设计公司薛定谔 (Schrödinger) 开始尝试利用自由能微扰 (free energy perturbation, FEP) 方法计算 pK_a , 说明蛋白质 pK_a 理论计算开始引起工业界的关注^[104]. 值得一提的是, 基于机器学习的 pK_a 预测模型虽处于起步的阶段 (2021 年至今), 却已表现出和物理模型同水平的预测精度, 例如本课题组开发的 DeepKa. 我们相信: AI 模型有可能突破先验知识, 在不久的将来提供更为高效的预测; 利用物理模型 CpHMD 建立的 pK_a 数据集 PHMD549 和基于 pK_a 数据库 PKAD 建立的测试集 EXP67S 将为基于机器学习的 pK_a 预测工具的研发奠定基础^[29]. 最近, 基于 DeepKa 本课题组开发了国内首个蛋白质 pK_a 在线计算平台 (<http://www.computbiophys.com/DeepKa/main>), 这对未来参与到人工智能驱动的新药研发产业具有重要意义^[105,106].

参考文献

- [1] Casey J R, Grinstein S, Orlowski J 2010 *Nat. Rev. Mol. Cell Biol.* **11** 50
- [2] Qian H, Wu X L, Du X M, Yao X, Zhao X, Lee J, Yang H Y, Yan N 2020 *Cell* **182** 98
- [3] Yang G H, Zhou R, Zhou Q, Guo X F, Yan C Y, Ke M, Lei J L, Shi Y G 2019 *Nature* **565** 192

- [4] Chung H S, Piana-Agostinetti S, Shaw D E, Eaton W A 2015 *Science* **349** 1504
- [5] Nasicca-Labouze J, Nguyen P H, Sterpone F, Berthoumieu O, Buchete N, Cote S, Simone A D, Doig A J, Faller P, Garcia A, Laio A, Li M S, Melchionna S, Mousseau N, Mu Y, Paravastu A, Pasquali S, Rosenman D J, Strodel B, Tarus B, Viles J H, Zhang T, Wang C, Derreumaux P 2015 *Chem. Rev.* **115** 3518
- [6] Morrow B H, Payne G F, Shen J 2015 *J. Am. Chem. Soc.* **137** 13024
- [7] Kumar A, Hossain R A, Yost S A, Bu W, Wang Y, Dearborn A D, Grakoui A, Cohen J I, Marcotrigiano J 2021 *Nature* **598** 521
- [8] Singharoy A, Maffeo C, Delgado-Magnero K H, Swainsbury D J K, Sener M, Kleinekathofer U, Vant J W, Nguyen J, Hitchcock A, Isralewitz B, Teo I, Chandler D E, Stone J E, Phillips J C, Pogorelov T V, Mallus M I, Chipot C, Luthey-Schulten Z, Tieleman D P, Hunter C N, Schulten K 2019 *Cell* **179** 1098
- [9] Shimizu H, Tosaki A, Kaneko K, Hisano T, Sakurai T, Nukina N 2008 *Mol. Cell Biol.* **28** 3663
- [10] Ellis C R, Shen J 2015 *J. Am. Chem. Soc.* **137** 9543
- [11] Thurlkill R L, Grimsley G R, Scholtz J M, Pace C N 2006 *Protein Sci.* **15** 1214
- [12] Jensen J H, Li H, Robertson A D, Molina P A 2005 *J. Phys. Chem. A* **109** 6634
- [13] Baptista A M, Martel P J, Petersen S B 1997 *Proteins* **27** 523
- [14] Shi C, Wallace J A, Shen J K 2012 *Biophys. J.* **102** 1590
- [15] Qing R, Hao S L, Smorodina E, Jin D, Zalevsky A, Zhang S G 2022 *Chem. Rev.* **122** 14085
- [16] Henderson J A, Liu R, Harris J A, Huang Y D, de Oliveira V M, Shen J D 2022 *Liv. J. Comput. Mol.* **4** 1563
- [17] Georgescu R E, Alexov E G, Gunner M R 2002 *Biophys. J.* **83** 1731
- [18] Anandakrishnan R, Aguilar B, Onufriev A V 2012 *Nucleic Acids Res.* **40** W537
- [19] Dolinsky T J, Nielsen J E, McCammon J A, Baker N A 2004 *Nucleic Acids Res.* **32** 665
- [20] Wang L, Li L, Alexov E 2015 *Proteins*. **83** 2186
- [21] Reis Pedro B P S, Vila-Viçosa D, Rocchia W, Machuqueiro M 2020 *J. Chem. Inf. Model.* **60** 4442
- [22] Huang Y D, Yue Z, Tsai C C, Henderson J A, Shen J 2018 *J. Phys. Chem. Lett.* **9** 1179
- [23] Li H, Robertson A D, Jensen J H 2005 *Proteins* **61** 704
- [24] Olsson Mats H M, Søndergaard C R, Rostkowski M, Jensen J H 2011 *J. Chem. Theory Comput.* **7** 525
- [25] Cai Z T, Luo F F, Wang Y X, Li E L, Huang Y D 2021 *ACS Omega* **6** 34823
- [26] Gokcan H, Lsayev O 2022 *Chem. Sci.* **13** 2462
- [27] Chen A Y, Lee J, Damjanovic Ana, Brooks B R 2022 *J. Chem. Theory Comput.* **184** 2673
- [28] Reis Pedro B P S, Bertolini M, Montanari F, Rocchia W, Machuqueiro M, Clevert D A 2022 *J. Chem. Theory Comput.* **18** 5068
- [29] Cai Z T, Liu T Z, Lin Q L, He J H, Lei X W, Luo F F, Huang Y D 2023 *J. Chem. Inf. Model.* **63** 2936
- [30] Baptista A M, Teixeira V H, Soares C M 2002 *J. Chem. Phys.* **117** 4184
- [31] Lee M S, Salsbury F R, Brooks III C L 2004 *Proteins* **56** 738
- [32] Mongan J, Case D A, McCammon J A 2004 *J. Comput. Chem.* **25** 2038
- [33] Meng Y, Roitberg A E 2010 *J. Chem. Theory Comput.* **6** 1401
- [34] Swails J M, York D M, Roitberg A E 2014 *J. Chem. Theory Comput.* **10** 1341
- [35] Machuqueiro M, Baptista A M 2006 *J. Phys. Chem. B* **110** 2927
- [36] Sequeira J G N, Rodrigues F E P, Silva T G D, Reis Pedro B P S, Machuqueiro M 2022 *J. Phys. Chem. B.* **126** 7870
- [37] Huang Y D, Chen W, Dotson D L, Beckstein O, Shen J 2016 *Nat. Commun.* **7** 12940
- [38] Stern H A 2007 *J. Chem. Phys.* **126** 164112
- [39] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H, Pedersen L G 1995 *J. Chem. Phys.* **103** 8577
- [40] Chen Y, Roux B 2015 *J. Chem. Theory Comput.* **11** 3919
- [41] Radak B K, Chipot C, Suh D, Jo S, Jiang W, Philips J C, Schulten K, Roux B 2017 *J. Chem. Theory Comput.* **13** 5933
- [42] Wang R X, Fang X L, Lu Y P, Yang C Y, Wang S M 2005 *J. Med. Chem.* **48** 4111
- [43] Pieri E, Ledentu V, Sahlin M, Dehez F, Olivucci M, Ferre N 2019 *J. Chem. Theory Comput.* **15** 4535
- [44] de Oliveria V M, Liu R, Shen J 2022 *Curr. Opin. Struct. Biol.* **77** 102498
- [45] Kong X, Brooks III C L 1996 *J. Chem. Phys.* **105** 2414
- [46] Khandogin J, Brooks III C L 2005 *Biophys. J.* **89** 141
- [47] Nguyen H, Maier J, Huang H, Perrone V, Simmerling C 2014 *J. Am. Chem. Soc.* **136** 13959
- [48] Huang Y D, Harris R C, Shen J 2018 *J. Chem. Inf. Model.* **58** 1372
- [49] Liu R, Yue Z, Tsai C C, Shen J 2019 *J. Am. Chem. Soc.* **141** 6553
- [50] Harris R C, Liu R, Shen, J 2020 *J. Chem. Theory Comput.* **16** 3689
- [51] Liu R, Zhan S, Che Y, Shen J 2022 *J. Med. Chem.* **65** 1525
- [52] Yao X, Chen C, Wang Y, Dong S, Liu Y, Li Y, Cui Z, Gong W, Perrett S, Yao L, Lamed R, Bayer E A, Cui Q, Feng Y 2020 *Sci. Adv.* **6** eabd7182
- [53] Verma N, Henderson J A, Shen J 2020 *J. Am. Chem. Soc.* **142** 21883
- [54] Arthur E J, Brooks III C L 2016 *J. Comput. Chem.* **37** 2171
- [55] Harris R C, Shen J 2019 *J. Chem. Inf. Model.* **59** 4821
- [56] Wallace J A, Shen J K 2011 *J. Chem. Theory Comput.* **7** 2617
- [57] Henderson J A, Huang Y D, Beckstein O, Shen J 2020 *Proc. Natl. Acad. Sci. U. S. A.* **117** 25517
- [58] Chen W, Huang Y D, Shen J 2016 *J. Phys. Chem. Lett.* **7** 3961
- [59] Yue Z, Li C, Voth G A, Swanson J M J 2019 *J. Am. Chem. Soc.* **141** 13421
- [60] Vo Q N, Mahinthichaichan P, Shen J, Ellis C R 2021 *Nat. Commun.* **12** 984
- [61] Li Z, Zhang X, Wang Q, Li C, Zhang N, Zhang X, Xu B, Ma B, Schrader T E, Coates L, Kovalevsky A, Huang Y D, Wan Q 2018 *ACS Catal.* **8** 8058
- [62] Tsai C C, Yue Z, Shen J 2019 *J. Am. Chem. Soc.* **141** 15092
- [63] Goh G B, Knight J L, Brooks III C L 2012 *J. Chem. Theory Comput.* **8** 36
- [64] Wallace J A, Shen J K 2012 *J. Chem. Phys.* **137** 184105
- [65] Chen W, Shen J K 2014 *J. Comput. Chem.* **35** 1986
- [66] Huang Y D, Chen W, Wallace J A, Shen J 2016 *J. Chem. Theory Comput.* **12** 5411
- [67] Harris J A, Liu R, de Oliveira V M, Vázquez-Montelongo E A, Henderson J A, Shen J 2022 *J. Chem. Theory Comput.* **18** 7510
- [68] Chen W, Wallace J A, Yue Z, Shen J K 2013 *Biophys. J.*

105 L15

- [69] Wallace J A, Shen J K 2009 *Methods Enzymol.* **466** 455
- [70] Ullmann G M 2003 *J. Phys. Chem. B* **107** 1263
- [71] Goh G B, Hulbert B S, Zhou H, Brooks III C L 2014 *Proteins* **82** 1319
- [72] Webb H, Tynan-Connolly B M, Lee G M, Farrell D, O'Meara F, Sondergaard C R, Teilum K, Hewage C, McIntosh L P, Nielsen J E 2010 *Proteins* **79** 685-702
- [73] Rocklin G J, Mobley D L, Dill K A, Hunenberger P H 2013 *J. Chem. Phys.* **139** 184103
- [74] Bignucolo O, Chipot C, Kellenberger S, Roux B 2022 *J. Phys. Chem. B* **126** 6868
- [75] Donnini S, Tegeler F, Groenhof G, Grubmüller H 2011 *J. Chem. Theory Comput.* **7** 1962
- [76] Aho N, Buslaev P, Jansen A, Bauer P, Groenhof G, Hess B 2022 *J. Chem. Theory Comput.* **18** 6148
- [77] Buslaev P, Aho N, Jansen A, Bauer P, Hess B, Groenhof G 2022 *J. Chem. Theory Comput.* **18** 6134
- [78] Knight J L, Brooks III C L 2011 *J. Comput. Chem.* **32** 3423
- [79] Donnini S, Ullmann R T, Groenhof G, Grubmüller H 2016 *J. Chem. Theory Comput.* **12** 1040
- [80] Huang Y D, Shuai J 2013 *J. Phys. Chem. B* **117** 6138
- [81] Lemkul J A, Huang J, Roux B, MacKerell A D 2016 *Chem. Rev.* **116** 4983
- [82] Khandogin J, Brooks III C L 2006 *Biochemistry* **45** 9363
- [83] Itoh S G, Damjanović A, Brooks B R 2011 *Proteins* **79** 3420
- [84] Dashti D S, Meng Y, Roitberg A E 2012 *J. Phys. Chem. B* **116** 8805
- [85] Swails J M, Roitberg A E 2012 *J. Chem. Theory Comput.* **8** 4393
- [86] Lee J, Miller B T, Damjanovic A, Brooks B R 2015 *J. Chem. Theory Comput.* **11** 2560
- [87] Lee J, Miller B T, Damjanovic A, Brooks B R 2014 *J. Chem. Theory Comput.* **10** 2738
- [88] Henderson J A, Verma N, Harris R, Shen J 2020 *J. Chem. Phys.* **153** 115101
- [89] Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid A E, Kolinski A 2016 *Chem. Rev.* **116** 7898
- [90] Bennett W D, Chen A W, Donnini S, Groenhof G, Tieleman D P 2013 *Can. J. Chem.* **91** 839
- [91] da Silva F L B, Sterpone F, Derreumaux P 2019 *J. Chem. Theory Comput.* **15** 3875
- [92] Crinewald F, Souza P C T, Abdizadeh H, Barnoud J, de Vries A H, Marrink S J 2020 *J. Chem. Phys.* **153** 024118
- [93] Reilley D J, Wang J, Dokholyan N V, Alexandrova A N 2021 *J. Chem. Theory Comput.* **17** 4583
- [94] Song Y, Mao J, Gunner M R 2009 *J. Comput. Chem.* **30** 2231
- [95] Wang L, Zhang M, Alexov E 2016 *Bioinformatics* **32** 614
- [96] Pahari S, Sun L, Basu S, Alexov E 2018 *Proteins* **86** 1277
- [97] Bas D C, Rogers D M, Jensen J H 2008 *Proteins* **73** 765
- [98] Sun Z, Wang X, Song J 2017 *J. Chem. Inf. Model.* **57** 1621
- [99] Stepniewska-Dziubinska M M, Zielenkiewicz P, Siedlecki P 2018 *Bioinformatics* **34** 3666
- [100] Pahari S, Sun L, Alexov E 2019 *Database* **2019** baz024
- [101] Ancona N, Bastola A, Alexov E 2023 *J. Comput. Biophys. Chem.* **22** 515
- [102] Reis Pedro B P S, Clevert D A, Machuqueiro M 2022 *Bioinformatics* **38** 297
- [103] Wei W, Hogues H, Sulea T 2023 *J. Chem. Inf. Model.* **63** 5169
- [104] Coskun D, Chen W, Clark A J, Lu C, Hardr E D, Wang L, Friesner R A, Miller E B 2022 *J. Chem. Theory Comput.* **18** 7193
- [105] Hagg A, Kirschner K N 2023 *J. Chem. Inf. Model.* **63** 4505
- [106] Bueschbell B, Caniceiro A B, Suzano P M S, Machuqueiro M, Rosário-Ferreira N, Moreira I S 2022 *Drug Resist. Updat.* **60** 100811

SPECIAL TOPIC—Machine learning in biomolecular simulations

Progress in protein pK_a prediction*Luo Fang-Fang Cai Zhi-Tao Huang Yan-Dong[†]*(College of Computer Engineering, Jimei University, Xiamen 361021, China)**(Received 20 August 2023; revised manuscript received 1 September 2023)*

Abstract

The pH value represents the acidity of the solution and plays a key role in many life events linked to human diseases. For instance, the β -site amyloid precursor protein cleavage enzyme, BACE1, which is a major therapeutic target of treating Alzheimer's disease, functions within a narrow pH region around 4.5. In addition, the sodium-proton antiporter NhaA from *Escherichia coli* is activated only when the cytoplasmic pH is higher than 6.5 and the activity reaches a maximum value around pH 8.8. To explore the molecular mechanism of a protein regulated by pH, it is important to measure, typically by nuclear magnetic resonance, the binding affinities of protons to ionizable key residues, namely pK_a values, which determine the deprotonation equilibria under a pH condition. However, wet-lab experiments are often expensive and time consuming. In some cases, owing to the structural complexity of a protein, pK_a measurements become difficult, making theoretical pK_a predictions in a dry laboratory more advantageous. In the past thirty years, many efforts have been made to accurately and fast predict protein pK_a with physics-based methods. Theoretically, constant pH molecular dynamics (CpHMD) method that takes conformational fluctuations into account gives the most accurate predictions, especially the explicit-solvent CpHMD model proposed by Huang and coworkers (2016 *J. Chem. Theory Comput.* **12** 5411) which in principle is applicable to any system that can be described by a force field. However, lengthy molecular simulations are usually necessary for the extensive sampling of conformation. In particular, the computational complexity increases significantly if water molecules are included explicitly in the simulation system. Thus, CpHMD is not suitable for high-throughput computing requested in industry circle. To accelerate pK_a prediction, Poisson-Boltzmann (PB) or empirical equation-based schemes, such as H++ and PropKa, have been developed and widely used where pK_a values are obtained via one-structure calculations. Recently, artificial intelligence (AI) is applied to the area of protein pK_a prediction, which leads to the development of DeepKa by Huang laboratory (2021 *ACS Omega* **6** 34823), the first AI-driven pK_a predictor. In this paper, we review the advances in protein pK_a prediction contributed mainly by CpHMD methods, PB or empirical equation-based schemes, and AI models. Notably, the modeling hypotheses explained in the review would shed light on future development of more powerful protein pK_a predictors.

Keywords: molecular dynamics, Poisson-Boltzmann equation, machine learning, pK_a prediction**PACS:** 87.15.ap, 87.14.E-, 87.10.Vg, 87.15.A-**DOI:** 10.7498/aps.72.20231356

* Project supported by the National Natural Science Foundation of China (Grant Nos. 11804114, 62006096), the Natural Science Foundation of Fujian Province, China (Grant Nos. 2023J01329, 2020J05146), the Natural Science Foundation of Xiamen, China (Grant No. 3502Z20227205), and the Scientific Starting Research Foundation of Jimei University, China (Grant No. ZQ2020027).

[†] Corresponding author. E-mail: yandonghuang@jmu.edu.cn

专题: 生物分子模拟中的机器学习

生物分子模拟中的机器学习方法*

管星悦¹⁾²⁾ 黄恒焱¹⁾²⁾ 彭华祺¹⁾²⁾ 刘彦航¹⁾ 李文飞^{1)†} 王炜^{1)‡}

1) (南京大学物理学院, 南京 210093)

2) (国科温州研究院, 温州生物物理重点实验室, 温州 325000)

(2023 年 10 月 8 日收到; 2023 年 11 月 1 日收到修改稿)

分子模拟技术已成为人们从分子层次探究生命原理的强有力工具. 经过近 50 年的发展, 生物分子模拟能够实现蛋白折叠、构象运动和蛋白-蛋白分子相互作用等复杂分子体系的生物过程的动力学和热力学性质进行定量表征. 近年来, 以深度学习为代表的机器学习算法的应用进一步推动了生物分子模拟技术的发展. 本文对生物分子模拟中的机器学习方法进行综述, 重点讨论机器学习算法在提高生物分子力场精度、分子模拟构象采样效率、以及高维生物分子模拟数据处理等方面取得的重要进展. 在此基础上, 对未来研究中基于机器学习技术进一步克服生物分子模拟的精度和效率瓶颈、扩展生物分子模拟适用范围、实现计算模拟与实验测量的深度融合做了展望.

关键词: 生物大分子, 分子模拟, 机器学习, 增强采样, 多尺度模型**PACS:** 87.15.ap, 87.15.Cc, 87.18.-h, 87.16.A-**DOI:** 10.7498/aps.72.20231624

1 引言

以分子动力学为代表的分子模拟技术在生物大分子结构与动力学研究中发挥着越来越重要的作用. 常规分子模拟技术用于复杂生物分子体系时, 不可避免地存在力场精度与构象采样效率瓶颈. 同时, 从高维分子模拟数据提取可解释的生物大分子结构与动力学特征也是一个挑战性难题. 生物分子模拟技术发展的核心任务便是解决以上难题, 扩展生物分子模拟的应用范围.

自从 20 世纪 70 年代 McCammon 等^[1]首次将分子动力学模拟用于生物大分子体系以来, 人们在生物分子力场发展、长程静电相互作用计算方法、增强采样与自由能计算等方面取得了多个突破^[2]. 分子模拟技术与高性能计算机等硬件技术的协同发展使得分子模拟能够覆盖的时间尺度以超过摩

尔定律的速度增加, 平均每 10 年增加约 3 个数量级^[3]. 这些进展使得人们能够直接模拟小蛋白分子毫秒时间尺度的折叠全过程^[4,5], 也能对固有无序蛋白 (intrinsically disordered protein, IDP) 的构象系综进行合理的分子模拟表征^[6,7], 甚至能够实现病毒颗粒、细胞质等超大分子体系进行分子模拟^[8,9]. 目前, 实验和模拟计算结合已成为生物大分子结构与动力学研究的基本范式. 另一方面, 对较大的分子体系, 目前的生物分子模拟能够达到的空间和时间尺度与实验测量仍有一定距离, 从而限制了其适用范围^[10]. 因此, 发展新的分子模拟技术, 扩展分子模拟技术的适用范围, 对基于生物分子模拟的基础和应用研究至关重要.

随着计算能力的提升和海量数据的积累, 机器学习算法被广泛应用于基础与应用科学的各个领域. 自然地, 人们也将机器学习算法应用于计算生物学与生物信息学研究, 如生物分子设计与结构预

* 国家自然科学基金 (批准号: 11974173) 资助的课题.

† 通信作者. E-mail: wfli@nju.edu.cn

‡ 通信作者. E-mail: wangwei@nju.edu.cn

测、分子模拟以及分子对接等. 机器学习概念诞生于 20 世纪 50 年代^[11], 并在曲折的发展中被多次重新理解与表述. 早期的机器学习算法多是对既有建模与优化方法的重新整理与表述, 如线性回归、多项式回归^[12] 以及 k -近邻算法^[13] 等. 尽管在早期历史中已初具雏形, 目前人们广泛使用的机器学习算法, 如决策树^[14]、神经网络^[15]、支持向量机^[16] 以及集成学习方法^[17,18] 等, 大多成型于 1980 年后, 并很快被应用于蛋白质二级结构预测^[19]、蛋白结构与功能分类^[20,21] 以及药物筛选^[22] 等问题. 在 20 世纪 90 年代, 人们也开始将神经网络用于构建简单分子体系 (如表面吸附气体分子) 的势能面并进行分子模拟^[23]. 在这些早期的应用中, 机器学习方法往往被视为可替代的工具, 且神经网络尚未表现出相对其他机器学习算法的显著优势, 因此相关算法在生物分子模拟领域的应用仍非常有限.

近年来, 以深度学习为代表的机器学习技术得到迅猛发展, 并在多个领域展现出惊人的能力. 特别是 AlexNet^[24] 的诞生, 展示了深度卷积神经网络对图像的强大识别能力, 宣布深度学习革命的到来. 之后出现的残差网络 (ResNet)^[25] 进一步推动了神经网络向深度发展, 也出现了如生成对抗网络 (GAN)^[26] 与 Transformer^[27] 等网络架构新范式. 这些新的机器学习算法开始广泛用于生物分子模拟、结构预测与设计等领域. 自 2017 年开始, 机器学习与生物分子模拟相结合的研究工作大幅增加, 成为势不可挡的学科交叉趋势. 这一趋势从近年来发表的相关研究论文数目的增长中可见一斑 (图 1).

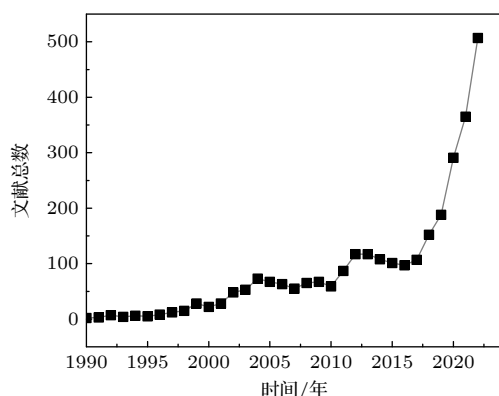


图 1 每年结合生物分子模拟与机器学习的文献数目随年份的变化, 数据来源于 Scopus

Fig. 1. Number of publications with the key words “molecular simulations” and “machine learning” published per year as a function of years. Data were taken from Scopus.

机器学习与生物分子模拟的结合为推进分子生物物理学研究提供了新的机会. 例如, 利用机器学习技术能够设计更准确的分子力场, 开发更高效灵活的增强采样算法, 发展更具普适性的复杂生物分子体系的结构与动力学预测算法, 并辅助药物分子的设计. 这一重要的交叉领域正在高速发展并持续产生具有突破性进展的研究成果^[28-35]. 因此对该领域的发展进行回顾与综述尤为重要. 关于机器学习在生物大分子结构预测与设计方面的进展, 已有非常全面的综述可供参考^[36-40], 本文不再过多讨论. 在机器学习与生物分子模拟交叉领域, 也有学者从不同角度进行了综述^[41-44]. 例如, Ramana-thand 等^[42] 在其综述论文中介绍了使用机器学习技术表征 IDP 系综以及进行多尺度模拟的方法, 并提出将模拟数据集与实验拟合的重要性及策略; Noé 等^[43] 详细介绍了机器学习算法在帮助解决生物分子模拟重要挑战中发挥的作用, 并探讨了将物理学原理融入机器学习算法的必要性及相关方法; Wang 等^[44] 详细总结了利用机器学习算法分析分子动力学模拟轨迹的方法, 以及利用机器学习与相关数据驱动方法进行增强采样的方案. 本文将在此基础上, 结合该领域的最新进展, 从生物分子力场构建、反应坐标的选取与增强采样、分子模拟数据处理等方面对机器学习与分子模拟交叉领域的代表性工作进行综述. 生物物理智识与机器学习技术迭代的融合已成为人们探索生命原理的有力手段, 而结合机器学习算法的生物分子模拟是借助神经网络的强大表达性与拟合能力分析复杂生命运动密码的重要实践. 期望本文对该领域的综述有助于读者综合了解机器学习算法在生物分子模拟中的重要应用, 共同思考和探索基于机器学习算法解决生物分子模拟领域关键难题的可能途径.

2 基于机器学习算法的生物分子力场构建

2.1 势能面与分子力场拟合

在生物分子模拟中, 精度和效率通常难以兼得. 不同的问题在精度和效率上有不同的偏重与要求, 因此需要针对性地选择能够平衡精度与效率要求的折中方案. 计算化学领域的“金标准”CCSD(T)方法能达到约 1 kcal/mol 的化学精度, 但代价

是计算效率低,通常适用于小体系的单点能计算.基于密度泛函理论(DFT)和Born-Oppenheimer绝热近似的方法在精度上作出妥协,从而提升了计算效率,能够将计算体系大小提升到数百个原子以上的规模.但是,对于绝大多数的生物大分子,计算体系通常包含上万个原子,并涉及微秒以上的时间尺度,因此进一步提升生物分子模拟的计算效率对扩展其应用范围十分关键.分子力场模型通过参数化力场的方式在原子坐标水平近似地描述绝热能量面,从而大幅提升计算模拟效率.这种逐级近似的框架之下,如何在提升计算模拟效率的同时尽可能减小精度的损失,成为构建分子力场的核心问题.全原子水平的分子力场可以看作是原子坐标和原子类型的高维空间上的多元函数.传统分子力场多使用基于经验的结构项和以单体、两体势表示的非键相互作用项的参数化方案^[45-47].这种预先设定的具体力场函数形式不可避免地力场精度带来限制.尽管人们可以通过进一步引入极化和多体效应等物理机制来提升参数化方案的表达能力^[48,49],但在精度上与DFT方法仍有较大差距.深度学习算法提供了一种表达能力强大的参数化方案(图2),可以降低对预设力场函数形式的依赖,因此原则上可以提升对分子力场的描述精度.需要注意的是,深度学习算法更强的参数化表达能力,需要由充足的计算能力和训练数据来作为支撑.近年来,计算能力与数据规模已经可以支持用于训练具有足够强表达能力的深度神经网络,因此使用深度学习算法构建生物分子力场,从而实现分子力场精度突破的条件已经成熟,且在此问题上已取得重要进展^[50-55].

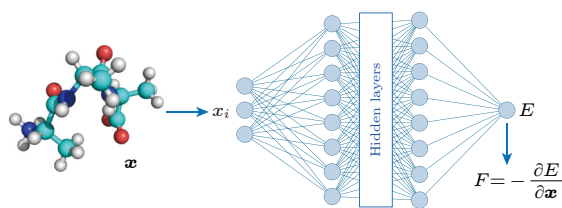


图2 神经网络用于生物分子构象能量面及力场的拟合
Fig. 2. Schematic diagram for representing the biomolecular force field by a neural network.

机器学习算法用于生物分子力场拟合的一个典型例子是Zhang等^[51,56]在2018年发表的DeePMD工作. DeePMD使用原子尺度的构象坐标以及量子力学精度的能量信息作为数据集,将系统构

象映射至其对应的能量与力(受益于神经网络组件的求导能力).给定系统构象坐标,可以通过网络的前向传播代替复杂的DFT计算,直接得到原子受力,从而在尽量保留DFT精度的前提下实现高效率分子动力学模拟. DeePMD的网络架构本身是深度前馈网络,由多个全连接网络的输出求和得到总能量. DeePMD使用分子构型的相对坐标来保证网络的输出不依赖于生物分子体系的平移与旋转变换.值得一提的是, DeePMD可以对接LAMMPS, Gromacs等传统分子动力学模拟软件,便于使用.

为了在神经网络训练中保持分子构型平移与旋转对称性,除使用相对坐标(或单个分子体系的内坐标)外,另一类方法是使用Behler与Parrinello^[57]在2007年提出的对称函数方法.对称函数方法将系统中每一个原子依次视为中心原子,计算其与附近原子的距离、夹角,得到对称函数值,并作为神经网络的输入特征量.例如,Artrith与Urban^[58]发展的Aenet神经网络模型以及Smith等^[59]发展的ANI-1神经网络模型均使用了该对称函数方法,并成功用于体相TiO₂等材料系统和有机物小分子系统的力场拟合. Fan等^[60]在基于进化策略算法构建用于原子模拟的机器学习势时也采用了类似的方法.该对称函数方法规避了笛卡尔坐标与内坐标的相互转换,从而提升深度网络的参数表达能力和训练效率.

以上DeepMD, Aenet, 以及ANI-1均采用了深度前馈网络构架.随着卷积神经网络(CNN)展示出其对图像特征提取与识别的强大能力并在机器学习领域带来革命,人们也尝试使用CNN处理图像的范式来处理分子构型并映射到能量面或力场.特别是残差网络构架的引入,使得人们可以在避免过拟合的前提下,构建足够深度的CNN网络,以增强其拟合效果.一个代表性的例子是Schütt等^[50]发展的SchNet. SchNet以残差卷积网络实现对分子构型特征的提取.不同于处理图像数据使用的网格状离散滤波器,为了保证能量面的光滑性与精确性, SchNet采用了连续滤波器.相对于深度前馈网络,基于CNN架构的SchNet能够显著提升在量子化学精度数据集QM9(包含有机小分子的构型、能量等)的预测精度,也在分子动力学数据集MD17^[61]上有更好的表现.

尽管 CNN 可以提取局域而抽象的特征,且相较于全连接神经网络在避免出现过拟合方面表现出色,但 CNN 最擅长的领域仍是处理规整的图像等数据.对于空间不规则且以共价链接为重要特征的分子构型,图 (graph) 是一种更为自然的表示.分子构型的图描述天然地拥有平移和旋转不变性,并且允许将距离、化学键等连接信息作为“边”数据存入图网络.因为这些优点,人们也尝试使用图神经网络来学习拟合分子力场. Park 等^[53]于 2021 年发表的 GNNFF 基于结合有向图与消息传递 (message passing) 的深度神经网络框架^[62],构建了神经网络分子力场模型,对有机小分子受力的预测精度超过 SchNet. Wang 等^[63]在同年发表的 sGNN,考虑了不同类型相互作用在空间尺度上的差异,对聚合物分子的主链共价作用和非键相互作用能量项分开建模,在空间尺度扩展性与对不同模拟体系的可迁移性方面表现良好.

2.2 粗粒化力场构建

相对于 DFT 等量子化学方法,基于分子力场的全原子分子动力学模型极大地扩展了计算模拟方法能够研究的生物分子体系的空间和时间尺度.目前,人们已经能够实现较小蛋白体系的完整折叠过程进行全原子分子动力学模拟.另外,通过结合增强采样算法,可以实现对较大生物分子体系构象变化的全原子分子动力学模拟和自由能计算.然而,对于更大的生物分子系统,如分子马达、核糖体、病毒颗粒以及染色质体系等,通常包含百万以上原子个数,并涉及毫秒以上时间尺度的动力学过程,远超出全原子分子动力学模拟能够达到的时间和空间尺度范围.为了突破全原子分子动力学模拟的计算效率瓶颈,人们通常采用粗粒化的近似方法^[64].在粗粒化模型中,将多个原子映射为 1 个虚拟粒子,从而很大程度上降低了体系的自由度,实现分子模拟效率的提升.然而,由于采用了虚拟粒子近似,构建具有合理精度的粗粒化分子力场是一个极具挑战性的难题.已有的粗粒化模型的力场参数主要通过“自下而上”和“自上而下”两种策略来优化得到.

“自下而上”策略的基本思路是基于高精度力场模型的计算结果来确定粗粒化力场参数,主要方法有玻尔兹曼反演法 (Boltzmann inversion method)^[65]、力匹配法 (force matching)^[66]、涨落匹配法

(fluctuating matching)^[67]以及能量分解法 (energy decomposition)^[68,69]等.例如,玻尔兹曼反演法主要通过全原子分子动力学模拟得到的径向分布函数 (radial distribution function) 来提取粗粒化层次的有效相互作用参数;而力匹配法的优化目标则是使粗粒化粒子的受力与其在高精度力场中对应粒子的受力尽可能一致.需要注意的是,由于粗粒化近似,粗粒化粒子所代表的原子体系的自由度被冻结,粗粒化力场需要包含所冻结自由度构象熵对能量面的贡献,因此是一种平均力势 (potential of mean force).

以上基于“自下而上”方案构建粗粒化力场的策略与前述基于 DFT 计算结果拟合全原子力场的思路相类似,都希望基于低精度模型拟合更高精度的数据 (能量或力),从而在提升计算效率的同时,尽可能保留足够的精确度.不同的是,从量子力学模型到全原子分子力场模型,由于原子自由度数目维持不变,因此分子力场不涉及构象熵的贡献,原子尺度力场的拟合可以直接使用能量或力作为目标;而在构建粗粒化分子力场模型时,需要在一定程度上体现被冻结自由度的熵效应,因此对分子构象的采样具有更高的要求,将力作为目标拟合力场参数是更常用的方法.另外,构建全原子力场模型的相关算法和构架,如神经网络架构、体现平移与旋转对称性的结构特征提取方法、激活函数的选择等,可以自然地迁移到基于力匹配的粗粒化力场拟合.近年来,基于深度学习构建粗粒化分子模型的工作越来越多地见诸于发表的论文中^[34,52,70-74].例如,DeePMD 团队同时开发出与 DeePMD 具有相似网络架构与结构特征提取策略的深度学习粗粒化力场方案——DeePCG^[52].其中力场参数的提取使用了力匹配法和逐级拟合的办法.同样是基于前馈神经网络架构和力匹配方法, Wang 等^[70]在 2019 年开发了 CGNet,并展示了用于丙氨酸二肽与多肽链的粗粒化模拟结果,能够很好地重现作为参考的全原子模拟得到的自由能面及其他统计性质.

以上例子均采用了基于“自下而上”思路的力匹配法作为粗粒化力场拟合方案.与其相对应的“自上而下”的思路追求粗粒化力场模拟结果与实验约束或高精度模型得到的宏观性质的相容性.然而,因为每一步优化都需要在当前参数下得到模拟轨迹并进行反向传播,自上而下的方法通常会给训练带来较大的计算负担,对拟合目标与参数优化方

案的选择具有更高要求^[75,76]. 近期 Clementi 和 Noé 等^[34,71] 提出了以 flow-matching 为例的一类新方法: 将标准化流 (normalizing flow, NF) 或去噪扩散模型 (denoising diffusion probabilistic model) 等生成模型与力匹配法相结合, 先利用高精度数据训练粗粒化构象的生成模型, 再从这种生成模型中提取粗粒化力场. 这些新的方法将生成模型描述的粗粒化构象偏好视作一种平衡采样, 从而与力场产生联系. 其他的生成模型, 如变分自编码器 (variational auto-encoder, VAE)^[72] 和使用对抗训练思想的 VADE^[73] 同样可以被用于描述粗粒化坐标下的构象分布.

另外, 在基于 C_α 的蛋白质粗粒化模型中, 由于侧链原子位置信息的缺失, 无法准确地体现蛋白质分子的表面积、静电势分布等蛋白质分子的基本性质. 但是这些信息对理解蛋白质分子的结构组装、构象动力学以及分子识别等过程至关重要. 因此, 如何在粗粒化模型框架下准确地计算蛋白质分子的表面积、静电势等蛋白质分子的基本性质是一个重要的技术挑战. 基于深度神经网络的机器学习算法为解决这一问题提供了一个可行的方案. 例如, 本文作者在最近的工作中, 构建了一套深度学习网络 DeepCGSA, 能够基于粗粒化模型结构高精度地估算蛋白质、核酸等生物大分子的溶剂可及性表面积 (图 3)^[74]. 尝试将类似的方法用于针对粗粒化蛋白质结构的静电势分布与 pK_a 值的预测也取得了很好的效果.

3 基于机器学习算法的分子模拟增强采样与数据处理

由于生物大分子具有庞大的自由度数和复杂的能量面特征, 全原子水平的分子模拟通常会遇到采样困难. 特别是在计算各种平衡统计性质时, 需要分子模拟的采样尽可能遍历重要的构象空间, 并在给定的系综条件下达到平衡. 尽管上述粗粒化模型提供了一种解决采样困难的有效方案, 但粗粒化近似不可避免地导致计算精度的损失. 特别是当特异性的氢键、盐桥等原子层次的相互作用起到主导作用时, 粗粒化模型通常无法显式地体现这类特异性相互作用特征, 从而限制了其应用范围. 因此, 发展增强采样算法是解决分子模拟采样困难的另一有效方案. 基于统计物理原理, 人们已经发展出多个有效的增强采样算法, 并广泛应用于生物大分子体系的蒙特卡罗模拟和分子动力学模拟^[78-88]. 目前常见的增强采样算法有伞形抽样 (umbrella sampling)^[78]、副本交换分子动力学 (replica exchange molecular dynamics)^[79]、元动力学 (metadynamics)^[80]、加速分子动力学 (accelerated molecular dynamics)^[81] 以及温度积分增强抽样方法 (integrated tempering sampling, ITS)^[82] 等. 这些增强采样算法多已通过外部插件 (如 PLUMED^[83]) 或直接整合到成熟的分子动力学模拟软件. 另外, 人们也发展了适用于研究构象转变路径的增强采样

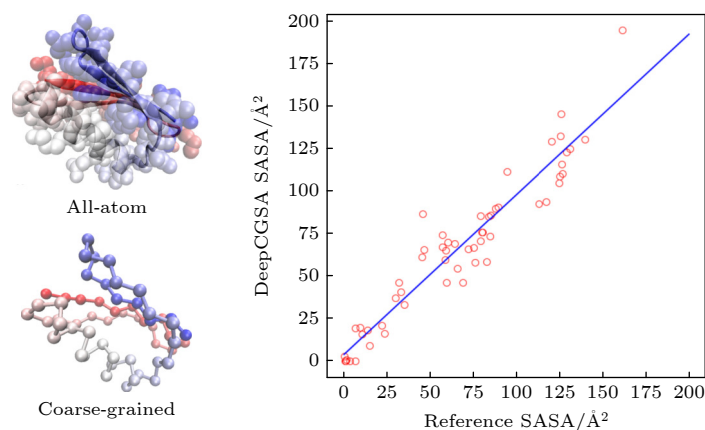


图 3 基于粗粒化结构的蛋白残基溶剂可及性表面积 (SASA) 计算. 左图: 蛋白分子 (protein G, PDB code: 1pgb) 的全原子结构图与粗粒化结构图; 右图: 使用 DeepCGSA 由粗粒化结构计算得到的 SASA 与参考值的对比. 其中参考值使用 Shrake-Rupley 算法由全原子结构计算得到^[77]. DeepCGSA 能够基于粗粒化结构给出接近参考值的 SASA 计算结果

Fig. 3. SASA estimation based on coarse-grained protein structure. Left: All-atom structure and coarse-grained structure of protein G (PDB code: 1 pgb). Right: Correlation plot between the SASA values from DeepCGSA based on one-bead coarse-grained structure and the reference values by Shrake-Rupley algorithm based on all-atom structure. The DeepCGSA can well reproduce the SASA values based on coarse-grained structure.

算法, 如 String 方法^[84]与 Transition path sampling 方法^[85]等. 最近, 人们将机器学习算法用于生物分子模拟的增强采样, 并取得了显著效果, 甚至还可以利用机器学习算法, 基于有限的构象采样数据实现高维自由能面的构建^[89,90].

3.1 基于机器学习算法提取反应坐标

常用的增强采样算法可分为两类: 依赖反应坐标的增强采样算法和不依赖反应坐标的增强采样算法. 例如, 伞形抽样、元动力学等增强采样算法依赖于预先定义的反应坐标, 这类算法的基本策略通常是沿预先定义的反应坐标方向添加偏置势, 从而避免在沿反应坐标的局部势阱中重复采样. 因此, 预先定义的反应坐标需对应所关注的生物分子体系最重要的运动方向, 而垂直于反应坐标方向的动力学具有更快的时间尺度. 然而, 定义合适的反应坐标本身就是一项极具挑战性的任务. 通常情况下, 反应坐标主要基于物理直觉来选取, 而机器学习等数据驱动的降维方法为反应坐标的选取给出了一个更为理性和可操作的方案.

常规的不使用神经网络的数据驱动降维方法主要基于如下思想设计: 在降维前后的空间里, 尽可能维持数据的某种结构信息不变. 这种“结构信息”可以分为全局信息和局域信息两类. 早在 20 世纪初就被开发的主元分析算法 PCA, 是一种典型的致力于维持全局结构信息的算法^[91]. PCA 将高维数据点相对于几何中心的欧式距离平方和视作需要保留的“结构信息”, 在通过线性变化降维过程中最小化该结构信息的损失, 并找到承担最大运动信息变化的反应坐标. PCA 方法的缺陷也在于此: 基于全局的欧式距离衡量信息并非总是一个合理的预设; 且 PCA 要求降维至超平面, 就只允许对数据做全局的线性变换, 很多时候这是一个过强的假设.

更一般地, 可以假设高维数据分布在一个黎曼流形 (或是几支黎曼流形) 上. 此时欧式距离只适用于描述数据点的局域结构, 即可以构建起离散数据点的近邻图, 而全局结构可视为由这些近邻图组合而成. 基于这一思想, Isomap 算法^[92]和 Diffusion Map 算法^[93]分别用测地线距离和模拟扩散距离衡量数据点的间距, 并希望降维映射前后这些距离尽量保持不变, 从而将流形“展平”以实现降维. 将 Isomap 与 Diffusion Map 用于分子模拟数据分析, 可以找到非线性地依赖于高维数据的反应坐标^[94-96].

在基于局域结构信息的降维方法中, 2008 年提出的 t-SNE 算法具有突出的表现^[97]. t-SNE 对数据点间的相似性做非线性变换, 使得降维过程中主要维护局部团簇 (cluster) 中两点相似性的分布不变, 而对相似性低的数据点的位置关系几乎没有约束. 因此, t-SNE 的降维尽量维持了数据点基于相似性簇团的内部结构, 而对簇团间的距离朝向则几乎没有要求, 从而带来了降维结果的随机性. t-SNE 使用梯度下降优化低维空间数据点的位置, 通常这是一个非凸优化, 每次得到的结果会有所差别. 相比于 2002 年提出的 SNE 算法^[98], t-SNE 构建对称的损失函数以代替 SNE 中不对称的 K-L 散度, 简化了基于梯度的优化过程; 同时 t-SNE 以更为长尾的 t-分布建立低维空间距离向概率的映射, 以更好应对高维数据点嵌入低维空间导致的拥挤问题. 图 4 给出了使用 PCA, t-SNE 以及 UMAP 对粗粒化分子动力学得到的蛋白折叠轨迹^[99]进行降维的效果对比: 相比于 PCA, t-SNE 和 UMAP 能更好地区分折叠态和解折叠态的结构. 在分子模拟中, 基于 t-SNE 的降维算法已被广泛应用于反应坐标的定义与高维动力学轨迹的可视化^[100-102]. 除 t-SNE 外, 基于局域结构信息的降维方法还有: 维持局域线性关系的 LLE (locally linear embedding)^[103]、维持局域邻近图的 Laplacian Eigenmaps^[104]、最小化局域曲率的 Hessian LLE^[105]等, 然而它们在分子模拟领域得到的关注和应用远不如 t-SNE. 2018 年 McInnes 等^[106]提出的 UMAP 降维算法采用了与 t-SNE 类似的、基于邻近图提取簇团信息的策略, 并同样用梯度下降方法优化得到低维嵌入. 不同的是, 相比于围绕着“点”进行的 t-SNE, UMAP 采用了以“边”为中心的优化策略, 使用交叉熵作为优化目标, 将边存在的概率映射为低维空间的距离. 在生物分子模拟中, UMAP 常被用于基因组、染色质和单细胞转录谱等数据^[107,108]. 在单细胞转录谱数据集与蛋白质动力学轨迹数据上的比较研究^[109-111]均表明: UMAP 具有不逊色于 t-SNE 的降维效果, 但是在计算成本上远低于 t-SNE, 对大规模的数据有良好的扩展性, 这与 UMAP 原始论文中指出其计算复杂度约为 $N^{1.4}$ 一致^[106].

如果认为降维算法的关键问题在于对信息的选择与度量, 那么以上非神经网络的机器学习降维算法都是通过引入某种预设 (或主观判断) 来解决

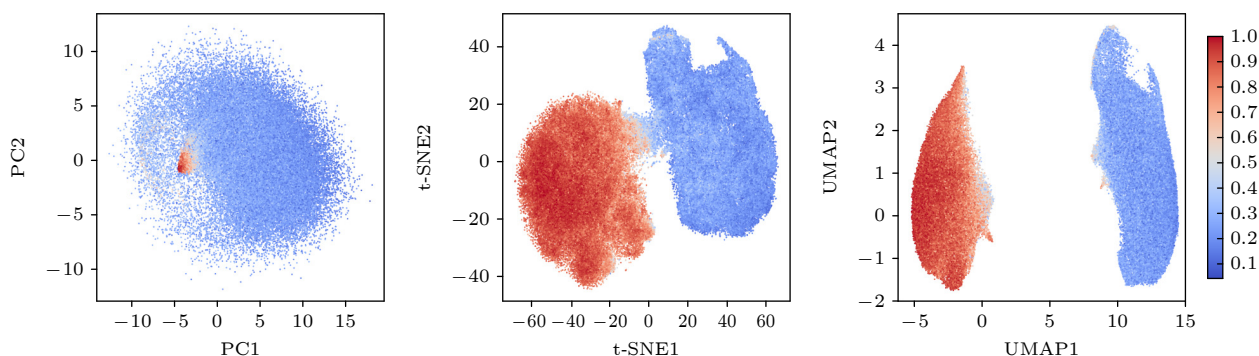


图4 用PCA(左)、t-SNE(中)和UMAP(右)对蛋白分子Protein G的基于粗粒化分子动力学的模拟轨迹^[99]降维效果对比. 蓝色到红色对应表征蛋白折叠程度的 Q 值; $Q=1$ (红色)为完全折叠结构, $Q=0$ (蓝色)为完全解折叠结构

Fig. 4. Projection of the sampled snapshots of the coarse-grained molecular dynamics simulations for protein G^[99] along the reaction coordinates constructed by PCA (left), t-SNE (middle), and UMAP (right), respectively. t-SNE and UMAP perform better than PCA in distinguishing the folded and unfolded structures. Colors from blue to red represent the structures with increasing folding extent: blue, fully unfolded; red, fully folded.

此问题,也因此降低了对降维变换的表达能力.借助于具有强大表达能力的神经网络,可以期待构建更有效的降维算法.

在2013年被开发的VAE,通过巧妙地设计神经网络架构,将原始数据通过编码器降维得到隐变量,再通过解码器升维,生成与原始数据同维度的高维数据^[112].如果生成数据具有和原始数据几乎相同的分布,则说明编码过程(即降维过程)几乎没有造成信息损失,低维的隐变量具有与原始数据相近的表达能力.就训练过程而言,VAE通过优化编码器和解码器参数,以最小化生成数据与原始数据分布上的差异.其中,隐变量的“信息”通过复现原始数据分布的能力衡量.相比于以上非神经网络的降维算法中预设信息为数据集上的某种结构的做法,VAE衡量信息的方式更具一般性与整体性.对于生物分子模拟系统,这一优势将有利于VAE通过降维找到整体性的反应坐标;而编码器、解码器所基于的深度神经网络架构保证了VAE强大的表达能力,降低了模型对预设信息的依赖,有利于增强降维的有效性.因此VAE常被用于生物分子模拟反应坐标的提取.另外,VAE寻找反应坐标的思路同样可以用于粗粒化模型的建立^[72]、反应路径搜索^[113]、甚至是药物分子设计^[114]等任务.

3.2 基于机器学习算法的增强采样

3.2.1 非生成模型

机器学习算法不仅可以用于寻找合适的反应坐标,还可以直接用于辅助分子模拟采样.例如,

在利用Metadynamics方法进行增强采样和自由能计算时,需要在分子体系的固有能量面添加一定形状的高斯形偏置势^[115],而确定高斯形偏置势的参数及其变化规律非常关键,直接影响采样效率.过强的高斯形偏置势可能会导致采样进入非物理的区域,而过弱的高斯形偏置势又难以遍历感兴趣的构象空间区域.2019年,Bonati等^[116]通过结合神经网络与变分增强采样思路,灵活地以变分形式在增强采样模拟过程中自适应地更新偏置势,使得反应坐标的实际分布能够逼近目标分布.相较于常规的Metadynamics方法,在灵活性、高效性与准确性方面得到了提升.

另一个使用神经网络给出偏置势用以增强采样的例子是Zhang等^[117]提出的TALOS(targeted adversarial learning optimized Ssampling).类似生成对抗网络GAN的思想(见下方关于生成模型描述),TALOS使用Wasserstein距离衡量真实分子模拟引擎生成的构型分布与目标构型分布的差异,将此距离的计算转化为对一个判别器网络的优化问题.TALOS的训练同样类似于GAN:对每个偏置势,通过优化判别器网络计算两分布Wasserstein距离的近似值,以之作为两分布差异的数值衡量;最小化此差异以优化偏置势,从而使偏置势下模拟产生的构型分布尽可能接近目标分布.

3.2.2 生成模型

在以上例子中,机器学习算法仅被用作提供构造反应坐标或设置偏置势的手段,即增强采样的

mics 的思想), 此后再进行 reweighting 操作, 得到正确的分布. 在将 Boltzmann Generator 用于 BPTI 蛋白的构象采样时, 成功得到了其“X”态到开放的“O”态之间的构象转变, 即使这种转变的过渡态并不存在于训练集中, 展现了 Boltzmann Generator 在用于生物分子构象采样的强大能力.

另外一类常见的增强采样算法采用了强化学习方法. 强化学习使用奖惩机制, 在不同的环境条件下强化学习器采取不同的动作时将给出一定的奖励或惩罚, 而训练的目标是使得强化学习的动作能够将奖励最大化. Shamsi 等^[119]提出了基于强化学习的 REAP 算法, 将奖惩机制与分子构象空间的探索绑定在一起, 寻找最利于在构象空间扩散的反应坐标. 该方法用于丙氨酸二肽和 Src 激酶体系时展示了出色的增强采样的效果. 基于类似的思想, 人们也可以基于强化学习, 在沿所设定反应坐标采样的不确定度 (uncertainty) 上施加奖惩来鼓励体系在未遍历的构象区域采样 (在已经遍历的方向上反应坐标的不确定度较低), 因此能对增强采样模拟施加一个自适应的灵活偏置势^[120], 达到增强采样的效果.

综上, 机器学习在增强采样领域表现出强大的功能和前景, 既可以在传统增强采样算法框架下通过构建反应坐标发挥作用, 也可以通过自适应的方式提供高效灵活的偏置势, 还可以直接利用生成模型作为采样核心. 随着新的机器学习算法的开发, 将机器学习用于辅助生物分子模拟增强采样是未来生物分子模拟领域的重要课题.

3.3 基于机器学习算法的生物分子模拟数据处理

生物分子模拟通常在高维空间中进行, 所得到的分子模拟轨迹包含了丰富的结构与动力学信息, 如何从这些高维的分子模拟轨迹提取出可解释的热力学与动力学数据, 并实现与实验结果的定量比较是分子模拟领域的另一个挑战性难题. 分子动力学模拟数据处理主要包括以下几个方面: 高维分子模拟数据特征提取、分子模拟轨迹降维与反应坐标构建、分子模拟微观状态粗粒化与马尔可夫状态模型构建, 以及低维自由能面构建等. 显然, 适合于处理复杂数据的各类机器学习算法在生物分子模拟数据处理中扮演着越来越重要的角色. 事实上, 前述关于增强采样算法部分介绍的基于机器学习

的反应坐标构建是机器学习用于分子模拟数据处理的重要方面. 除此之外, 人们也发展了深度网络模型, 用于提取生物分子体系的动力学与自由能信息. 例如, Mardt 等^[121]设计了 VAMPNet, 能够端到端地直接实现从分子模拟数据轨迹得到马尔可夫状态模型 (Markov state model) 的映射. 以马尔可夫过程的变分法 (VAMP) 为基础, 深度网络用于表达特征变换的形式, 通过变换后的特征空间内近似得到弛豫时间 τ 范围内的状态转移矩阵, 从而用于提取生物分子的力学信息. 另外, Schneider 等^[90]通过训练神经网络, 实现了高维自由能的计算以及典型系综平均性质的计算.

4 总结与展望

本文对机器学习方法在生物分子模拟领域的应用进行了综述. 借助其突出的特征提取和参数拟合能力, 机器学习方法 (特别是神经网络算法) 在全原子/粗粒化分子力场构建、分子模拟数据降维与反应坐标提取、以及生物分子构象采样等方面已经开始发挥重要作用. 随着以深度神经网络为代表的机器学习算法的迭代更新, 结合机器学习算法的生物分子模拟技术将成为人们在分子层次探索生命原理的重要研究范式. 需要指出的是, 目前机器学习算法大多作为辅助工具在生物分子模拟中发挥作用. 即使整合了机器学习算法, 对较大的生物分子体系能够达到的分子模拟时间尺度仍与真实生物学相关时间尺度有较大差距. 完全解决生物分子模拟精度与效率瓶颈, 实现生物分子模拟与实验测量的定量比较, 需要在分子模拟的理论框架与算法方面同时进行探索. 近年来, 整合全原子模型和粗粒化模型的多尺度生物分子模拟技术越来越受到人们的重视^[67,122-124], 是解决生物分子模拟精度与效率瓶颈的一个值得重点尝试的思路. 神经网络等机器学习算法的发展将成为进一步推动多尺度分子模拟技术发展的新突破口.

尽管本文将机器学习用于生物分子模拟的工作分为力场构建、增强采样以及数据处理等不同的主题来进行综述, 近年来突破性的工作通常打破了主题分类的边界, 并依赖于多个步骤的集成耦合. 因此, 实现机器学习在生物分子模拟多方面的融合应用, 需要开发能够集成机器学习算法与生物分子模拟的软件平台. 例如机器学习与生物物理交叉领

域代表工作——AlphaFold2 与 ESM 大语言模型, 均得益于对多模态数据与算法的集成整合能力^[30,125]. 国内在相关领域的集成软件平台开发方面也取得了很大进展. 由深势科技开发的 RiDYMO 平台集成了神经网络、分子动力学引擎、增强采样方法, 不仅能进行分子动力学模拟, 分析蛋白质构象空间、还能探索药物结合位点并计算药效相关动力学参数, 适合药物的设计与开发工作^[126]. 北京大学与华为等团队开发的 MindSPONGE^[127] 在华为昇思 MindSpore 框架下整合了多种分子模拟、结构预测设计以及全面的神经网络支持. 这些集成平台将降低新算法的开发和使用门槛, 促进生物分子模拟技术的应用范围.

关于机器学习与生物分子模拟融合应用的研究进展给我们带来一个重要的启示: 生物物理知识与机器学习发展是相辅相成的. 例如, AlphaFold2 的架构借鉴了由序列比对得到的共进化信息, 而 AlphaFold2 的成功又是机器学习推进生物分子结构预测领域的代表例子. 生物分子模拟与神经网络结合的需求也同样在推进机器学习领域的发展. SchNet 为了拟合光滑连续力场而在卷积神经网络架构下提出的连续滤波器, 可以被推广到其他机器学习任务情景; 而主要受分子结构拓扑相关研究驱动而发展的图神经网络, 也被推广到诸如社交网络等应用情景中. 机器学习架构的每一次突破性进展都会为生物分子研究领域带来难以估量的灵感与启发. 如何借助神经网络的成功进一步反哺生物物理知识与经验将是未来生物物理与人工智能交叉领域的重点研究课题.

参考文献

- [1] McCammon J A, Gelin B R, Karplus M 1977 *Nature* **267** 585
- [2] Schlick T, Portillo-Ledesma S 2021 *Nat. Comput. Sci.* **1** 321
- [3] Vendruscolo M, Dobson C M 2011 *Curr. Biol.* **21** R68
- [4] Shaw D E, Maragakis P, Lindorff-Larsen K, et al. 2010 *Science* **330** 341
- [5] Zhou C Y, Jiang F, Wu Y D 2015 *J. Phys. Chem. B* **119** 1035
- [6] Zerze G H, Zheng W, Best R B, Mittal J 2019 *J. Phys. Chem. Lett.* **10** 2227
- [7] Robustelli P, Piana S, Shaw D E 2018 *Proc. Natl. Acad. Sci. U.S.A.* **115** E4758
- [8] Perilla J R, Schulten K 2017 *Nat. Commun.* **8** 15959
- [9] Yu I, Mori T, Ando T, Harada R, Jung J, Sugita Y, Feig M 2016 *eLife* **5** e19274
- [10] Li W F, Zhang J, Wang J, Wang W 2015 *Acta Phys. Sin.* **64** 098701 (in Chinese) [李文飞, 张建, 王骏, 王炜 2015 物理学报 **64** 098701]
- [11] Samuel A L 1959 *IBM J. Res. Dev.* **3** 210
- [12] Stigler S M 1974 *Hist. Math.* **1** 431
- [13] Fix E, Hodges J L 1951 *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties* (Randolph Field, Texas: USAF School of Aviation Medicine) Tech. Rep. 4
- [14] Breiman L, Friedman J H, Olshen R A, Stone C J 1984 *Biometrics* **40** 874
- [15] Runelhart D E, Hinton G E, Williams R J 1986 *Nature* **323** 533
- [16] Cortes C, Vapnik V 1995 *Mach. Learn.* **20** 273
- [17] Ho T K 1995 *Proceedings of 3rd International Conference on Document Analysis and Recognition* Montreal, QC, Canada, August 14–16, 1995 p278
- [18] Freund Y, Schapire R E 1996 *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning* San Francisco, CA, USA, July 1996 p148
- [19] Holley L, Karplus M 1989 *Proc. Natl. Acad. Sci. U.S.A.* **86** 152
- [20] Cai Y, Liu X, Xu X, Zhou G 2001 *BMC Bioinf.* **2** 1
- [21] Cai C, Wang W, Sun L, Chen Y 2003 *Math. Biosci.* **185** 111
- [22] Zernov V V, Balakin K V, Ivaschenko A A, Savchuk N P, Pletnev I V 2003 *J. Chem. Inf. Comput. Sci.* **43** 2048
- [23] Blank T B, Brown S D, Calhoun A W, Doren D J 1995 *J. Chem. Phys.* **103** 4129
- [24] Krizhevsky A, Sutskever I, Hinton G E 2017 *Commun. ACM* **60** 84
- [25] He K, Zhang X, Ren S, Sun J 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Las Vegas, NV, USA, June 27–30, 2016 p770
- [26] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y 2020 *Commun. ACM* **63** 139
- [27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I 2017 *Proceedings of the 31st International Conference on Neural Information Processing Systems* New York, USA, December 4–9, 2017 p6000
- [28] Noé F, Olsson S, Köhler J, Wu H 2019 *Science* **365** eaaw1147
- [29] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D 2020 *Proc. Natl. Acad. Sci. U.S.A.* **117** 1496
- [30] Jumper J, Evans R, Pritzel A, et al. 2021 *Nature* **596** 583
- [31] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee G R, Wang J, Cong Q, Kinch L N, Schaeffer R D, Millán C, Park H, Adams C, Glassman C R, DeGiovanni A, Pereira J H, Rodrigues A V, Van Dijk A A, Ebrecht A C, Opperman D J, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy M K, Dalwadi U, Yip C K, Burke J E, Garcia K C, Grishin N V, Adams P D, Read R J, Baker D 2021 *Science* **373** 871
- [32] Huang B, Xu Y, Hu X, Liu Y, Liao S, Zhang J, Huang C, Hong J, Chen Q, Liu H 2022 *Nature* **602** 523
- [33] Liu Y, Zhang L, Wang W, Zhu M, Wang C, Li F, Zhang J, Li H, Chen Q, Liu H 2022 *Nat. Comput. Sci.* **2** 451
- [34] Köhler J, Chen Y, Krämer A, Clementi C, Noé F 2023 *J. Chem. Theory Comput.* **19** 94216
- [35] Watson J L, Juergens D, Bennett N R, Trippe B L, Yim J, Eisenach H E, Ahern W, Borst A J, Ragotte R J, Milles L F, Wicky B I M, Hanikel N, Pellock S J, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres S V, Lauko

- A, Bortoli V D, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola T S, DiMaio F, Baek M, Baker D 2023 *Nature* **620** 1089
- [36] Kuhlman B, Bradley P 2019 *Nat. Rev. Mol. Cell Biol.* **20** 681
- [37] Jisna V, Jayaraj P 2021 *Protein J.* **40** 522
- [38] AlQuraishi M 2021 *Curr. Opin. Chem. Biol.* **65** 1
- [39] Xu Y, Verma D, Sheridan R P, Liaw A, Ma J, Marshall N M, McIntosh J, Sherer E C, Svetnik V, Johnston J M 2020 *J. Chem. Inf. Model.* **60** 2773
- [40] Huang B, Du Y, Zhang S, Li W, Wang J, Zhang J 2020 *Chin. Phys. B* **29** 108704
- [41] Zhang J, Chen D, Xia Y, et al. 2023 *J. Chem. Theory Comput.* **19** 4338
- [42] Ramanathan A, Ma H, Parvatikar A, Chennubhotla S C 2021 *Curr. Opin. Struct. Biol.* **66** 216
- [43] Noé F, Tkatchenko A, Müller K R, Clementi C 2020 *Annu. Rev. Phys. Chem.* **71** 361
- [44] Wang Y, Ribeiro J M L, Tiwary P 2020 *Curr. Opin. Struct. Biol.* **61** 139
- [45] Sambasivarao S V, Acevedo O 2009 *J. Chem. Theory Comput.* **5** 1038
- [46] Brooks B R, Brooks III C L, Mackerell Jr. A D, Nilsson L, Petrella R J, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caffisch A, Cavas L, Cui Q, Dinner A R, Feig M, Fischer S, Gao J, Hodoseck M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor R W, Post C B, Pu J Z, Schaefer M, Tidor B, Venable R M, Woodcock H L, Wu X, Yang W, York D M, Karplus M 2009 *J. Comput. Chem.* **30** 1545
- [47] Wang J, Wolf R M, Caldwell J W, Kollman P A, Case D A 2004 *J. Comput. Chem.* **25** 528
- [48] Peng X, Zhang Y, Chu H, Li Y, Zhang D, Cao L, Li G 2016 *J. Chem. Theory Comput.* **12** 2973
- [49] Liu C, Qi R, Wang Q, Piquemal J P, Ren P 2017 *J. Chem. Theory Comput.* **13** 2751
- [50] Schütt K T, Kindermans P J, Saucedo H E, Chmiela S, Tkatchenko A, Müller K R 2017 *Proceedings of the 31st International Conference on Neural Information Processing Systems* New York, USA, December 4–9, 2017 p992
- [51] Zhang L, Han J, Wang H, Car R, Weinan E 2018 *Phys. Rev. Lett.* **120** 143001
- [52] Zhang L, Han J, Wang H, Car R, Weinan E 2018 *J. Chem. Phys.* **149** 034101
- [53] Park C W, Kornbluth M, Vandermause J, Wolverson C, Kozinsky B, Mailoa J P 2021 *npj Comput. Mater.* **7** 73
- [54] batznerzner S, Musaelian A, Sun L, Geiger M, Mailoa J P, Kornbluth M, Molinari N, Smidt T E, Kozinsky B 2022 *Nat. Commun.* **13** 2453
- [55] Wang Y, Li S, He X, Li M, Wang Z, Zheng N, Shao B, Wang T, Liu T Y 2022 arXiv: 2210.16518 [cs.LG]
- [56] Zhang L F, Han J Q, Wang H, Saidi W, Car R, E W H 2018 *Advances in Neural Information Processing Systems* Montreal, Canada, Decembe 3–8, 2018 p4441
- [57] Behler J, Parrinello M 2007 *Phys. Rev. Lett.* **98** 146401
- [58] Artrith N, Urban A 2016 *Comput. Mater. Sci.* **114** 135
- [59] Smith J S, Isayev O, Roitberg A E 2017 *Chem. Sci.* **8** 3192
- [60] Fan Z, Wang Y, Ying P, et al. 2022 *J. Chem. Phys.* **157** 114801
- [61] Chmiela S, Tkatchenko A, Saucedo H E, Poltavsky I, Schütt K T, Müller K R 2017 *Sci. Adv.* **3** e1603015
- [62] Gilmer N M P, Schoenholz S S, Riley P F, Vinyals O, Dahl G E 2017 *Proceedings of the 34th International Conference on Machine Learning* Sydney, Australia, August 6–11, 2017 p1263
- [63] Wang X, Xu Y, Zheng H, Yu K 2021 *J. Phys. Chem. Lett.* **12** 7982
- [64] Takada S, Kanada R, Tan C, Terakawa T, Li W, Kenzaki H 2015 *Acc. Chem. Res.* **48** 3026
- [65] Reith D, Pütz M, Müller-Plathe F 2003 *J. Comput. Chem.* **24** 1624
- [66] Izvekov S, Voth G A 2005 *J. Phys. Chem. B* **109** 2469
- [67] Chu J W, Ayton G, Izvekov S, Voth G 2007 *Mol. Phys.* **105** 167
- [68] Li W, Wolynes P G, Takada S 2011 *Proc. Natl. Acad. Sci. U.S.A.* **108** 3504
- [69] Gohlke H, Kiel C, Case D A 2003 *J. Mol. Biol.* **330** 891
- [70] Wang J, Olsson S, Wehmeyer C, Pérez A, Charron N E, De Fabritiis G, Noé F, Clementi C 2019 *ACS Cent. Sci.* **5** 755
- [71] Arts M, Satorras V G, Huang C W, Zuegner D, Federici M, Clementi C, Noé F, Pinsler R, van den Berg R 2023 arXiv: 2302.00600 [cs.LG]
- [72] Wang W, Gómez-Bombarelli R 2019 *Npj Comput. Mater.* **5** 125
- [73] Zhang J, Lei Y K, Yang Y I, Gao Y Q 2020 *J. Chem. Phys.* **153** 174115
- [74] Dong T, Gong T, Li W 2021 *J. Phys. Chem. B* **125** 9490
- [75] Marrink S J, Risselada H J, Yefimov S, Tieleman D P, de Vries A H 2007 *J. Phys. Chem. B* **111** 7812
- [76] Souza P C T, Alessandri R, Barnoud J, Thallmair S, Faustino I, Grünewald F, Patmanidis I, Abdizadeh H, Bruininks B M H, Wassenaar T A, Kroon P C, Meler J, Nieto V, Corradi V, Khan H M, Domański J, Javanainen M, Martinez-Seara H, Reuter N, Best R B, Vattulainen I, Monticelli L, Periolo X, Tieleman D P, de Vries A H, Marrink S J 2021 *Nat. Methods* **18** 382
- [77] Shrake A, Rupley J A 1973 *J. Mol. Biol.* **79** 351
- [78] Torrie G M, Valleau J P 1977 *J. Comput. Phys.* **23** 187
- [79] Sugita Y, Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141
- [80] Laio A, Parrinello M 2002 *Proc. Natl. Acad. Sci. U.S.A.* **99** 12562
- [81] Hamelberg D, Mongan J, McCammon J A 2004 *J. Chem. Phys.* **120** 11919
- [82] Yang L, Liu C W, Shao Q, Zhang J, Gao Y Q 2015 *Acc. Chem. Res.* **48** 947
- [83] Tribello G A, Bonomi M, Branduardi D, Camilloni C, Bussi G 2014 *Comput. Phys. Commun.* **185** 604
- [84] E W, Ren W, Vanden-Eijnden E 2002 *Phys. Rev. B* **66** 052301
- [85] Dellago C, Bolhuis P G, Csajka F S, Chandler D 1998 *J. Chem. Phys.* **108** 1964
- [86] Chen C, Huang Y, Xiao Y 2013 *J. Biomol. Struct. Dyn.* **31** 206
- [87] Zhang J, Gong H 2020 *J. Chem. Theory Comput.* **16** 4813
- [88] Zhu W, Zhang J, Wang J, Li W, Wang W 2021 *Phys. Rev. E* **103** 032404
- [89] Zheng S, He J, Liu C, et al. 2023 arXiv: 2306.05445 [physics.chem-ph]
- [90] Schneider E, Dai L, Topper R Q, Drechsel-Grau C, Tuckerman M E 2017 *Phys. Rev. Lett.* **119** 150601
- [91] Jolliffe I T 2002 *Principal Component Analysis for Special Types of Data* (New York: Springer) pp338–372
- [92] Tenenbaum J B, de Silva V, Langford J C 2000 *Science* **290** 2319
- [93] Lafon S, Lee A B 2006 *IEEE Trans. Pattern Anal. Mach. Intell.* **28** 1393
- [94] Das P, Moll M, Stamati H, Kavradi L E, Clementi C 2006

- Proc. Natl. Acad. Sci. U.S.A.* **103** 9885
- [95] Plaku E, Stamati H, Clementi C, Kaviraki L E 2007 *Proteins Struct. Funct. Bioinf.* **67** 897
- [96] Trstanova Z, Leimkuhler B, Lelièvre T 2020 *Proc. R. Soc. A* **476** 20190036
- [97] van der Maaten L, Hinton G 2008 *J. Mach. Learn. Res.* **9** 2579
- [98] Hinton G, Roweis S 2002 *Proceedings of the 15th International Conference on Neural Information Processing Systems* Vancouver, British Columbia, Canada, December 9–14, 2002 p857
- [99] Li W, Terakawa T, Wang W, Takada S 2012 *Proc. Natl. Acad. Sci. U.S.A.* **109** 17789
- [100] Rydzewski J, Nowak W 2016 *J. Chem. Theory Comput.* **12** 2110
- [101] Zhou H, Wang F, Tao P 2018 *J. Chem. Theory Comput.* **14** 5499
- [102] Spiwok V, Kříž P 2020 *Front. Mol. Biosci.* **7** 132
- [103] Roweis S T, Saul L K 2000 *Science* **290** 2323
- [104] Belkin M, Niyogi P 2001 *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* Vancouver, British Columbia, Canada, December 3–8, 2001 p585
- [105] Donoho D L, Grimes C 2003 *Proc. Natl. Acad. Sci. U.S.A.* **100** 5591
- [106] McInnes L, Healy J, Melville J 2018 arXiv: 1802.03426 [stat.ML]
- [107] Chen S, Lake B B, Zhang K 2019 *Nat. Biotechnol.* **37** 1452
- [108] Mimitou E P, Lareau C A, Chen K Y, et al 2021 *Nat. Biotechnol.* **39** 1246
- [109] Becht E, McInnes L, Healy J, Dutertre C A, Kwok I W, Ng L G, Ginhoux F, Newell E W 2019 *Nat. Biotechnol.* **37** 38
- [110] Trozzi F, Wang X, Tao P 2021 *J. Phys. Chem. B* **125** 5022
- [111] Do V H, Canzar S 2021 *Genome Biol.* **22** 130
- [112] Kingma D P, Welling M 2013 arXiv:1312.6114 [stat.ML]
- [113] Ramaswamy V K, Musson S C, Willcocks C G, Degiacomi M T 2021 *Phys. Rev. X* **11** 011052
- [114] Gómez-Bombarelli R, Wei J N, Duvenaud D, Hernández-Lobatzner J M, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel T D, Adams R P, Aspuru-Guzik A 2018 *ACS Cent. Sci.* **4** 268
- [115] Barducci A, Bussi G, Parrinello M 2008 *Phys. Rev. Lett.* **100** 020603
- [116] Bonati L, Zhang Y Y, Parrinello M 2019 *Proc. Natl. Acad. Sci. U.S.A.* **116** 17641
- [117] Zhang J, Yang Y I, Noé F 2019 *J. Phys. Chem. Lett.* **10** 5791
- [118] Rezende D J, Mohamed S 2015 *Proceedings of the 32nd International Conference on International Conference on Machine Learning* **37** 1530
- [119] Shamsi Z, Cheng K J, Shukla D 2018 *J. Phys. Chem. B* **122** 8386
- [120] Zhang L, Wang H, E W 2018 *J. Chem. Phys.* **148** 12411
- [121] Mardt A, Pasquali L, Wu H, Noé F 2018 *Nat. Commun.* **9** 5
- [122] Li W, Yoshii H, Hori N, Kameda T, Takada S 2010 *Methods* **52** 106
- [123] Li W, Wang J, Zhang J, Wang W 2015 *Curr. Opin. Struct. Biol.* **30** 25
- [124] Li G H 2023 *Chemical Theory and Multiscale Simulation in Biomolecules: From Principles to Case Studies (1st Ed.)* (Elsevier)
- [125] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A 2021 *Language Models Enable Zero-shot Prediction of the Effects of Mutations on Protein Function (35th Conference on Neural Information Processing Systems (NeurIPS 2021))*
- [126] Wang D, Wang Y, Chang J, Zhang L, Wang H, E W 2021 *Nat. Comput. Sci.* **2** 20
- [127] Huang Y P, Xia Y, Yang L, Wei J, Yang Y I, Gao Y Q 2022 *Chin. J. Chem.* **40** 160

SPECIAL TOPIC—Machine learning in biomolecular simulations

Machine learning in molecular simulations of biomolecules*

Guan Xing-Yue¹⁾²⁾ Huang Heng-Yan¹⁾²⁾ Peng Hua-Qi¹⁾²⁾

Liu Yan-Hang¹⁾ Li Wen-Fei^{1)†} Wang Wei^{1)‡}

1) (*School of Physics, Nanjing University, Nanjing 210093, China*)

2) (*Wenzhou Key Laboratory of Biophysics, Wenzhou Institute, University of Chinese Academy of Sciences,
Wenzhou 325000, China*)

(Received 8 October 2023; revised manuscript received 1 November 2023)

Abstract

Molecular simulation has already become a powerful tool for studying life principles at a molecular level. The past 50-year researches show that molecular simulation has been able to quantitatively characterize the kinetic and thermodynamic properties of complex molecular processes, such as protein folding and conformational changes. In recent years, the application of machine learning algorithms represented by deep learning has further promoted the development of molecular simulation. This work reviews machine learning methods in biomolecular simulation, focusing on the important progress made by machine learning algorithms in improving the accuracy of molecular force fields, the efficiency of molecular simulation conformation sampling, and also the processing of high-dimensional simulation data. The future researches to further overcome the bottleneck of accuracy and efficiency of molecular simulation, expand the scope of molecular simulation, and realize the integration of computational simulation and experimental based on machine learning technique is prospected.

Keywords: bio-molecules, molecular simulations, machine learning, enhanced sampling, multiscale model

PACS: 87.15.ap, 87.15.Cc, 87.18.-h, 87.16.A-

DOI: [10.7498/aps.72.20231624](https://doi.org/10.7498/aps.72.20231624)

* Project supported by the National Natural Science Foundation of China (Grant No. 11974173).

† Corresponding author. E-mail: wfli@nju.edu.cn

‡ Corresponding author. E-mail: wangwei@nju.edu.cn

专题: 生物分子模拟中的机器学习

靶向 PD-L1 蛋白的计算机辅助药物筛选*

林开东¹⁾ 林晓倩¹⁾²⁾ 林绪波^{1)†}

1) (北京航空航天大学, 医学科学与工程学院/生物与医学工程学院, 北京市生物医学工程高精尖创新中心, 北京 100191)

2) (北京航空航天大学沈元学院, 北京 100191)

(2023 年 6 月 29 日收到; 2023 年 8 月 10 日收到修改稿)

针对 PD-1/PD-L1 免疫检查点的单克隆抗体抑制剂逐渐进入市场并在多种类型的肿瘤治疗中取得一定的积极效果. 然而, 随着应用范围的不断扩展, 抗体药物的局限性以及过多同质化研究等问题逐渐显现出来, 小分子化合物抑制剂成为了研究者们关注的新焦点. 本文旨在利用基于配体和基于结构的结合活性预测方法实现针对 PD-L1 靶点的小分子化合物虚拟筛选, 从而帮助加速小分子药物的开发. 通过从相关研究文献及专利收集 PD-L1 小分子抑制活性数据集, 根据不同分子表征方法和算法构建机器学习活性判定分类模型和活性强度预测回归模型, 两类模型从大型类药小分子库 (ZINC15) 中筛选获得 68 种高 PD-L1 抑制活性候选化合物. 其中 10 种化合物不仅具备良好的药物相似性和药代动力学, 还在分子对接中与已报道的热点化合物表现出同等水平的结合强度和相似的作用机制, 这一现象在后续分子动力学模拟和结合自由能估计中得到进一步验证. 本文提出了一个融合基于配体方法和基于结构方法的计算机辅助药物研发工作流程, 其在大型化合物数据库中有效筛选出有潜力的 PD-L1 小分子抑制剂, 有望助力加速肿瘤免疫治疗的应用.

关键词: PD-1/PD-L1, 虚拟筛选, 机器学习, 分子动力学模拟**PACS:** 05.10.-a, 02.70.-c**DOI:** 10.7498/aps.72.20231068

1 引言

阻断免疫检查点蛋白与其配体的结合作为肿瘤治疗的方法之一, 近年来在临床应用中迅速发展. 正常生理状态下, 部分负调节因子作为免疫检查点来抑制 T 细胞的过度激活, 确保免疫反应保持自我耐受^[1]. 然而不幸的是, 肿瘤细胞可以利用这种机制诱导 T 细胞衰竭, 形成促进肿瘤生长和侵袭的微环境, 从而逃避免疫系统的攻击^[2-4]. 为了重新激活并增强 T 细胞介导的抗肿瘤功能, 前人已经设计了一系列疗法来阻断免疫检查点蛋白与其配体的结合^[5,6]. 其中, 细胞程序性死亡蛋白 1 (programmed cell death protein 1, PD-1) 是最受关注的免疫检查点之一, 其通常表达于活化的

T 细胞、自然杀伤性细胞、B 淋巴细胞和其他免疫细胞的表面. PD-1 与其在肿瘤细胞上高度表达的配体 PD-L1 相互作用后, 发生一定的构象变化, 介导胞内信号通路从而抑制 T 细胞的增殖、活化和细胞杀伤性功能^[7-12]. 前人的研究已表明, PD-1 或 PD-L1 的基因敲除或抗体抑制可以增强小鼠免疫系统的抗肿瘤功能, 这表明阻断 PD-1 和 PD-L1 之间的相互作用可能为肿瘤免疫治疗提供一种有效的策略^[13,14].

PD-1/PD-L1 抗体抑制剂药物的研发目前取得非常显著的进展. 2014 年, 美国食品药品监督管理局批准了第一款 PD-1 抗体 Pembrolizumab 用于治疗晚期黑色素瘤之后, 一系列 PD-1/PD-L1 单克隆抗体被应用在了非小细胞肺癌、肝癌和食管胃交界癌等多种肿瘤疾病的临床治疗中^[15-17]. 然

* 国家自然科学基金 (批准号: 21903002) 和北京航空航天大学沈元学院卓越研究基金 (批准号: 230121202) 资助的课题.

† 通信作者. E-mail: linxbseu@buaa.edu.cn

而,随着应用的不断深入,抗体半衰期长、慢性免疫毒性、组织渗透有限、存储和运输成本高等难以避免的缺点逐渐暴露出来^[18-20].另一方面,抗体的同质化研究过多,造成了较大的资源浪费.小分子抑制剂具备较好的肿瘤组织渗透性、相对稳定的生物安全性和良好的口服利用性等优势,已成为 PD-1/PD-L1 抑制剂药物的下一个研发热点^[18,21-23].

百时美施贵宝 (Bristol-Myers Squibb, BMS) 于 2015 年公开了一系列非肽基联苯类小分子抑制剂,对 PD-1/PD-L1 结合具有强大的阻断抑制活性^[24]. Holak 团队^[25]曾报道, BMS 化合物通过与 PD-L1 结合诱导其发生二聚化,以间接的方式抑制 PD-1/PD-L1 相互作用.基于这一机制,其他的公司和学术团队开发了一系列不同骨架的衍生物,如 Incyte 公司的 INCB086550 (NCT04629339/NCT03762447)^[26]、红日药业的 IMM-010 (NCT04343859)^[27,28]、再极药业的 MAX-10181 (NCT04122339)、贝达药业的 BPI-371153 (NCT05341557)、歌礼药业的 ACS61 (NCT05287399) 以及和誉生物医药的 ABSK043 (NCT04964375) 等化合物目前已进入到了临床试验阶段.由于目前 PD-L1 小分子抑制剂的市场空白,寻找更多样骨架且具备良好用药性质的化合物仍具备重要意义.

数据驱动的计算方法已成为药物开发的重要工具, PD-1/PD-L1 小分子抑制剂亦不例外^[29].基于结构的方法,如药效团分析、分子对接和分子动力学 (molecular dynamics, MD) 模拟,在先前的研究中被广泛应用于靶向 PD-L1 二聚体的小分子抑制剂的虚拟筛选^[29-33].本文基于各种机器学习算法和分子描述符或指纹构建了一系列基于配体方法的分类和回归模型,以预测 ZINC15^[34] 中类药物化合物对 PD-1/PD-L1 相互作用的抑制活性.具有高预测活性的化合物将继续通过药物相似性、药代动力学筛选并以分子对接和 MD 模拟进行基于结构方法上的验证,以最终获得具备 PD-L1 抑制潜力的小分子化合物.

2 研究方法

2.1 数据获取及整理

本文用于训练及测试的数据集来源于与 PD-L1 小分子抑制剂相关的 37 篇研究性论文及 16 项专利 (见补充材料表 S1 ([online](#))), 为避免因实验技

术手段不同而导致化合物对 PD-L1 抑制活性测定数据水平不一致的情况,本文仅收录了以半抑制浓度 (half-maximal inhibitor concentration, IC_{50}) 为指标的均相时间分辨荧光 (homogeneous time-resolved fluorescence, HTRF) 的实验数据.

参考研究性论文及专利中的结构示意图,本文利用 ChemDraw 获取并检查化合物的简化分子线性输入规范 (simplified molecular input line entry system, SMILES) 字符串. IC_{50} 不高于 1 $\mu\text{mol/L}$ 的化合物定义为阳性样本 (即对 PD-L1 具备抑制活性), 而 IC_{50} 高于 10 $\mu\text{mol/L}$ 的化合物定义为阴性样本 (即对 PD-L1 不具备抑制活性), 为减小用于分类模型构建的两类样本数量不平衡问题, 一项细胞水平的高通量筛选 (high throughput screening, HTS) 实验记录 (PubChem BioAssay AID: 2316) 部分数据被引入补充阴性样本. 最后, 仅具备明确 IC_{50} 测定值而非测定范围的化合物被收录于回归模型的数据集中, IC_{50} 的对数转化值 pIC_{50} (即 $-\lg IC_{50}$) 作为样本标签.

2.2 数据集聚类

为了确保机器学习模型尽可能均匀地获得数据集中不同结构的信息, 本文对数据集中的化合物先进行了聚类处理. 首先, 2130 个分类模型阳性样本和 1099 个回归模型样本分别转化为 600 和 350 种仅保留环形结构和连接环形结构的最短路径的 Mureko 骨架^[35]. 分子骨架以 2 为最大半径, 2048 为向量长度转化为圆形扩展指纹 (extended-connectivity finger print, ECFP)^[36] 后, 以平均为链接算法、欧氏距离为计算方式对两类骨架的 ECFP 进行分层聚类, 簇的数量由 5—20 之间对应的最佳轮廓系数决定, 后续模型训练将从簇内分层抽样进行数据集划分 (图 1).

2.3 分子表征

由 ChemDraw 得到的化合物 SMILES 字符串分别转化为三类分子描述符 (RDKit, PaDEL 1D&2D, Mordred) 和三类分子指纹 (ECFP, MACCS, PubChem) 以作为化合物的特征向量. 开源 Java 软件 PaDEL-Descriptor v2.0 用于计算 PaDEL 1D&2D 描述符^[37]、MACCS 分子指纹^[38] 和 PubChem 分子指纹^[37]. RDKit 描述符和 ECFP 分子指纹^[36] 均由化学信息处理程序包 RDKit 转化,

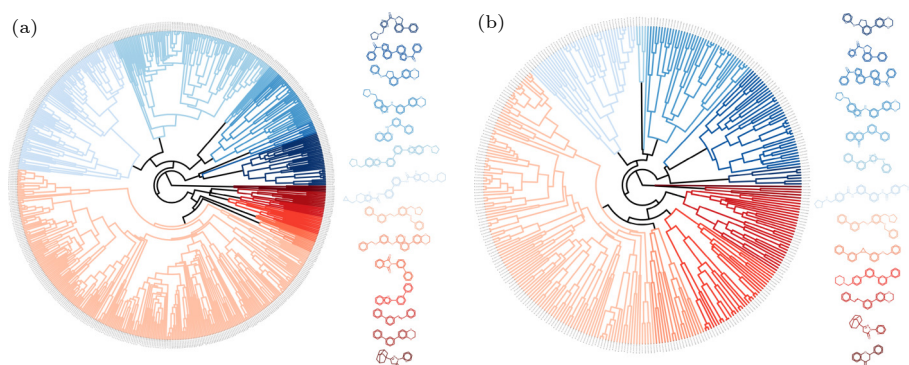


图1 数据集聚类 (a) 分类模型阳性样本骨架的 14 个簇; (b) 回归模型样本骨架的 13 个簇. 环形树状图的颜色代表不同的分子结构骨架簇, 颜色越接近, 骨架越相似, 一旁的化学结构图是每个簇中最具代表性的骨架

Fig. 1. Dataset clustering: (a) 14 clusters of scaffolds of active compounds in the classification models; (b) 13 clusters of scaffolds of compounds in the regression models. The colors of the circular dendrograms represent different clusters of scaffolds, and the closer the colors are, the more similar the scaffolds are. The chemical structure diagrams on the side are the most representative scaffolds of each cluster.

其中 ECFP 分子指纹为 2048 维向量, 指示最大以两个原子为半径的结构碎片存在与否. 此外, 描述符计算软件 Mordred^[39] 被用于最后一类特征向量的转化.

2.4 机器学习

分别使用五种机器学习算法构建 PD-L1 小分子抑制剂分类模型和抑制活性 IC_{50} 预测回归模型, 其中逻辑回归 (logistic regression, LR)、邻近算法 (K-nearest neighbor, KNN)、支持向量机 (support vector machine, SVM)、随机森林 (random forest, RF) 和多层感知机 (multilayer perceptron, MLP) 用于分类任务模型构建, SVM、岭回归 (ridge regression)、高斯过程回归 (Gaussian process regression, GPR)、RF 和 MLP 用于回归任务模型构建. 两类模型数据集的 80% 为训练数据集, 另外 20% 为测试数据集. 所有分子描述符及分子指纹的特征值删除空缺之后, 基于训练集进行 min-max

标准化处理. 各类算法的最佳超参数由基于 Scikit-learn 网格搜索方法的五重交叉验证 (5-fold cross validation) 确定, 此过程中, 分类模型以马修斯相关系数 (matthews correlation coefficient, MCC), 回归模型以预测值的平均绝对误差 (mean absolute error, MAE) 为超参数选择标准.

2.5 模型评价标准

分类模型性能由灵敏度 (sensitivity, SE)、特异度 (specificity, SP)、准确度 (accuracy, ACC)、马修斯相关系数 (Matthews correlation coefficient, MCC) 进行泛化评估, 其计算公式如下:

$$SE = \frac{TP}{TP + FN}, \quad (1)$$

$$SP = \frac{TN}{TN + FP}, \quad (2)$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}, \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \quad (4)$$

其中, TP(true positive) 为分类模型判定为具备抑制活性且实验结果亦为具备抑制活性的化合物样本数, FP(false positive) 为分类模型判定为具备抑制活性但实验结果为不具备抑制活性的化合物样本数, TN(true negative) 为分类模型判定为不具备抑制活性且实验结果亦为不具备抑制活性的化

合物样本数, FN(false negative) 为分类模型判定为不具备抑制活性但实验结果为具备抑制活性的化合物样本数.

回归模型性能由平均绝对误差 (mean absolute error, MAE)、均方根误差 (root mean-square error, RMSE) 和决定系数 (correlation coefficient, R^2)

进行泛化评估, 其计算公式如下:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_{\text{true}} - Y_{\text{pred}}|, \quad (5)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_{\text{true}} - Y_{\text{pred}})^2}, \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_{\text{true}} - Y_{\text{pred}})^2}{\sum_{i=1}^n (Y_{\text{true}} - \bar{Y}_{\text{true}})^2}, \quad (7)$$

其中, n 为回归模型待评估数据集样本数, Y_{true} 为化合物 pIC_{50} 实验测定值, Y_{pred} 为回归模型对化合物 pIC_{50} 的预测估计值.

2.6 药物相似性及 ADMET 检验

本文选取来自 ZINC15 数据库^[34] 的 7400926 个可商业购买 (截至 2023 年 2 月 24 日)、电中性且收录三维结构信息的类药小分子作为候选化合物筛选池, 尽管这些分子已通过 ZINC15 的药物相似性检验, 但由于本文所筛选的化合物针对蛋白质相互作用, 传统的药物相似性指标 (quantitative estimate of drug-likeness, QED) 已经不再适用^[40]. Kosugi 和 Ohue^[41] 提出了一种更加适合于针对蛋白质相互作用抑制剂筛选的药物相似性指标 (quantitative estimate index for compounds targeting protein-protein interactions, QEPPi), 化合物的 QEPPi 分数为一个介于 0—1 之间的值, 数值越大代表化合物的蛋白质相互作用抑制剂药物相似性越高, 本文以 0.7 为阈值筛选高类药性化合物.

ADMET 代表化合物的吸收 (absorption)、分配 (distribution)、代谢 (metabolism)、排泄 (excretion) 和毒性 (toxicity) 等重要用药性质, 在早期药物筛选中十分重要, 本文采用 ADMETlab 2.0 对化合物进行 ADMET 筛选^[42].

2.7 分子对接

基于结构的活性预测方法需要蛋白质靶点的晶体结构信息, 本文选取的 PD-L1 二聚体结构来源于蛋白质数据库 PDB (ID: 7DY7). 本文选择 PD-L1 二聚体中的结合界面区域为对接结合口袋, 基于如下理由: 1) 前期的研究表明 PD-L1 二聚体

中的结合界面区域是很有潜力的药物结合位点^[43,44]; 2) 分子对接结果显示小分子在该区域的结合亲和力最强.

分子对接利用 AutoDock Vina 进行^[45], AutoDockTools 将 PD-L1 二聚体转化为 pdbqt 格式, 对接网格盒子尺寸为 $40 \text{ \AA} \times 30 \text{ \AA} \times 40 \text{ \AA}$, 中心坐标为 $(144.799 \text{ \AA}, -9.364 \text{ \AA}, 16.163 \text{ \AA})$. 每个化合物各随机生成 50 种对接姿态和位置, 根据对接得分进行排名, 最佳得分前十位小分子对应的结合构象将用于分子动力学模拟的初始构象. 此外, 为验证通过上述流程得到的候选化合物作用机制是否与前人的研究结果相近, 本文利用 LigPlot+ 研究小分子与 PD-L1 二聚体的相互作用^[46].

2.8 分子动力学模拟

本文所有的分子动力学模拟采用 CHARMM 36 全原子力场^[47,48], 化合物配体的力场参数通过 CGenFF 程序获取^[47-50]. 十个高对接分数的候选化合物和两个对照化合物 (BMS202 以及 INCB086550) 与 PD-L1 二聚体复合系统搭建于 $10 \text{ nm} \times 10 \text{ nm} \times 10 \text{ nm}$ 盒子内, 每个盒子填充水分子并以 NaCl 中和体系电荷数, 整个搭建过程由 GROMACS^[51] 工具 gmx solvate 和 gmx genion 实现. 所有的分子动力学模拟工作均使用 GROMACS 2019.6 程序包运行, 模拟步长为 2.0 fs, 采用等温等压 (NPT) 系综, 模拟时长为 100 ns. 模拟盒在 x 轴、 y 轴和 z 轴 3 个方向上均设定了周期性边界条件. 采用半各向同性的 Parrinello-Rahman^[52] 方法将系统压强维持在 1 bar ($1 \text{ bar} = 10^5 \text{ Pa}$), 压缩系数设定为 4.5×10^{-5} , 弛豫时间为 5 ps. 此外, 系统采用 Nose-Hoover^[53,54] 控温方法将配体-蛋白质复合物和溶剂进行耦合, 温度维持在 310 K, 弛豫时间为 1 ps. 长距离静电相互作用使用 Particle-Mesh Ewald (PME)^[55] 方法计算, 短距离静电相互作用的截止距离设置为 1.2 nm, LINCS (LINear constraint solver)^[56] 算法用于约束含 H 原子的键长.

2.9 MM/PBSA 结合自由能计算

本文以 gmx_MMPBSA^[57] 为工具利用分子力学泊松-玻尔兹曼表面积法 (molecular mechanics Poisson-Boltzmann surface area, MM/PBSA) 来计算各个体系最后 10 ns 轨迹的平均结合自由能

(ΔG), 其计算公式如下:

$$\Delta G = G_{\text{complex}} - (G_{\text{protein}} + G_{\text{ligand}}), \quad (8)$$

其中 G_{complex} 为蛋白质-配体复合物的自由能, G_{protein} 和 G_{ligand} 分别为蛋白质和配体各自的自由能. 各组的自由能计算公式如下:

$$G = E_{\text{MM}} + G_{\text{sol}} - T\Delta S, \quad (9)$$

$$E_{\text{MM}} = E_{\text{vdw}} + E_{\text{ele}}, \quad (10)$$

$$G_{\text{sol}} = E_{\text{PB}} + E_{\text{SA}}, \quad (11)$$

其中 E_{MM} 为气相结合能量, 由范德瓦耳斯项 E_{vdw} 和静电项 E_{ele} 构成; G_{sol} 为溶剂化自由能, 由极性 E_{PB} 和非极性 E_{SA} 贡献构成; $T\Delta S$ 为熵变, 因其计算成本较高且未必能够提高自由能的精度, 本文暂不对其进行计算.

3 研究结果

3.1 分类模型性能

30 种分类模型分别基于 LR, KNN, SVM, RF 和 MLP 五种算法以及 RDKit, PaDEL 1D&2D, Mordred 分子描述符和 ECFP, PubChem, MACCS

分子指纹 6 种分子表征方式所建立. 图 2 为各个模型经五折验证并设置最佳超参数后在测试集上的性能表现, 其中以 ECFP 为输入的 KNN 模型表现出最佳性能, SE(阳性样本预测正确率) = 0.9937, SP(阴性样本预测正确率) = 0.9781, ACC(全部样本预测正确率) = 0.9859, MCC(综合评价两类样本预测性能指标) = 0.9720. 考虑到分子描述符或分子指纹的计算较为耗时, 同时获取 740 万余个类药小分子的全部六种输入向量的计算量更大, 因此, 仅采用以 ECFP 为输入用于后续的分类筛选. 另一方面, 相对于 LR 和 MLP 两种模型, KNN, SVM 和 RF 模型对于可能具备噪声数据的任务具有较好的鲁棒性, 因此, 选择 KNN, SVM 和 RF 模型作为分类筛选的模型; 当有两种或以上的模型支持小分子为活性时, 则认定该化合物对 PD-L1 具备抑制活性.

3.2 分类模型解释

为了进一步探索输入向量的特征与分类之间的相关性, 本文用活性和非活性化合物之间的 RDKit, PaDEL 和 Mordred 三种描述符来分析基于特征重要性而构建的 RF 模型中权重最高的

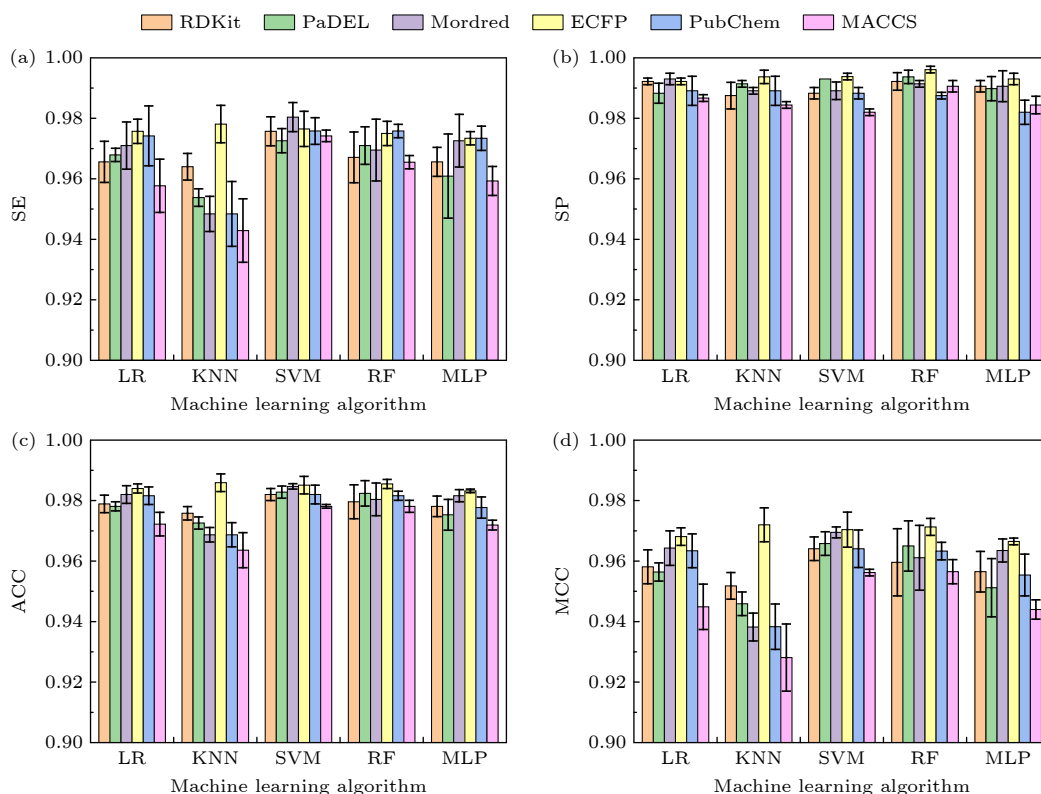


图 2 分类模型性能表现 (a) 灵敏度; (b) 特异度; (c) 准确度; (d) 马修斯相关系数

Fig. 2. Performance of binary models for classification tasks: (a) SE; (b) SP; (c) ACC; (d) MCC.

5个特征的分布差异(见补充材料图 S1 (online)). 尽管部分特征在两类样本间的分布差异是十分显著的,但由于分子描述符的特征信息难以解释,特征值根据理化性质及矩阵运算得出,其大小本身不具备具象含义,我们的认知只能停留在较浅薄的层面.

与描述符不同的是,分子指纹 ECFP 的位点指示局部亚结构的存在或不存,这可能揭示分子结构与对 PD-1/PD-L1 的抑制活性之间的关系. 分别统计了前十位活性化合物中比例显著高于非活性化合物的子结构和前十位活性化合物中比例显著低于非活性化合物的子结构(图 3). 带有黄色芳香性原子的环状结构在活性化合物中的计数远多于其在非活性化合物中的计数,这意味着芳香性结构片段可能对化合物的 PD-L1 抑制活性有正向贡献,相反,在非活性化合物中常出现的带灰色原

子的脂肪链片段则对活性没有正向贡献.

值得注意的是, ECFP985 和 ECFP1161 片段清楚地表示了化合物的联苯特性,这也是 BMS 化合物的核心结构特征之一 [24,58]. 此外,苯甲基 (ECFP253)、吡啶 (ECFP1453) 和与苯环相连的醚键 (ECFP1971) 通常存在于多数表现出 PD-L1 抑制活性的小分子结构中. 前人的分子对接分析表明,苯甲基能够与 PD-L1 的 Ala121 和 Met115 产生强烈的疏水相互作用 [59]. Lu 等 [60] 发现 PD-L1-BMS202 复合物中,吡啶环的氮原子周围有很大的空间,因此他们在该位点附加了一系列的取代基,以提高配体的结合亲和力. 由于该片段的重要性,许多其他研究团队也将工作重点放在吡啶的修饰上 [59,61,62]. 除了环状结构,连接环状结构的路径可能也是值得关注的组成部分,其中,以醚键连接六元环或五元环的活性化合物约占 50% [43,63-65].

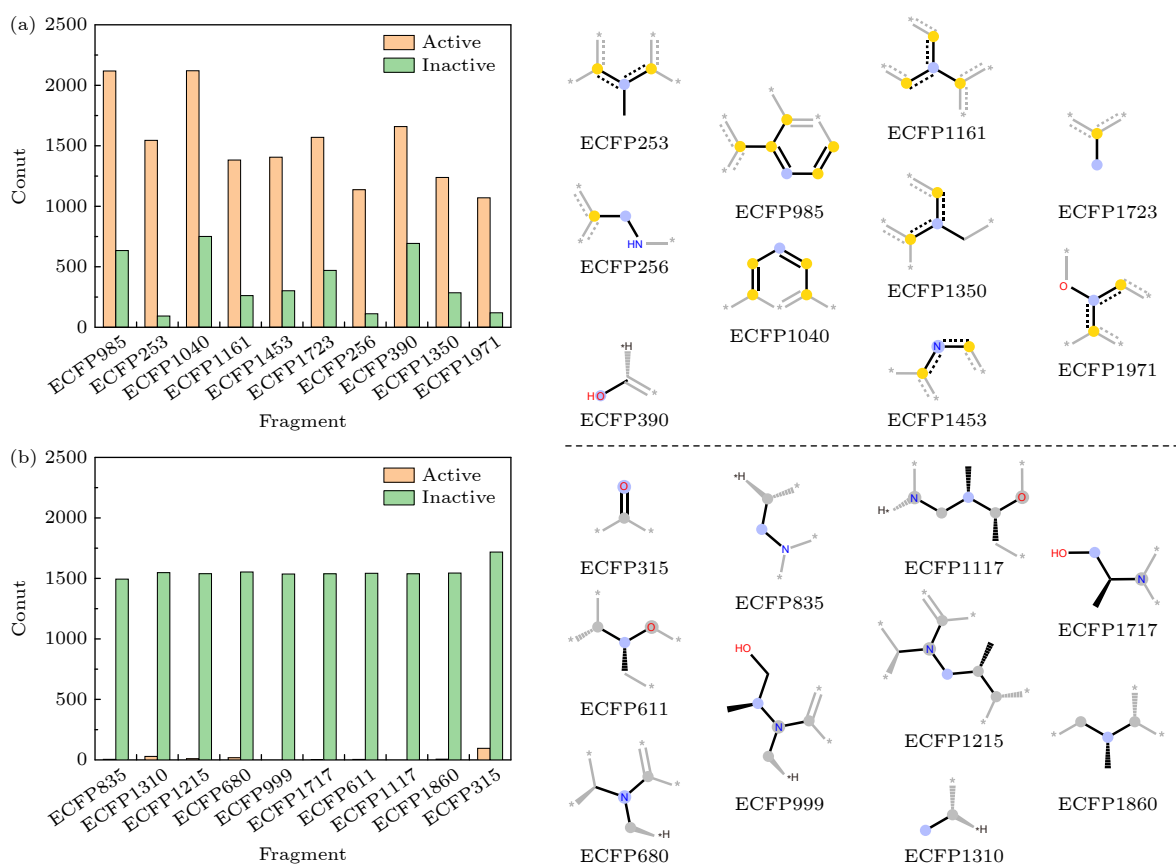


图 3 分子结构片段在两类样本间的计数差异 (a) 活性化合物计数占优的结构片段; (b) 非活性化合物计数占优的结构片段. 结构片段的中心原子以蓝色突出显示; 芳香性原子被着色为黄色, 而脂肪烃链原子则被着色为灰色

Fig. 3. Count difference of fragments of structures between active and inactive compounds: (a) Fragments in active compounds with a proportion higher than inactive compounds; (b) fragments in active compounds with a proportion lower than inactive compounds. The center atoms of substructure fragments are highlighted in blue; aromaticity atoms are colored in yellow and aliphatic hydrocarbon atoms are colored in gray.

3.3 回归模型性能

30种回归模型分别基于SVM, Ridge Regression, GPR, RF和MLP五种算法以及RDKit, PaDEL 1D&2D, Mordred分子描述符和 ECFP, PubChem, MACCS分子指纹6种分子表征方式所建立. 图4为各个模型经五折验证后、最佳超参数设置下在测试集上的性能表现, 其中以 ECFP为输入的 SVM模型以 $MAE = 0.4503$, $RMSE =$

0.6375 , GPR模型以 $MAE = 0.4557$, $RMSE = 0.6375$ 的性能表现显著优于其他模型. 图5展示了两种模型在训练集及测试集所有样本点的预测结果及偏差, 绝大部分的样本点预测偏差在 ± 1 范围内, 以 ECFP为输入的 SVM模型在测试上的决定系数 $R^2 = 0.782$, 以 ECFP为输入的 GPR模型在测试上的决定系数 $R^2 = 0.787$, 两类模型预测值的平均值将作为候选化合物的 PD-L1 抑制活性预测值.

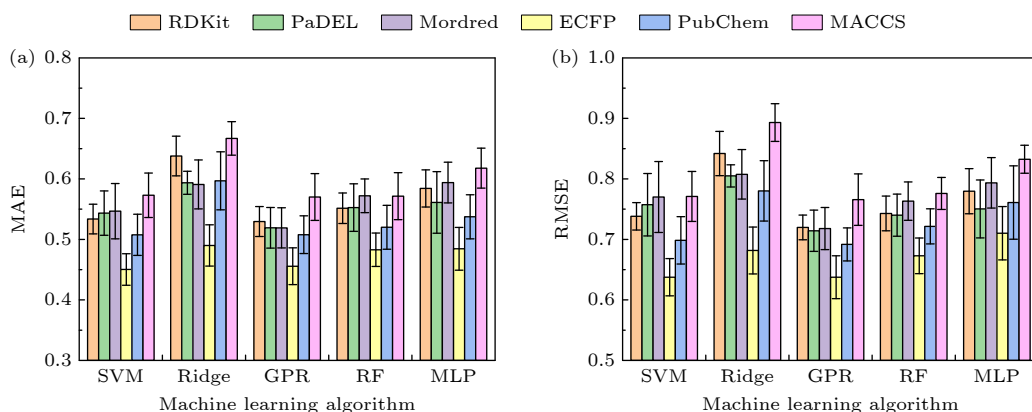


图4 回归模型性能表现 (a) 平均绝对误差; (b) 均方根误差

Fig. 4. Performance of continuous models for regression tasks: (a) MAE; (b) RMSE.

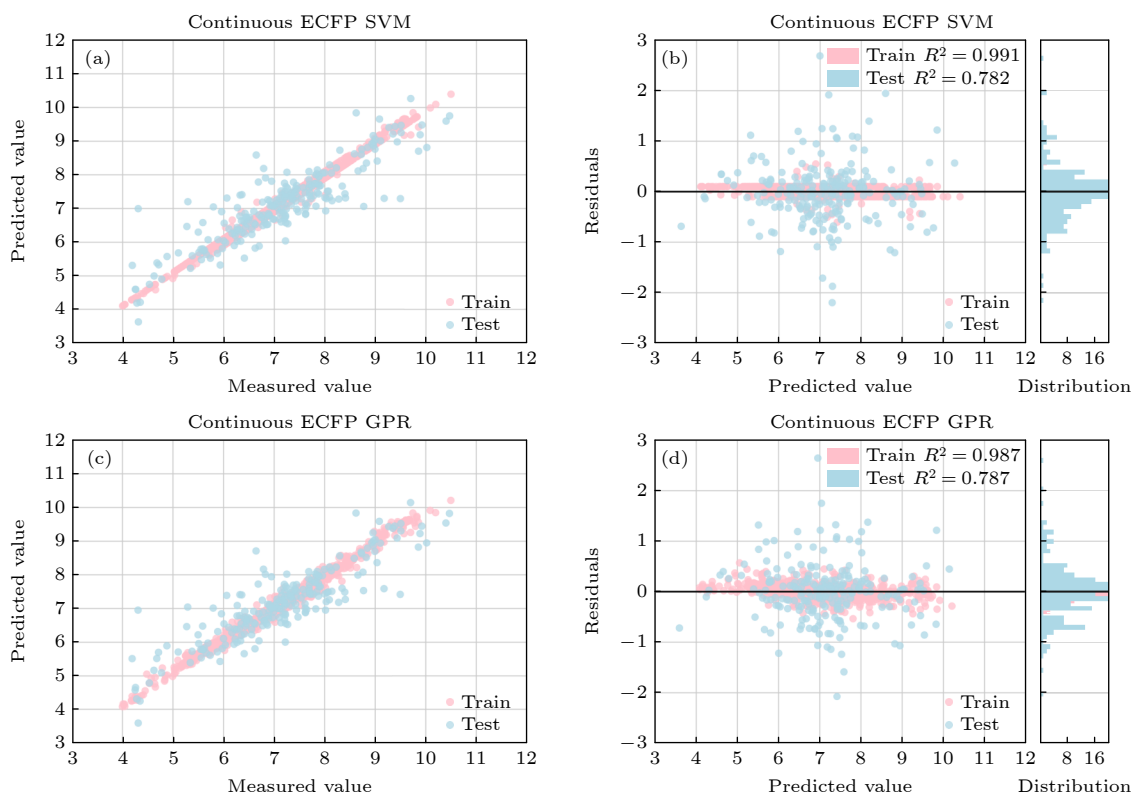


图5 两种最佳性能回归模型预测结果 (a), (c) 样本预测值与标签值分布; (b), (d) 样本预测偏差

Fig. 5. Prediction results of two best-performing continuous models: (a), (c) Distribution of predicted values and label values; (b), (d) prediction residuals.

3.4 虚拟筛选

ZINC15 数据库中电中性且收录三维结构信息的类药物小分子构成本文的化合物筛选池 (7400926 个小分子). 整个虚拟筛选过程包含 4 个环节: 1) 同时使用以 ECFP 为输入的 KNN, SVM 和 RF 等 3 种分类模型, 其中至少有 2 种判定结果为对 PD-L1 蛋白具有抑制活性, 则认定该分子具有抑制活性. 2) 同时使用以 ECFP 为输入的 SVM 和 GPR 回归模型预测 pIC₅₀ 值, 两模型 pIC₅₀ 的平均值小于 7 (IC₅₀ > 100 nmol/L) 的化合物将被剔除. 3) 计算小分子化合物的 QEPPPI 分数 (见补充材料表 S2 (online)), 并保留 QEPPPI 得分超过 0.7 的候选化合物. 尽管初始筛选池的化合物在 ZINC15 中根据 Lipinski 五原则^[66] 被定义为类药化合物, 但其中部分化合物因分子量过小可能难以充分结合到 PD-L1 二聚体的相互作用界面, 因此, 重新评估他们的药物相似性仍具有重要意义.

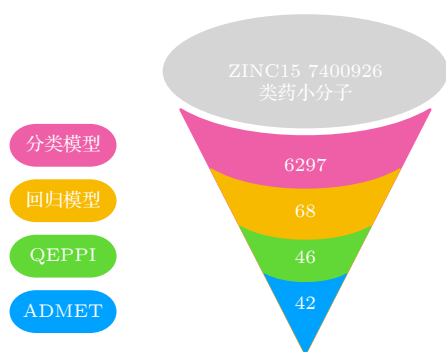


图 6 虚拟筛选流程

Fig. 6. Workflow of virtual screening.

义. 4) 使用 ADMETlab 2.0^[42] 计算小分子的 ADMET 性质, 以进一步筛选具有应用潜能的小分子 (见补充材料表 S3 (online)). 最终, 通过整个虚拟筛选流程, 42 个候选化合物被认为对 PD-L1 具有较强的抑制活性并具备良好的药用性能 (图 6).

3.5 分子对接

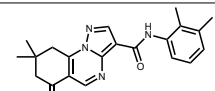
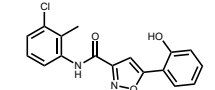
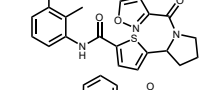
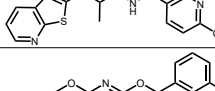
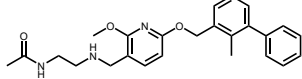
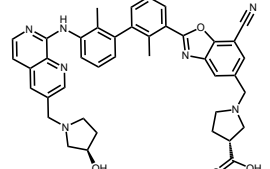
为了验证筛选结果并通过基于结构的方法进一步识别有抑制潜力的化合物, 利用 AutoDock Vina 将 42 个候选化合物以及 BMS202 和 INCB-086550 对接到 PD-L1 二聚体的 IgV 结构域 (PDB ID:7DY7). 具有最低对接评分的 10 种化合物被认为有潜力的 PD-L1 抑制剂并继续用于之后的分析 (表 1), 其中值得注意的是, 10 种候选化合物对接分数均低于临床 II 期试验化合物 INCB086550 的对接分数, 因此可以认为先前构建的分类和回归模型是筛选 PD-L1 小分子抑制剂的有效工具. 分析 12 种化合物与 PD-L1 二聚体的相互作用可知 (见补充材料图 S2 (online)), TYR56(A), TYR56(B), MET115(A), MET115(B), ALA121(A) 和 TYR123(A) (括弧中 A 或 B 分别表示 PD-L1 二聚体中的单体 A 或 B) 等氨基酸残基与所有配体均发生了较强的疏水相互作用, 这意味着这些小分子的结合位置和结合模式可能是近似的. 此外, 除了 ZINC000021874692 和 ZINC000021874694 这一对化合物为旋光异构体, 其余化合物的结构差异较大, 相似度较低, 意味着其可为未来的湿实验筛选提供较高的容错空间 (见补充材料图 S3 (online)).

表 1 最低对接评分的 10 种候选化合物及 BMS202 和 INCB086550 的对接结果

Table 1. Docking results for the top 10 hits with BMS202 and INCB086550.

序号	化合物	化学结构式	对接分数/(kcal·mol ⁻¹)
Hit 1	ZINC000021723762		-11.8
Hit 2	ZINC000021874692		-10.7
Hit 3	ZINC000175468610		-10.7
Hit 4	ZINC000019770413		-10.6
Hit 5	ZINC000021874694		-10.6
Hit 6	ZINC000952973550		-10.5

表 1 (续) 最低对接评分的 10 种候选化合物及 BMS202 和 INCB086550 的对接结果
 Table 1 (continued). Docking results for the top 10 hits with BMS202 and INCB086550.

序号	化合物	化学结构式	对接分数/(kcal·mol ⁻¹)
Hit 7	ZINC000064987401		-10.5
Hit 8	ZINC000003908573		-10.4
Hit 9	ZINC000020538424		-10.4
Hit 10	ZINC000004063088		-10.3
Ctr 1	BMS202		-11.1
Ctr 2	INCB086550		-10.0

3.6 分子动力学模拟

为验证配体与 PD-L1 二聚体结合的稳定性, 计算了 12 种配体-蛋白质复合物 100 ns 轨迹中蛋白质骨架和配体小分子构象基于蛋白质骨架叠合的均方根偏差 (root mean square deviation, RMSD) (图 7). 所有体系的 PD-L1 二聚体骨架和配体的 RMSD 在 50 ns 后均稳定在 0.8 nm 以下, 模拟轨迹达到平衡, 蛋白质与配体结合稳定. 值得注意的是, 以 INCB086550 为例的分子量较大的小分子与以 ZINC000019770413 为例的分子量较小的小分子

相比, 配体构象 RMSD 波动较大. 由图 8 的蛋白质-配体结合模式可知, 起到关键作用的结构片段集中在配体的一端, 而另一端暴露于结合口袋之外. 若配体的分子量较大, 分子骨架较长, 其在结合区域外的部分更多, 这部分运动更为自由, 这可能是部分小分子构象 RMSD 波动相对较大的原因.

3.7 MM/PBSA 结合自由能计算

分子动力学模拟的最后 10 ns 轨迹被用于 MM/PBSA 计算以评估配体小分子与 PD-L1 二聚

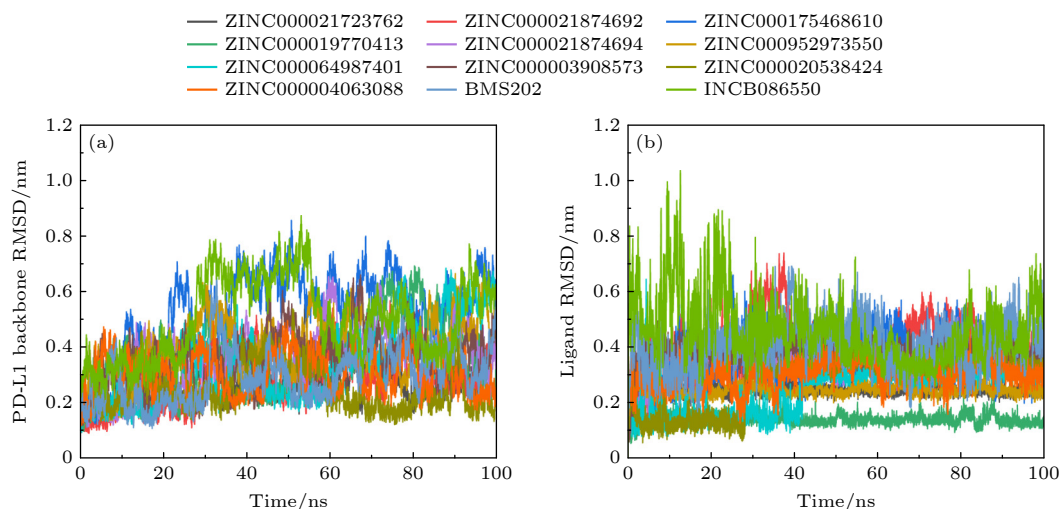


图 7 RMSD (a) PD-L1 二聚体骨架; (b) 配体
 Fig. 7. RMSD: (a) Backbone of PD-L1 dimer; (b) ligand.

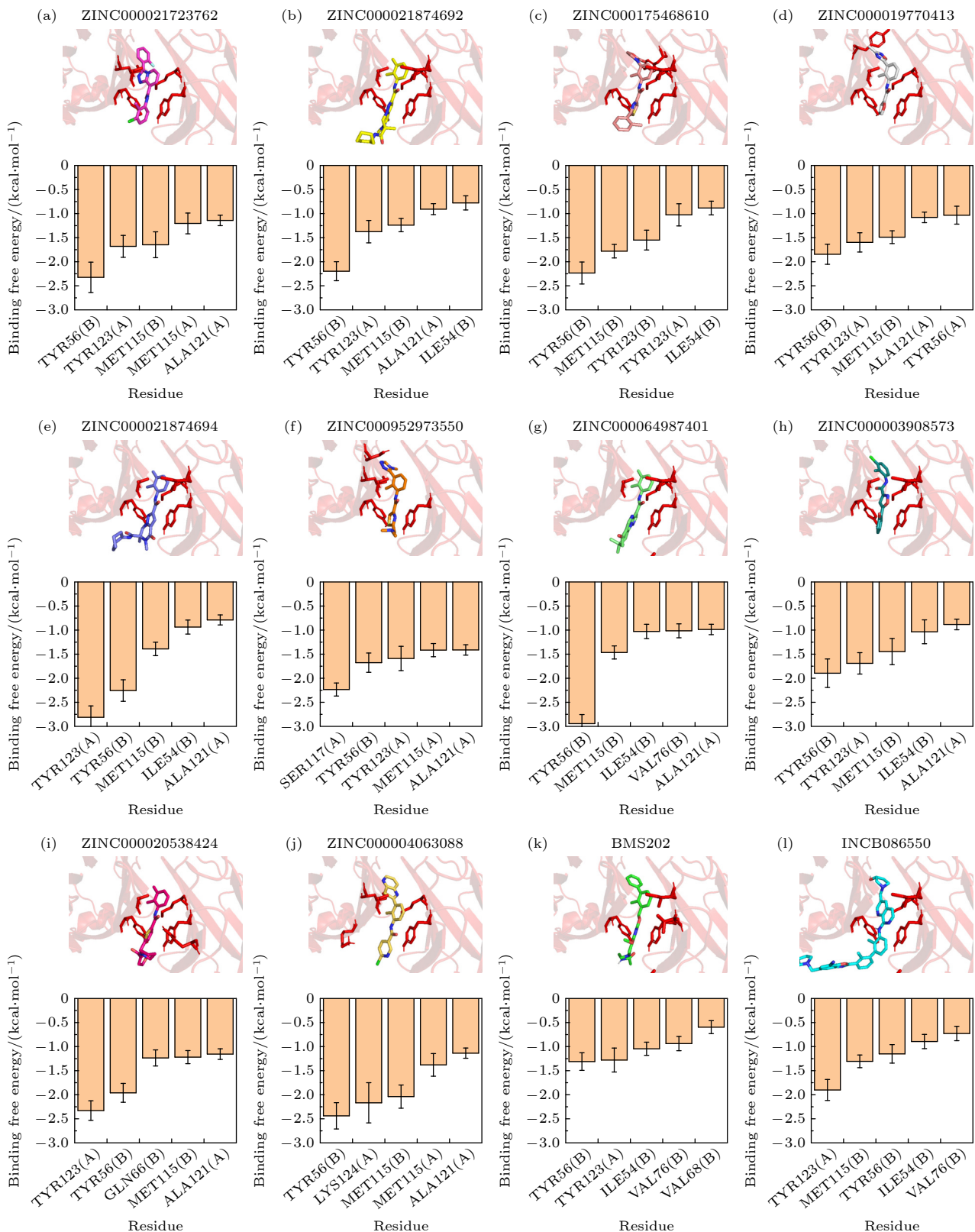


图 8 配体与 PD-L1 结合模式和关键残基 (a) ZINC000021723762; (b) ZINC000021874692; (c) ZINC000175468610; (d) ZINC000019770413; (e) ZINC000021874694; (f) ZINC0000952973550; (g) ZINC000064987401; (h) ZINC000003908573; (i) ZINC000020538424; (j) ZINC000004063088; (k) BMS202; (l) INCB086550

Fig. 8. Ligand binding mode with PD-L1 and key residues: (a) ZINC000021723762; (b) ZINC000021874692; (c) ZINC000175468610; (d) ZINC000019770413; (e) ZINC000021874694; (f) ZINC0000952973550; (g) ZINC000064987401; (h) ZINC000003908573; (i) ZINC000020538424; (j) ZINC000004063088; (k) BMS202; (l) NCB086550.

表 2 结合自由能计算结果
Table 2. Results of MMPBSA.

化合物	$E_{vdw}/(\text{kcal}\cdot\text{mol}^{-1})$	$E_{ele}/(\text{kcal}\cdot\text{mol}^{-1})$	$E_{PB}/(\text{kcal}\cdot\text{mol}^{-1})$	$E_{SA}/(\text{kcal}\cdot\text{mol}^{-1})$	$\Delta G/(\text{kcal}\cdot\text{mol}^{-1})$
ZINC000021723762	-58.86 ± 0.12	-8.76 ± 0.10	35.53 ± 0.09	-4.03 ± 0.00	-36.12 ± 0.12
ZINC000021874692	-50.42 ± 0.11	-6.50 ± 0.14	35.65 ± 0.16	-4.50 ± 0.01	-25.78 ± 0.12
ZINC000175468610	-42.46 ± 0.09	-14.34 ± 0.08	36.20 ± 0.13	-4.02 ± 0.00	-24.61 ± 0.10
ZINC000019770413	-47.31 ± 0.08	-3.45 ± 0.08	26.32 ± 0.07	-3.48 ± 0.00	-27.91 ± 0.08
ZINC000021874694	-55.17 ± 0.09	-17.87 ± 0.13	45.67 ± 0.14	-4.59 ± 0.00	-31.97 ± 0.12
ZINC000952973550	-55.82 ± 0.08	-13.47 ± 0.11	45.83 ± 0.11	-4.19 ± 0.00	-27.65 ± 0.11
ZINC000064987401	-48.50 ± 0.08	-14.71 ± 0.12	38.59 ± 0.11	-4.06 ± 0.00	-28.68 ± 0.12
ZINC000003908573	-44.75 ± 0.07	4.59 ± 0.11	20.20 ± 0.13	-3.74 ± 0.00	-23.70 ± 0.12
ZINC000020538424	-49.50 ± 0.09	-8.50 ± 0.13	32.61 ± 0.16	-4.17 ± 0.01	-29.57 ± 0.10
ZINC000004063088	-64.14 ± 0.09	-5.38 ± 0.11	36.75 ± 0.10	-4.21 ± 0.00	-36.98 ± 0.10
BMS202	-40.23 ± 0.08	-3.57 ± 0.07	22.35 ± 0.10	-4.31 ± 0.01	-25.75 ± 0.09
INCB086550	-50.65 ± 0.14	-9.87 ± 0.17	44.44 ± 0.27	-5.13 ± 0.02	-21.22 ± 0.15

体间的结合作用强度. 由表 2 可知, 大部分候选化合物的结合自由能均与对照组的结合自由能处于同一水平, 而 ZINC000021723762, ZINC000021874694 和 ZINC000004063088 甚至显著优于对照组水平, 范德瓦耳斯相互作用为驱动这三类化合物区别于其他化合物的主要因素.

将结合自由能分解至每个氨基酸残基与配体分子的相互作用上, 提取了每个体系中对结合强度贡献度最大的 5 个关键残基. 如图 8 所示, TYR56(B), TYR123(A), MET115(B), ALA121(A) 在绝大部分体系中都起到关键作用, 这意味着 12 种配体结合于 PD-L1 二聚体相互作用界面的模式是近似的. 前人研究揭示出, MET115(B) 和 ALA121(A) 与化合物中的芳香环能够发生强烈疏水相互作用, 而 TYR56(B) 能够与化合物的苯环形成 π - π 堆积以增强结合稳定性^[32,33,59,67], 这些作用在结合自由能贡献中均得到了体现.

4 结论

通过从相关研究文献及专利收集 PD-L1 小分子抑制活性数据集, 并对数据集内化合物以结构相似性进行分层抽样, 本文构建了各 30 种基于不同分子表征方法和算法的机器学习活性判定分类模型和活性强度回归预测模型, 其中性能最佳分类模型分类正确率可达 98% 以上, 回归模型预测平均绝对误差在 0.5 以下, 具备良好的应用价值. 将以上模型并结合药用性质筛选工具构成的虚拟筛选方法应用于 ZINC15 大型小分子数据库, 获得了 42

种有潜力的 PD-L1 小分子抑制剂, 其中的 10 种通过分子对接、分子动力学模拟和结合自由能估计验证发现, 候选化合物的作用机制与对照化合物相近, 部分化合物的结合强度甚至优于对照化合物, 这意味着本文前期建立的机器学习模型是可以帮助加速 PD-L1 小分子抑制剂虚拟筛选的有效工具.

然而, 完整的药物研发并不是一个简单的工程, 本文仅使用计算方法帮助加速 PD-L1 小分子抑制剂的研发, 在完整的研发产业链中处于较上游的位置. 为尽可能使得本文的筛选结果得到更进一步的有效认证, 后续分子水平、细胞水平以及动物模型水平的湿实验验证仍是必不可少的, 计算机技术目前仅能扮演锦上添花的角色. 相信随着生物计算领域与生物技术领域的合作加深, 药物研发将能够向一个尽可能高效、低成本的方向发展.

参考文献

- [1] Waldman A D, Fritz J M, Lenardo M J 2020 *Nat. Rev. Immunol.* **20** 651
- [2] Wherry E J, Kurachi M 2015 *Nat. Rev. Immunol.* **15** 486
- [3] Zou W, Wolchok J D, Chen L 2016 *Sci. Transl. Med.* **8** 328rv4
- [4] Jiang X, Wang J, Deng X, Xiong F, Ge J, Xiang B, Wu X, Ma J, Zhou M, Li X, Li Y, Li G, Xiong W, Guo C, Zeng Z 2019 *Mol. Cancer* **18** 10
- [5] Topalian S L, Drake C G, Pardoll D M 2015 *Cancer Cell* **27** 450
- [6] Postow M A, Callahan M K, Wolchok J D 2015 *J. Clin. Oncol.* **33** 1974
- [7] Tang Q, Chen Y, Li X, Long S, Shi Y, Yu Y, Wu W, Han L, Wang S 2022 *Front. Immunol.* **13** 964442
- [8] Ai L, Xu A, Xu J 2020 *Adv. Exp. Med. Biol.* **1248** 33
- [9] Dermani F K, Samadi P, Rahmani G, Kohlan A K, Najafi R

- 2019 *J. Cell. Physiol.* **234** 1313
- [10] Parry R V, Chemnitz J M, Frauwirth K A, Lanfranco A R, Braunstein I, Kobayashi S V, Linsley P S, Thompson C B, Riley J L 2005 *Mol. Cell. Biol.* **25** 9543
- [11] Patsoukis N, Brown J, Petkova V, Liu F, Li L, Boussiotis V A 2012 *Sci. Signal.* **5** ra46
- [12] Hui E, Cheung J, Zhu J, Su X, Taylor M J, Wallweber H A, Sasmal D K, Huang J, Kim J M, Mellman I, Vale R D 2017 *Science* **355** 1428
- [13] Iwai Y, Ishida M, Tanaka Y, Okazaki T, Honjo T, Minato N 2002 *Proc. Natl. Acad. Sci. U.S.A.* **99** 12293
- [14] Iwai Y, Terawaki S, Honjo T 2005 *Int. Immunol.* **17** 133
- [15] Gong J, Chehrizi-Raffle A, Reddi S, Salgia R 2018 *J. Immunother. Cancer* **6** 8
- [16] Akinleye A, Rasool Z 2019 *J. Hematol. Oncol.* **12** 92
- [17] Shiravand Y, Khodadadi F, Kashani S M A, Hosseini-Fard S R, Hosseini S, Sadeghirad H, Ladwa R, O'Byrne K, Kulasingham A 2022 *Curr. Oncol.* **29** 3044
- [18] Zhang J, Zhang Y, Qu B, Yang H, Hu S, Dong X 2021 *Eur. J. Med. Chem.* **218** 113356
- [19] Johnson D B, Nebhan C A, Moslehi J J, Balko J M 2022 *Nat. Rev. Clin. Oncol.* **19** 254
- [20] Mould D R, Meibohm B 2016 *BioDrugs* **30** 275
- [21] Wu Q, Jiang L, Li S C, He Q J, Yang B, Cao J 2021 *Acta Pharmacol. Sin.* **42** 1
- [22] Zhan M M, Hu X Q, Liu X X, Ruan B F, Xu J, Liao C 2016 *Drug Discov. Today* **21** 1027
- [23] Liu C, Seeram N P, Ma H 2021 *Cancer Cell Int.* **21** 239
- [24] Chupak L S, Zheng X 2015 WO Patent 2015034820 A1
- [25] Zak K M, Grudnik P, Guzik K, Zieba B J, Musielak B, Dömling A, Dubin G, Holak T A 2016 *Oncotarget* **7** 30323
- [26] Koblisch H K, Wu L, Wang L S, Liu P C C, Wynn R, Rios-Doria J, Spitz S, Liu H, Volgina A, Zolotarjova N, Kapilashrami K, Behshad E, Covington M, Yang Y O, Li J, Diamond S, Soloviev M, O'Hayer K, Rubin S, Kanellopoulou C, Yang G, Rupar M, DiMatteo D, Lin L, Stevens C, Zhang Y, Thekkat P, Geschwindt R, Marando C, Yeleswaram S, Jackson J, Scherle P, Huber R, Yao W, Hollis G 2022 *Cancer Discov.* **12** 1482
- [27] Jiang J, Zou X, Liu Y, Liu X, Dong K, Yao X, Feng Z, Chen X, Sheng L, Li Y 2021 *Front. Pharmacol.* **12** 677120
- [28] Wang Y, Liu X, Zou X, Wang S, Luo L, Liu Y, Dong K, Yao X, Li Y, Chen X, Sheng L 2021 *Pharmaceutics* **13** 598
- [29] Sobral P S, Luz V C C, Almeida J, Videira P A, Pereira F 2023 *Int. J. Mol. Sci.* **24** 5908
- [30] Luo L, Zhong A, Wang Q, Zheng T 2021 *Mar. Drugs* **20** 29
- [31] Acúrcio R C, Pozzi S, Carreira B, Pojo M, Gómez-Cebrián N, Casimiro S, Fernandes A, Barateiro A, Farricha V, Brito J, Leandro A P, Salvador J A R, Graça L, Puchades-Carrasco L, Costa L, Satchi-Fainaro R, Guedes R C, Florindo H F 2022 *J. Immunother. Cancer* **10** e004695
- [32] Lung J, Hung M S, Lin Y C, Hung C H, Chen C C, Lee K D, Tsai Y H 2020 *Molecules* **25** 5293
- [33] Guo Y, Liang J, Liu B, Jin Y 2021 *Int. J. Mol. Sci.* **22** 10924
- [34] Sterling T, Irwin J J 2015 *J. Chem. Inf. Model.* **55** 2324
- [35] Bemis G W, Murcko M A 1996 *J. Med. Chem.* **39** 2887
- [36] Rogers D, Hahn M 2010 *J. Chem. Inf. Model.* **50** 742
- [37] Yap C W 2011 *J. Comput. Chem.* **32** 1466
- [38] Durant J L, Leland B A, Henry D R, Nourse J G 2002 *J. Chem. Inf. Comput. Sci.* **42** 1273
- [39] Moriwaki H, Tian Y S, Kawashita N, Takagi T 2018 *J. Cheminform.* **10** 4
- [40] Bickerton G R, Paolini G V, Besnard J, Muresan S, Hopkins A L 2012 *Nat. Chem.* **4** 90
- [41] Kosugi T, Ohue M 2021 *Int. J. Mol. Sci.* **22** 10925
- [42] Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, Yin M, Zeng X, Wu C, Lu A, Chen X, Hou T, Cao D 2021 *Nucleic Acids Res.* **49** W5
- [43] Jing T, Zhang Z, Kang Z, Mo J, Yue X, Lin Z, Fu X, Liu C, Ma H, Zhang X, Hu W 2023 *J. Med. Chem.* **66** 6811
- [44] Qin M, Meng Y, Yang H, Liu L, Zhang H, Wang S, Liu C, Wu X, Wu D, Tian Y, Hou Y, Zhao Y, Liu Y, Xu C, Wang L 2021 *J. Med. Chem.* **64** 5519
- [45] Trott O, Olson A J 2010 *J. Comput. Chem.* **31** 455
- [46] Laskowski R A, Swindells M B 2011 *J. Chem. Inf. Model.* **51** 2778
- [47] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell Jr. A D 2010 *J. Comput. Chem.* **31** 671
- [48] Yu W, He X, Vanommeslaeghe K, MacKerell Jr. A D 2012 *J. Comput. Chem.* **33** 2451
- [49] Vanommeslaeghe K, MacKerell Jr. A D 2012 *J. Chem. Inf. Model.* **52** 3144
- [50] Vanommeslaeghe K, Raman E P, MacKerell Jr. A D 2012 *J. Chem. Inf. Model.* **52** 3155
- [51] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B, Lindahl E 2015 *SoftwareX* **1** 19
- [52] Parrinello M, Rahman A 1981 *J. Appl. Phys.* **52** 7182
- [53] Nosé S 1984 *Mol. Phys.* **52** 255
- [54] Hoover W G 1985 *Phys. Rev. A* **31** 1695
- [55] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H, Pedersen L G 1995 *J. Chem. Phys.* **103** 8577
- [56] Hess B, Bekker H, Berendsen H J C, Fraaije J G E M 1997 *J. Comput. Chem.* **18** 1463
- [57] Valdés-Tresanco M S, Valdés-Tresanco M E, Valiente P A, Moreno E 2021 *J. Chem. Theory Comput.* **17** 6281
- [58] Chupak L S, Ding M, Martin S W, Zheng X, Hewawasam P, Connolly T P, Xu N, Yeung K S, Zhu J, Langley D R, Tenney D J, Scola P M 2015 WO Patent 2015160641 A3
- [59] Wang T, Cai S, Wang M, Zhang W, Zhang K, Chen D, Li Z, Jiang S 2021 *J. Med. Chem.* **64** 7390
- [60] Lu L, Qi Z, Wang T, Zhang X, Zhang K, Wang K, Cheng Y, Xiao Y, Li Z, Jiang S 2022 *ACS Med. Chem. Lett.* **13** 586
- [61] Dai X, Wang K, Chen H, Huang X, Feng Z 2021 *Bioorg. Chem.* **114** 105034
- [62] Le Biannic R, Magnez R, Klupsch F, Leleu-Chavain N, Thiroux B, Tardy M, El Bouazzati H, Dezitter X, Renault N, Vergoten G, Bailly C, Quesnel B, Thuru X, Millet R 2022 *Eur. J. Med. Chem.* **236** 114343
- [63] Russomanno P, Assoni G, Amato J, D'Amore V M, Scaglia R, Brancaccio D, Pedrini M, Polcaro G, La Pietra V, Orlando P, Falzoni M, Cerofolini L, Giuntini S, Fragai M, Pagano B, Donati G, Novellino E, Quintavalle C, Condorelli G, Sabbatino F, Seneci P, Arosio D, Pepe S, Marinelli L 2021 *J. Med. Chem.* **64** 16020
- [64] Liu L, Yao Z, Wang S, Xie T, Wu G, Zhang H, Zhang P, Wu Y, Yuan H, Sun H 2021 *J. Med. Chem.* **64** 8391
- [65] Cheng B, Ren Y, Cao H, Chen J 2020 *Eur. J. Med. Chem.* **199** 112377
- [66] Lipinski C A 2004 *Drug Discov. Today Technol.* **1** 337
- [67] Liang J, Wang B, Yang Y, Liu B, Jin Y 2023 *Int. J. Mol. Sci.* **24** 1280

SPECIAL TOPIC—Machine learning in biomolecular simulations

Virtual screening of drugs targeting PD-L1 protein*Lin Kai-Dong¹⁾ Lin Xiao-Qian¹⁾²⁾ Lin Xu-Bo^{1)†}

1) (*Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, School of Engineering Medicine, Beihang University, Beijing 100191, China*)

2) (*Shen Yuan Honors College, Beihang University, Beijing 100191, China*)

(Received 29 June 2023; revised manuscript received 10 August 2023)

Abstract

Monoclonal antibody inhibitors targeting PD-1/PD-L1 immune checkpoints are gradually entering the market and have achieved certain positive effects in the treatments of various types of tumors. However, with the expansion of application, the limitations of antibody drugs and problems such as excessive homogenization of research gradually appear, making small-molecule inhibitors the new focus of researchers. This study aims to use ligand-based and structure-based binding activity prediction methods to achieve virtual screening of small-molecule inhibitors targeting PD-L1, thereby helping to accelerate the development of small molecule drugs. A dataset of PD-L1 small-molecule inhibitory activity from relevant research literature and patents is collected and activity judgment classification models with intensity prediction regression models are constructed based on different molecular featurization methods and machine learning algorithms. The two types of models filter 68 candidate compounds with high PD-L1 inhibitory activity from a large drug-like small molecule screening pool (ZINC15). Ten of these compounds not only have good drug similarities and pharmacokinetics, but also exhibit comparable binding affinities and similar mechanisms of action with previous reported hotspot compounds in molecular docking. This phenomenon is further verified in subsequent molecular dynamics simulation and the estimation of binding free energy. In this study, a virtual screening workflow integrating ligand-based method and structure-based method is developed, and potential PD-L1 small-molecule inhibitors are effectively screened from large compound databases, which is expected to help accelerate the application and expansion of tumor immunotherapy.

Keywords: PD-1/PD-L1, virtual screening, machine learning, molecular dynamics simulation

PACS: 05.10.-a, 02.70.-c

DOI: 10.7498/aps.72.20231068

* Project supported by the National Nature Science Foundation of China (Grant No. 21903002) and the Excellence Research Fund of Shen Yuan Honors College, Beihang University, China (Grant No. 230121202).

† Corresponding author. E-mail: linxbseu@buaa.edu.cn

专题: 生物分子模拟中的机器学习

高分子塌缩相变和临界吸附相变的 计算机模拟和机器学习

罗启睿¹⁾ 沈一凡²⁾ 罗孟波^{2)†}

1) (杭州链坊科技有限公司, 杭州 310013)

2) (浙江大学物理学院, 杭州 310027)

(2023年6月28日收到; 2023年7月23日收到修改稿)

高分子的塌缩和临界吸附是高分子科学中的两个重要相变现象, 两者均伴随着高分子构象的显著变化. 本文利用朗之万动力学方法和动力学 Monte Carlo 方法分别模拟了高分子的塌缩和临界吸附, 同时获得了不同温度下大量的高分子构象数据. 机器学习方法利用模拟得到的大量伸展无规线团态和塌缩液滴态、脱附态和吸附态构象数据训练神经网络, 学习高分子不同状态的特征, 快速准确地分析不同温度的高分子构象信息, 得到对应的塌缩相变温度和临界吸附温度. 结果表明机器学习能正确给出高分子体系的相变温度, 这为机器学习技术研究高分子的相变提供了新的思路和方法.

关键词: 高分子, 塌缩, 临界吸附, 机器学习**PACS:** 05.70.Jk, 36.20.Ey, 64.70.km**DOI:** 10.7498/aps.72.20231058

1 引言

机器学习是一种人工智能技术, 其基本思想是通过建立多层神经网络模型来实现对数据的学习和识别^[1,2]. 机器学习使用大量的数据进行训练, 可以自动从数据中提取出最优的特征表示, 并在多个层次上逐步抽象数据的特征, 从而实现高效的模式识别和分类任务. 机器学习已成为图像识别、语音识别、自然语言处理、材料信息等领域的关键技术, 并在多个领域得到广泛应用^[3-6]. 近年来, 机器学习也开始应用于材料研究和性能预测、高分子相变和玻璃态转变的研究中^[7-9].

高分子相变是指高分子材料在不同温度及不同条件下发生的相变现象, 包括熔融、结晶、凝胶化等过程, 也包括高分子在溶液中的塌缩相变和在平面上的临界吸附相变. 这些相变现象与高分子材

料的物理性质密切相关, 对高分子材料的制备和应用具有重要作用. 塌缩相变是高分子稀溶液中一个重要的现象, 随温度的变化高分子发生从伸展无规线团态到塌缩液滴态的转变, 从而引发相分离. 塌缩相变在纳米材料制备、药物传递等很多领域有广泛的应用^[10]. 此外, 高分子链在界面的吸附是高分子科学和生物物理的重点研究领域之一. 高分子在表面的吸附可以改变表面的性质, 在制备高分子复合材料、改善材料表面性能、制备生物医用材料, 以及印刷电路等许多技术和生物应用中有重要作用^[11-13]. 因此, 聚合物的塌缩和吸附相变得到了实验、理论和模拟的广泛研究^[14-17]. 其中, 研究塌缩相变温度和临界吸附温度是高分子科学研究中的重要基础问题.

高分子的构象统计性质可以用均方末端距 $\langle R^2 \rangle$ 和均方回转半径 $\langle R_G^2 \rangle$ 等统计物理量来描述. 高分子的塌缩相变和临界吸附相变伴随着高分子

† 通信作者. E-mail: luomengbo@zju.edu.cn

构象的显著变化, 因此可以利用机器学习来智能化分析高分子的不同构象, 实现高分子状态的自动分析和判断. 机器学习从大量的高分子构象数据中学习得到高分子构象的特征, 从而通过构象数据快速准确地判断高分子所处的状态. 目前, 机器学习在分子构象预测、分类、聚类等方面都取得了不错的成果^[6,18]. 例如, 使用卷积神经网络 (CNN) 可以有效地预测高分子的二级结构和三级结构^[19]. 这些方法为高分子材料的设计和 optimization 提供了新的思路 and 手段.

本文用朗之万动力学方法产生了稀溶液中不同温度的高分子构象, 利用均方回转半径 $\langle R_G^2 \rangle$ 随温度的变化确定了高分子塌缩相变的温度, 机器学习则通过计算伸展无规线团态和塌缩液滴态的概率得到高分子塌缩相变的温度. 本文也用动力学 Monte Carlo 方法模拟了低接枝密度的接枝高分子的临界吸附现象, 利用吸附链节数的涨落确定了高分子临界吸附温度, 同时获得了不同温度的大量高分子构象, 然后机器学习通过计算高分子处于脱附态和吸附态的概率得到高分子临界吸附相变的温度. 研究发现模拟和机器学习得到的高分子塌缩相变温度和临界吸附相变温度几乎相同. 动力学 Monte Carlo 和朗之万动力学是研究高分子热力学性质的两个重要的模拟方法, 本文分别用这两种模拟方法模拟了高分子的塌缩和临界吸附相变, 并分别与机器学习进行了比较.

2 模型和研究方法

利用计算机模拟和机器学习的方法研究稀溶液中高分子的塌缩相变温度和低接枝密度的接枝高分子的临界吸附温度. 对于稀溶液中高分子的塌缩相变, 用朗之万动力学方法模拟了高分子均方回转半径 $\langle R_G^2 \rangle$ 随温度的变化, 估算了塌缩相变温度; 对于接枝高分子在吸引平面上的临界吸附, 采用动力学 Monte Carlo 方法模拟了吸附链节数随温度的变化, 估算了临界吸附温度. 然后利用大量不同状态的高分子构象数据对神经网络进行训练, 完成训练后的神经网络从不同温度的高分子构象中计算出高分子处于塌缩液滴态及吸附态的概率. 最后, 利用从机器学习得到的高分子状态概率变化的极大值估算高分子塌缩相变温度和临界吸附温度, 并与模拟结果进行了比较.

2.1 模型

采用粗粒化的珠簧高分子模拟, 链长为 N 的高分子链由直径为 σ 的链节组成. 为简化模拟系统并加快模拟速度, 将溶剂看成背景. 溶剂分子与高分子的随机相互作用提供模拟系统的随机力, 而高分子运动带动溶剂分子的运动则表现为高分子链受到溶剂的黏滞力. 图 1 给出了两个模拟系统的高分子示意图: 稀溶液中高分子和低接枝密度的孤立接枝高分子. 在这两个系统中, 高分子链之间相互作用可以忽略, 因此模拟系统内只考虑一条高分子链. 高分子链内的相互作用包括成键链节之间的相互作用和非成键链节之间的相互作用, 接枝高分子链还存在链节与平面的相互作用.

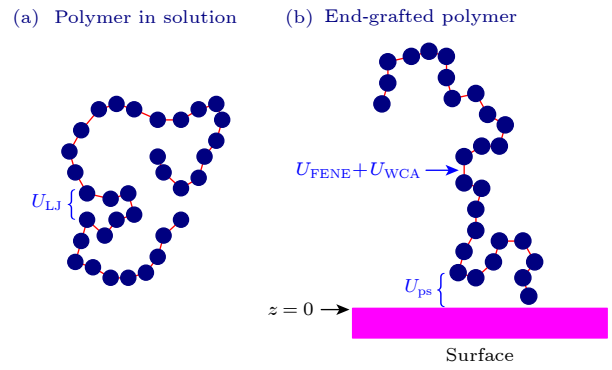


图 1 模拟高分子的示意图 (a) 稀溶液中的高分子; (b) 低接枝密度的孤立接枝高分子
Fig. 1. Schematic diagram of simulated polymers: (a) A polymer in dilute solution; (b) an isolated grafted polymer at low grafting density.

成键链节之间的相互作用包含有限伸展的非线性弹性 (finitely extensible nonlinear elastic, FENE) 势:

$$U_{\text{FENE}}(b) = \begin{cases} -\frac{1}{2}KR_0^2 \ln \left[1 - \left(\frac{b}{R_0} \right)^2 \right], & b < R_0, \\ \infty, & b \geq R_0; \end{cases} \quad (1)$$

以及链节-链节之间成对的纯排斥 Weeks-Chandler-Andersen (WCA) 势^[20]:

$$U_{\text{WCA}}(b) = \begin{cases} 4\epsilon \left[\left(\frac{\sigma}{b} \right)^{12} - \left(\frac{\sigma}{b} \right)^6 + \frac{1}{4} \right], & b < 2^{1/6}\sigma, \\ 0, & b \geq 2^{1/6}\sigma. \end{cases} \quad (2)$$

这里 b 是键长, $K = 30\epsilon/\sigma^2$ 是键的弹性系数, $R_0 = 1.5\sigma$ 是最大键长, ϵ 是 WCA 势的相互作用强度.

FENE 势和 WCA 势的共同作用决定了键的平均长度约为 1σ . 高分子中非键连的链节之间的相互作用采用截断的 Lennard-Jones (LJ) 势:

$$U_{\text{LJ}}(r) = \begin{cases} 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \\ -4\epsilon \left[\left(\frac{\sigma}{r_{\text{cut}}} \right)^{12} - \left(\frac{\sigma}{r_{\text{cut}}} \right)^6 \right], & r < r_{\text{cut}}, \\ 0, & r \geq r_{\text{cut}}, \end{cases} \quad (3)$$

其中 r 是链节之间的空间距离. 为了加快计算速度, LJ 相互作用的计算在 r_{cut} 处截断. 如果取 $r_{\text{cut}} = 2^{1/6}\sigma$, LJ 势演变为 WCA 势, 链节之间只考虑短程的排斥作用. 如果取 $r_{\text{cut}} > 2^{1/6}\sigma$, 模拟还考虑链节之间的相互吸引作用. 在塌缩相变的研究中, 取 $r_{\text{cut}} = 2.5\sigma$, 链节之间的吸引作用在低温下引起链的塌缩; 而在临界吸附的研究中, 取 $r_{\text{cut}} = 2^{1/6}\sigma$, 高分子链节之间为纯排斥作用, 高分子总是处于伸展无规线团状态.

在临界吸附的研究中, 高分子的一端接枝在平面上. 平面位于 $z = 0$, 接枝链节中心位于 $z = 1$ 的位置. 假定一个厚的无限大平面, 高分子链节与平面的相互作用势取 [21]:

$$U_{\text{ps}}(z) = \begin{cases} \epsilon_{\text{ps}} (3/2) (2/5)^{1/2} \\ \times \left[\frac{2}{15} \left(\frac{\sigma}{z} \right)^9 - \left(\frac{\sigma}{z} \right)^3 \right] + U_{\text{c}}, & z < z_{\text{c}}, \\ 0, & z \geq z_{\text{c}}. \end{cases} \quad (4)$$

这里 z 是链节离开平面的垂直距离, ϵ_{ps} 是平面的吸引强度, 取截断距离 $z_{\text{c}} = 4\sigma$ 和

$$U_{\text{c}} = -\epsilon_{\text{ps}} \left(\frac{3}{2} \right) \left(\frac{2}{5} \right)^{1/2} \left[\frac{2}{15} \left(\frac{\sigma}{z_{\text{cut}}} \right)^9 - \left(\frac{\sigma}{z_{\text{cut}}} \right)^3 \right].$$

当 $z_{\text{min}} = 0.8585\sigma$, U_{ps} 取极小值 $-\epsilon_{\text{ps}}$. 当链节位于 $z_{\text{min}} < z < 1.22\sigma$ 时, 势能值小于 $-0.5\epsilon_{\text{ps}}$, 认为这样的链节为吸附链节. 模拟中平面的吸引强度 ϵ_{ps} 固定为 ϵ .

模拟中的物理量均是约化的无量纲量, 取长度单位 $\sigma = 1$, 能量单位 $\epsilon = 1$, 和链节的质量为质量单位 $m = 1$. 温度的单位为 ϵ/k_{B} , 其中 k_{B} 是玻尔兹曼常数.

2.2 朗之万动力学方法

高分子链节的运动方程采用朗之万 (Langevin) 方程:

$$m \frac{d^2 \mathbf{r}_i}{dt^2} = -\nabla U + \mathbf{F}^{(T)} - \eta \mathbf{v}_i, \quad (5)$$

其中 $-\nabla U$ 代表相互作用力, $\mathbf{F}^{(T)}$ 是热运动随机力, $-\eta \mathbf{v}_i$ 是黏滞力. $\mathbf{F}^{(T)}$ 具有高斯分布, 其平均值为 0, 涨落为 $6\eta k_{\text{B}} T$. 链节的位置和速度的演化过程采用修正 velocity-Verlet 算法.

模拟的时间单位为 $\tau_0 = (m\sigma^2/\epsilon)^{1/2}$. 模拟中黏滞系数取 $\eta = 1$, 模拟的时间步长取 $\delta t = 0.01\tau_0$, 每间隔 δt 时间高分子链节的位置和速度同步演化一次.

采用模拟退火的方法得到高分子不同温度的构象. 从高温的初始无序态高分子构象出发, 缓慢地降低模拟温度, 前一个温度的高分子构象作为后一个温度的高分子初始构象. 在每个温度, 长时间模拟得到平衡态, 然后用更长的模拟时间统计高分子链的模拟数据. 由于温度的改变非常小, 高分子构象的平衡通常都比较快. 为了选择正确的模拟时间, 会先对少量样本做一个预模拟, 观察与高分子构象相关的物理量的收敛过程, 估算出平衡时间和弛豫时间.

2.3 动力学 Monte Carlo 方法

在动力学 Monte Carlo 方法中, 高分子链的整体运动通过高分子链节的局域移动来实现. 高分子链节的局域运动导致高分子链构象的变化, 这种构象变化可以通过构造一个 Markov 过程来实现, 即假定高分子链原来处于构象 $\{\mathbf{r}\}$, 通过链节运动以一定的转移概率 $P(\{\mathbf{r}\} \rightarrow \{\mathbf{r}'\})$ 得到新的构象 $\{\mathbf{r}'\}$. 链节局域运动成功与否采用 Metropolis 算法决定, 即该转移概率 P 取 $P = \min[1, \exp(-\Delta E/k_{\text{B}}T)]$, 其中 ΔE 是构象转变前后的能量差, 即 $\Delta E = E\{\mathbf{r}'\} - E\{\mathbf{r}\}$.

链节的每次局域运动对应于链节在 x , y 和 z 方向分别移动 dx , dy 和 dz , 其中 dx , dy 和 dz 是介于 $(-\Delta, \Delta)$ 之间均匀分布的随机量, 模拟中取 $\Delta = 0.1\sigma$. Monte Carlo 的时间单位为 Monte Carlo 步 (MCS), 1 个 MCS 中每个高分子链节平均运动一次. 与朗之万动力学模拟相同, 动力学 Monte Carlo 方法也采用退火模拟的方法得到高分子构象随温度的变化.

2.4 机器学习

采用机器学习中的监督学习模式, 建立了基于

神经网络的分类器. 高分子链的空间构象通过链节序列的数据来表示, 每个数据点代表每个链节所处的空间位置. 在塌缩相变的研究中, 链节序列的数据反映了高分子链处于伸展无规线团态或塌缩液滴态; 而在临界吸附相变的研究中, 链节序列的数据反映了高分子链属于吸附或者非吸附两种状态的一种. 机器学习的任务是: 神经网络分析输入的高分子构象的链节序列数据, 正确输出该高分子构象所属的状态概率. 经典的机器学习方法有循环神经网络 (recurrent neural network, RNN) 及长短期记忆 (long-short term memory, LSTM) 等神经网络结构 [22]. 但是对于高分子链来说, 循环神经网络可能存在一定问题. 循环神经网络会认为后时刻输入的内容与前面时刻输入的内容完全无关, 因此后输入的链节数据可能会赋予极高的判断权重, 而早期的链节数据会被“遗忘”, 这与所有链节等效的物理事实不符. 而长短期记忆神经网络可以克服这个问题. 与一般的前馈神经网络不同, 长短期记忆神经网络可以利用前后数据的时间序列对输入进行分析, 在自然语言处理方面有广泛的应用 [22].

考虑到高分子链节的顺序排列性, 长短期记忆神经网络可以有效地处理高分子构象的长数据特征, 从而正确得到高分子链节数据与高分子链构象类型的映射关系. 因此, 本文处理高分子链的构象数据的核心神经网络的是长短期记忆网络. 图 2 给出了机器学习进行数据处理的流程图.

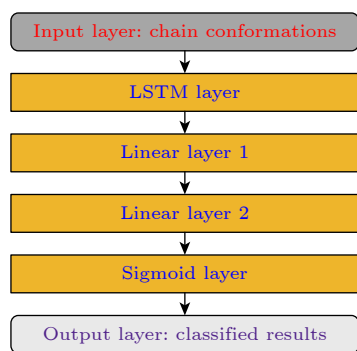


图 2 机器学习的流程图
Fig. 2. Flowchart of machine learning.

长短期记忆神经网络的 LSTM 层接收高分子链各个链节的空间坐标信息, 并和当前的 LSTM 层状态进行计算, 输出新的状态. 完成高分子链构象所有链节数据的计算后, LSTM 层最后输出的状态向量是该高分子链构象在高维嵌入空间中的一

个表达. 接下来的两个线性层都做数据的空间降维工作. 第一个线性层负责将数据从高维嵌入空间变换到一个较低维的隐藏空间, 第二个线性层再将隐藏空间变换到最后的标签空间. 激活层用 sigmoid 函数将数据转换成 (0, 1) 之间的数, 然后输出层输出该值表示该高分子构象中处于所属状态的概率.

监督学习需要有一个训练过程, 因此先对神经网络进行训练和验证. 收集高温和低温的高分子构象, 假设在这两种温度下, 几乎只会生成对应两种状态的高分子链构象, 取其中 75% 的数据作为训练集, 剩余 25% 的数据作为验证集. 我们设置验证的成功概率大于 0.9999 以保证我们的学习效果. 本文的神经网络模型使用二分类交叉熵损失 (binary cross entropy loss) 作为损失函数, 使用 AdamW 优化器 [23]. 在训练时, 二分类交叉熵损失结果意味着神经网络输出结果和期望结果的差异, 重复训练过程直到损失收敛. 图 3 给出了损失随训练次数的变化. 由图 3 可以看到, 损失随训练次数的增加先快速下降, 然后快速收敛到一个稳定值. 根据图 3 的结果, 在训练次数达到 40 的时候, 认为该机器学习模型已经收敛. 因此在实际训练中, 设置训练次数为 40, 在确保模型效果的情况下节约计算时间.

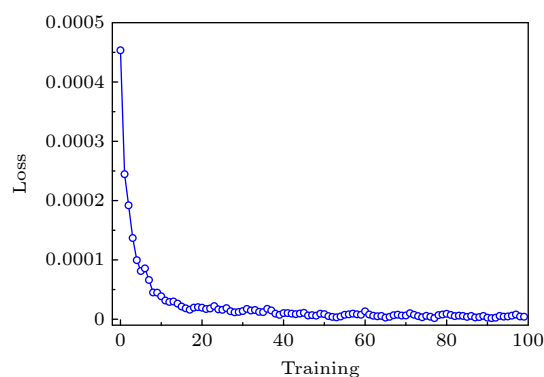


图 3 机器学习中的损失随训练次数的变化
Fig. 3. Loss in machine learning varies with the number of trainings.

完成长短期记忆神经网络训练以后, 把高温和低温之间的其他温度的高分子链构象作为测试集, 判断各个温度下的高分子构象处于给定状态的概率. 最后对同一温度的所有高分子构象的状态概率求平均, 得到该温度下高分子处于某一状态的统计平均值.

本文的神经网络模型中 LSTM 层使用双向

LSTM 结构, 内部为三层双向 LSTM, 每层单方向有 200 个隐藏神经元, 共有 1200 个神经元. 第一线性层有 400 个隐藏神经元, 第二线性层有 80 个隐藏神经元, 最后的激活层有一个 sigmoid 神经元. 因此处理 1 个数据的神经元数目为 1681. 输入层使用 1000 的批大小, 即一次同时输入和处理 1000 条高分子构象数据. 这样, 神经网络模型中总神经元数量达到 1681000. 机器学习使用 Python 语言, 基于 PyTorch 框架搭建了机器学习程序, 在支持 CUDA 的 Nvidia GPU 上运行. 为了能处理不同长度的高分子链的数据, 把每条数据长度固定为 80, 长度不足的数据会自动用 0 补齐. 对不同链长的高分子构象均得到相同的结论, 因此本文给出的数据是模拟中所用的最大链长的计算结果.

3 高分子的塌缩相变

在塌缩相变的研究中, 还考虑了高分子链节之间的短程屏蔽库仑势^[24], 即把高分子链视作聚电解质. 引入静电相互作用大幅降低了塌缩相变温度, 也同时降低了模拟的温度区间, 减小了热运动的无序涨落, 从而可以使用较大的模拟时间步长, 加快模拟的速度. 用朗之万动力学方法模拟了链长 $N = 64$ 的高分子构象性质随温度的变化, 图 4 给出了高分子的均方回转半径 $\langle R_G^2 \rangle$ 对温度的依赖关系. 模拟在一个考虑周期性边界条件的三维立方系统中进行, 系统的尺寸 L 取约为最高模拟温度 ($T = 3.2$) 高分子的方均根回转半径的 10 倍, 即 $L \approx 10 \langle R_G^2 \rangle^{1/2}$. 每个高分子构象的平方回转半径定义为高分子链节距离质心的平均平方距离, 即

$$R_G^2 = \frac{1}{N} \sum_{j=1}^N |\mathbf{r}_j - \mathbf{r}_c|^2, \quad (6)$$

其中 \mathbf{r}_j 是链节的位矢; \mathbf{r}_c 是高分子质心的位矢. $\langle R_G^2 \rangle$ 是 76000 个独立高分子构象的平均. 高分子的 $\langle R_G^2 \rangle$ 随着温度的降低而减小, 表明高分子构象发生了从高温的伸展无规线团 (coil state, C 态) 态到低温的紧缩液滴 (globule state, G 态) 态的转变, 即塌缩相变. 塌缩相变温度 T_c 通常定义为 $\langle R_G^2 \rangle$ 随温度变化最快时对应的温度, 即 $d \langle R_G^2 \rangle / dT$ 极值处. 从图 4 的插图, 得到 $T_c = 0.5$, 这与之前的模拟结果接近^[24].

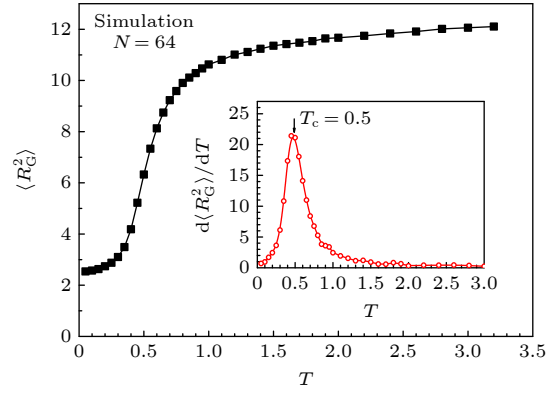


图 4 高分子均方回转半径 $\langle R_G^2 \rangle$ 与温度 T 的关系的朗之万动力学模拟结果. 插图给出了 $\langle R_G^2 \rangle$ 对 T 的导数与 T 的关系
Fig. 4. Simulation results of the mean square radius of gyration $\langle R_G^2 \rangle$ versus temperature T . The inset presents the $\langle R_G^2 \rangle$ derivative of T in relation to T .

比较了不同温度下高分子构象的差别, 图 5 给出了 3 个特殊温度下 (极低温、塌缩相变温度和极高温) 高分子 R_G^2 的归一化概率分布 $P(R_G^2)$. 在极低温 $T = 0.01$, 从图 5(a) 的分布可以看到不同构象的 R_G^2 差别极小, R_G^2 分布在 2.5 到 3 之间很小的区间范围内. 如图 5(a) 插图所示, 高分子的构象是一个紧缩的液滴状. 而在远高于塌缩相变温度的高温, 如图 5(c) 的温度 $T = 3.2$, 高分子 R_G^2 的变化范围很宽, R_G^2 主要变化范围为从 10 到 60, 远大于紧缩液滴状的 R_G^2 . 这表明, 虽然高分子的构象很多, 但基本上都处于伸展的无规线团状, 如图 5(c) 的插图. 在塌缩相变温度 $T_c = 0.5$, 如图 5(b) 所示, 高分子 R_G^2 的变化范围从 4 到 30, 涵盖了近紧缩液滴状构象和伸展无规线团状构象. 我们看到随着温度的变化, 高分子构象的分布也随之发生变化, 这正是机器学习的基础.

然后用基于长短期记忆神经网络的机器学习方法计算了高分子链处于塌缩态的平均概率 P_G . 每个温度下高分子的构象数目均为 76000. 在机器学习的训练阶段, 假定高分子在模拟的最低温 ($T = 0.01$) 均处在 G 态而在模拟的最高温 ($T = 3.2$) 都处于 C 态. 这种假定的合理性通过成功率大于 0.9999 的构象验证得到保证. 对长短期记忆神经网络训练以后, 机器学习自动计算不同温度的高分子处于塌缩态的概率. 高分子链处于塌缩态的平均概率 P_G 随温度的变化结果见图 6. 由图 6 可以看到, P_G 随温度的升高而下降, 表明高分子的构象发生了从 G 态到 C 态的转变, 在 $T = 0.5$ 附近构象的

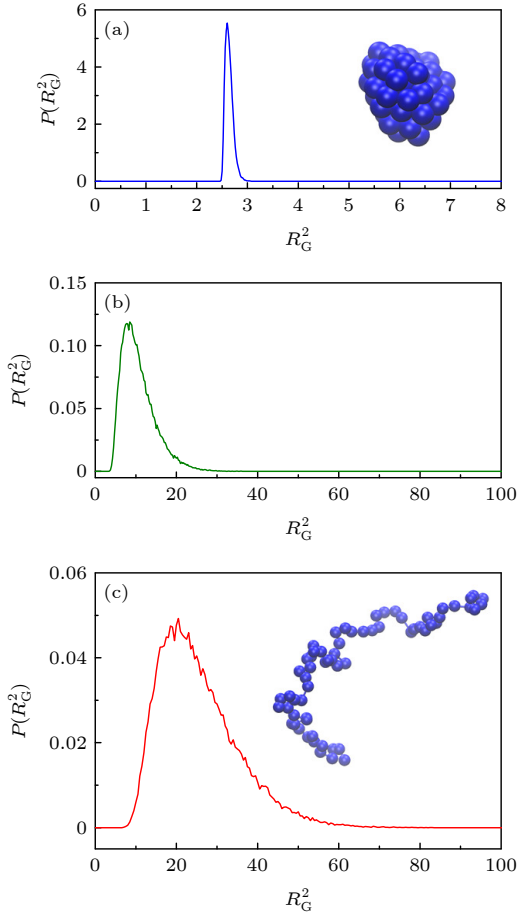


图 5 高分子平方回转半径 R_G^2 在温度 $T = 0.01$ (a), 0.5 (b) 和 3.2 (c) 的概率分布. 插图分别给出了 $T = 0.01$ 和 3.2 时高分子的典型构象

Fig. 5. Plots of the probability distribution of square radius of gyration R_G^2 for polymer at temperatures $T = 0.01$ (a), 0.5 (b), and 3.2 (c). The insets show the typical polymer conformations at $T = 0.01$ and 3.2.

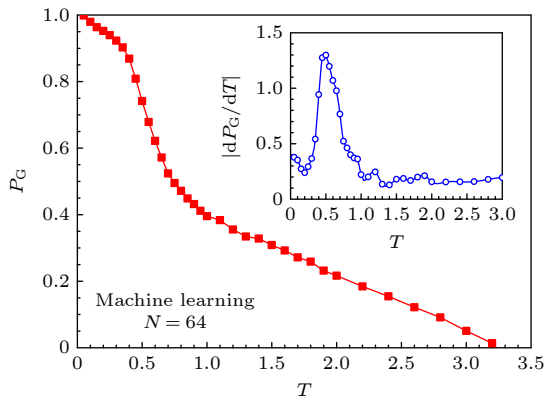


图 6 高分子处于塌缩态的平均概率 P_G 与温度 T 关系的机器学习结果. 插图给出了 $|dP_G/dT|$ 随 T 的变化.

Fig. 6. Machine learning results of the mean probability of polymer being in the compact globule state, P_G , versus temperature T . The inset shows the change of $|dP_G/dT|$ with temperature T .

转变迅速增加. 图 6 的插图给出了概率随温度的变化率 $|dP_G/dT|$, $|dP_G/dT|$ 的峰值位置在 $T = 0.5$, 表明在 $T = 0.5$ 高分子的状态变化最快, 即机器学习得到的塌缩相变温度为 0.5, 与模拟结果一致. 这表明机器学习通过学习 C 态和 G 态的高分子构象, 能有效地判断其他高分子的构象特征, 从而给出符合模拟结果的塌缩相变温度.

这里需要指出模拟结果只给出了高分子构象数据和构象性质的统计平均值, 如图 4 所示, 并不能给出高分子处于 G 态和 C 态的概率, 而机器学习则从高分子构象判断出高分子的状态, 如图 6 所示. 虽然模拟和机器学习的一致性不能直接通过状态的概率来比较, 但可以通过临界温度的数值的一致性来比较.

4 高分子的临界吸附相变

高分子的吸附伴随着吸附能量或者吸附链节数的变化, 高温脱附态的吸附能量低而构象熵大, 低温吸附态的吸附能量大而构象熵小. 高分子吸附过程是一个能量和熵的变化和竞争过程, 临界吸附点定义为能量和熵的竞争平衡点. 吸附过程中平均吸附链节数 $\langle M \rangle$ 随着温度的下降而持续增大. 高分子的临界吸附相变通常被认为是连续相变, 在临界吸附相变温度, 由于高分子不断的吸附和脱附, 吸附链节数的涨落 (类似于热容):

$$\sigma_M^2 = \langle M^2 \rangle - \langle M \rangle^2 \quad (7)$$

达到最大^[15,25]. 模拟研究中常利用吸附链节数的涨落来标定高分子链的临界吸附温度^[25].

用动力学 Monte Carlo 方法模拟了链长 $N = 65$ 的接枝高分子链的吸附. 无穷大平面位于 $z = 0$, 平面的边长 L 也取约为最高模拟温度 ($T = 4$) 高分子的方均根回转半径的 10 倍, 即 $L \approx 10 \langle R_G^2 \rangle^{1/2}$, 系统的高度大于高分子完全伸直的长度. 在平行平面的方向考虑周期性边界条件. 高分子的第一个链节中心固定在平面中心位于 $z = 1$ 的位置, 模拟中定义 $z < 1.22$ 的其他链节 (除接枝链节) 为吸附链节. 图 7 给出了链长 $N = 65$ 的高分子吸附链节数的涨落 σ_M^2 与温度 T 的关系. 模拟得到高分子的临界吸附温度约为 $T_{\text{ads}} = 0.9$. 注意到因为本工作中平面对高分子的吸引势 ((4) 式) 是文献^[26] 的 0.95 倍, 我们的模拟结果与之前朗之万动力学的模拟

结果相近^[26]. 图 7 的插图显示了高温的非吸附态 (non-adsorption state 或 desorption state, D 态) 和低温的吸附态 (adsorption state, A 态) 两种典型的高分子构象: 高温时高分子呈现为蘑菇状的脱附态, 低温时高分子呈现为吸附态. 可见, 高分子在吸附前后的构象也发生了明显的变化.

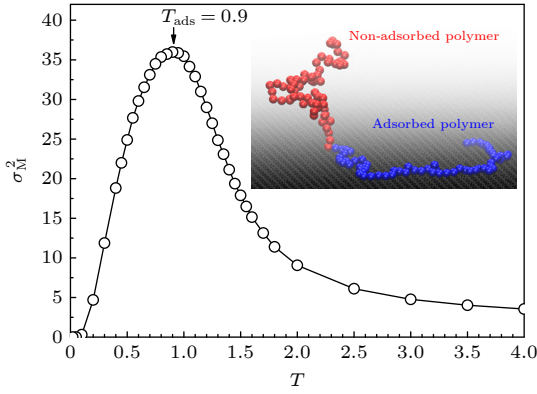


图 7 高分子吸附链节数涨落 σ_M^2 与温度 T 的关系的动力学 Monte Carlo 模拟结果. 插图给出了高温的非吸附态和低温的吸附态高分子构象

Fig. 7. Plot of the fluctuation of adsorption monomer number σ_M^2 of polymer chain versus temperature T . The inset presents non-adsorbed polymer at high temperature and adsorbed polymer at low temperature.

基于长短期记忆神经网络的机器学习通过学习脱附态和吸附态的高分子三维构象, 然后自动给出了不同温度下高分子处于吸附态的平均概率 P_A , 结果如图 8 所示. 这里每个温度的高分子构象数为 248640. 机器学习的结果也显示高分子从低温的完全吸附态过渡到高温的完全脱附态, 与模拟结果一致. 从温度变化率 $|dP_A/dT|$ 随温度的变化可看到, 高分子状态变化最激烈的温度在 $T = 0.9$ 附近, 与模拟得到的临界吸附温度 $T_{\text{ads}} = 0.9$ 一致.

从图 7 的插图可以看到, 吸附链与脱附链构象的最大区别体现为链节离开平面的距离 (z 坐标) 的区别, 即吸附链节数的区别, 而描述高分子构象尺寸的平方回转半径 R_g^2 的差别并不是很明显, 因此我们认为机器学习主要通过区分高分子构象的链节 z 坐标来实现的. 为此, 只让机器学习分析高分子构象的链节 z 坐标, 而忽略它们的 x 和 y 坐标. 图 9 给出了分别利用高分子构象的三维坐标和 z 坐标进行机器学习得到的高分子处于吸附态的平均概率 P_A 与温度 T 的关系. 发现两种方法得到的差别比较小, 说明机器学习在研究临界吸附

时主要分析了构象的 z 坐标. 但在临界吸附温度 T_c 附近, P_A 的值有一些差别, 这说明在 T_c 附近, 三维构象还是对机器学习有一定的影响. 这可能是在 T_c 附近, 高分子构象的变化非常大, 这个时候 z 坐标不能唯一决定 P_A , 高分子状态可能还跟链节的 z 的变化序列相关.

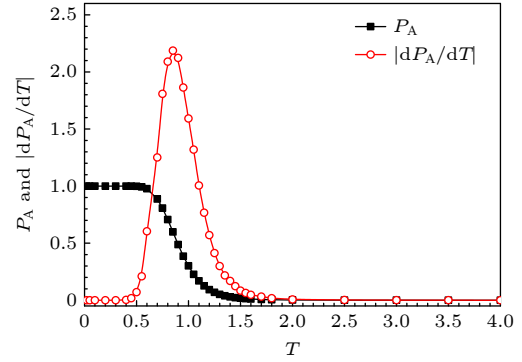


图 8 高分子处于吸附态的平均概率 P_A 和温度变化率 $|dP_A/dT|$ 与温度 T 的关系的机器学习结果

Fig. 8. Plot of the mean adsorption probability P_A of polymer and its temperature change rate dP_A/dT versus temperature T .

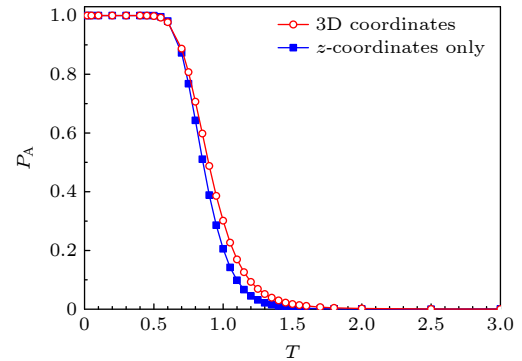


图 9 利用高分子构象的三维坐标和 z 坐标进行机器学习得到的高分子处于吸附态的概率 P_A 与温度 T 的关系

Fig. 9. Relationship between adsorption probability P_A and temperature T obtained by machine learning using the three-dimensional coordinates and z -coordinates only of polymer conformations.

进一步分析了每个高分子构象的机器学习结果与构象性质之间的关联. 关联了总共 248640 个构象的吸附数 M 和链质心高度 z_c 与机器学习得到的每个构象可能处于吸附态的概率 P_A , 找出 P_A 在 $(0, 0.2)$, $(0.2, 0.8)$ 和 $(0.8, 1)$ 三个范围内高分子构象的分布. 图 10 给出了临界吸附温度 $T_c = 0.9$ 时的结果. 发现小的 P_A 对应于小的 M 和大的 z_c , 而大的 P_A 对应于大的 M 和小的 z_c . 这说明机器学习的结果是符合物理预期的. 在 $P_A \in (0.2, 0.8)$

区域有一段重叠区, 在这段重叠区内, 虽然高分子的 M 和 z_c 相同, 但高分子构象变化范围大, 因此具有各种不同的状态和 P_A 值. 这说明 P_A 不是 M 和 z_c 的单值函数, 机器学习对高分子构象的判断还与构象的其他因素有关, 这也与图 9 的结论相符.

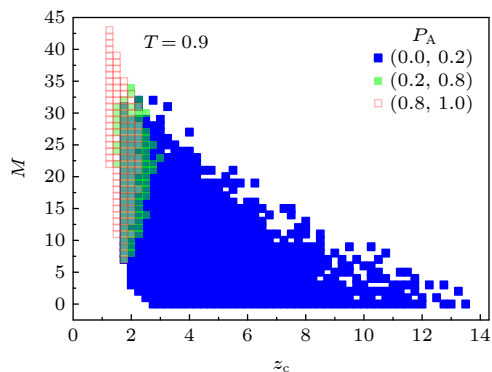


图 10 机器学习得到的吸附态概率 P_A 在 $(0, 0.2)$, $(0.2, 0.8)$ 和 $(0.8, 1)$ 三个范围内高分子构象相对于构象的吸附数 M 和链质心高度 z_c 的分布

Fig. 10. Distribution of polymer conformation relative to the adsorbed number M and the mean height z_c for the adsorption probability P_A obtained by machine learning in three ranges of $(0, 0.2)$, $(0.2, 0.8)$ and $(0.8, 1)$.

5 结 论

本文模拟研究了高分子的塌缩相变和临界吸附相变, 得到了大量的高分子构象数据. 机器学习采用长短期记忆网络作为核心神经网络, 对高分子链构象按塌缩态和吸附态进行了分类统计. 结果表明: 模拟和机器学习得到了相同的塌缩相变温度和临界吸附相变温度, 机器学习的结果符合物理预期. 本文的研究扩展了机器学习在高分子构象识别方面的应用, 有望应用到高分子材料在不同温度下

的相变行为的智能自动分析中.

参考文献

- [1] Hinton G, Deng L, Yu D, et al. 2012 *IEEE Signal Process. Mag.* **29** 82
- [2] Silver D, Huang A, Maddison C J, et al. 2016 *Nature* **529** 484
- [3] Umehara M, Stein H S, Guevarra D, et al. 2019 *NPJ Comput. Mater.* **5** 34
- [4] Iwasaki Y, Takeuchi I, Stanev V, et al. 2019 *Sci. Rep.* **9** 2751
- [5] Chen J Z, Yang C W, Ren J 2021 *Acta Phys. Sin.* **70** 144204 (in Chinese) [陈江芷, 杨晨温, 任捷 2021 物理学报 **70** 144204]
- [6] Cencer M M, Moore J S, Assary R S 2022 *Polym. Int.* **71** 537
- [7] Zhang Y, Xu X 2021 *J. Mol. Graphics Modell.* **103** 107796
- [8] Liang Z, Li Z, Zhou S, et al. 2022 *Cell Reports Physical Science* **3** 100931
- [9] Zhang K, Li X, Jin Y, Jiang Y 2022 *Soft Matter* **18** 6270
- [10] Xu Y, Wang Z G 2021 *Macromolecules* **54** 10984
- [11] Milner S T 1991 *Science* **251** 905
- [12] Besteman K, Lee J O, Wiertz F G M, Heering H A, Dekker C 2003 *Nano Lett.* **3** 727
- [13] Duan X, Zhang R, Ding M, Huang Q, Xu W S, Shi T, An L 2017 *Polymer* **122** 125
- [14] Sumithra K, Brandau M, Straube E 2009 *J. Chem. Phys.* **130** 234901
- [15] Li Y W, Wüst T, Landau D P 2013 *Phys. Rev. E* **87** 012706
- [16] Yang Q H, Wu F, Wang Q, Luo M B 2016 *J. Polym. Sci. Part B: Polym. Phys.* **54** 2359
- [17] Ziebarth J D, Gardiner A A, Wang Y M, Jeong Y, Ahn J, Jin Y, Chang T 2016 *Macromolecules* **49** 8780
- [18] Bhattacharya D, Patra T K 2021 *Macromolecules* **54** 3065
- [19] Nguyen T, Bavarian M 2022 *Ind. Eng. Chem. Res.* **61** 12690
- [20] Weeks J D, Chandler D, Andersen H C 1971 *J. Chem. Phys.* **54** 5237
- [21] Chremos A, Glynos E, Koutsos V, Camp P J 2009 *Soft Matter* **5** 637
- [22] Hochreiter S, Schmidhuber J A 1997 *Neural Comput.* **9** 1735
- [23] Loshchilov I, Hutter F 2017 *arXiv:1711.05101 [cs.LG]*
- [24] Luo M B, Tsehay D A, Sun L Z 2017 *J. Chem. Phys.* **147** 034901
- [25] Yang X, Wu F, Hu D D, Zhang S, Luo M B 2019 *Chin. Phys. Lett.* **36** 098202
- [26] Qi H K, Yang X, Yang Q H, Luo M B 2022 *Polymer* **259** 125330

SPECIAL TOPIC—Machine learning in biomolecular simulations

Computer simulation and machine learning of polymer collapse and critical adsorption phase transitions

Luo Qi-Rui¹⁾ Shen Yi-Fan²⁾ Luo Meng-Bo^{2)†}

¹⁾ (*NFTGo, Hangzhou 310013, China*)

²⁾ (*School of Physics, Zhejiang University, Hangzhou 310027, China*)

(Received 28 June 2023; revised manuscript received 23 July 2023)

Abstract

Collapse and critical adsorption of polymers are two crucial phase transitions in polymer science, both are accompanied by significant changes in polymer conformation. In this paper, Langevin dynamics and dynamic Monte Carlo methods are used to simulate the collapse and critical adsorption of polymer, respectively, and corresponding phase transition temperatures are estimated. Meanwhile, a large number of polymer conformations at different temperatures are obtained. In the machine learning method, a large number of extended random coil and collapsed spherical, desorption and adsorption conformations are used to train the neural network, so that the neural network can learn the characteristics of different states of the polymer, and it can quickly and accurately analyze the polymer conformations at different temperatures and obtain the corresponding collapse phase transition temperature and critical adsorption temperature. The results demonstrate that machine learning can correctly calculate the phase transition temperature of polymer system, which provides new ideas and methods for machine learning technology in the study of polymer phase transitions.

Keywords: polymer, collapse, critical adsorption, machine learning

PACS: 05.70.Jk, 36.20.Ey, 64.70.km

DOI: [10.7498/aps.72.20231058](https://doi.org/10.7498/aps.72.20231058)

† Corresponding author. E-mail: luomengbo@zju.edu.cn

专题: 生物分子模拟中的机器学习

RNA 扭转角预测的深度学习方法*

欧秀娟 肖奕†

(华中科技大学物理学院, 武汉 430074)

(2023 年 6 月 29 日收到; 2023 年 8 月 2 日收到修改稿)

RNA 分子三级结构建模是分子生物物理学研究的基本问题之一, 对理解 RNA 的功能和设计新的结构有重要意义. RNA 三级结构主要由主链和侧链上的 7 个扭转角确定, 准确预测这些扭转角是 RNA 分子三级结构建模的基础. 目前只有个别采用深度学习模型预测 RNA 分子扭转角的方法, 要用于建模 RNA 分子的三级结构其预测精度还有待进一步提高. 本文提出了一种预测 RNA 分子扭转角的深度学习模型 1dRNA, 采用了考虑相邻核苷酸的卷积模型 (DRCNN) 和考虑全链核苷酸的超长短期记忆模型 (DHLSTM) 两种不同的深度学习模型. 结果显示, 与现有方法相比, 这两种模型都能提高 RNA 分子大部分扭转角的预测精度, DRCNN 预测精度提高在 5% 到 28% 之间, DHLSTM 预测精度提高在 6% 到 15% 之间. 结果还显示, α 和 γ 角是最难预测的, 环区扭转角比螺旋区的扭转角难预测, 模型对预测序列长度的变化不敏感, 模型预测角度与 decoys 的角度偏差可用于模型质量评估.

关键词: RNA 结构, 扭转角预测, 深度学习

PACS: 87.14.gn, 87.15.A-, 87.15.bg

DOI: 10.7498/aps.72.20231069

1 引言

RNA 分子三级结构建模是分子生物物理学研究的基本问题之一, 对理解 RNA 的功能和设计新的结构有重要意义^[1-3]. RNA 分子三级结构建模是给出 RNA 分子的核苷酸序列构建其三级结构^[4-10]. RNA 三级结构可以分为主链结构和侧链结构, 主链结构由螺旋区和环区构成, 由 6 个扭转角 ($\alpha, \beta, \gamma, \delta, \epsilon, \zeta$) 确定, 侧链方向由扭转角 χ 确定 (图 1). RNA 分子主链和侧链结构还涉及共价键键长和键角, 但这些键长和键角会相对平衡位置进行微振动, 在生理温度这些参数的变化关于平衡位置对称, 影响将相互抵消^[11]. 因此, 扭转角被认为是 RNA 分子三级结构的决定因素, 预测这些扭转角可以帮助建模 RNA 分子的三级结构.

扭转角预测在蛋白质分子三级结构建模中已

经有深入的研究. 与 RNA 分子不同, 蛋白质分子三级结构主要由主链上的 2 个扭转角 ψ 和 ϕ 确定. 从 2007 年以来, 人们提出了不同的神经网络模型预测扭转角 ψ 和 ϕ . 2007 年, Real-SPINE1.0 使用一层全连接神经网络预测蛋白质主链 ψ 角, 角度的平均绝对误差 (mean absolute error, MAE) 为 54° ^[12]; 2008 年, Real-SPINE2.0 使用同样神经网络和输入特征, 角度标签 $[0^\circ, 180^\circ]$ 不变, $[-180^\circ, 0^\circ]$ 加上 360° 做一个平移, 同时预测蛋白质主链 ψ 和 ϕ 角, 角度的 MAE 分别为 38° 和 25° ^[13]; 2009 年, Real-SPINE2.0 使用两层全连接网络, ψ 和 ϕ 角预测精度进一步改进, MAE 分别为 36° 和 22° ^[14]; 2009 年和 2012 年, SPINE XI 和 SPINE X 使用多步神经网络, ψ 角预测的 MAE 分别为 33° ^[15] 和 35° ^[16]; 2015 年 SPIDER2 使用深度学习 3 层全连接神经网络预测角度的正弦和余弦值, ψ 角预测的 MAE 降低到 30° ^[17]; 2017 年, SPIDER3 使用 4 层双向

* 国家自然科学基金 (批准号: 32071247) 资助的课题.

† 通信作者. E-mail: yxiao@hust.edu.cn

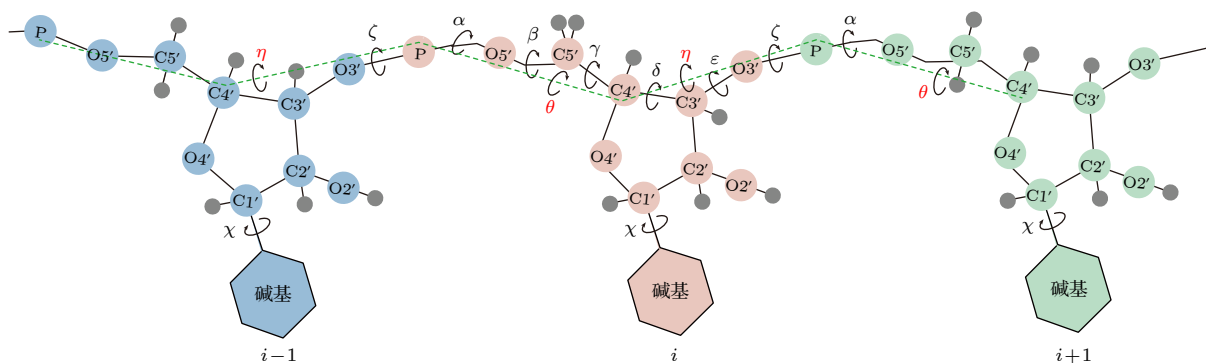


图 1 RNA 分子主链和侧链 7 个扭转角和 2 个伪角的示意图

Fig. 1. Diagram of RNA seven backbone torsion and two pseudo-torsion angles.

LSTM 模型使 ψ 角预测的 MAE 进一步下降为 27° ^[18]; 2019 年, SPOT-1D 使用 10 层以上的 LSTM (long short-term memory) 残差网络预测角度的正弦和余弦值, ψ 角预测的 MAE 为 23° ^[19]; 2020 年, 使用 3 层全连接网络, 滑动窗口特征, ψ 角预测的 MAE 仅为 18° ^[20]. 对于 RNA 分子, 2021 年, SPOT-RNA-1D 首次使用 1 层普通卷积和 2 层膨胀卷积预测 RNA 的 7 个扭转角和 2 个自定义伪角 (η, θ) (图 1) 的正弦和余弦值, $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \chi, \eta, \theta$ 的平均绝对误差分别为 $43.94^\circ, 21.94^\circ, 32.98^\circ, 14.61^\circ, 20.69^\circ, 33.27^\circ, 19.59^\circ, 30.25^\circ$ 和 32.91° ^[21]. 可以看到, 相对于蛋白质分子, RNA 分子扭转角预测的精度还有待提高.

本文提出了一种基于时序网络深度学习模型预测 RNA 分子扭转角的方法 1dRNA, 分别使用深度残差卷积模型 (deep residual CNN, DRCNN) 和深度超长短期记忆模型 (deep HyperLSTM, DHLSTM) 预测 RNA 分子的 7 个扭转角和 2 个伪角, 以此分析抓取相邻核苷酸特征的卷积网络和抓取全局核苷酸特征的循环网络, 哪种网络更合适扭转角预测问题, 并将两个模型的结果和抓取间隔核苷酸特征的 SPOT-RNA-1D 比较. DRCNN 模型基于只能看到相邻核苷酸特征的一维卷积, 卷积过程不改变序列长; DHLSTM 模型基于能看到全部核苷酸的特征、并能改变常规长短期记忆 (LSTM) 网络权重共享范式的超 LSTM 网络. 结果表明, 本文采用的两个深度学习模型都可以进一步提高 RNA 分子扭转角的预测精度, 不同模型在不同角度上各有优势, $\delta, \zeta, \chi, \eta$ 和 θ 角的预测更适合卷积网络, β 和 ϵ 角的预测更适合循环网络, 而在 α 和 γ 角中, 抓取间隔核苷酸的膨胀网络更好.

2 深度学习模型

2.1 深度学习模型

DRCNN 模型架构如图 2 所示, 由一个一维卷积层^[22]开始, 输入通道为 4, 输出通道为 512 (卷积输出通道超参数 512 比 256 效果好和 1024 效果类似), 训练批次为 8 (本文模型在一张 11G 显存 GTX 1080 Ti 显卡上能容下的最大样本数), 卷积核为 15 (卷积核超参数 15 比 7 和 30 效果好), 填充方式为“same”, 其他为默认值. 初始卷积层之后, 是 4 个残差块的依次叠加 (残差块的数目 1 到 6 测试显示 4 个残差块效果最好), 每个残差块^[23]依次包含: 一维批归一化层 BatchNorm1d^[24] (特征维度为 512, 添加在卷积网络中, 有助于模型训练的的稳定, 效果比 LayerNorm 样本归一化要好), ReLU 激活函数^[25] (对本文模型激活函数 ReLU 比 tanh 和 Leaky ReLU 效果好), 一维卷积层 (输入通道维度为 512, 输出通道维度为 512, 卷积窗口一次能看到的核苷酸数目为 15, 填充方式为“same”, 其他为默认值), 再一维批归一化层, ReLU 激活函数和一维卷积层, 最后将此层卷积的输出和残差块的输入相加, 相加的结果再输入下一个残差块中, 重复 4 次. 数据流出残差块后, 经过一个 ReLU 激活函数 (激活函数放在残差块外训练效果更好), 一维批归一化层 (特征维度为 512), dropout 层 (和全连接层连用, 减少网络的过拟合, 采样概率 0.4, 比 0.2 和 0.5 效果好), 全连接层 (输入维度 512, 输出维度 18), tanh 激活函数 (输出区间在 $[-1, 1]$, 和预测角度的正弦和余弦值区间一致) 得到输出.

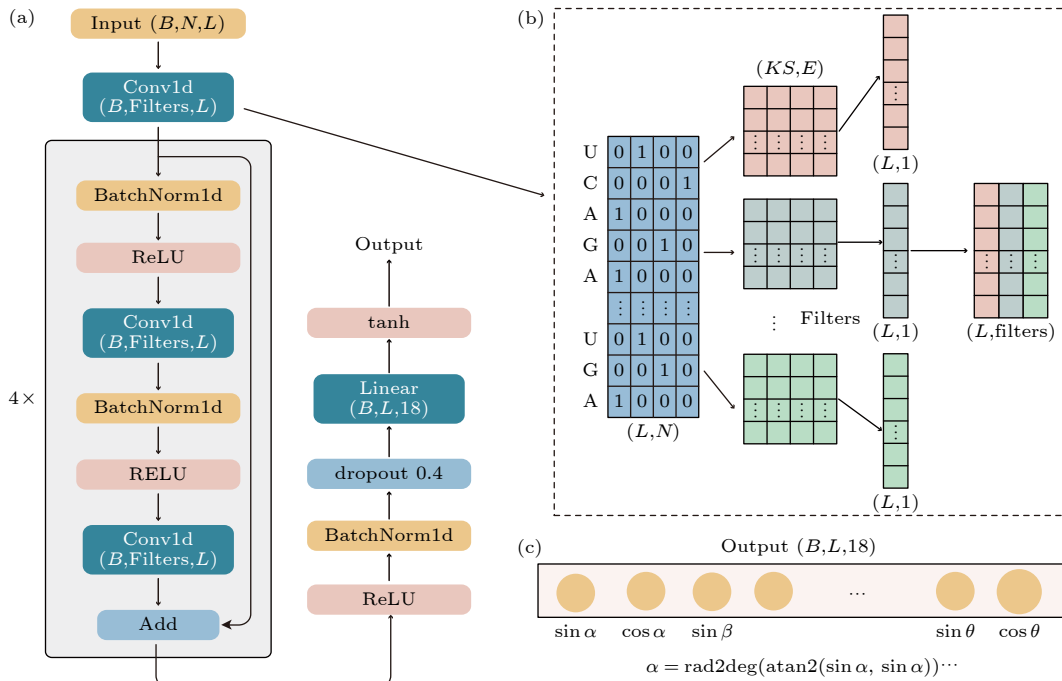


图 2 DRCNN (a) 模型架构; (b) 模型中一维卷积层的原理; (c) 输出层. B, L, N, KS 和 $Filters$ 分别为训练中更新一次模型参数选择的序列数目、序列的长度、输入特征维度、卷积核的小大 (卷积窗口一次能看到的相邻核苷酸数目)、卷积核的数目 (卷积层的输出维度)

Fig. 2. DRCNN: (a) Network architecture; (b) Conv1d layer; (c) output layer. B, L, N, KS and $Filters$ are batch size, sequence length, the size of the input, the size of the filter (the filter can see the number of nucleotides at one time), the number of filters.

DHLSTM 模型结构如图 3 所示, 里面的 HyperLSTM 层原理来自于文献 [26], 输入数据的维度是 (512, 8, 4), 模型更新一次参数选取的样本批次数目为 8, 描述一个核苷酸的初始特征向量维度为 4; 然后经过一个 HyperLSTM 层 (这里的超参数, 外部大 LSTM 层 [27] 的输出维度 Hidden 取 64、内部小 LSTM 层的输出维度和改变 LSTM 层权重的 Hypercell 单元里线性投影的维度 Hyper 都取 16; Hidden 超参数 64 比 16, 32 和 128 效果好, Hyper 超参数 16 比 32 和 64 效果好), 具体来说, 第 t 个核苷酸特征向量和两类隐藏态进入 HyperLSTM cell 单元, 得到第 $t + 1$ 个核苷酸新的特征向量和两类隐藏态, 这里每个核苷酸使用不同的 HyperLSTMcell 权重参数, 依次算完所有核苷酸, 得到描述一个批次每个核苷酸新特征数据维度 (512, 8, 64); 接着经过另一个 HyperLSTM 层 (这里三层 HyperLSTMcell 单元的超参数 Hidden 都取 64, Hyper 都取 16), 具体来说, 上一层输出的第 t 个核苷酸特征向量和两类隐藏态 (维度 (8, 64)) 依次进入三个 HyperLSTMcell 单元, 得到第 $t + 1$ 个核苷酸新的特征向量 (维度 (8, 64)) 和两类隐藏态输出 (维度分别为 (8, 64), (8, 16)), 依次算完所有核

苷酸, 得到描述一个批次每个核苷酸的新特征数据维度 (512, 8, 64); 最后将第二层 HyperLSTM 的输出和第一层的 HyperLSTM 输出相加, 作为一个残差块; 数据流出残差块后, 进入全连接层 (输入维度 512, 输出维度 18), tanh 激活函数得到输出.

DHLSTM 和 DRCNN 训练都使用 MSE 损失函数和 RMSprop 优化器 [28] 训练 (优化器学习率取 0.001、正则化系数取 0.0001, 此优化器比 Adam 和 AdamW 优化器效果好, 学习率 0.01 比 0.1, 0.001, 0.0001 和 0.00001 效果好, 正则化系数经过尝试取学习率的百分之一 0.0001 比较好); 同时预测 9 个角和单独预测一个角, 预测结果基本一致, 故 DHLSTM 和 DRCNN 都同时预测 9 个角; DHLSTM 模型在训练过程中, 训练损失随着 epoch 的增大一直下降, 验证损失在第 85 个 epoch 后开始逐步上升, 如图 4(a) 所示, 故取第 85 个 epoch 的模型为最终模型; DRCNN 模型在训练过程中, 训练损失随着 epoch 的增大一直下降, 验证损失在第 109 个 epoch 后开始逐步上升, 如图 4(b) 所示, 故取第 109 个 epoch 的模型为最终模型. DHLSTM 和 DRCNN 的实现都使用 Facebook 的 PyTorch 深度学习框架 [29].

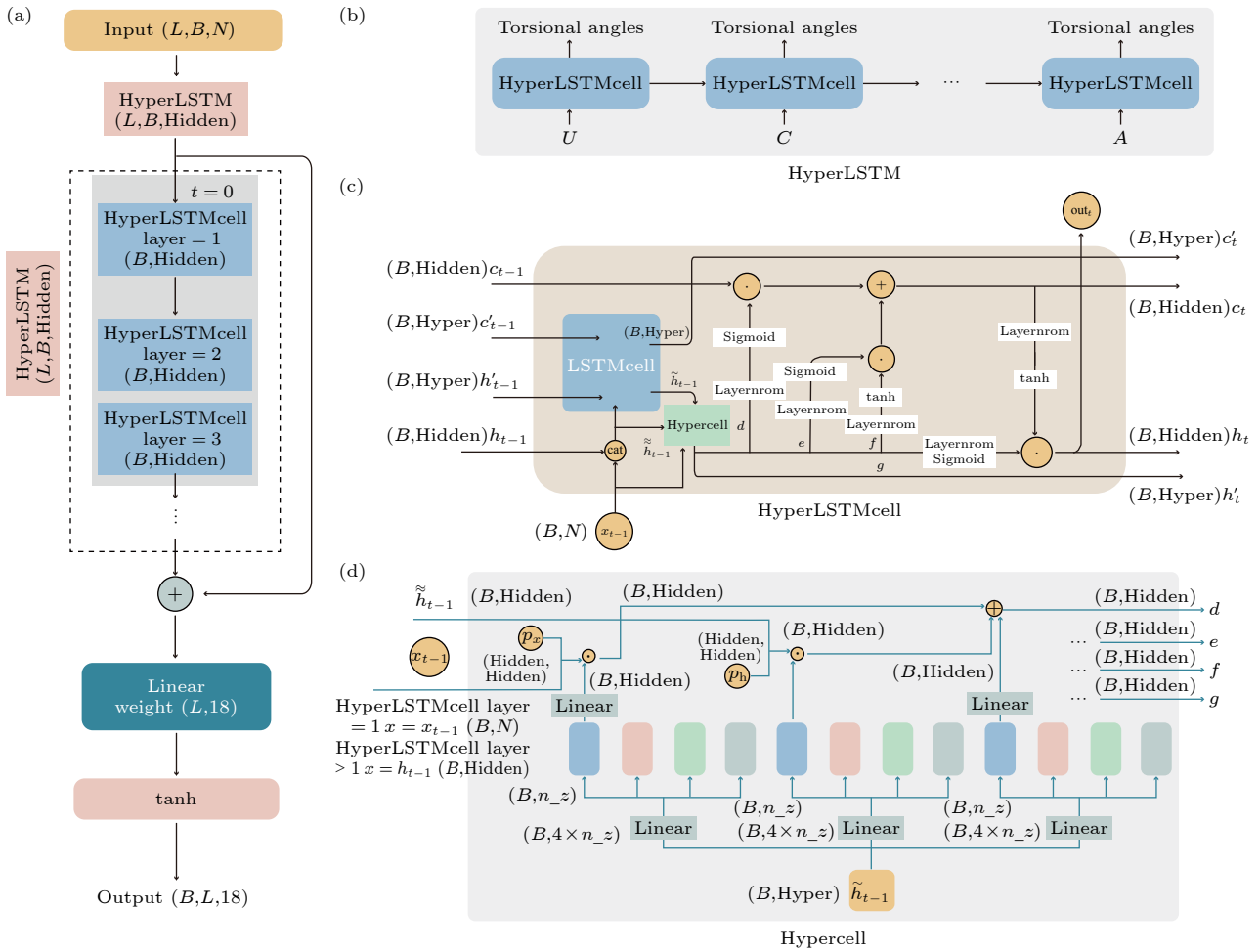


图 3 DHLSTM (a) 模型架构; (b) HyperLSTM 层; (c) 对每个核苷酸的处理单元 HyperLSTMcell, 其中 h_t , c_t 和 h_{t-1} , c_{t-1} 分别是外部更大的 LSTM 在 t 和 $t-1$ 时刻的隐藏态; h'_t , c'_t 和 h'_{t-1} , c'_{t-1} 分别是更小的 LSTM 在 t 和 $t-1$ 时刻的隐藏态; (d) Hypercell 单元. L , B , N , Hidden , Hyper 和 n_z 分别为序列的长度、训练中更新一次模型参数选择的序列数目、输入特征维度、大 LSTM 层的输出维度、内部 LSTM 层的输出维度和改变大 LSTM 层权重的 Hypercell 单元里线性投影的维度, P_x 和 P_h 为动态可训练参数, 绑定在内部超网络里, 作用在输入态 x_{t-1} 和隐藏态, 初始值为全零张量

Fig. 3. DHLSTM: (a) Network architecture; (b) HyperLSTM layer; (c) HyperLSTMcell; h_t , c_t and h_{t-1} , c_{t-1} are the states of the larger outer LSTM at time t and $t-1$, respectively; h'_t , c'_t and h'_{t-1} , c'_{t-1} are the states of the smaller LSTM at time t and $t-1$. (d) Hypercell. L , B , N , Hidden are sequence length, batch size, the size of the input, the size of the LSTM, and Hyper is the size of the smaller LSTM that alters the weights of the larger outer LSTM, n_z is the size of the feature vectors used to alter the larger LSTM weights, P_x and P_h are dynamically trainable parameters, bound in the internal hypernetwork, acting on the input state x_{t-1} and the hidden state, and the initial value is an all-zero tensor.

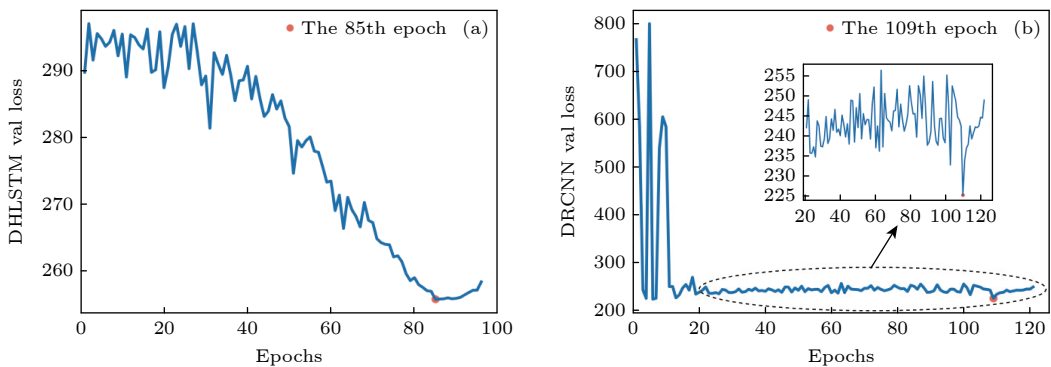


图 4 (a) DHLSTM 模型和 (b) DRCNN 模型验证损失 (MAE) 随 epoch 的变化
Fig. 4. Validation loss curve with the epoch by (a) DHLSTM and (b) DRCNN.

2.2 数据集

为了比较, 采用了 SPOT-RNA-1D 使用的训练集、验证集和测试集 (<https://github.com/jaswinder-singh2/SPOT-RNA-1D/tree/main/datasets>)^[21]. 训练集含有 286 个结构, 从 PDB 结构数据库^[30] 目前可以下载到 284 个结构 (6N5R_A, 6N5L_A 下架), 本文训练集为这 284 个结构; 验证集含有 30 个结构, 都可从 PDB 下载; 测试集有 3 个分别含有 63, 30 和 54 个结构, 从 PDB 数据库分别下载到 62 (5Y85_B 内含脱氧核苷酸下架)、30 和 54 个结构.

SPOT-RNA-1D 数据集来自于 2020 年 10 月 3 日 PDB 数据库中所有 X 衍射分辨率小于 3.5 Å 的 RNA 结构; 用 CD-HIT-EST^[31] 软件对所有这些结构的序列设置相似度 0.8 进行聚类, 多簇类中的代表序列构成训练集; 然后将训练集和单簇类利用 BLAST-N^[32] 软件设置截断值为 10 处理, 训练集与单簇类有命中的序列被删除, 单簇类中有命中的序列也被删除; 经过这些处理, 训练集剩下的序列作为最终训练集, 单簇类剩下的序列随机分为验证集、测试集 I 和测试集 II; 另外, 对 2021 年 4 月 5 日 PDB 数据库中所有 NMR 结构, 使用相同方法, 去除和训练集、验证集、测试集 I 和测试集 II 的冗余, 作为测试集 III. 数据集的长度和二级结构分布信息如表 1 所列.

2.3 输入和输出

模型的输入为核苷酸序列特征, 大小为 $L \times 4$ 的 one-hot 编码, 四个核苷酸 (A, U, G 和 C) 分别用 (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) 和 (0, 0, 0, 1) 表示, L 为序列长度, 序列长度最长为 512, 长度不够的补 0. 数据集中最长序列为 414, 常规做法是

将所有序列用 0 补齐到最长序列长度. 在预测时, 模型预测的目标序列长度应不大于最长序列长度. 这里取 512 是借鉴很多蛋白质模型中取值 512, 又观察到所有序列长度补齐到 414 和 512 的预测结果类似, 故为了模型能预测更长的序列, 取值 512. 在训练中测试过将所有序列补 0 区域采用 mask 机制, 补 0 区域值虽然被计算但不参与下层值的计算, 模型性能改善不明显. 输出具体如图 2(c) 所示, 有 18 个节点用于预测 9 个角的正弦和余弦值, 然后利用 atan2 函数将角度的正弦和余弦值转化为角度的弧度值, 再利用 rad2deg 函数将角度的弧度值转化为角度值. 这种变换在蛋白质扭转角预测里也常用.

2.4 评估

使用 MAE 评估整体性能, 具体如 (1) 式, 预测角度值和实验确定的角度值的绝对差, 360° 和这个绝对差的差值, 取两者的小值:

$$\text{MAE} = \sum_i \min \left(|\text{torsion}_{\text{pred}} - \text{torsion}_{\text{true}}|, (360 - |\text{torsion}_{\text{pred}} - \text{torsion}_{\text{true}}|) \right). \quad (1)$$

3 计算结果和讨论

本文两个深度学习模型使用上面的训练集、验证集和 3 个独立的测试集进行训练、验证和测试. 为了了解模型每个角度在每个测试集的总体表现, 表 2 列出了 DRCNN, DHLSTM 和 SPOT-RNA-1D^[21] 在验证集和 3 个测试集上整体的性能评估. 在含有 62 个 RNA 的测试集 I 上, DRCNN 预测的 β , δ , ζ , χ , η 和 θ 角的 MAE 比 SPOT-RNA-1D 分别减小了 5%, 28%, 17%, 16%, 24% 和 20%, α , γ 和 ε 角的 MAE 比 SPOT-RNA-1D 分别增大

表 1 训练集、验证集和 3 个测试集的长度和二级结构信息 (百分数是数据集不同配对类型的核苷酸数目占比)

Table 1. Length and secondary-structure information of training, validation and test sets. The number mentioned along with the base pairing type is the percentage of total nucleotides in the region.

数据集	序列长度区间数目						二级结构		
	20—50	50—100	100—200	200—300	300—400	400—512	括号	假结	不配对
训练集	50	179	46	1	7	1	55.10%	5.63%	39.36%
验证集	20	10	0	0	0	0	52.19%	9.8%	38.01%
测试集I	11	41	10	0	0	0	57.58%	2.81%	39.61%
测试集II	8	16	6	0	0	0	58.42%	5.25%	36.33%
测试集III	40	13	1	0	0	0	65.02%	2.67%	32.31%

表 2 DHLSTM, DRCNN 和 SPOT-RNA-1D 在验证集和 3 个测试集上的 MAE
Table 2. Performance comparison in terms of MAE on validation sets and three test sets by three models.

数据集	7个标准扭转角							伪角		
	$\alpha/(^\circ)$	$\beta/(^\circ)$	$\gamma/(^\circ)$	$\delta/(^\circ)$	$\varepsilon/(^\circ)$	$\zeta/(^\circ)$	$\chi/(^\circ)$	$\eta/(^\circ)$	$\theta/(^\circ)$	
DHLSTM	验证集	47.91	20.22	37.18	16.57	18.23	35.02	19.85	28.09	32.85
	测试集I	48.20	20.66	37.13	13.08	18.82	30.27	17.33	25.74	29.22
	测试集II	47.95	19.89	35.30	15.19	17.87	30.99	17.67	27.20	31.49
	测试集III	45.45	22.30	40.80	13.51	21.43	30.69	16.96	23.87	29.84
DRCNN	验证集	44.67	19.96	35.31	13.86	22.20	31.62	19.49	24.77	30.22
	测试集I	44.84	20.74	36.27	10.51	21.48	27.53	16.39	23.12	26.34
	测试集II	43.41	19.55	35.45	12.19	22.71	28.13	17.16	24.28	28.12
	测试集III	27.14	15.81	25.20	9.73	14.51	17.98	11.58	13.67	17.77
SPOT-RNA-1D [21]	验证集	45.18	20.58	33.88	17.99	20.72	37.50	23.01	33.55	37.02
	测试集I	43.94	21.94	32.98	14.61	20.69	33.27	19.59	30.25	32.91
	测试集II	39.50	18.92	29.47	16.01	17.46	28.91	18.20	28.14	30.25
	测试集III	37.89	21.04	34.68	13.83	22.32	27.87	17.01	25.31	27.22

了 2%, 10% 和 4%; DHLSTM 预测的 $\beta, \delta, \varepsilon, \zeta, \chi, \eta$ 和 θ 角的 MAE 比 SPOT-RNA-1D 分别减小了 6%, 10%, 9%, 9%, 12%, 15% 和 11%, α 和 γ 角的 MAE 比 SPOT-RNA-1D 分别增大了 10% 和 13%, 这表明在 $\delta, \zeta, \chi, \eta$ 和 θ 角这些角中, 每层考虑相邻核苷酸特征的 DRCNN 比每层考虑全部核苷酸特征的 DHLSTM 要好, 在 β 和 ε 角中, 每层考虑全部核苷酸特征的 DHLSTM 比每层考虑相邻核苷酸特征的 DRCNN 要好, 在 α 和 γ 角中, 每层考虑间隔核苷酸的 SPOT-RNA-1D 比 DRCNN 和 DHLSTM 都要好. MAE 值越大预测难度越大, 在 DRCNN 中角度预测难度 $\delta, \chi, \varepsilon, \beta, \eta, \theta, \zeta, \gamma$ 和 α 依次递增, 在 DHLSTM 中角度预测难度 $\delta, \chi, \beta, \varepsilon, \eta, \theta, \zeta, \gamma$ 和 α 依次递增, 在 SPOT-RNA-1D 中角度预测难度 $\delta, \chi, \varepsilon, \beta, \eta, \theta, \gamma, \zeta$ 和 α 依次递增, 可以看到 $\delta, \chi, \eta, \theta$ 和 α 角在 3 个模型里预测难度的排序一致, 考虑相邻核苷酸的 DRCNN 和考虑间隔核苷酸的 SPOT-RNA-1D 都表明 ε 比 β 容易预测, 而对于 DHLSTM, ε 比 β 难预测, DRCNN 和 DHLSTM 都表明 ζ 比 γ 容易预测, 而对于 SPOT-RNA-1D, ζ 比 γ 难预测. 这 3 种方法都认为 α 是最难预测的, 表明 3 个模型在角度预测难度方面有一定相似性, 也各有特点. 在测试集 II 和测试集 III 观察到类似的性能趋势, 表明模型对不同类型的测试集具有鲁棒性.

为了了解模型在单个序列上的表现, 图 5 给出了 DRCNN, DHLSTM 和 SPOT-RNA-1D 在 3 个测试集上单个 RNA 分子扭转角预测的 MAE 分布

图, 其中 SPOT-RNA-1D 绘制每个盒子需要五类值 (最大值、最小值、中位数、上下四分位数和异常值), 由论文图形数据获取工具 WebPlotDigitizer^[33] 得到. 每个模型在 3 个数据集 9 个角度的 27 个 MAE 最小值上, DRCNN 占 18 次, DHLSTM 占 3 次, SPOT-RNA-1D 占 6 次, 而在 27 个 MAE 最大值上, DRCNN 占 4 次, DHLSTM 占 8 次, SPOT-RNA-1D 占 15 次, 表明考虑相邻核苷酸特征的卷积模型 DRCNN 最有可能预测到最小的 MAE 值, DHLSTM 次之, SPOT-RNA-1D 很难预测相比比较小的 MAE 值. 箱子越窄意味着每次预测 MAE 变化更小, 模型预测更稳定, 每个模型在 3 个测试集 9 个角度的 27 个箱子中, DRCNN 出现 9 次, DHLSTM 出现 15 次, SPOT-RNA-1D 出现 3 次, 表明预测最稳定的模型是考虑全部核苷酸特征的 DHLSTM, 且性能中规中矩, 其次是 DRCNN, 对样本反应比较敏感的是 SPOT-RNA-1D. 在 27 个盒子相对较小的中位数上, DRCNN 占 18 次, DHLSTM 占 2 次, SPOT-RNA-1D 占 7 次, 表明 DRCNN 预测的一半数目链的总 MAE 比其他两个模型值要低. 在异常值方面, 3 个测试集 9 个角度上, DRCNN, DHLSTM 和 SPOT-RNA-1D 出现的异常值的数目分别为 24, 21 和 38, 且 DRCNN 和 DHLSTM 出现的异常值本身是比较小, 同样表明 DHLSTM 预测比较稳定. 以上说明, 考虑相邻核苷酸特征的 DRCNN 模型性能整体更强大, 考虑全部核苷酸特征的 DHLSTM 模型预测更稳定.

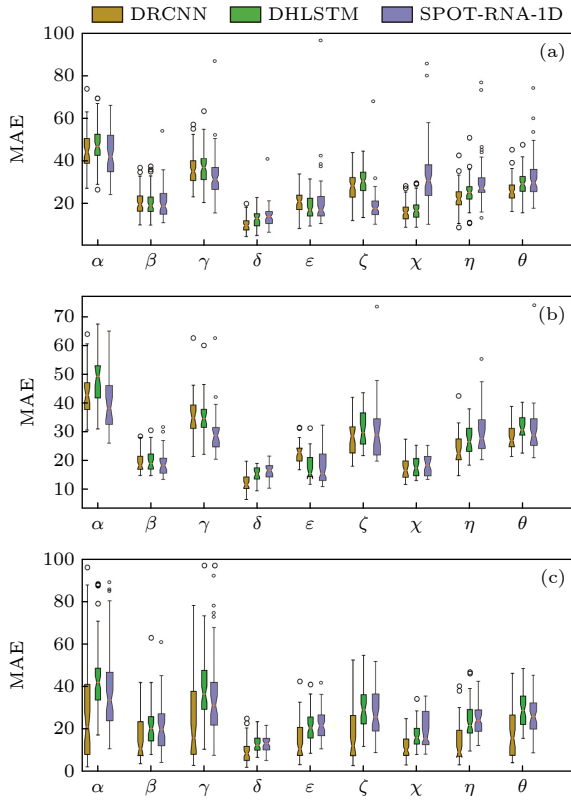


图5 DRCNN(黄色)、DHLSTM(绿色)和SPOT-RNA-1D(紫色)在测试集 I (a)、测试集 II (b)和测试集 III (c)上单个 RNA 链的 MAE 分布图. 每个盒子显示出一组数据的最大值、最小值、中位数、上下四分位数和异常值

Fig. 5. Distribution of MAE for individual RNA chains on test set I (a), test set II (b) and test set III (c) by DRCNN predictor (yellow), by DHLSTM (in green) and SPOT-RNA-1D (in purple). Each box shows the minimum, the maximum, the sample median, the first and third quartiles and outlier.

另外绘制了角度的实验值分布, 如图 6 橙色虚线所示, 可以看出每个角度的实验值的分布是比较陡峭的, 大部分角度都集中在跨度在 40° 左右的角度空间, 有少部分角度值分布在跨度在 360° 的角度空间中, 最容易预测的 δ 角跨度也是最窄的, 最难预测的 α 角分布有 3 个峰, 跨度是最广的. 为了了

解本文模型在预测分布上的能力, 绘制了 DRCNN 和 DHLSTM 在测试集 I 的预测分布如图 6 黄色和绿色虚线所示, DRCNN 预测所有的角度分布都比 DHLSTM 好; 在测试集 II 和测试集 III 上, DRCNN 在 β 和 γ 角上预测的分布比 DHLSTM 要好, 两个模型在预测其他 7 个角的分布类似.

二级结构对 RNA 建模起着重要作用, 根据 DSSR 软件^[34]输出的 RNA 二级结构, 可将 RNA 二级结构分为三种类型, 括号 (['(', ')']), 假结 (['[', ']', '{', '}', '<', '>', 'A', 'a']), 环区 ['.', '.]. 比较了 DRCNN 和 DHLSTM 在测试集 III 中对 3 种二级结构类型的整体预测性能 (表 3), 可以看出, 对 DRCNN 和 DHLSTM 来说括号类型的核苷酸的扭转角最容易预测的, 处于环区的核苷酸的扭转角是最难预测; 还可以观察到, DRCNN 预测 3 种类型的 MAE 误差都比相应的 DHLSTM 预测的要低; 在其他两个测试集观察到同样结果, 因此, 扭转角预测的误差主要来自于环区和假结区域, 在预测括号、假结和环区区域的扭转角上 DRCNN 都比 DHLSTM 好.

表 1 统计了训练集、验证集和 3 个测试集的序列长度分布. 由表 1 可以看出, 在训练集和验证集中各个长度分布并不均匀, 长度在 50 到 100 区间的有 179 个结构, 在 100 到 200 区间的只有 46 个. 为了了解这种差异是否会导致 DRCNN 和 DHLSTM 对长 RNA 扭转角预测性能较差, 图 7 绘制了两个模型在 9 个角度上的表现与序列长度的关系. 观察 DHLSTM 和 DRCNN 的预测结果, 9 个角的 MAE 值在数目少的长度区间 [78, 94], [155, 171] 和 [171, 186] 并不大; 还观察到 DRCNN 在短长度区间 [1, 47] 结果比 DHLSTM 结果好; 因此, 虽然训练集和验证集对不同长度的 RNA 数目分布不均匀, 但并没有造成 DRCNN 和 DHLSTM 在预测上的长度偏好.

表 3 DHLSTM 和 DRCNN 在测试集 III 不同配对类型中扭转角预测的 MAE

Table 3. Performance according to mean absolute error by DHLSTM and DRCNN for nucleotides in different pairing type on test set III.

配对类型	七个标准扭转角							伪角		
	$\alpha/(\circ)$	$\beta/(\circ)$	$\gamma/(\circ)$	$\delta/(\circ)$	$\epsilon/(\circ)$	$\zeta/(\circ)$	$\chi/(\circ)$	$\eta/(\circ)$	$\theta/(\circ)$	
DHLSTM	括号	34.08	16.48	30.21	9.76	17.98	21.38	11.23	18.03	21.91
	假结	34.20	14.98	27.06	6.80	14.25	20.29	10.98	27.41	18.02
	环区	66.77	32.60	60.72	21.05	27.54	47.85	28.52	35.41	46.16
DRCNN	括号	19.43	11.40	18.54	6.65	11.84	12.0	8.30	10.90	12.94
	假结	20.42	14.25	16.75	6.73	12.86	13.54	10.25	16.14	13.52
	环区	40.84	23.26	37.44	15.59	19.07	29.07	18.44	19.25	27.08

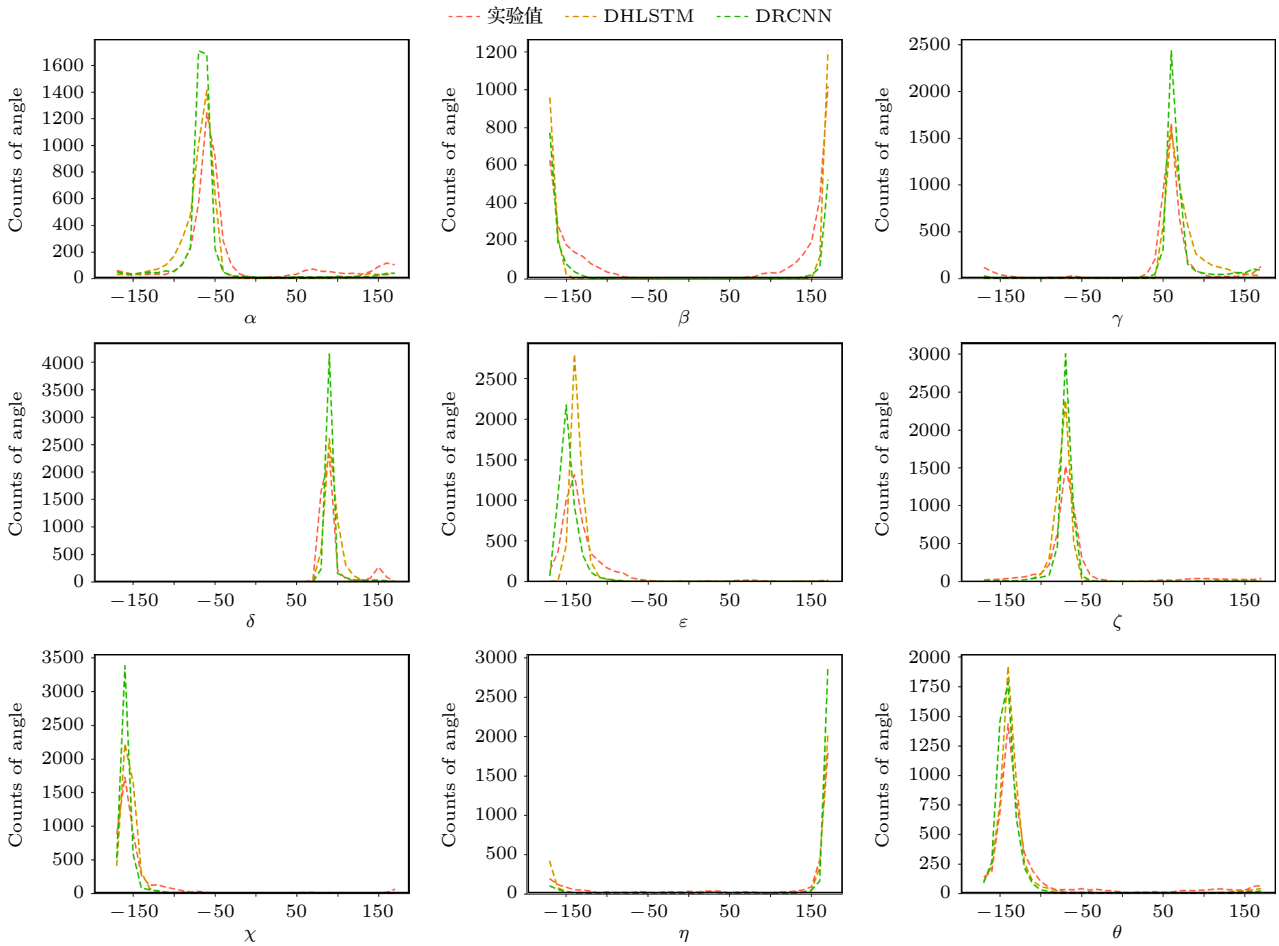


图 6 测试集 I 扭转角的实验值 (橙色)、DHLSTM 预测值 (黄色) 和 DRCNN 预测值 (绿色) 分布图

Fig. 6. Distribution plots of native (in orange), DHLSTM predicted (in yellow), and DRCNN predicted (in green) nine torsion angles on test set I.

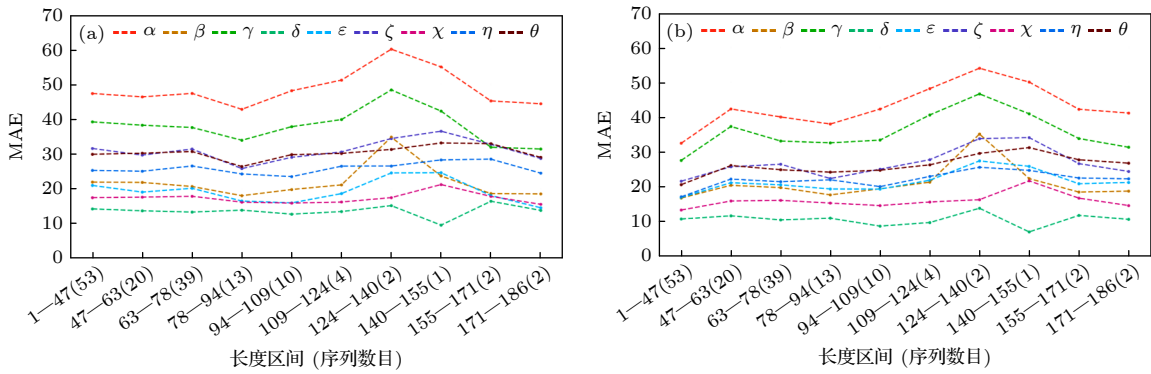


图 7 (a) DHLSTM 和 (b) DRCNN 分别在 3 个测试集 (147 个 RNA) 的 9 个扭转角的 MAE 与 RNA 序列长度的函数

Fig. 7. On 147 RNAs in the three test sets, the MAE is measured as a function of the length for the nine torsion angles by (a) DHLSTM and (b) DRCNN.

和 SPOT-RNA-1D 方法一样, 为了了解扭转角之间的相关性, 在测试集 I 上绘制了如图 8 所示的扭转角相关矩阵. 一般情况下, 相邻扭转角之间高度相关, 而较远扭转角相关性较小, 但是矩阵显示, 对于 DRCNN 和 DHLSTM, α 和 γ 角有很强的相关性, 两者也是模型预测难度最大的两个角,

ζ 和 θ 有最强的相关性, 两者预测难度排名也是相邻的. 在其他两个测试集的结果相同.

观察一条链中预测的每个角度, 预测的大部分扭转角比一些天然态或者类天然态结构的扭转角更接近天然态结构扭转角的值. 和 SPOT-RNA-1D 方法一样, 也测试了 DRCNN 和 DHLSTM 这

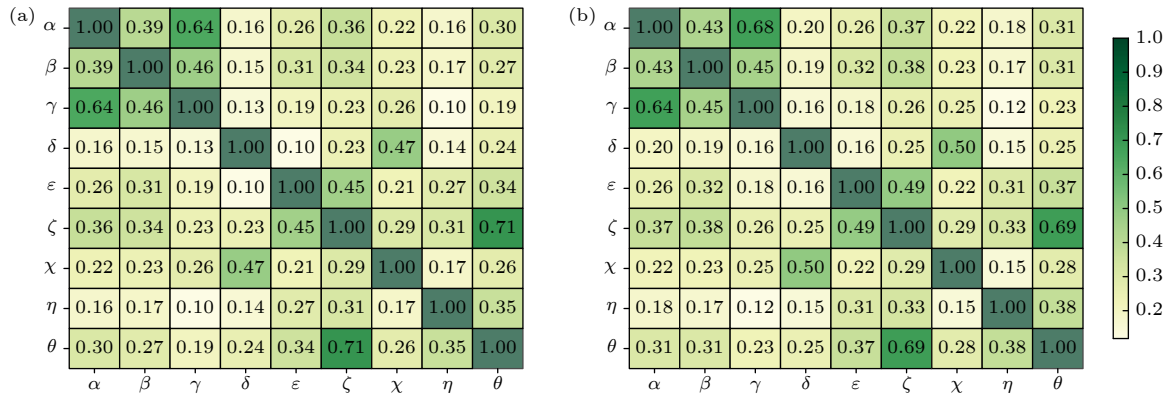


图 8 (a) DHLSTM 和 (b) DRCNN 分别在测试集 I 上扭转角的 MAE 的相关系数 (CCs), 值越大表示两个角度越相关
 Fig. 8. Correlation coefficient (CCs) for MAE of between the nine torsion angles of test set I by (a) DHLSTM and (b) DRCNN, the larger the CC value, the more correlated between the two torsions.

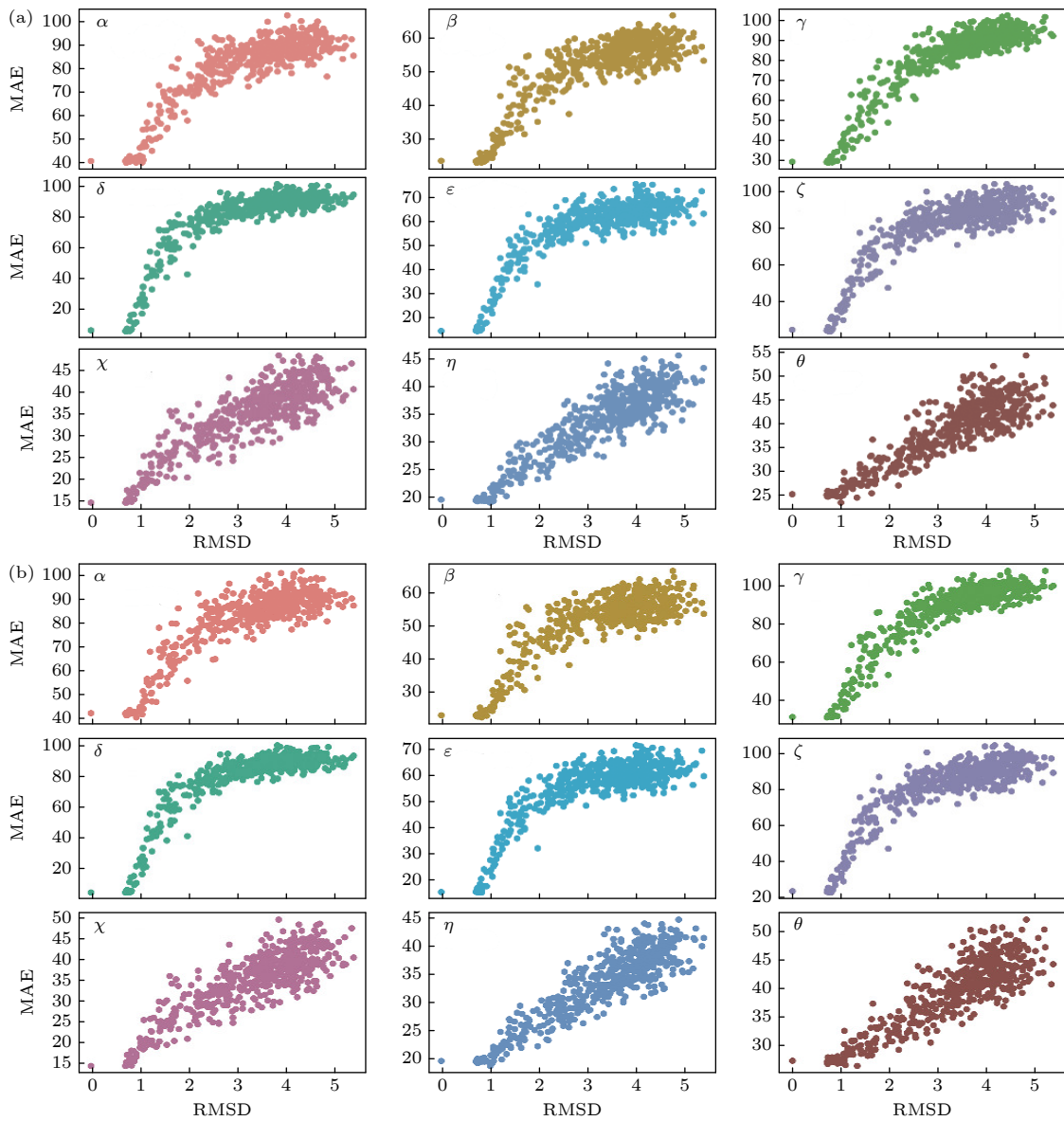


图 9 (a) DRCNN 和 (b) DHLSTM 分别在 RNA 1Y69(链 9) 上预测角度与 decoys 结构角度之间的 MAE 与 RMSD 的关系
 Fig. 9. On RNA 1Y69 (chain 9), the MAE is measured as a function of RMSD for the nine torsion angles by (a) DRCNN and (b) DHLSTM.

两种深度学习模型预测的角度和不同 RMSD 结构的角度之间的差异是否可以用于结构的质量评估. 为此, 使用 3dRNA^[3] 测试集 85 个 RNA 和它们的 decoys 进行了测试. 图 9 绘制了 DRCNN 和 DHLSTM 在其中一个 RNA (PDB ID 号 1Y69, 链 9) 在预测角度与诱饵模型结构角度之间的 MAE 和结构精度的函数关系, MEA 随 RMSD 持续增加. 在 85 个数据集中的其余 84 个 RNA 中也观察到类似的趋势, 这表明与模型预测角度的偏差或结合其他参量可用于模型质量评估.

4 结 论

本文提出了一种预测 RNA 分子扭转角的深度学习方法 1dRNA, 采用了 DRCNN 和 DHLSTM 两个基于时序网络的模型去预测 RNA 的 7 个扭转角 ($\alpha, \beta, \gamma, \delta, \varepsilon, \zeta$ 和 χ) 和 2 个伪角 (η 和 θ), 并和现有方法 SPOT-RNA-1D 进行了比较. 结果表明不同网络在不同角度上各有优势, 当序列长度不超过 50 时, 在预测 9 个角时, 考虑相邻核苷酸特征的 DRCNN 比考虑全部核苷酸特征的 DHLSTM 和考虑间隔核苷酸特征的 SPOT-RNA-1D 都好; 当序列长度超过 50, 在 $\delta, \zeta, \chi, \eta$ 和 θ 角这些角中, DRCNN 预测的结果整体上比 DHLSTM 和 SPOT-RNA-1D 要好, 在 β 和 ε 角中, DHLSTM 预测的结果整体上比 DRCNN 和 SPOT-RNA-1D 要好, 在 α 和 γ 角中, SPOT-RNA-1D 预测的结果整体上比 DHLSTM 和 DRCNN 要好; 3 个模型在 9 个角度的预测难度上类似, 角度的实验值和预测值分布可以看出角度预测的难度主要在于角度分布的复杂程度, 分布越复杂越难预测, DRCNN 和 SPOT-RNA-1D 预测出来的角度分布比 DHLSTM 丰富; 序列环区的角度分布比配对区域复杂, 角度预测难度也比配对区域大很多; 每个模型在链长度集中在非长链区的训练集和验证集上训练, 但在预测时对长链预测效果也不错; 在模型预测稳定性上, 考虑全链核苷酸的 DHLSTM 比考虑相邻核苷酸的 DRCNN 和考虑间隔核苷酸的 SPOT-RNA-1D 要稳定很多, 异常值少; 模型的各个结果在 3 个测试集上表现类似, 表明模型性能对不同数据集稳定. 从结果来看, 面对比较短序列, 9 个角度都用考虑相邻核苷酸特征的卷积网络更好, 当序列长时, 在预测 $\delta, \zeta, \chi, \eta$ 和 θ 角用考虑相邻核苷酸特征的卷积网络更好,

预测 β 和 ε 用考虑全链核苷酸特征的超循环网络更好, 预测 α 和 γ 角用考虑间隔核苷酸特征的膨胀卷积网络更好. 在数据集方面, 尝试过加入新发表的 RNA 结构增大数据集训练, 精度能提高但不明显; 可以设计其他类型的网络, 尝试使用单纯的全连接网络和 Transformer^[35] 网络训练, 角度预测整体 MAE 比 DRCNN 和 DHLSTM 更好, 但预测的角度分布很差, 很难预测出角度分布峰值之外的区域; 尝试过在 DRCNN 和 DHLSTM 这个两个模型上改进, 精度能提高但不明显; 在加入新特征方面, 加入二级结构特征, 能提高精度但也不明显. 在改进角度预测方面, 从结果可以看出角度分布决定了预测难度, 在预测前如何预先处理这种分布, 和如何把这种分布加入损失函数, 应该可以很大提高预测精度; 另外直接预测角度实值难度大, 可以考虑将跨度 360° 的角度分布分成 36 个 bin 去预测.

参考文献

- [1] Jiao K, Hao Y Y, Wang F, et al. 2021 *Biophys. Rep.* **7** 21
- [2] Sun S, Chen X Z, Chen J, et al. 2021 *Biophys. Rep.* **7** 8
- [3] You Y L, Tang Z M, Lin H, Shi J L 2021 *Biophys. Rep.* **7** 159
- [4] Zhang Y, Wang J, Xiao Y 2022 *J. Mol. Biol.* **434** 167452
- [5] Zhang Y, Wang J, Xiao Y 2020 *Comput. Struct. Biotechnol. J.* **18** 2416
- [6] Wang J, Wang J, Huang Y Z, Xiao Y 2019 *Int. J. Mol. Sci.* **20** 4116
- [7] Wang J, Xiao Y 2017 *Curr. Protoc. Bioinf.* **57** 5.9.1
- [8] Wang J, Zhao Y J, Zhu C Y, Xiao Y 2015 *Nucleic Acids Res.* **43** e63
- [9] Zhao Y J, Huang Y Y, Gong Z, et al. 2012 *Sci. Rep.* **2** 734
- [10] Wang J, Mao K K, Zhao Y J, Zeng C, Xiang J J, Zhang Y, Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
- [11] Olson W K 1982 *Topics in Nucleic Acid Structures* (Part 2) (London: Macmillan Press) pp1-79
- [12] Dor O, Zhou Y Q 2007 *Proteins* **68** 76
- [13] Xue B, Dor O, Faraggi E, Zhou Y Q 2008 *Proteins* **72** 427
- [14] Faraggi E, Xue B, Zhou Y Q 2009 *Proteins* **74** 847
- [15] Faraggi E, Yang Y D, Zhang S H, Zhou Y Q 2009 *Structure* **17** 1515
- [16] Faraggi E, Zhang T, Yang Y D, Kurgan L, Zhou Y Q 2012 *J. Comput. Chem.* **33** 259
- [17] Heffernan R, Paliwal K, Lyons J, et al. 2015 *Sci. Rep.* **5** 11476
- [18] Heffernan R, Yang Y D, Paliwal K, Zhou Y Q 2017 *Bioinformatics* **33** 2842
- [19] Hanson J, Paliwal K, Litfin T, Yang Y D, Zhou Y Q, Valencia A 2019 *Bioinformatics* **35** 2403
- [20] Mataeimoghdam F, Newton M A H, Dehjangi A, Karim A, Jayaram B, Ranganathan S, Sattar A 2020 *Sci. Rep.* **10** 19430
- [21] Singh J, Paliwal K, Singh J, Zhou Y Q 2021 *J. Chem. Inf. Model.* **61** 2610
- [22] Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman D J 2021 *Mech. Sys. Signal Proc.* **151** 107398

- [23] He K M, Zhang X Y, Ren S Q, Sun J 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Las Vegas, NV, USA, June 27–30, 2016 p770
- [24] Nam H, Kim H E 2018 arXiv: 1805.07925v3 [cs.CV]
- [25] Clevert D A, Unterthiner T, Hochreiter S 2015 arXiv: 1511.07289v5 [cs.LG]
- [26] Jayasiri V, Wijerathne N 2020 <https://nn.labml.ai/> [2023-04-02]
- [27] Hochreiter S, Schmidhuber J 1997 *Neural Comput.* **9** 1735
- [28] Tieleman T, Hinton G 2012 *Lecture 6.5-RMSProp: Divide the Gradient by a Running Average of its Recent Magnitude (COURSERA: Neural Networks for Machine Learning)*
- [29] Paszke A, Gross S, Massa F, et al. 2019 *33rd Conference on Neural Information Processing Systems* Vancouver, Canada, December 8, 2019 pp8026-8037
- [30] Burley S K, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow G V, et al 2021 *Nucleic Acids Res.* **49** D437
- [31] Fu L M, Niu B F, Zhu Z W, Wu S T, Li W Z 2012 *Bioinformatics* **28** 3150
- [32] Altschul S F, Gish W, Miller W, Myers E W, Lipman D J 1990 *J. Mol. Biol.* **215** 403
- [33] Rohatgi A 2022 Software available at <https://automeris.io/WebPlotDigitizer> Version 4.6[software]
- [34] Lu X J, Bussemaker H J, Olson W K 2015 *Nucleic Acids Res.* **43** e142
- [35] Vaswani A, Shazeer N, Parmar N, et al. 2017 arXiv: 1706.03762v7 [cs.CL]

SPECIAL TOPIC—Machine learning in biomolecular simulations

Deep learning methods of predicting RNA torsion angle*

Ou Xiu-Juan Xiao Yi †

(School of Physics, Huazhong University of Science and Technology, Wuhan 430074, China)

(Received 29 June 2023; revised manuscript received 2 August 2023)

Abstract

Modeling of RNA tertiary structure is one of the basic problems in molecular biophysics, and it is very important in understanding the biological function of RNA and designing new structures. RNA tertiary structure is mainly determined by seven torsions of main-chain and side-chain backbone, the accurate prediction of these torsion angles is the basis of modeling RNA tertiary structure. At present, there are only a few methods of using deep learning to predict RNA torsion angles, and the prediction accuracy needs further improving if it is used to model RNA tertiary structure. In this study, we also develop a deep learning method, 1dRNA, to predict RNA backbone torsions and pseudotorsion angles, including two different deep learning models, the convolution model (DRCNN) that considers the features of adjacent nucleotides and the Hyper-long-short-term memory model (DHLSTM) that considers the features of all the nucleotides. We then empirically show that DRCNN and DHLSTM outperform existing state-of-the-art methods under the same datasets, the prediction accuracy of DRCNN model is improved by 5% to 28% for β , δ , ζ , χ , η , and θ angle, and the prediction accuracy of DHLSTM model is improved by 6% to 15% for β , δ , ζ , χ , η , θ angle. The DRCNN model predicts better results than the DHLSTM model and the existing models in the δ , ζ , χ , η , θ angle, and the DHLSTM model predicts better results than the DRCNN model and the existing model in the β and ε angles, and the existing models predicted better results than the DRCNN model and DHLSTM model in the α and γ angles. The DRCNN model and the existing models predict a richer distribution of angles than the DHLSTM model. In terms of model stability, the DHLSTM model is much more stable than the DRCNN model and the existing models, with fewer outliers. The results also show that the α angle and γ angle are the most difficult to predict, the angles of the ring region is more difficult to predict than the angles of the helix region, the model is also not sensitive to the change of the target sequence length, and the deviation of the model prediction angle from the decoys can also be used to evaluate the RNA tertiary structures quality.

Keywords: RNA structure, torsional angle prediction, deep learning

PACS: 87.14.gn, 87.15.A–, 87.15.bg

DOI: 10.7498/aps.72.20231069

* Project supported by the National Natural Science Foundation of China (Grant No. 32071247).

† Corresponding author. E-mail: yxiao@hust.edu.cn

专题: 生物分子模拟中的机器学习 • 封面文章

使用中间层受监督的自编码器探索蛋白质的构象空间*

陈光临 张志勇†

(中国科学技术大学物理系, 合肥 230026)

(2023年6月28日收到; 2023年7月29日收到修改稿)

蛋白质的功能往往与其结构和动态变化密切相关. 分子动力学模拟是研究蛋白质结构变化的有效方法, 然而使用分子动力学模拟对蛋白质的构象空间进行采样需要花费很长的时间. 近年来的一些研究表明, 使用简单的机器学习模型——自编码器及其改进型, 可以在有限采样的情况下, 快速完成对蛋白质构象空间的探索. 该模型通过训练神经网络, 完成对隐变量的提取, 同时根据其产生构象, 但是由于提取出的隐变量没有直观的含义, 探索构象空间的方向会受到影响. 本工作通过引入反应坐标 (如质心距离等), 建立了一个中间层受监督的自编码器模型, 以解决上述问题. 该模型应用于噬菌体 T4 溶菌酶和腺苷酸激酶两个蛋白质分子, 结果表明, 仅使用短时间分子动力学模拟作为训练数据, 就可以探索到这两种蛋白分子的多种典型构象. 有监督 (合理的反应坐标或者实验数据等) 的自编码器模型有望成为探索蛋白质构象空间的有效工具.

关键词: 蛋白质构象空间, 分子动力学模拟, 机器学习, 自编码器**PACS:** 87.15.ap, 87.15.hp**DOI:** 10.7498/aps.72.20231060

1 引言

蛋白质的功能与其结构和动态构象变化密切相关^[1]. 为了获得蛋白质分子的结构, 人们开发了各种实验和预测技术. X 射线晶体衍射^[2]和冷冻电镜技术^[3]可以解析高分辨率的蛋白质分子结构, 而核磁共振^[4]可以提供分子中的原子距离等信息. 此外, 小角 X 射线散射^[5]、化学交联^[6]和荧光共振能量转移^[7]等技术可以从不同的角度给出蛋白质分子的各种结构信息. 基于人工智能的结构预测方法, 如 AlphaFold2^[8]和 RoseTTAFold^[9], 可以直接根据氨基酸序列预测蛋白质的结构. 这些方法在获取蛋白质静态结构时十分有效, 但是不易得到蛋白质的动态变化信息.

计算模拟方法, 例如分子动力学 (molecular dynamics, MD) 模拟, 是研究蛋白质分子动态变化的重要工具^[10]. MD 方法用半经验的能量函数来描述原子间的相互作用, 在经典力学的框架下对蛋白质分子进行模拟. 从一个已知的分子结构出发, 通过迭代求解运动方程, 得到分子动态变化的过程. 为了确保结果的可靠性, 通常要求对整个构象空间充分采样. 但由于分子模拟的结果服从玻尔兹曼统计, 在生理条件下, 对高能构象的采样十分困难, 这一问题通常需要引入增强采样等方法来解决^[11]. 模拟的另一个问题来自分子力场, 它是对分子间相互作用的一种近似描述, 因而必然存在一定的误差. 力场选择不合适可能会导致模拟结果表现出与实际情况不同的倾向^[12], 即使经过大量计算后达到了充分采样的要求, 也无法正确描述生物大分子

* 国家重点研发计划 (批准号: 2021YFA1301504)、国家自然科学基金 (批准号: 91953101) 和中国科学院战略性先导科技专项 (B类)(批准号: XDB37040202) 资助的课题.

† 通信作者. E-mail: zzyzhang@ustc.edu.cn

的动态变化. 这种情况下, 可以先尽可能多地产生不同的构象, 再验证其合理性.

近年来, 机器学习方法的快速发展为解决分子模拟中的采样和力场问题提供了新思路^[13,14]. 自编码器是一种生成神经网络, 最初用于计算机图形领域^[15], 目前也应用于探索蛋白质分子的构象空间^[16]. 自编码器由编码器和解码器组成, 高维的蛋白质结构信息经过编码器压缩得到低维空间的隐变量, 再经过解码器重构出蛋白质结构, 同时要求重构的结构与输入的结构尽可能一致. 训练完成后, 只需要向解码器输入随机数据, 就可以构建出不同的蛋白质构象. 由于自编码器在训练过程中只要求数据成功重构, 中间层的隐变量没有明确的含义, 而构象生成是从中间层的数据开始的, 因此探索构象空间的方向也是不确定的, 有时可以找到各种不同的构象, 有时只能得到不感兴趣或不合理的构象. 为了解决上述问题, 一种常用的方案是对中间层的结果进行一些限制.

本研究设计了一个有监督的自编码器模型. 将一些反应坐标引入到自编码器中, 要求其在重构蛋白质结构的同时, 中间层的数据要与给定的反应坐标接近, 从而使构象空间的探索在给定的方向上进行. 将该模型运用到两个多结构域蛋白, 噬菌体 T4 溶菌酶和腺苷酸激酶, 探索得到的蛋白质构象空间覆盖了目前已知的实验结构. 通过引入合理的反应坐标和实验数据, 建立有监督的自编码器模型, 有望成为探索蛋白质构象空间的有效工具.

2 方法

2.1 中间层受监督的自编码器模型

为了实现在给定方向的构象空间探索, 使用 Pytorch2.0 设计了一个中间层受监督的自编码器(图 1). 该模型的整体结构与普通的自编码器相似, 由编码器和解码器组成. 其中编码器是一个多层的全连接神经网络, 在输入层之后每一层的维数分别是 2048, 512, 128, 32, 8, 2, 解码器也是多层全连接神经网络, 其结构与编码器对称, 每一层的维数依次是 2, 8, 32, 128, 512, 2048, 输出层的维数与编码器输入层相同. 除了最后一层外, 编码器和解码器的每一层都使用了 ReLU 作为激活函数, 而最后一层则使用 Sigmoid 激活函数, 以控制输出结果的范围. 这一模型的参数量很少, 对计算资源的要求较低.

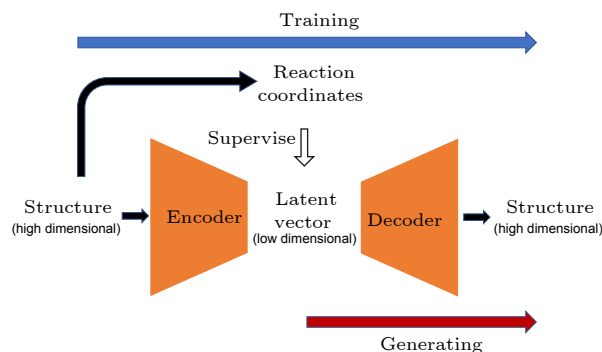


图 1 中间层受监督的自编码器示意图

Fig. 1. Schematic of supervised-AE.

不同于无监督的自编码器, 将监督学习引入自编码器的中间层, 训练时使用的损失函数形式如下:

$$L = \mathcal{L}_{\text{output}} + \omega \mathcal{L}_{\text{middle}}, \quad (1)$$

其中 $\mathcal{L}_{\text{output}}$ 为输出层的损失函数, 用来描述重构后的结构与输入结构之间的差距; $\mathcal{L}_{\text{middle}}$ 为中间层的损失函数, 描述中间层数据与输入结构对应的反应坐标之间的差距. 只使用反应坐标往往不能准确地描述和重构整个分子结构, 只能反映结构的某些特征, 因此模型需要在正确提取反应坐标和成功重构分子结构之间找到平衡. 引入了权重因子 ω 来调整两者对损失函数的贡献, ω 较大时, 中间层对损失函数的贡献更大, 模型会倾向得到给定的反应坐标, 而重构分子结构的效果会变差, 反之, ω 较小时, 模型可以完成分子结构的重构, 但中间层的数值不一定接近给定的反应坐标. 本文中, 该因子的值设定为 100.

2.2 数据获取

训练模型的数据来自 MD 模拟. 模拟的体系分别是噬菌体 T4 溶菌酶 (T4 lysozyme, T4L) 和大肠杆菌腺苷酸激酶 (adenylate kinase, AdK). T4L 及其突变体在 PDB 数据库中有大量晶体结构, 其结构变化主要体现在 N 端结构域和 C 端结构域之间口袋的打开和关闭(图 2(a)). AdK 可以分为 CORE, LID 以及 AMPbd 三个结构域, 分别在 CORE 和 LID, 以及 CORE 和 AMPbd 之间形成两个口袋. 在酶的催化过程中, 口袋的打开和关闭十分重要(图 2(b)). 这两个蛋白分子的动态构象变化已经研究得比较充分, 适合用来验证我们的模型.

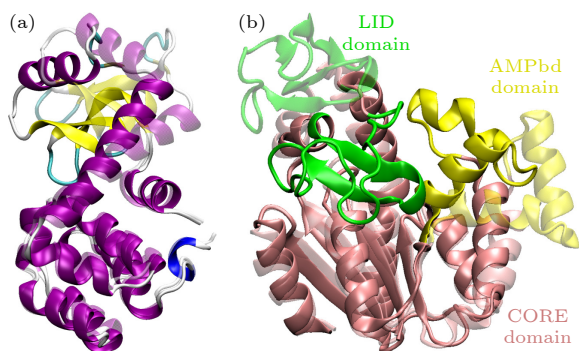


图2 本研究中使用的两种蛋白质分子的不同结构 (a) T4L的闭合(不透明)和打开(透明)结构,紫色为 α 螺旋,黄色为 β 折叠; (b) AdK的闭合(不透明)和打开(透明)结构,不同颜色表示不同的结构域

Fig. 2. Different structures of the two proteins in the work. (a) The close (opaque) and open (transparent) state of T4L. α -helix is colored in purple and β -sheet is colored in yellow. (b) The close (opaque) and open (transparent) state of AdK. Different domains are colored in different colors.

根据蛋白质分子的结构变化特征,计算相应的反应坐标作为监督引入到自编码器的中间层.从T4L及其突变体的晶体结构中选取能够反映其构象变化的41个结构,消除它们之间的平动和转动后,使用 C_{α} 原子的坐标进行主成分分析.特征值最大的2个主成分分别对应T4L的开闭和扭转运动,其占比分别为86%和6%.因此使用这2个主成分作为反应坐标,可以较好地描述T4L分子的运动^[17].AdK的结构变化主要表现为结构域的相对运动,因此可以选取CORE-LID和CORE-AMPbd结构域的质心距离作为反应坐标^[18].

分子动力学模拟使用GROMACS-2023版本进行^[19].从PDB数据库中分别选取T4L的打开(PDB编号2LZM^[20])和关闭(PDB编号178L^[21])结构,以及AdK的打开(PDB编号1AKE^[22])和关闭(PDB编号4AKE^[23])结构作为模拟的初始构象.为了验证模型是否受分子力场的影响,每一组模拟都分别使用了AMBER99SB力场/OPC水模型的组合^[24,25]以及CHARMM36m力场/TIP3P水模型的组合^[26].将分子放入正十二面体的周期性盒子中,同一分子的不同体系使用同样大小的盒子,以避免盒子尺寸对模拟结果的影响.向体系中填充水分子,并加入离子直到电荷平衡.先后用2000步最速下降法和1000步共轭梯度法进行能量最小化,然后在NPT系综下进行100 ps的位置约束MD模拟,以平衡系统的温度和压强,随后进行NPT模拟以获取训练模型的数据.AdK在没有

结合配体时无法维持关闭状态,因此在模拟中额外加入了结构域距离的位置限制.所有模拟的步长均为2 fs,使用LINCS算法约束氢原子参与的化学键,分别用V-rescale^[27]和C-rescale算法控制系统的温度和压强,非键相互作用中静电相互作用通过PME^[28]算法计算,范德瓦耳斯力则做截断处理,截断距离为1 nm.

由于不要求充分采样,每组用于产生训练数据的模拟仅进行100 ns,每10 ps保存一个结构,共保存10000个.消除不同结构之间的平动和转动变化后,提取主链部分的原子,即N, C_{α} , C, O的笛卡尔坐标作为模型的输入,同时计算出每个结构的反应坐标作为标签.在开始训练之前,还需要对数据进行归一化处理,数据的每一个维度都分别被放缩到0.2与0.8之间,这一区间Sigmoid函数的斜率较大,有利于模型训练更快达到收敛.

2.3 利用有监督的自编码器探索蛋白质构象空间

将模拟轨迹整理为数据集,从中随机取出80%作为训练集,剩余的20%作为测试集.以平方误差作为损失函数,用Adam优化器^[29]进行训练,遍历训练集500次,初始学习率为 1×10^{-4} ,并随着遍历次数以 1×10^{-8} 的速率减小.完成训练后,在 $[0.05, 0.95] \times [0.05, 0.95]$ 的范围内均匀选取10000个点作为自编码器中间层隐空间的数据点,将这些点输入解码器构建出对应的蛋白质分子主链结构.模型训练和数据生成的相关运算在RTX 3090Ti上运行.

由于生成的结构并不总是合理的,通过两种判据对其进行筛选.其一是蛋白质的主链二面角取值需要满足一定的规律,这一规律通常用Ramachandran图来描述,将大量已知蛋白质结构的Ramachandran图的统计结果^[30]作为参考,与模型生成的蛋白质结构的Ramachandran图进行比较,若90%以上处于合理区间,则认为该结构的主链二面角分布是合理的.其二是不同原子之间不能存在空间冲突,使用分子模拟工具OpenMM^[31]对分子结构进行一小段能量最小化,如果最终原子间的力比较小,就可以认为该分子不存在空间冲突.考虑到这一步需要频繁进行,与其他分子模拟工具相比,使用直接运行在Python中的OpenMM可以节省大量用于初始化模拟引擎的时间.由于模型仅

产生主链部分的原子坐标,并非完整的分子,用 ParmEd 工具^[32]将力场参数中非主链的部分删去,同时将所有原子的电荷设置为 0,在能量最小化时仅保留化学键和范德瓦耳斯力.能量最小化不仅可以筛选掉明显不合理的结构,还可以对结构中的一些键长键角的错误进行修正.

模拟得到的构象空间分布十分有限,在此基础上进行构象空间探索也因此受到限制.为了进一步扩大构象空间探索的范围,将模型生成的结构与原有数据集的一半合并成新的数据集,并重复进行模型训练和构象空间探索.随着探索范围逐渐扩大,模型生成的不合理结构逐渐增加,构象空间的探索效率也随之下降,因此只重复上述流程 3 次.

3 结果与讨论

3.1 T4L 构象空间探索结果

以 T4L 的模拟轨迹作为训练集,进行训练以及构象空间探索,整个流程耗时仅 20 min.探索结果如图 3 所示,由于使用不同力场得到的模拟轨迹不同,构象空间探索的区域也有所不同,整体上看使用 AMBER99SB 力场/OPC 水模型的探索范围更大.不过使用两种力场得到的探索范围都可以覆盖包括所有参考晶体结构在内的训练集附近的区域,例如可以找到与 PDB 编号为 173L 晶体结构^[21]十分相似的构象(图 4(a)),RMSD 为 0.7 Å.此外,探索结果中还可以看到大幅度的构象变化,例如闭合状态与打开状态的不同(图 4(b)),以及两个结构域的相对转动角度不同(图 4(c)).

虽然模型生成的结构都通过了二面角分布的检验,以及键长键角和空间冲突的修正,但依然存在一些不合理的情况,如生成的结构中二级结构含量显著低于晶体结构和模拟轨迹中二级结构的含量.为了验证模型产生结构的合理性,我们使用 kmeans 算法,根据反应坐标将探索结果分为 50 组,取每一组最靠近中心的构象作为代表,用 tleap 补全侧链,然后进行 100 ns 约束 C_α 原子的 MD 模拟,从而在不改变反应坐标的情况下修复二级结构.除少数情况由于侧链存在空间冲突而失败外,大部分代表构象的二级结构得到修复(图 5(a)和图 5(b)),DSSP^[33]计算表明修复后二级结构含量基本可以接近模拟轨迹的水平(图 5(c)).还计算了每个代表构象与同组各构象的主链 RMSD,

所有 RMSD 数值都小于 2 Å(图 5(a)和图 5(b)),这说明二级结构的缺失只是由一些局部的偏差

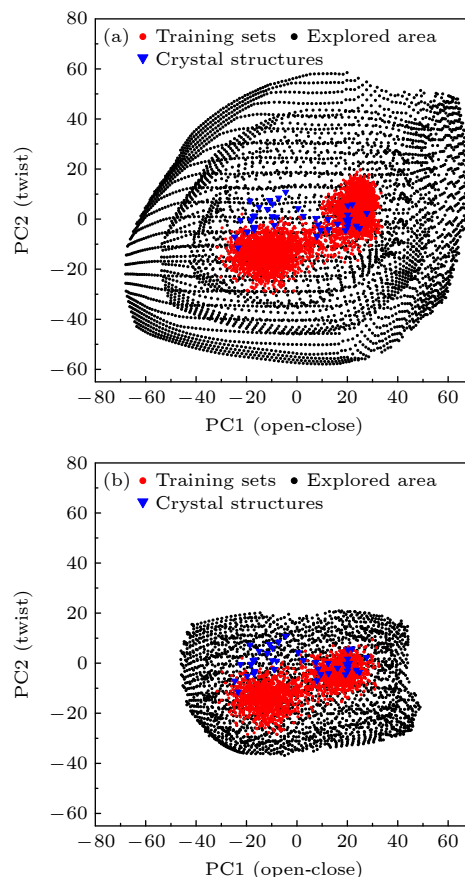


图 3 T4L 的构象空间探索结果 (a) 使用 AMBER99SB 力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型

Fig. 3. Results of conformational space exploration of T4L: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P.

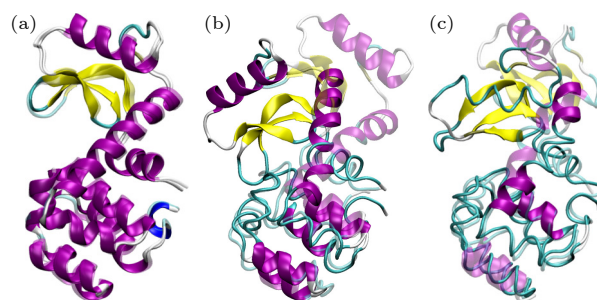


图 4 探索到的不同 T4L 构象 (a) PDB:173L 的晶体结构(不透明)与探索到的相似结构(透明); (b) 开合程度不同的两个构象; (c) 扭曲情况不同的两个构象; 紫色为 α 螺旋, 黄色为 β 折叠

Fig. 4. Different T4L conformations explored: (a) PDB:173L (opaque) and a similar structure explored; (b) two conformations with different degrees of opening and closing; (c) two conformations with different degrees of twisting. α -helix is colored in purple and β -sheet is colored in yellow.

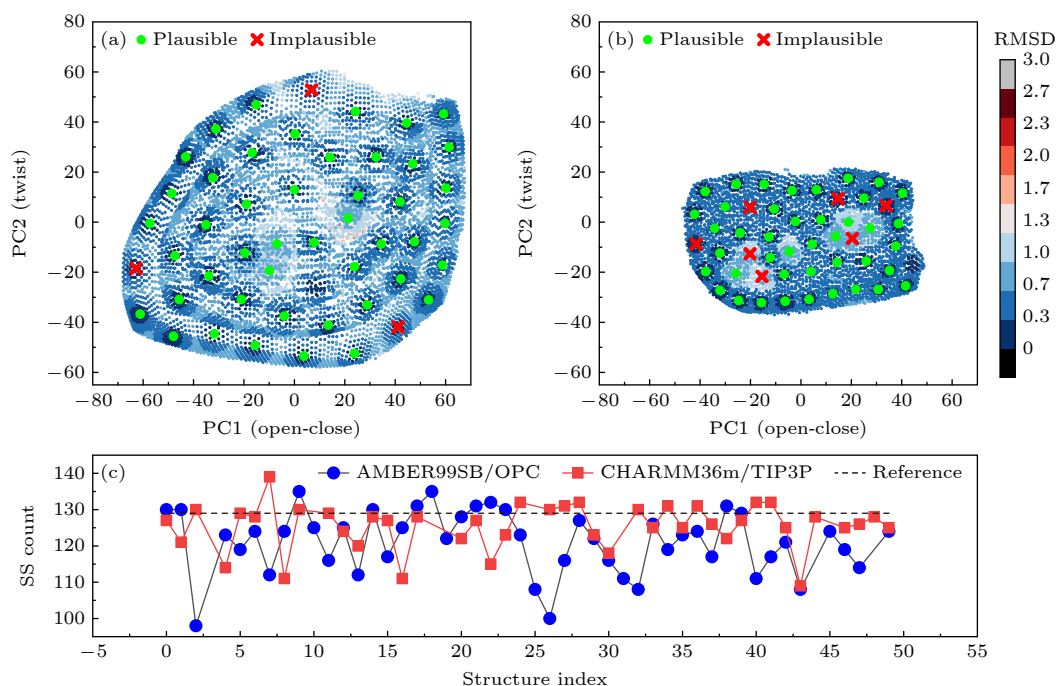


图 5 T4L 构象探索结果的合理性检验 (a) 使用 AMBER99SB 力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型; (c) 修复后各代表构象的二级结构含量, 参考值为模拟轨迹的平均值
 Fig. 5. Plausibility check of T4L conformational exploration results: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P; (c) secondary structure counts of each representative conformation after fixing, the reference is the average value of the simulated trajectory.

导致的, 模型生成的大多数结构都可以通过简单修正得到合理的结果, 而侧链可能存在空间冲突的情况则需要进一步改进模型来解决。

在上述流程中, 闭合与打开两段模拟轨迹都被用于模型的训练. 还测试了仅使用打开状态的模拟轨迹训练的情况 (图 6), 虽然探索区域由于训练集减少而缩小, 但是仍然可以覆盖包括闭合状态在内的各个晶体结构。

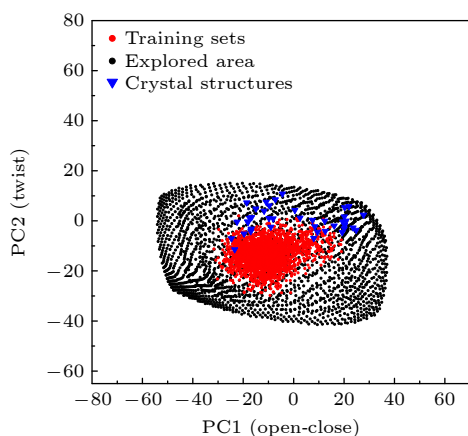


图 6 仅从打开状态出发的 T4L 构象探索结果
 Fig. 6. Results of T4L conformational exploration from the open state only.

3.2 AdK 构象空间探索结果

以 AdK 的模拟轨迹作为训练集, 进行训练以及构象空间探索. 结果如图 7 所示, 除了训练集中包含的完全关闭和完全打开状态外, 还可以从中找到 LID 结构域单独打开 (图 8(a)) 和 AMPbd 结构域单独打开的结构 (图 8(b)).

对 AdK 构象探索结果的合理性进行了检验, 结果如图 9 所示. 在使用 CHARMM36m 力场/TIP3P 水模型时, 修复后二级结构含量与模拟轨迹相当, 而使用 AMBER99SB 力场/OPC 水模型时, 虽然也能修复到较高的水平, 但与前者相比显著偏低. 这表明与 CHARMM36m 相比, AMBER99SB 力场/OPC 水模型的组合使蛋白质结构更加容易发生变化, 探索构象空间的范围更大, 同时二级结构也会有一定的破坏, 更适用于柔性较强的蛋白质分子。

值得注意的是, 大部分构象与其所在组的中心构象之间的 RMSD 较小, 除少数不合理构象外, 大部分 RMSD 较大的构象都在模拟产生的训练集中. 这意味着模型产生的构象仅包含反应坐标相关的信息, 而在与反应坐标正交的自由度上没有表现

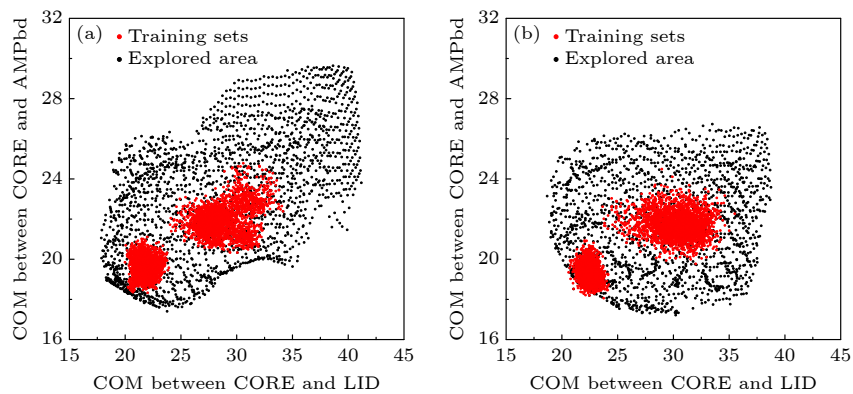


图 7 AdK 的构象空间探索结果 (a) 使用 AMBER99SB 力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型
Fig. 7. Results of conformational space exploration of AdK: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P.

出差异. 这是由自编码器自身的性质决定的, 对于相同的输入总是会给出相同的输出, 而实际上如模拟轨迹反映的一样, 相同的反应坐标下, 构象仍应该有一定的变化空间, 这些空间是自编码器无法探索的. 因此, 反应坐标的选取对该模型的效果至关重要. 若要解决这一问题, 可以将自编码器换成变分自编码器, 学习构象系综而非单个分子的特征, 从而体现相同反应坐标下的差异.

以上结果是使用常规的自编码器难以获得的. 将引入反应坐标监督的自编码器换成无监督的自

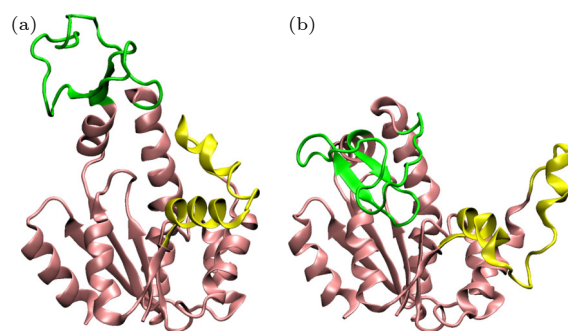


图 8 探索到的不同 AdK 构象
Fig. 8. Different AdK conformations explored.

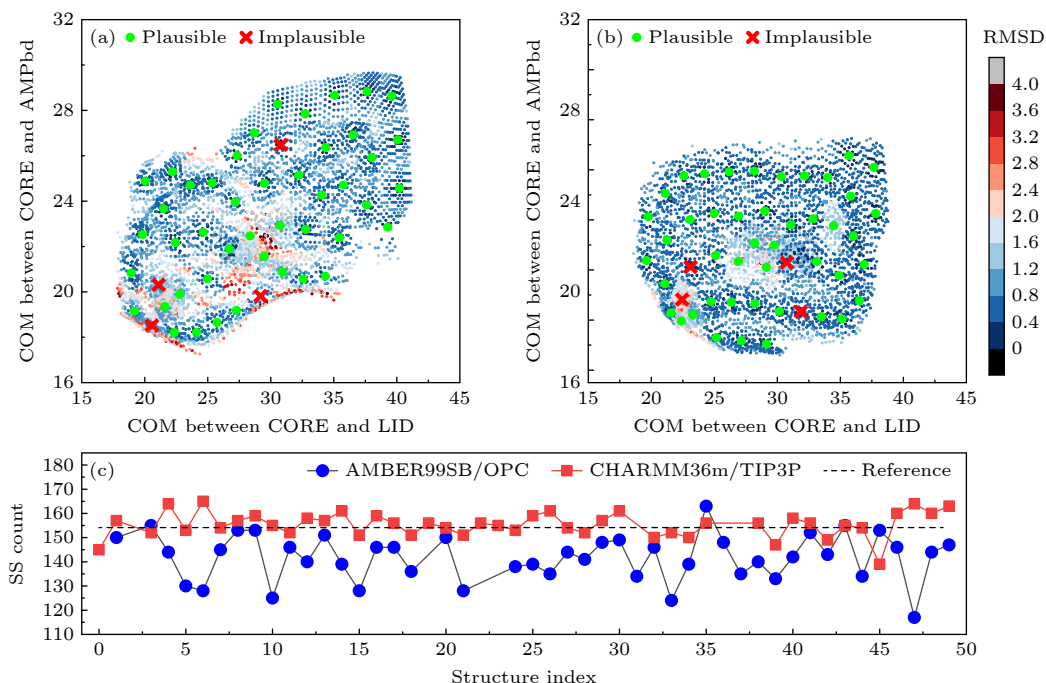


图 9 AdK 构象探索结果的合理性检验 (a) 使用 AMBER99SB 力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型; (c) 修复后各代表构象的二级结构含量, 参考值为模拟轨迹的平均值

Fig. 9. Plausibility check of AdK conformational exploration results: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P; (c) secondary structure counts of each representative conformation after fixing, the reference is the average value of the simulated trajectory.

编码器, 对 AdK 的构象空间进行探索, 结果如图 10 所示. 自编码器需要从训练集中学习反应坐标, 这在采样不足的情况下非常困难. 通常情况下, 自编码器只能提取两组轨迹的差异, 并完成对两种状态之间的构象空间探索, 但是无法探索其他区域, 例如图 8 所示的单个结构域打开的构象. 引入反应坐标作为监督的改进, 使得自编码器不再需要提取反应坐标, 从而可以在采样不足的情况下工作.

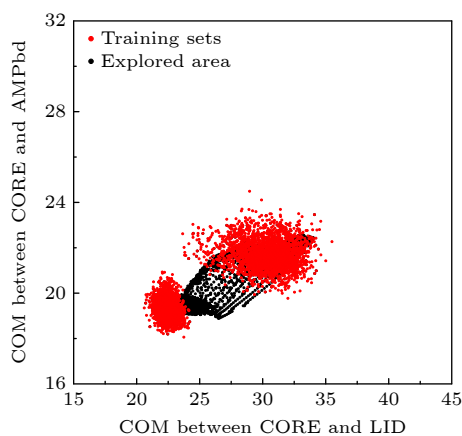


图 10 使用普通自编码器探索 AdK 的构象空间

Fig. 10. Exploring the conformational space of AdK with a common self-encoder.

4 结 论

本文对使用自编码器探索蛋白质构象空间的方法进行了改进, 将监督学习引入自编码器的中间层, 并使用改进后的方法对 T4L 和 AdK 的构象空间进行探索, 达到了预期的效果. 结果表明这一改进使该方法可以在有限采样的情况下, 仅使用很少的计算资源, 就可以大范围探索蛋白质的构象空间.

虽然模型只能生成构象, 并不能给出构象的生物学意义以及动力学过程, 但是如果对特定体系引入实验信息, 就可以筛选出具有生物学意义的构象, 以便进行下一步的研究. 对于实验信息较少的蛋白质分子, 可以直接通过模型生成有潜在研究价值的构象, 然后从这些构象出发进行 MD 模拟, 研究蛋白质分子的动态过程, 进而预测可能的生物学意义. 这种策略与仅依靠 MD 模拟的构象空间采样相比, 效率更高.

在测试模型时, 发现了进一步的改进空间. 通过对模型生成构象的筛选和修正, 可以确保构象的

合理性, 但同时也降低了生成构象的效率. 考虑直接将对构象合理性的要求引入模型的损失函数中, 从而省去筛选和修正的过程. 由于模型中只有蛋白质的主链部分, 有可能出现侧链不合理情况, 需要对不同氨基酸残基做不同修正或在模型中使用完整的蛋白质分子. 对于模型生成的构象无法表现出反应坐标之外变化的问题, 可以尝试使用变分自编码器. 最后, 反应坐标决定了构象空间探索的方向, 结合实验数据选取合适的反应坐标对模型的效果十分重要. 基于这些思路, 将继续对该模型进行发展和完善.

感谢中国科学技术大学超算中心张运动提供的硬件和软件技术支持.

参考文献

- [1] Chu X, Gan L, Wang E, Wang J 2013 *Proc. Natl. Acad. Sci. U.S.A.* **110** E2342
- [2] Smyth M S, Martin J H 2000 *Mol. Pathol.* **53** 8
- [3] Danev R, Yanagisawa H, Kikkawa M 2019 *Trends Biochem. Sci.* **44** 837
- [4] Vincenzi M, Mercurio F A, Leone M 2021 *Curr. Med. Chem.* **28** 2729
- [5] Kachala M, Valentini E, Svergun D I 2015 *Adv. Exp. Med. Biol.* **870** 261
- [6] Chu F, Thornton D T, Nguyen H T 2018 *Methods* **144** 53
- [7] Bhaumik S R 2021 *Emerg. Top Life Sci.* **5** 49
- [8] Junper J, Evans R, Pritzl A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, Bridgland A, Meyer C, Kohl S A A, Ballard A J, 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 A W, Kavukcuoglu K, Kohli P, Hassabis D 2021 *Nature* **596** 583
- [9] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee G R, Wang J, Cong Q, Kinch L N, Schaeffer R D, Millán C, Park H, Adams C, Glassman C R, DeGiovanni A, Pereira J H, Rodrigues A V, van Dijk A A, Ebrecht A C, Opperman D J, Sagmeister T, Buhheller C, Pavkov-Keller T, Rathinaswamy M K, Dalwadi U, Yip C K, Burke J E, Garcia K C, Grishin N V, Adams P D, Read R J, Baker D 2021 *Science* **373** 871
- [10] Karplus M, Kuriyan J 2005 *Proc. Natl. Acad. Sci.* **102** 6679
- [11] Bernardi R C, Melo M C R, Schulten K 2015 *Biochim. Biophys. Acta* **1850** 872
- [12] Mu J, Liu H, Zhang J, Luo R, Chen H F 2021 *J. Chem. Inf. Model.* **61** 1037
- [13] Lemke T, Peter C 2019 *J. Chem. Theory Comput.* **15** 1209
- [14] Zhu J, Wang J, Han W, Xu D 2022 *Nat. Commun.* **13** 1661
- [15] Hinton G E, Salakhutdinov R R 2006 *Science* **313** 504
- [16] Degiacomi M T 2019 *Structure* **27** 1034
- [17] Wen B, Peng J, Zuo X, Gong Q, Zhang Z 2014 *Biophysical J.* **107** 956

- [18] Giri Rao V V H, Gosavi S 2014 *PLOS Computational Biology* **10** e1003938
- [19] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B, Lindahl E 2015 *SoftwareX* **1–2** 19
- [20] Weaver L H, Matthews B W 1987 *J. Mol. Biol.* **193** 189
- [21] Zhang X J, Wozniak J A, Matthews B W 1995 *J. Mol. Biol.* **250** 527
- [22] Müller C W, Schulz G E 1992 *J. Mol. Biol.* **224** 159
- [23] Müller C W, Schläuderer G J, Reinstein J, Schulz G E 1996 *Structure* **4** 147
- [24] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C 2006 *Proteins Struct. Funct. Bioinf.* **65** 712
- [25] Izadi S, Anandakrishnan R, Onufriev A V 2014 *J. Phys. Chem. Lett.* **5** 3863
- [26] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot B L, Grubmüller H, MacKerell A D 2017 *Nat. Methods* **14** 71
- [27] Bussi G, Donadio D, Parrinello M 2007 *J. Chem. Phys.* **126** 014101
- [28] Essmann U, Perera L E, Berkowitz M L, Darden T A, Lee H C, Pedersen L G 1995 *J. Chem. Phys.* **103** 8577
- [29] Kingma D P, Ba J 2014 arXiv:1412.6980 [cs.LG]
- [30] Lovell S C, Davis I W, Arendall III W B, de Bakker P I W, Word J M, Prisant M G, Richardson J S, Richardson D C 2003 *Proteins Struct. Funct. Bioinf.* **50** 437
- [31] Eastman P, Swails J, Chodera J D, McGibbon R T, Zhao Y, Beauchamp K A, Wang L P, Simmonett A C, Harrigan M P, Stern C D, Wiewiora R P, Brooks B R, Pande V S 2017 *PLoS Comput. Biol.* **13** e1005659
- [32] Shirts M R, Klein C, Swails J M, Yin J, Gilson M K, Mobley D L, Case D A, Zhong E D 2017 *J. Comput. -Aided Mol. Des.* **31** 147
- [33] Touw W G, Baakman C, Black J, te Beek T A, Krieger E, Joosten R P, Vriend G 2015 *Nucleic Acids Res.* **43** D364

SPECIAL TOPIC—Machine learning in biomolecular simulations • COVER ARTICLE

Exploring protein's conformational space by using encoding layer supervised auto-encoder*

Chen Guang-Lin Zhang Zhi-Yong[†]*(Department of Physics, University of Science and Technology of China, Hefei 230026, China)*

(Received 28 June 2023; revised manuscript received 29 July 2023)

Abstract

Protein function is related to its structure and dynamic change. Molecular dynamics simulation is an important tool for studying protein dynamics by exploring its conformational space, however, conformational sampling is a nontrivial issue, because of the risk of missing key details during sampling. In recent years, deep learning methods, such as auto-encoder, can couple with MD to explore conformational space of protein. After being trained with the MD trajectories, auto-encoder can generate new conformations quickly by inputting random numbers in low dimension space. However, some problems still exist, such as requirements for the quality of the training set, the limitation of explorable area and the undefined sampling direction. In this work, we build a supervised auto-encoder, in which some reaction coordinates are used to guide conformational exploration along certain directions. We also try to expand the explorable area by training through the data generated by the model. Two multi-domain proteins, bacteriophage T4 lysozyme and adenylate kinase, are used to illustrate the method. In the case of the training set consisting of only under-sampled simulated trajectories, the supervised auto-encoder can still explore along the given reaction coordinates. The explored conformational space can cover all the experimental structures of the proteins and be extended to regions far from the training sets. Having been verified by molecular dynamics and secondary structure calculations, most of the conformations explored are found to be plausible. The supervised auto-encoder provides a way to efficiently expand the conformational space of a protein with limited computational resources, although some suitable reaction coordinates are required. By integrating appropriate reaction coordinates or experimental data, the supervised auto-encoder may serve as an efficient tool for exploring conformational space of proteins.

Keywords: protein conformational space, molecular dynamics simulation, machine learning, auto-encoder**PACS:** 87.15.ap, 87.15.hp**DOI:** [10.7498/aps.72.20231060](https://doi.org/10.7498/aps.72.20231060)

* Project supported by the National Key Research and Development Program of China (Grant No. 2021YFA1301504), the National Natural Science Foundation of China (Grant No. 91953101), and the Strategic Priority Research Program (B) of the Chinese Academy of Sciences (Grant No. XDB37040202).

[†] Corresponding author. E-mail: zzyzhang@ustc.edu.cn

专题: 生物分子模拟中的机器学习

蛋白质计算中的机器学习*

张嘉晖†

(中国科学技术大学生命科学学院, 合肥 230027)

(2023 年 10 月 7 日收到; 2024 年 1 月 4 日收到修改稿)

蛋白质计算一直以来都是科学领域中的重要课题, 而近年来其与机器学习的结合, 更是极大地推进了相关学科的发展. 本综述主要讨论了机器学习在四个重要的蛋白质计算领域内的研究进展, 这四个领域包括: 分子动力学模拟、结构预测、性质预测和分子设计. 分子动力学模拟依赖于力场参数, 准确的力场参数是分子动力学模拟的必需品, 而机器学习可以帮助研究者得到更加准确的力场参数. 在分子动力学模拟中, 机器学习也可以从复杂的体系中以较小的代价计算出所需求解的自由能. 结构预测一般是给定蛋白质序列预测其结构. 结构预测复杂度高、数据量大, 而这恰恰是机器学习所擅长的. 在机器学习的协助下, 近年来科研人员已经在单个蛋白质三维结构预测上取得了不错的成果. 性质预测则是指通过给定的已知蛋白质信息, 推断其可能拥有的性质, 这对于蛋白质的研究也是至关重要的. 更具挑战性的是分子设计, 虽然近年来机器学习在蛋白质设计上取得突破, 但这一领域还有很大空间值得探索. 本综述将针对以上四点分别展开论述, 并对蛋白质计算中的机器学习研究进行展望.

关键词: 蛋白质, 机器学习, 分子动力学模拟, 结构预测, 性质预测, 分子设计

PACS: 93.85.Bc, 31.15.-p, 87.19.Pp

DOI: 10.7498/aps.73.20231618

1 引言

蛋白质 (protein) 是生命的关键物质基础之一. 研究它们对理解生命体系、探究生命进程和治疗疾病有着重大意义^[1-3]. 由于时间与空间尺度、复杂度和可控性以及实验成本等原因, 只依靠实验方法对蛋白质进行研究是不够的, 用计算方法对蛋白质的研究可弥补实验研究的不足^[4,5]. 对蛋白质实施计算研究主要有四种目的: 研究蛋白质的结构、运动或相互作用细节 (通常是通过分子动力学模拟)^[6]; 给定蛋白质的序列来预测其空间结构^[7]; 给定蛋白质的序列等信息来预测某些重要性质^[8]; 以及设计满足一定条件或功能的蛋白质^[9]. 这四个领域在近年来彼此融合, 相辅相成, 使得蛋白质计算研究达到了一个新的高度^[10,11], 被人们寄予了厚望. 然而,

因其具有时间与空间尺度大、复杂度高和数据量大等特点, 发展计算蛋白质研究仍然是一项具有挑战性的任务^[12-16].

另一方面, 近年来机器学习 (machine learning) 的迅速崛起已对许多领域产生了深远的影响^[17-19]. 机器学习是人工智能 (artificial intelligence, AI) 的一个重要分支, 通过使用算法让计算机系统从数据中学习和改进, 而无需明确编程^[17]. 机器学习利用模型对输入数据的解析和理解, 从而进行预测、决策或生成, 而不仅仅是按照严格定义的任务指令执行^[17]. 机器学习任务有多种类型, 包括监督学习、无监督学习、半监督学习和强化学习. 在监督学习中, 算法从标记的训练数据中学习, 然后将所学知识应用于新的、未见过的数据^[20]. 在无监督学习中, 算法通过在没有事先标签的数据中寻找隐藏的结构或关系来进行学习^[21]. 半监督学习介

* 国家自然科学基金 (批准号: 22177107) 资助的课题.

† 通信作者. E-mail: jhzhang@ustc.edu.cn

于这两者之间,当部分数据被标记时就会使用^[22]. 强化学习涉及到一个智能体,它通过与环境的交互和反馈来学习最佳行为策略^[23]. 深度学习是机器学习的一种特殊形式,它基于人工神经网络,并借鉴了人脑神经元连接的方式^[24]. 深度学习可以处理大规模、高维度的数据,包括图片、音频和文本等,已广泛应用于图像识别、自然语言处理、语音识别以及许多其他领域^[25]. 机器学习正在计算蛋白质研究领域内发挥着越来越重要的作用,这是因为机器学习是一种数据驱动的方法,它具有处理大规模、复杂性和高维度数据的独特能力,这使得机器学习在解决传统蛋白质计算中的一些问题方面具有优势^[26]. 机器学习与蛋白质计算的结合可以加速人类理解生命、改造生命的过程.

本综述介绍机器学习在蛋白质的分子动力学模拟(第2节)、蛋白质的结构预测(第3节)、蛋白质的性质预测(第4节)和蛋白质的分子设计(第5节)四方面的研究进展,并对机器学习与蛋白质计算结合进行了总结与展望(第6节). 首先讨论如何使用机器学习技术优化和解析分子动力学模拟,这可以帮助人们更加深入地了解蛋白质的动态结构. 随后,探讨如何利用机器学习进行准确的蛋白质结构预测,这对于理解蛋白质的空间结构和功能至关重要. 接下来,探究机器学习在给定蛋白序列情况下对蛋白性质的预测. 第5节则聚焦于如何在复杂的蛋白质分子设计工程问题上应用机器学习. 蛋白质的功能通常通过其动态结构决定,而不仅仅依赖于静态结构. 因此,结构预测与动力学模拟的融合正在成为一个重要的研究方向^[10]. 例如,预测出的蛋白质结构可以作为动力学模拟的初始结构,以探索蛋白质的动态行为和活性状态. 借助分子动力学模拟,科学家们可以更直观地了解分子间的相互作用,从而优化新设计的蛋白质分子. 同时,机器学习方法也被用于动力学模拟的数据分析,以指导新分子的设计^[27]. 而理解蛋白质的结构是设计新药物或调控其功能的关键,将结构预测与分子设计相结合,可以帮助我们更好地理解靶点分子的结构特性,并据此设计出高效的候选药物^[28]. 最后,设计出的蛋白序列必须满足一些必要的性质要求,例如水溶性和免疫原性^[29,30]. 因此机器学习在这四个领域内的应用不仅促进了各自领域的发展,也促进了这四个领域走向融合,协同发展. 结构预测、性质预测、分子设计和动力学模拟之间的交叉融合为我

们提供了在原子分辨水平全面解析生物现象的可能,使我们能够在多个层次上理解和操纵生物系统. 第6节总结并展望了机器学习与蛋白质计算结合的未来,强调了跨领域融合的重要性,并展望了未来可能的研究方向和挑战. 笔者认为,机器学习算法的进步和生物大数据的快速增长,将在更深、更广泛的层面上推动这四个领域的融合与协同发展,从而开启新的科学发现和应用的可能.

2 分子动力学模拟中的机器学习

分子动力学模拟是一种通过计算遵从牛顿运动定律的多粒子系统(如蛋白质体系)的时间演化,以了解其物理性质的重要方法^[6]. 在分子动力学模拟中,分子被视为一组相互作用的粒子,通过数值仿真这些粒子随时间变化的轨迹,可以分析系统的宏观性质. 给定恰当的初始条件和相应的相互作用势能后,可通过数值求解牛顿运动方程实现模拟. 分子动力学模拟在多个领域有广泛的应用,包括但不限于物理、化学、生物学及材料科学. 例如,化学家可以利用分子动力学模拟预测反应途径^[31]; 物理学家则可能深入探究固态物理的世界^[32]; 生命科学研究人员能更好地理解蛋白质折叠和其他生物大分子的动态行为^[6,13,33]. 尽管分子动力学模拟拥有巨大的潜力,但也需要注意其局限性. 首先,分子动力学模拟的可信度取决于力场参数的准确性,而实际上人们很难用传统方法获取相对准确的力场参数. 机器学习的介入,对这些问题的解决起到了极大的帮助^[34,35]. 其次,对体系进行准确的自由能计算是一个很具挑战性的任务. 本节将针对机器学习与上述两点的结合,逐条展开论述,介绍相应的研究进展.

2.1 力场生成

在分子动力学模拟中,力场(force field)是一个至关重要的概念. 力场指的是一种用于描述和计算分子系统内各原子间相互作用力的数学模型^[36-38]. 具体来说,力场包含了各种类型的相互作用项,如键长、键角、二面角、范德瓦耳斯作用和静电作用等. 每种相互作用项都对应一个能量函数. 力场的总能量为所有相互作用项能量之和. 而在分子动力学模拟中,正是通过对力场给定的能量函数求导,而得到系统在这一时刻受的力,并据此得出分子系

统在下一时刻的位置和速度,从而模拟出分子的动态行为.传统的力场参数通常由第一性原理 (first principles)^[39] 计算和实验数据^[40] 得到,但由于复杂性、灵活性、适应性、时间效率等因素的制约,越发地需要机器学习帮助我们获取和优化力场参数^[35,41].

首先,我们指出,数据驱动的学习方法在蛋白质等生物分子研究领域内的核心思想和基于

第一性原理的量子力学方法是非常相似的^[42].如图1所示,机器学习和量子力学都经历了从准确而难以求解到近似而容易求解的蜕变.实际上,无论是量子力学,还是机器学习,如图1的上半部分所示,都在致力于应用数学工具对所需预测的量进行一个尽可能准确的预测,然而那将导致不可承受的计算量,于是人们分别对量子力学和机器学习做了近似,使它们能胜任复杂体系的计算(图1).而量

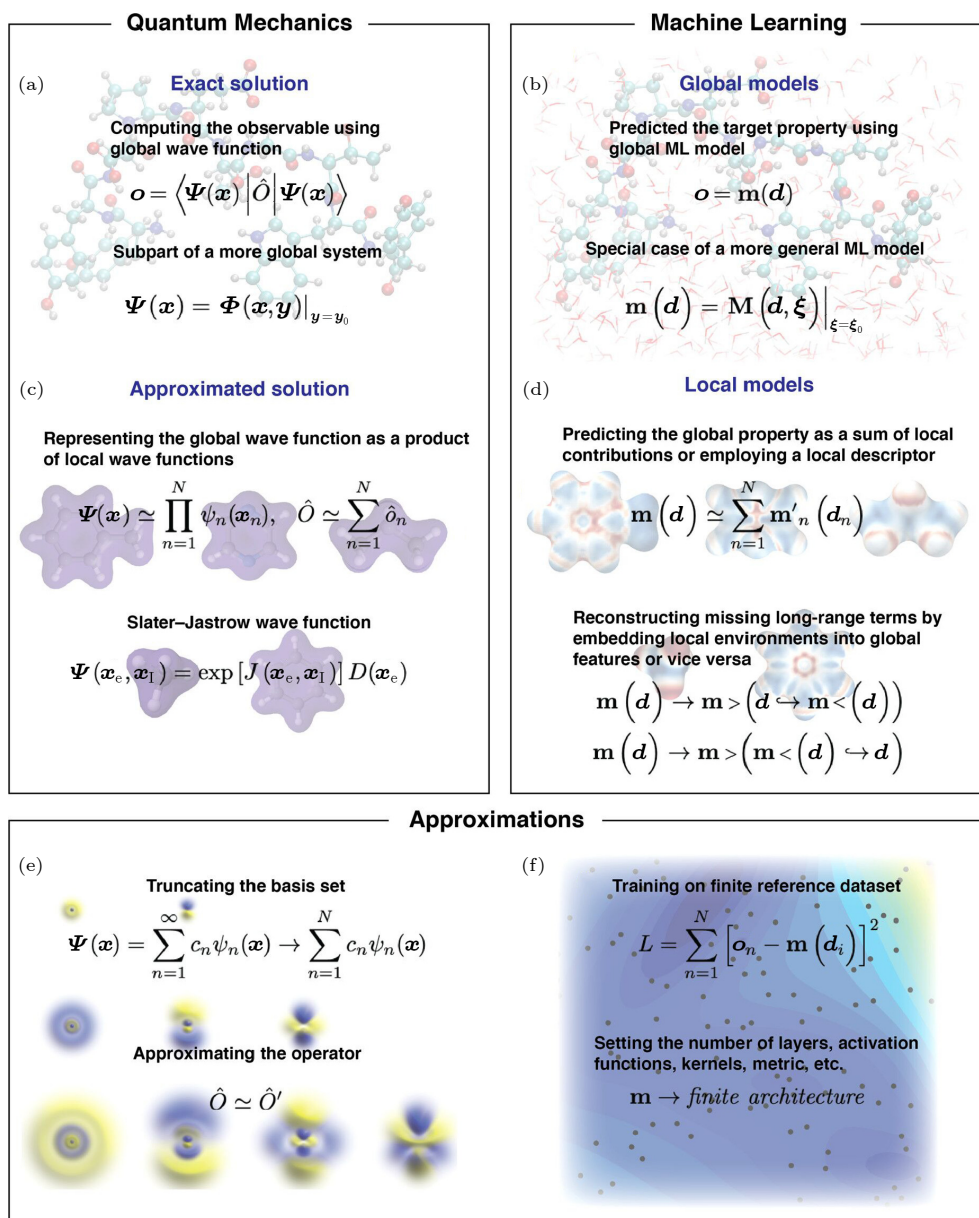


图1 量子力学与机器学习间的相似性. 从左到右, 从上到下的图片分别是: Chignolin 蛋白质在 (a) 无水环境和 (b) 有水环境下的情况, 使用 SchNet 模型得到的 (c) 可视化电荷密度和 (d) 局部化学势, (e) 氢原子的波函数以及 (f) Müller-Brown 势能. 图片引自文献^[42] (版权属于美国化学会)

Fig. 1. Similarity between quantum mechanics and machine learning. Images from left to right from top to bottom: Chignolin protein (a) without and (b) with the water environment, (c) visualized total charge densities and (d) local chemical potentials obtained using the SchNet model, (e) wave functions for hydrogen atom and (f) Müller-Brown potential. Reprinted with permission from Ref. ^[42] (Copyright 2021 American Chemical Society).

子力学和机器学习具体的近似法则, 都是从无限到有限, 从复杂到简单, 这说明了第一性原理计算和机器学习计算在原理和方法上的相关性. 具体而言, 如果取图中的 m 为能量, 那么训练出来的神经网络便可以作为一个力场使用. 用这种方法所生成的力场一般是平滑可微的, 这就使得原子受的力可求, 从而为机器学习生成的力场在分子动力学模拟中的应用提供了保障. 然而, 需要注意的是, 机器学习生成的力场有时是不满足能量守恒约束的, 使用机器学习生成能量守恒的分子力场目前仍是一个具有挑战性的课题^[35].

使用机器学习生成分子力场的一般步骤如下. 首先, 需要获取或生成一组训练数据. 这些数据应包含各种可能的分子构型和对应的能量及力. 数据可能来自实验测量、第一性原理计算或已有的经验力场模拟. 然后, 需要选择一种特征描述符来表示分子系统. 特征描述符应能够唯一且有效地描述分子的结构. 常见的特征描述符包括原子间距离、键角、二面角等. 接下来, 选择合适的机器学习模型(例如神经网络)并用前两步获得的数据进行训练. 在模型训练好之后, 进行优化和验证以确保其泛化能力. 优化可能涉及调整模型超参数、增加训练数据等. 验证通常通过将模型预测结果与独立的测试数据集进行比较来完成. 最后, 可以使用训练好的机器学习模型来生成新的力场. 这个力场将被用于更大规模或更长时间尺度的分子动力学模拟.

2.2 自由能计算

分子动力学模拟用于定量预测的一个核心任务是计算自由能^[31,43,44]. 自由能的定义式为

$$F(s) = -\frac{1}{\beta} \ln \left(\int dx \delta[s - s(x)] e^{-\beta U(x)} \right). \quad (1)$$

由(1)式可知, 自由能可以理解为反应路径上的加权平均势能. 研究体系的自由能或自由能变化对理解体系的状态和反应路径有举足轻重的作用^[45].

对于生物大分子体系, 结合自由能是一个经典而具有挑战性的课题^[46]. Bitencourt-Ferreira 和 de Azevedo^[47] 通过机器学习的方法, 对蛋白质-配体的结合吉布斯自由能 (Gibbs free energy) 进行了预测. 训练一个神经网络, 直接从复合物的原子坐标预测出结合自由能是极其困难的, 因此在该项研究工作中, 他们采用了 AutoDock Vina^[48] 的评分作为起点来预测蛋白质-配体复合物的吉布斯自

由能, 即训练一个神经网络, 输入 AutoDock Vina 的评分, 输出预测结合吉布斯自由能. 这篇工作的思路虽然简单, 但极大地提高了蛋白质-配体结合吉布斯自由能预测的准确性, 为结合蛋白的设计与筛选提供了一个更优的平台.

除了结合自由能之外, 反应自由能也是非常重要的研究方向^[49]. Pan 等^[50] 完成了一项运用机器学习预测酶反应自由能的工作. 该工作中, 研究者们结合了量子力学与分子动力学 (QM/MM)^[51], 通过构建一个神经网络, 将两者计算出的体系属性(电势、受力与坐标)输入至神经网络中, 并以此还原出体系能量和受力. 这么做的好处是, 通过少量相对昂贵的 QM/MM 计算, 使用神经网络拟合出能反映体系的动力学要素的量, 并在后续的工作中以计算成本较低的神经网络为基础进行化学反应的模拟. 该项工作中, 他们使用了雨伞采样 (umbrella sampling)^[43] 的方法构建反应路径并计算体系沿着反应路径的自由能.

机器学习在蛋白质相关的分子体系的自由能计算中还有着许多其他的应用. 2017 年 Riniker^[52] 提出了一种新的端点方法来预测溶解自由能和分配系数, 主要思路是: 对分子进行分子动力学模拟, 在不同环境(真空和溶剂)中提取一些属性, 如势能、体积等; 将每个属性的分布表示成指纹, 使用平均值、标准差和中位数. 2020 年 Bennett 等^[53] 结合分子动力学模拟和机器学习来预测小分子的自由能变化, 他们使用 MD 模拟计算了 15000 个小分子从水到环己烷的转移自由能变化, 作为机器学习模型的训练数据. 2021 年 Bertazzo 等^[54] 提出了一个结合增强采样、机器学习和定制算法的半自动化 workflow, 以计算配体-受体结合的平均势能和标准结合自由能, 该方法在主客体系和 GSK-3 β 蛋白-配体复合物上得到了验证. 这些应用不仅在各自所在的特定的科学研究领域做出了重要贡献, 更是推进了机器学习在自由能计算这一大方向的发展.

3 结构预测中的机器学习

在给定初始结构的情况下, 第 2 节中讨论的分子动力学模拟可以在蛋白质的研究中起到强大的作用. 然而, 在很多情况下, 我们仅仅知道蛋白质的序列, 而并不知道它们的结构. 这种现象主要被归结于检测技术的成熟度、条件苛刻度和对应的时间成本^[55]. 事实上, 我们知道的蛋白质序列信息要

远远多于蛋白质结构信息^[56]. 这时, 为了通过计算研究已知序列、未知结构的蛋白质的性质和行为, 就需要对具有该序列的蛋白质进行结构预测. 由于蛋白质的复杂度高, 使用机器学习预测其结构成为近年来一个潮流^[57]. 本节针对机器学习预测蛋白质的二级、三级和四级结构分别展开讨论.

3.1 二级结构预测

蛋白质的二级结构是由氢键稳定的规则结构, 这些氢键是在蛋白质的主链之间形成的. 研究生物大分子的二级结构具有重要的意义, 因为二级结构是构成三级和四级结构的基本元素, 且往往与生物大分子的功能密切相关. 而通过已知的一级结构信息, 可以预测其可能的二级结构, 这对于理解生物大分子的功能和进行分子设计都非常重要.

对于蛋白质分子, 尽管目前很多三级结构预测模型已经表现得足够好^[58-60], 但专注于二级结构预测仍然有其重要性和必要性. 与三级结构预测相比, 二级结构预测的计算成本较低. 对于大规模或复杂的蛋白质系统, 二级结构预测可能是更实用的选择; 二级结构是蛋白质功能的重要决定因素之一. 对二级结构的研究可以帮助我们更好地理解蛋白质的功能机制; 通过二级结构预测, 可以更好地理解蛋白质氨基酸序列与其结构之间的关系, 这对于蛋白质设计和工程也非常重要.

在蛋白质分子的二级结构机器学习预测中, 人们主要选取三种模式的神经网络: 循环神经网络 (recurrent neural network, RNN)^[61]、卷积神经网络

(convolutional neural network, CNN)^[62]与混合神经网络^[63](即结合了循环神经网络和卷积神经网络). 循环神经网络方法充分利用了一级结构的序列特征, 通过学习序列之间的先后次序, 发现其与蛋白质二级结构间的复杂关系, 从而进行蛋白质二级结构预测^[64,65]. 而卷积神经网络则专注于提取序列的局部信息, 并对其进行分析、整合, 以此来提取所关注的一段序列与二级结构间的对应关系^[66]. 混合神经网络方法则是在神经网络中同时使用了循环神经网络结构和卷积神经网络结构, 这使得预测的准确性有所提升^[67,68].

3.2 三级结构预测

蛋白质的三级结构预测至关重要, 因为蛋白质的三级结构往往决定了其功能、稳定性、与其他分子间的相互作用以及与其某些疾病的相关性等^[69]. 目前主流的机器学习蛋白质三级结构预测软件 (例如 AlphaFold2^[58]) 的实际工作流程较为复杂, 这里只介绍其核心思想. AlphaFold2 的结构示意图如图 2 所示. 从图 2 可以看出, 当把序列输入给模型后, 模型首先会做两件事情: 从基因数据库中获取多序列比对以及从结构数据库中获取成对信息模版. 在生物信息学中, 多序列比对^[70] (multiple sequence alignment, MSA) 是一种常用的方法, 它可以将 3 个或更多的生物序列 (通常是蛋白质或核酸) 对齐, 以识别这些序列之间的相似性. 通过多序列比对, 研究人员能够识别保守的序列区域、协变区域, 这些区域在物种间或者基因家族成员间具

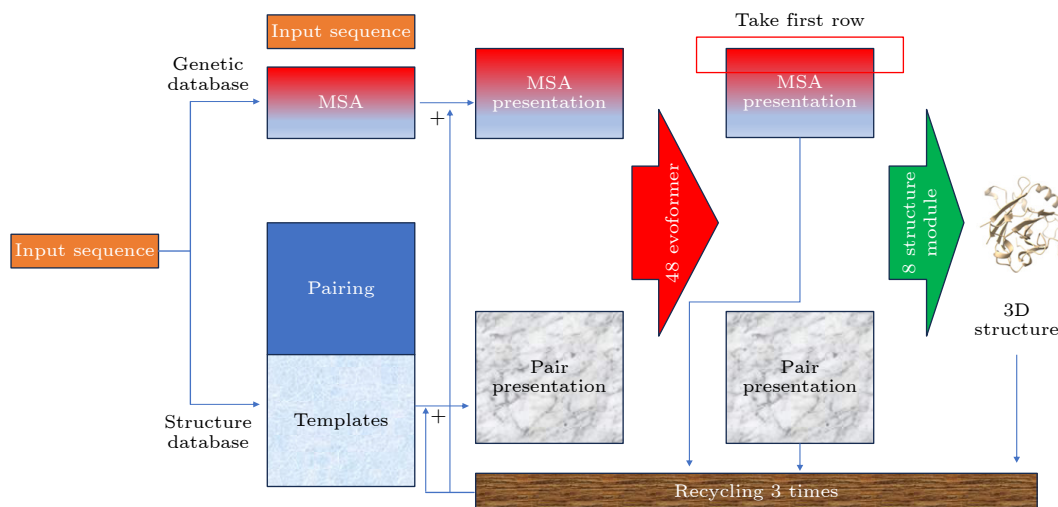


图 2 AlphaFold2 的结构图

Fig. 2. Architecture of AlphaFold2.

有高度的相似性、共进化性,可能对蛋白质的结构和功能有着至关重要的意义.简而言之,多序列比对作为输入,相比于单个序列而言,多出了额外的与蛋白结构相关的信息,可以帮助对蛋白质的三维结构进行推断.在图 2 中,输入的序列与多序列比对信息被转化为了一个多序列比对表象的矩阵,这个矩阵可以被粗略地理解为包含了序列进化信息.

另一方面,可以看到二维的成对矩阵和成对信息模版被模型转化成了成对表象矩阵.这个矩阵包含着丰富的残基间信息,如残基间的距离和相对方向.然后,模型通过基于注意力机制^[71]的 evoformer 模块将多序列比对表象矩阵和成对表象矩阵的信息结合起来,反复更新两者.最后两者通过结构模块,从每个残基的局部信息和残基间信息中通过学习提取关键数据,生成最终的蛋白质的每个原子的三维坐标.注意,生成过程并不是一次完成的,而是需要反复迭代三次.

3.3 四级结构预测

蛋白质的四级结构研究至关重要,因为它们对生物体的正常运作有着重要影响,这有助于深入研究生物大分子的功能和调控,并对药物设计做出必要的指导^[72,73].蛋白质分子间的相互作用主要由以下几种非共价作用组成:氢键、离子键、范德瓦耳斯力和疏水相互作用^[74].生物大分子间的相互作用主要取决于表面基团的化学性质、几何结构、动态结构等因素.要想正确地预测蛋白质的四级结构,就必须处理大量高维信息,而这正是机器学习所擅长的.

传统的蛋白质对接预测软件大多是基于分数,例如 ZDOCK^[75],是使配体遍历受体附近的每一个位置和自身的每一个方向,通过经验公式对每一个构象进行打分,最终选定分数最高的几个构象作为备选答案.然而,这种方法具有着一定的劣势,例如打分的机制往往存在很多经验项,用于拟合的实验数据过少以及计算速度过慢等.目前虽然已有关于 RNA-蛋白质复合物的四级结构预测软件 Open Complex^[76],但相关文章尚未发表,因此本小节主要介绍著名的蛋白质四级结构预测软件 AlphaFold-Multimer^[77].

由于极高的复杂度和更大的搜索空间,蛋白质的四级结构预测远比三级结构预测要困难.有学者曾调整过 AlphaFold 的输入,增加了虚拟的空位

或者连接基团,多链蛋白质强行转化成单链蛋白质,再进行结构预测^[78-81].其道理在于,虽然四级结构中的链与链之间失去了骨架的连接,但蛋白质链间残基之间相互作用的物理本质和同一条链上距离较远的残基之间的相互作用的物理本质是一样的.而 AlphaFold-Multimer 也是采用了同样的思想,只不过摒弃了空位和连接基团的引入^[77].

AlphaFold-Multimer 基本框架和 AlphaFold 是一样的,但主要做了如下几点改变:第一,对输入进行了改变,采用了一种针对多链蛋白更加科学的构建多序列比对的方法,其主要原理是分别生成不同序列的多序列比对,再在此基础上生成基于基因组的和基于系统发育的多链多序列比对^[82](如图 3 所示),并对结果进行整合.第二,对损失函数(表征机器学习中预测值与真实值之间的差距)进行了修改,考虑了含有相同链的蛋白中链与链之间的交换效应;修正了 AlphaFold 中的帧对齐点误差损失的上限以优化训练时的梯度信号;额外增加了链质心损失以防不同的链被预测到重叠的位置上.第三,对训练流程进行了改进,为了缓解计算资源的局限性,AlphaFold-Multimer 对蛋白质进行剪裁,并训练 AlphaFold 系统来处理全长蛋白质的裁剪片段,这些裁剪区域最多可达 384 个残基的连续块.

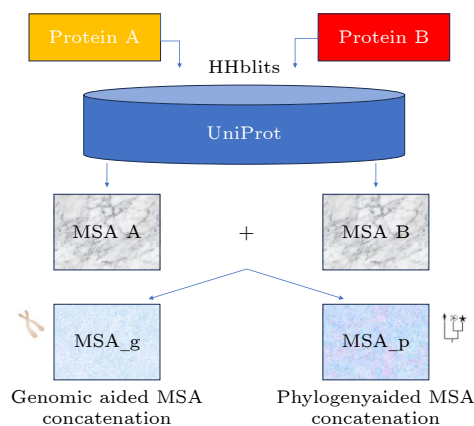


图 3 AlphaFold-Multimer 的多序列比对构建方法
Fig. 3. Construction of MSA used in AlphaFold-Multimer.

4 性质预测中的机器学习

生物分子的结构决定了它们的性质^[83],但绝大多数情况下,仅凭人类的推理,很难从复杂的结构信息中提取到重要的依据来判定生物分子的性

质, 因此需要借助机器学习的力量^[8,83,84]从复杂的序列等信息中提取出所需的性质信息. 由于实验成本的原因, 仅从序列信息推理得到蛋白质分子的性质, 是人们长久以来希望实现的. 在蛋白质的种种性质中, 水溶性、免疫原性和热稳定性尤为重要. 本节将针对这三点性质的预测逐一讨论.

4.1 蛋白质水溶性预测

蛋白质的水溶性主要取决于其自身的氨基酸组成和空间结构^[85]. 一般来说, 富含亲水性氨基酸残基(如赖氨酸、精氨酸、谷氨酸等)的蛋白质, 水溶性较好, 这些亲水性残基能与水分子形成氢键, 提高蛋白质的溶解度; 含有较多疏水性氨基酸残基(如缬氨酸、异亮氨酸、苯丙氨酸等)的蛋白质, 水溶性较差, 这些疏水性残基难以与水分子接触, 使蛋白质不溶于水; 蛋白质的空间结构也影响其溶解性, 紧密折叠的球状蛋白较易溶解, 而松散的随机卷曲蛋白溶解度较低, 这是因为紧密结构能使更多亲水基团暴露于水分子之间. 蛋白质溶解时, 也会发生构象变化, 一些原本隐藏在内部的亲水基团会暴露出来, 提升蛋白质的溶解度. 虽然以上经验会为预测蛋白质的水溶性提供一些帮助, 但由于蛋白质自身的复杂性, 依然需要借助机器学习的力量来完成蛋白质水溶性预测工作.

DeepSol^[86]是一款基于卷积神经网络的蛋白质水溶性预测软件, 在这个软件中, 蛋白质序列被当作唯一的输入传递给卷积神经网络, 而模型的输出则是一个大于0小于1的实数, 分数越大表示模型认为该序列越有可能来自一个可溶的蛋白质. EPSOL^[87]是近年来另一款具有代表性的蛋白质水溶性预测软件, 它比DeepSol的结果更加准确, 但是也需要输入更多的信息以帮助其进行判断, 例如蛋白质的二级结构和溶剂可及性(solvent accessibility).

预测蛋白质的水溶性可以帮助我们: 解释蛋白质的物理化学性质; 指导蛋白质的提取和纯化; 为蛋白质的功能研究提供参考; 辅助蛋白质药物的药效学研究; 指导蛋白质工程设计以及分析蛋白质的稳定性和折叠行为. 这些对于蛋白质研究都是极其重要的.

4.2 蛋白质免疫原性

蛋白质的免疫原性^[88]指的是某种蛋白质所具

有的诱导免疫反应并激活免疫系统的能力. 简单来说, 就是某些蛋白质能够被人体免疫系统识别为“外来抗原”, 并触发体液免疫和细胞免疫反应以清除这种抗原. 虽然研究表明, 蛋白质的免疫原性与密码子(codon)^[89]和翻译后修饰(post-translational modification, PTM)^[90]都有关系, 但其与蛋白质本身的关系依然有迹可循^[91], 而机器学习正是一个解释这种复杂关系的极好工具.

2019年Smith等^[92]训练了一个机器学习模型(基于线性回归), 基于肿瘤抗原的免疫原性本质特征, 来预测新抗原的免疫原性. 在该研究中, 学者在两种肿瘤小鼠模型中验证了该预测模型的效果, 证明了它可以用于选择有治疗作用的抗原表位, 并在TCGA全癌症数据集中分析了高免疫原性新抗原与肿瘤微环境免疫特征的关联, 发现在结肠腺癌和肺腺癌中存在显著关联. 最后提供了证据支持一种预测的移码新抗原能够驱动抗肿瘤的细胞免疫反应, 提示移码抗原也可能成为潜在的治疗靶点. 另一方面, 针对疫苗的免疫原性研究也同样重要. 2020年Gonzalez-Dias等^[93]总结和讨论了使用系统疫苗学和机器学习方法来预测疫苗免疫原性和不良反应的技术, 并概述了不同的机器学习算法在这个框架中的应用, 如支持向量机、神经网络、随机森林等, 还探讨了一些目前在该领域的挑战, 如变量混杂的处理、获取更多高质量数据的需要等.

通过对蛋白质的免疫原性的预测可以评估蛋白质作为候选疫苗、药物的潜力. 对于代替性蛋白质药物, 需要在设计的过程中降低其免疫原性, 避免集体产生抗体促使药物失效, 也避免机体产生不必要的免疫反应. 但对于疫苗, 需要提高其免疫原性, 以最大程度激发机体的免疫反应. 总之, 免疫原性的预测对医用蛋白质有着举足轻重的作用.

4.3 蛋白质的热稳定性

蛋白质的热稳定性由很多因素共同决定^[94]. 通常情况下, α -螺旋和 β -折叠通常较之无规律卷曲更热稳定. 疏水相互作用也能提高蛋白质的热稳定性; 氢键和离子键的数量越多, 越有利于热稳定性; 蛋白质表面暴露的非极性残基越多, 热稳定性越低; 多聚体的形成有利于提高蛋白质的热稳定性; 蛋白质本身的残基比例也会影响其热稳定性, 例如富含脯氨酸、苏氨酸的蛋白质热稳定性较差. 虽然有着

很多简单的经验可以推断蛋白质的热稳定性, 鉴于蛋白质序列、结构的高度复杂性, 依然需要机器学习来辅助预测蛋白质的热稳定性.

TemStaPro 是近年来被公开的一款基于深度学习预测蛋白质热稳定性的软件^[95]. 在这款软件的架构中, 开发者们巧妙地使用了迁移学习 (transfer learning), 直接从复杂的蛋白质语言模型 (protein language models, PLM)^[96,97] 获得被解码的信息, 并构建一个小型的神经网络用于预测最终的序列热稳定性. 该模型可以判断给定序列在一定温度以上是否依然具有热稳定性, 预测结果是一个大于 0 小于 1 的实数, 数值越大, 代表越可能具有热稳定性.

预测蛋白质在体温环境下的稳定性和降解情况对蛋白药物的设计很重要, 提高热稳定性可以延长其体内半衰期. 除此之外, 预测和改善工业用酶的热稳定性, 以扩展其在工业生产过程中的适用温度范围和使用寿命, 可以减少酶的更换和处理成本.

5 分子设计中的机器学习

生物分子设计是一个涉及修改自然存在的生物分子或创建新分子以实现特定功能的科学领域, 而其中最受人瞩目的方向之一便是蛋白质设计^[98]. 分子设计的一般流程如下: 第 1 步, 确定目标, 明确并理解所期望的分子的功能或性质; 第 2 步, 选取适当算法和模型; 第 3 步, 生成候选分子, 这一步会产生大量备选分子; 第 4 步, 筛选和评估, 即通过计算方法来评估分子的功能和性质, 筛选出最可能成功的几个分子; 第 5 步, 验证和测试, 对选中的分子进行实验, 评估实验结果是否达到预期; 第 6 步, 优化和修改, 即基于实验结果, 对分子或算法进行进一步优化, 必要时, 将对所设计的分子进行迭代改进. 本节将从几个不同方面介绍蛋白质设计.

5.1 蛋白质的结构设计

要对蛋白质进行从头设计不是一件容易的事, 因为蛋白质本身结构复杂, 而功能与结构的关系也复杂^[98]. 而蛋白质设计, 实际上就是一个优化问题:

designed protein

$$= \operatorname{argmax}_{\text{protein}} P(\text{protein}|\text{condition}). \quad (2)$$

因为我们把骨架结构和序列设计进行了拆分, 因此可以认为它们是最终设计出的蛋白质的两个因素:

$P(\text{protein}|\text{condition})$

$$= P(\text{sequence, structure}|\text{condition}). \quad (3)$$

因为功能直接由结构决定, 因此在蛋白质从头设计中, 人们通常从设计蛋白质的骨架结构开始^[99,100], 即在给定的条件下找到最有可能符合该条件的骨架结构:

designed structure

$$= \operatorname{argmax}_{\text{structure}} P(\text{structure}|\text{condition}). \quad (4)$$

不是所有的骨架都可以被自然氨基酸生成的, 要想生成符合自然规律的骨架, 就必须遵守一定的规则^[99]. 因此, 一个直观的想法便是, 如果能以某种方式, 通过机器学习的力量, 学习到自然存在的蛋白质骨架应该具有什么样的特征, 那么就可以不断地向应有的特征的方向调整所生成骨架的相应特征, 这样就会得到符合自然法则的蛋白质骨架结构. 进一步地, 如果能把自然存在的蛋白质统计意义上的特征表征成一种基于统计 (而非物理) 的能量项, 那么理论上以这个能量项为基础, 就可以通过动力学模拟的方法自发生成符合自然规律的蛋白质骨架结构. SCUBA 模型^[99] 正是基于此思想.

SCUBA 的核心功能是在与序列无关的骨架结构空间中, 通过寻找能量最低点的方法找到预测的最优骨架结构, 而后续的基于结构的序列设计工作则交给其他模型. 在 SCUBA 这项工作中, 研究者们将统计能量进行了拆分, 并逐项通过临近点计数-神经网络的方法进行训练以获得相应的连续可微分的能量函数^[99]. 临近点计数-神经网络方法的训练是基于有监督学习的, 其核心思想就是通过神经网络的强大泛化性将粗糙的统计散点数据转化为连续可微的能量函数.

另一方面, 扩散模型 (diffusion models)^[101] 作为一款生成模型, 近年来在众多领域都做出了突出的贡献^[102,103]. 于是, 基于扩散模型的蛋白质骨架结构从头设计模型也应运而生^[100,104]. 扩散是一个自发的熵增过程, 在机器学习中的扩散, 通常是指在训练过程中逐步地为原始数据添加噪音, 最终将

得到一个纯粹的噪音. 而扩散模型所做的便是通过学习每一步扩散过程中增加的那一部分噪音与数据分布之间的关系, 从而生成一个逆向的神经网络, 逐步预测被注入噪音后的数据最可能的原来的样子. 这样, 只给定随机噪音, 逆向神经网络就能自发地生成一个与训练数据高度相似的数据.

RFdiffusion 的核心思想是对 RoseTTAFold^[60]进行了微调, 使之能完成图中所示的特殊的三维结构预测任务. 初始时刻, 骨架原子坐标是随机的. 在每一步中, RFdiffusion 会根据本步的骨架坐标, 通过微调后的 RoseTTAFold 生成一个虚拟的预测结果, 然后根据这个虚拟的预测结果推测出上一个扩散步骤中被加入的噪音, 依此推测出上一个扩散步骤的骨架坐标. 如此, 最终可以得到扩散尚未开始时的骨架原子坐标. 另一方面, 人们也一直在尝试不需要在结构预测模型的基础上进行微调的基于扩散模型的蛋白质结构生成模型^[104,105]. 其中 SCUBA-D^[104]模型结合了生成对抗模型和扩散模型各自的生成质量高、创新性大等优势, 在蛋白从头设计领域做出了突出的贡献.

5.2 蛋白质的序列设计

在设计好蛋白质的骨架结构之后, 就需要找到可以满足该骨架结构的序列. 需要做的实际上便是最大化如下概率:

$$\text{designed sequence} = \operatorname{argmax}_{\text{sequence}} P(\text{sequence} | \text{designed structure, condition}). \quad (5)$$

由于蛋白质的空间结构复杂, 且序列空间很大, 因此借助机器学习的力量对给定骨架结构的蛋白质进行序列设计是一个很好的选择.

在 ABACUS^[106,107]模型中, 学者们通过遍历大量已知结构的蛋白, 学习到了统计意义上的在特定结构下, 某个位置上是某个氨基酸的概率以及某两个位置上是某两个氨基酸的联合概率, 再通过 $e = -\ln P$ 的方法将统计意义上的概率转化为统计意义上的能量. 随后, 学者们将统计意义上的能量与经验化的物理意义上的能量 (原子间相互作用等) 进行加和, 得到了最终的能量表达式. 初始的蛋白序列是一条完全随机的序列, 随后 ABACUS 对序列在序列空间进行蒙特卡罗模拟, 以能量函数的变化来判断是否保留每一步的突变, 最终在进行足够多步后, 得到一个足够好的序列. 目前, 基于

ABACUS 的工作依然在继续, 研究人员正在试图通过解码与残基自身和该残基相邻的所有残基空间结构、相对位置信息, 来还原位置序列的蛋白质结构中每一个残基的氨基酸类型.

而在 ProteinMPNN^[108]中, 研究者们则使用了图神经网络 (graph neural networks, GNN)^[109]的框架, 如图 4 所示. 在该模型中, 一个蛋白质骨架结构被理解为一张图, 其中图的节点代表着蛋白质中的每一个氨基酸, 而每一条边则代表着氨基酸对之间的空间信息, 这里选用了 N, C_α, C, O, C_β 之间的距离. 模型由两部分组成, 骨架编码器负责读取骨架的空间信息, 而序列解码器则负责将编码器处获得的信息解码成序列.

5.3 结构序列协同设计

传统的蛋白质设计方案先对骨架结构进行设计, 再对蛋白序列进行设计, 得到的蛋白序列如 (5) 式所示, 而实际上, 总的结果相当于:

$$\begin{aligned} & \text{designed protein} \\ & = \operatorname{argmax}_{\text{sequence}} P(\text{sequence, structure} | \\ & \quad \text{designed structure, condition}). \end{aligned} \quad (6)$$

对比 (2) 式和 (3) 式可以发现, 这里的搜索空间变少了, 而限制条件变多了, 因此有

$$P_{\max}^{\text{traditional}} \leq P_{\max}^{\text{co-design}}, \quad (7)$$

其中 $P_{\max}^{\text{co-design}}$ 是协同设计时蛋白质满足条件的概率, 只有在传统的设计方案得到的骨架结构刚好等于协同设计得到的骨架结构时, (7) 式中的等号才成立.

上述讨论说明, 比起传统的先设计蛋白质骨架结构, 再对蛋白的序列进行设计的方案, 直接对蛋白质的骨架结构和序列信息进行协同设计往往更能设计出符合要求的蛋白质. 另一方面, 结构序列协同设计也更加灵活, 如当需要固定被设计的蛋白中的某部分骨架结构或某些氨基酸类型时, 就可以在协同设计中直接将这些变量固定. 而这种任务常常是在设计分子间相互作用下的蛋白质^[110,111]时所面对的.

2022 年, Shi 等^[112]提出了一款基于协同设计思想的蛋白质从头设计机器学习模型. 模型结构如图 5 所示, 在该模型中, 通过输入初始被设计蛋白的每个残基的性质 (例如二级结构) 和残基间性质

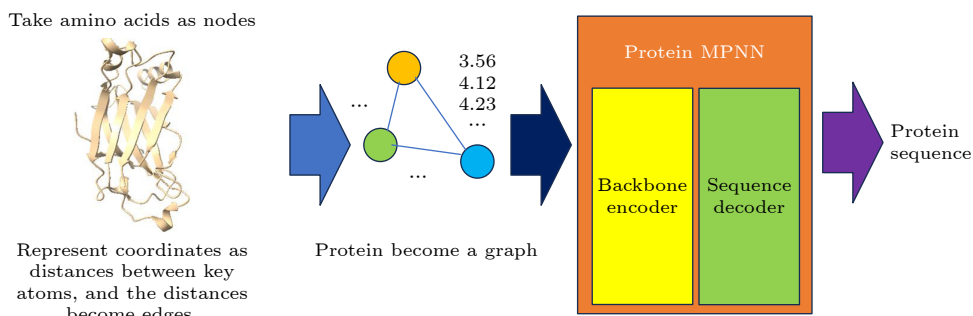


图 4 ProteinMPNN 模型核心思想示意图

Fig. 4. Main idea of ProteinMPNN.

(例如是否接触)的信息,使用基于注意力机制^[71]的算法进行不断迭代,最终设计出符合要求的蛋白质.在该模型中,初始序列和骨架结构都是未知的,而模型通过学习自然存在的蛋白质的结构和序列,可以做到生成最可能在自然界中稳定存在的满足设计要求的蛋白质.然而,Shi 等指出该模型最大的问题是,目前还不确定该模型能否自发设计出超越现有蛋白质拓扑结构的蛋白.该模型的输入是一串指定序列局部信息的数组和一个指定序列连接信息的矩阵,而这通常就包含了蛋白质足够的信息.这样就使得模型有点不那么像是一个生成模型,反而有些像一个回归模型.但毫无疑问的是,这项工作为蛋白质结构序列协同设计提供了很好的理论支持.在设计蛋白-蛋白相互作用的蛋白质时,很多时候需要协同地考虑一些接触位点的空间结构和氨基酸类型,这时,协同设计便会发挥其强大的功能.

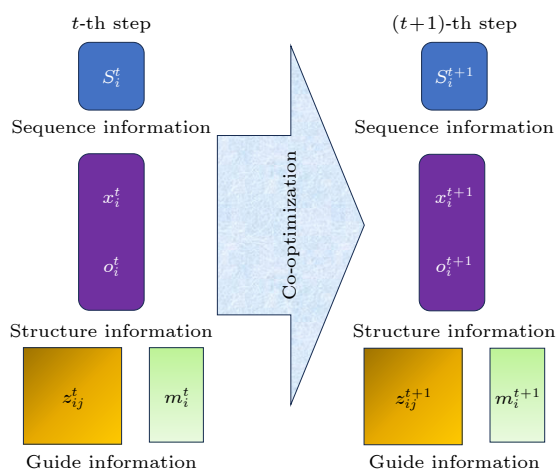


图 5 蛋白质结构序列协同设计的一种机器学习模型示意图

Fig. 5. Illustration of a machine learning model of protein structure-sequence co-design.

6 总结与展望

蛋白质计算与机器学习的结合在近年来取得了飞速的发展^[113,114],这使得生物学本身与生物信息学、生物物理学和生物化学等交叉学科获得了极大的突破.机器学习对蛋白质计算领域的介入,使我们可以更好地认识自然,理解自然,进而改造自然.本综述的第 2 节、第 3 节和第 4 节体现了对自然生命分子和生命过程的认识和理解,而第 5 节则体现了对自然生命分子和生命过程的改造.正如第 1 节中讨论的那样,认识自然和改造自然不是彼此独立的,而是相互交汇的.在认识和理解了一个生物现象之后,便要对其向好的方向进行改造,而这往往会让我们发现更多需要被认识的新的生物现象.

然而,机器学习在蛋白质计算,尤其是蛋白质分子设计领域还有着许多需要解决的问题.首先,我们观察到,通过现有的蛋白质骨架从头设计软件设计出的骨架非常倾向于生成刚性结构域,而较少生成对调节蛋白动态性质至关重要的环(loop)区.另一方面,现有的序列设计软件通常会极大程度考虑结构的静态稳定性而不是动态性质.因此最终设计出的蛋白大多都非常刚性,很难满足一些特定的要求,例如设计出有活性的酶,因为酶的活性是与其动态性质息息相关的^[115].未来蛋白质设计的发展趋势将会更加注重设计蛋白的柔性和活性,尽可能地设计出柔软的“器官”,而不是坚硬的“零件”.

放眼未来,人们会利用机器学习设计出更多经济实用的药物.例如,由于 mRNA 易于合成且在人体内可以长期地表达特定蛋白,在近年来已成为最受关注的新兴药物之一^[116].而在分别理解了蛋白质结构预测、蛋白质设计、RNA 结构预测和密

码子优化^[117]等 mRNA 设计后,便可以考虑蛋白-mRNA 协同设计,即根据需要的蛋白的功能,将蛋白的功效率和 mRNA 的翻译效率协同考虑,直接设计出相应的药用 mRNA 序列.虽然这比独立设计蛋白质和 RNA 都要困难很多,但在机器学习的帮助下,这个难题终将被攻克.

比起单个生物分子,人们往往更加关注生物分子体系,尤其是生物大分子间的相互作用^[57,118].在未来,随着机器学习算法的提升和硬件性能的提高,人们将可以研究更加细节化的生物大分子间相互作用,也能预言尺度更大、数量更多的生物大分子间相互作用,从而渐渐实现从分子到分子间,再从分子间到体系的突破,最终实现精准快速的细胞尺度模拟.

目前机器学习与蛋白质计算的结合已取得了众多突破性的进展,本综述主要总结了机器学习在蛋白质的分子动力学模拟、结构预测、性质预测和分子设计中的实现,希望能以此为相关领域研究者提供参考并激发广大科研工作者对本领域的兴趣.

感谢中国科学技术大学生命科学学院刘海燕老师在写作过程中给予我充分的帮助和支持.

参考文献

- [1] Baltoumas F A, Zafeiropoulou S, Karatzas E, et al. 2021 *Biomolecules* **11** 1245
- [2] Wolf Y I, Katsnelson M I, Koonin E V 2018 *Proc. Natl. Acad. Sci. USA* **115** E8678
- [3] Fusco A, Fedele M 2007 *Nat. Rev. Cancer* **7** 899
- [4] Noble D 2002 *Nat. Rev. Mol. Cell Biol.* **3** 459
- [5] Markowitz F 2017 *PLoS Biology* **15** e2002050
- [6] Hollingsworth S A, Dror R O 2018 *Neuron* **99** 1129
- [7] Zhang Y 2008 *Curr. Opin. Struct. Biol.* **18** 342
- [8] Agostini F, Vendruscolo M, Tartaglia G G 2012 *J. Mol. Biol.* **421** 237
- [9] Chen L, Fan Z, Chang J, et al. 2023 *Nat. Commun.* **14** 4217
- [10] Geng H, Chen F, Ye J, Jiang F 2019 *Computat. Struct. Biotechnol. J.* **17** 1162
- [11] Salo-Ahen O M, Alanko I, Bhadane R, et al. 2020 *Processes* **9** 71
- [12] Norberg J, Nilsson L 2003 *Q. Rev. Biophys.* **36** 257
- [13] van der Kamp M W, Shaw K E, Woods C J, Mulholland A J 2008 *J. R. Soc. Interface* **5** 173
- [14] Dror R O, Dirks R M, Grossman J, Xu H, Shaw D E 2012 *Annu. Rev. Biophys.* **41** 429
- [15] Lin X, Li X, Lin X 2020 *Molecules* **25** 1375
- [16] Pearce R, Zhang Y 2021 *Curr. Opin. Struct. Biol.* **68** 194
- [17] Jordan M I, Mitchell T M 2015 *Science* **349** 255
- [18] Butler K T, Davies D W, Cartwright H, Isayev O, Walsh A 2018 *Nature* **559** 547
- [19] Liakos K G, Busato P, Moshou D, Pearson S, Bochtis D 2018 *Sensors* **18** 2674
- [20] Jiang T, Gradus J L, Rosellini A J 2020 *Behav. Ther.* **51** 675
- [21] Hastie T, Tibshirani R, Friedman J, Hastie T, Tibshirani R, Friedman J 2009 *Unsupervised Learning. In: The Elements of Statistical Learning. Springer Series in Statistics* (New York: Springer) pp485–585
- [22] Van Engelen J E, Hoos H H 2020 *Machine Learning* **109** 373
- [23] Wiering M A, Van Otterlo M 2012 *Reinforcement Learning* (Heidelberg, Berlin: Springer) p729
- [24] LeCun Y, Bengio Y, Hinton G 2015 *Nature* **521** 436
- [25] Deng L, Yu D 2014 *Deep Learning: Methods and Applications* (Now Foundations and Trends) p197
- [26] Jones D T 2019 *Nat. Rev. Mol. Cell Biol.* **20** 659
- [27] Das P, Sercu T, Wadhawan K, et al. 2021 *Nat. Biomed. Eng.* **5** 613
- [28] Kuhlman B, Bradley P 2019 *Nat. Rev. Mol. Cell Biol.* **20** 681
- [29] Trevino S R, Scholtz J M, Pace C N 2008 *J. Pharm. Sci.* **97** 4155
- [30] Kelley K W, Weigent D A, Kooijman R 2007 *Brain Behav. Immun.* **21** 384
- [31] Babin V, Roland C, Sagui C 2008 *J. Chem. Phys.* **128**
- [32] Morozov I V, Kazennov A M, Bystryi R, Norman G E, Pisarev V V, Stegailov V V 2011 *Comput. Phys. Commun.* **182** 1974
- [33] Karplus M, McCammon J A 2002 *Nat. Struct. Biol.* **9** 646
- [34] Wang Y, Ribeiro J M L, Tiwary P 2020 *Curr. Opin. Struct. Biol.* **61** 139
- [35] Chmiela S, Tkatchenko A, Sauceda H E, Poltavsky I, Schütt K T, Müller K R 2017 *Sci. Adv.* **3** e1603015
- [36] Ponder J W, Case D A 2003 *Adv. Protein Chem.* **66** 27
- [37] Monticelli L, Tieleman D P 2013 *Biomolecular Simulations: Methods and Protocols* 197
- [38] Wang J, Wolf R M, Caldwell J W, Kollman P A, Case D A 2004 *J. Comput. Chem.* **25** 1157
- [39] Hughes Z E, Wright L B, Walsh T R 2013 *Langmuir* **29** 13217
- [40] Cesari A, Bottaro S, Lindorff-Larsen K, Banáš P, Šponer J, Bussi G 2019 *J. Chem. Theory Comput.* **15** 3425
- [41] Unke O T, Chmiela S, Sauceda H E, Gastegger M, Poltavsky I, Schütt K T, Tkatchenko A, Müller K R 2021 *Chem. Rev.* **121** 10142
- [42] Poltavsky I, Tkatchenko A 2021 *J. Phys. Chem. Lett.* **12** 6551
- [43] Kästner J 2011 *WIREs Comput. Mol. Sci.* **1** 932
- [44] Izrailev S, Stepaniants S, Israilewitz B, Kosztin D, Lu H, Molnar F, Wriggers W, Schulten K 1999 *Computational Molecular Dynamics: Challenges, Methods, Ideas: Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling* Berlin, May 21–24, 1997 p39
- [45] Moradi M, Babin V, Roland C, Sagui C 2013 *Nucleic Acids Res.* **41** 33
- [46] Simonson T, Archontis G, Karplus M 2002 *Acc. Chem. Res.* **35** 430
- [47] Bitencourt-Ferreira G, de Azevedo W F 2018 *Biophys. Chem.* **240** 63
- [48] Trott O, Olson A J 2010 *J. Comput. Chem.* **31** 455
- [49] Besora M, Vidossich P, Lledos A, Ujaque G, Maseras F 2018 *J. Phys. Chem. A* **122** 1392
- [50] Pan X, Yang J, Van R, Epifanovsky E, Ho J, Huang J, Pu J, Mei Y, Nam K, Shao Y 2021 *J. Chem. Theory Comput.* **17** 5745
- [51] Senn H M, Thiel W 2009 *Angew. Chem. Int. Ed.* **48** 1198

- [52] Riniker S 2017 *J. Chem. Inf. Model.* **57** 726
- [53] Bennett W D, He S, Bilodeau C L, Jones D, Sun D, Kim H, Allen J E, Lightstone F C, Ingólfsson H I 2020 *J. Chem. Inf. Model.* **60** 5375
- [54] Bertazzo M, Gobbo D, Decherchi S, Cavalli A 2021 *J. Chem. Theory Comput.* **17** 5287
- [55] Eswar N, John B, Mirkovic N, et al. 2003 *Nucleic Acids Research* **31** 3375
- [56] Asara J M, Schweitzer M H, Freimark L M, Phillips M, Cantley L C 2007 *Science* **316** 280
- [57] Greener J G, Kandathil S M, Moffat L, Jones D T 2022 *Nat. Rev. Mol. Cell Biol.* **23** 40
- [58] Jumper J, Evans R, Pritzel A, et al. 2021 *Nature* **596** 583
- [59] Wu R, Ding F, Wang R, et al. 2022 *bioRxiv* 2022.07.21.500999
- [60] Baek M, DiMaio F, Anishchenko I, et al. 2021 *Science* **373** 871
- [61] Medsker L R, Jain L 1999 *Recurrent Neural Networks: Design and Applications* (1st Ed.) (CRC Press) p2
- [62] Kim P 2017 *Convolutional Neural Network. In: MATLAB Deep Learning* (Berkeley, CA: Apress) p121
- [63] Wardah W, Khan M G, Sharma A, Rashid M A 2019 *Comput. Biol. Chem.* **81** 1
- [64] Mirabella C, Pollastri G 2013 *Bioinformatics* **29** 2056
- [65] Heffernan R, Yang Y, Paliwal K, Zhou Y 2017 *Bioinformatics* **33** 2842
- [66] Wang S, Peng J, Ma J, Xu J 2016 *Sci. Rep.* **6** 1
- [67] Li Z, Yu Y 2016 *arXiv: 1604.07176 [q-bio.BM]*
- [68] Wang Y, Mao H, Yi Z 2017 *Knowledge-Based Systems* **118** 115
- [69] Nishikawa K, Ooi T, Isogai Y, Saitō N 1972 *J. Phys. Soc. JPN* **32** 1331
- [70] Edgar R C, Batzoglou S 2006 *Curr. Opin. Struct. Biol.* **16** 368
- [71] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I 2017 *Advances in Neural Information Processing Systems 30* Long Beach, USA, December 4–9, 2017 p30
- [72] Janin J, Bahadur R P, Chakrabarti P 2008 *Q. Rev. Biophys.* **41** 133
- [73] Zafferani M, Hargrove A E 2021 *Cell Chem. Biol.* **28** 594
- [74] Hunter C A 2004 *Angew. Chem. Int. Ed.* **43** 5310
- [75] Chen R, Li L, Weng Z 2003 *Proteins Struct. Funct. Bioinf.* **52** 80
- [76] Jingcheng Y, Zhaoming C, Zhaoqun L, Mingliang Z, Wenjun L, He H, Qiwei Y 2022 *Code of Open Complex* <https://github.com/baaihealth/OpenComplex>.
- [77] Evans R, O' Neill M, Pritzel A, et al. 2021 *bioRxiv* 2021.10.04.463034
- [78] Moriwaki Y 2021 *Twitter* https://twitter.com/Ag_smith/status.
- [79] Ko J, Lee J 2021 *bioRxiv* 2021.07.27.453972
- [80] Tsaban T, Varga J K, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O 2022 *Nat. Commun.* **13** 176
- [81] Bryant P, Pozzati G, Elofsson A 2022 *Nat. Commun.* **13** 1265
- [82] Zhou T M, Wang S, Xu J 2017 *bioRxiv* 240754
- [83] Cang Z, Wei G W 2017 *PLoS Comput. Biol.* **13** e1005690
- [84] Yagi K, Re S, Mori T, Sugita Y 2022 *Curr. Opin. Struct. Biol.* **72** 88
- [85] Vendruscolo M, Knowles T P, Dobson C M 2011 *CSH Perspect. Biol.* **3** a010454
- [86] Khurana S, Rawi R, Kunji K, Chuang G Y, Bensmail H, Mall R 2018 *Bioinformatics* **34** 2605
- [87] Wu X, Yu L 2021 *Bioinformatics* **37** 4314
- [88] Schellekens H 2003 *Nephrology Dialysis Transplantation* **18** 1257
- [89] Ternette N, Tippler B, Überla K, Grunwald T 2007 *Vaccine* **25** 7271
- [90] Jefferis R 2016 *J. Immunol. Res.* 2016
- [91] Schellekens H 2005 *Nephrology Dialysis Transplantation* **20** vi3
- [92] Smith C C, Chai S, Washington A R, et al. 2019 *Cancer Immunol. Res.* **7** 1591
- [93] Gonzalez-Dias P, Lee E K, Sorgi S, de Lima D S, Urbanski A H, Silveira E L, Nakaya H I 2020 *Hum. Vacc. Immunother.* **16** 269
- [94] Timr S, Madern D, Sterpone F 2020 *Prog. Mol. Biol. Transl. Sci.* **170** 239
- [95] Pudžiuvelytė I, Olechnovič K, Godliauskaite E, Sermokas K, Urbaitis T, Gasiunas G, Kazlauskas D 2023 *bioRxiv* 2023.03.27.534365
- [96] Rives A, Meier J, Sercu T, et al. 2021 *Proc. Natl. Acad. Sci. U.S.A.* **118** e2016239118
- [97] Elnaggar A, Heinzinger M, Dallago C, et al. 2022 *IEEE Trans. Pattern Anal. Mach. Intell.* **44** 7112
- [98] Huang P S, Boyken S E, Baker D 2016 *Nature* **537** 320
- [99] Huang B, Xu Y, Hu X, Liu Y, Liao S, Zhang J, Huang C, Hong J, Chen Q, Liu H 2022 *Nature* **602** 523
- [100] Watson J L, Juergens D, Bennett N R, et al. 2023 *Nature* **620** 1089
- [101] Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, Shao Y, Zhang W, Cui B, Yang M H 2022 *arXiv: 2209.00796 [cs.LG]*
- [102] Croitoru F A, Hondru V, Ionescu R T, Shah M 2023 *IEEE Trans. Pattern Anal. Mach. Intell.* **45** 10850
- [103] Kong Z, Ping W, Huang J, Zhao K, Catanzaro B 2020 *arXiv: 2009.09761 [eess.AS]*
- [104] Liu Y, Chen L, Liu H 2022 *bioRxiv* 2022.12.17.52084
- [105] Watson J L, Juergens D, Bennett N R, et al. 2022 *bioRxiv* 2022.12.09.519842
- [106] Xiong P, Wang M, Zhou X, Zhang T, Zhang J, Chen Q, Liu H 2014 *Nat. Commun.* **5** 5330
- [107] Xiong P, Hu X, Huang B, Zhang J, Chen Q, Liu H 2020 *Bioinformatics* **36** 136
- [108] Dauparas J, Anishchenko I, Bennett N, et al. 2022 *Science* **378** 49
- [109] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M 2020 *AI open* **1** 57
- [110] Chen Y, Chen Q, Liu H 2022 *J. Chem. Inf. Model.* **62** 971
- [111] Marchand A, Van Hall-Beauvais A K, Correia B E 2022 *Curr. Opin. Struct. Biol.* **74** 102370
- [112] Shi C, Wang C, Lu J, Zhong B, Tang J 2022 *arXiv: 2210.08761 [q-bio. BM]*
- [113] Dixit R, Khambhati K, Supraja K V, Singh V, Lederer F, Show P L, Awasthi M K, Sharma A, Jain R 2022 *Bioresour. Technol.* 128522
- [114] Kaptan S, Vattulainen I 2022 *Adv. Phys.: X* **7** 2006080
- [115] Casadevall G, Duran C, Osuna S 2023 *JACS Au* **3** 1554
- [116] Webb C, Ip S, Bathula N V, et al. 2022 *Mol. Pharmaceutics* **19** 1047
- [117] Mauro V P, Chappell S A 2014 *Trends Mol. Med.* **20** 604
- [118] Sarkar D, Saha S 2019 *J. Biosci.* **44** 104

SPECIAL TOPIC—Machine learning in biomolecular simulations

Machine learning for *in silico* protein research^{*}

Zhang Jia-Hui[†]

(School of Life Sciences, University of Science and Technology of China, Hefei 230027, China)

(Received 7 October 2023; revised manuscript received 4 January 2024)

Abstract

In silico protein calculation has been an important research subject for a long time, while its recent combination with machine learning promotes the development greatly in related areas. This review focuses on four major fields of the *in silico* protein research that combines with machine learning, which are molecular dynamics, structure prediction, property prediction and molecule design. Molecular dynamics depend on the parameters of force field, which is necessary for obtaining accurate results. Machine learning can help researchers to obtain more accurate force field parameters. In molecular dynamics simulation, machine learning can also help to perform the free energy calculation in relatively low cost. Structure prediction is generally used to predict the structure given a protein sequence. Structure prediction is of high complexity and data volume, which is exactly what machine learning is good at. By the help of machine learning, scientists have gained great achievements in three-dimensional structure prediction of proteins. On the other hand, the predicting of protein properties based on its known information is also important to study protein. More challenging, however, is molecule design. Though machine learning has made breakthroughs in drug-like small molecule design and protein design in recent years, there is still plenty of room for exploration. This review focuses on summarizing the above four fields and looks forward to the application of machine learning to the *in silico* protein research.

Keywords: protein, machine learning, molecular dynamics simulation, structural prediction, properties prediction, molecular design

PACS: 93.85.Bc, 31.15.-p, 87.19.Pp

DOI: [10.7498/aps.73.20231618](https://doi.org/10.7498/aps.73.20231618)

^{*} Project supported by the National Natural Science Foundation of China (Grant No. 22177107).

[†] Corresponding author. E-mail: jhzhang@ustc.edu.cn

专题: 生物分子模拟中的机器学习

分子体系自由能地貌图的变分分析及 AI 算法实现*

杜泊船¹⁾ 田圃^{1)2)†}

1) (吉林大学生命科学学院, 长春 130012)

2) (吉林大学人工智能学院, 长春 130012)

(2023 年 11 月 14 日收到; 2024 年 1 月 18 日收到修改稿)

精确描述复杂分子体系的自由能地貌图是理解和操控其行为, 并进一步实现分子设计制造工业化的重要基础. 刻画高维空间自由能地貌图的主要挑战是其往往在不同空间尺度上具有多个层次, 每个层次都可能不止一个亚稳态被相应的自由能垒分开, 且跨越路径有可能不止一条. 另外很多体系涉及非线性行为, 这使得理论解析和直接使用分子模拟都有很大困难. 针对这些挑战, 多年来研究者们发展了多种多样的增强采样方法, 但往往需要很多经验选择和操作, 从而一方面使得研究进程较为缓慢, 另一方面也让误差控制成为困难. 变分虽然在物理、统计和工程中已经被广泛应用并取得巨大成功, 但在复杂分子体系中的应用却随着神经网络的发展刚刚开始. 本文将对这些探索性工作的主要方向、进展和局限进行简要总结, 也对将来的可能发展给出展望, 希望能够激发更多对基于变分的分子体系自由能地貌图人工智能算法的关注和努力, 促进大分子药物、分子生物机器等实践应用的发展.

关键词: 变分, 神经网络, 复杂分子体系, 自由能地貌图**PACS:** 87.80.-y, 87.15.A-**DOI:** 10.7498/aps.73.20231800

1 引言

大多数复杂分子, 尤其是生物大分子体系, 都是通过构象变化或者在一定尺度上的相变实现其功能的^[1-6]. 和诸多分子的实验合成与表征测试过程相比较, 一方面分子模拟的代价往往更低廉; 另一方面很多生物大分子复合体的大量合成非常困难甚至不可能, 或者在能够获取的前提下动态表征很难实现. 因此分子模拟被广泛用于研究复杂分子体系^[7-9]. 决定分子体系各种行为的基础是对应的自由能地貌图, 因此对其准确刻画成为必要. 实现这一目标的主要挑战是复杂分子体系一般不止一个亚稳态并且相互之间有较高的自由能垒. 所以对典型的复杂分子体系 (如核糖体), 想要从全原子分子模拟中完成所有亚稳态的充分采样, 观察对应的

构象变化过程往往需要生成毫秒级甚至更长时间的模拟轨迹^[10]. 这对百万或更多原子的分子体系一方面算力需求很难满足, 另一方面在高维空间中理解所生成的轨迹也很不容易. 因此人们发展了各种各样的增强采样方法^[11-26]和轨迹降维分析方法^[27]. 增强采样方法大致可以分为两大类, 一类是保持分子体系的玻尔兹曼分布不变, 通过改变温度加速分子体系跨越能垒的方法^[12,13]. 另外一类则是通过加持偏置力/势 (bias force/potential) (如元动力学方法^[8]、自适应偏置力方法^[28]), 这类方法的主要依据是虽然一般分子体系的总自由度数目成千上万甚至更多, 但在跨越能垒的时间尺度上很多局部的原子运动都由于时间尺度的分离而成为近似白噪声, 使得体系在对应时空间尺度的运动可以用较少的反应坐标 (reaction coordinates, RC) 或者集合自由度 (collective variable, CV) 成功描述, 下

* 吉林大学“学科交叉融合创新”项目 (批准号: JLUXKJC2021ZZ05) 资助的课题.

† 通信作者. E-mail: tianpu@jlu.edu.cn

文中统称集合自由度 (CV). 这类采样算法的主要困难是集合自由度的构建设没有系统的方法和步骤, 研究者往往依靠物理直觉选择部分体系自由度进行组合尝试. 由于我们生活中感受到的都是三维空间中的物理存在, 所以在体系维度升高后直觉判断的准确性会大打折扣. 如何准确地构建有效的 CV 是目前复杂分子体系模拟中尚未解决的重大挑战之一. 集合自由度空间中主要有 3 类互相关联的问题, 其一是准确描述体系的集合自由度的构建; 其二是绘制出该空间内主要亚稳态所在的构象空间位置和统计权重, 并计算不同亚稳态之间的转化速率; 其三是构建不同亚稳态之间的过渡路径. 这几类问题的传统应对策略已经被多个优秀综述覆盖 [14, 29–41], 本文主要简述变分及其神经网络实现在这些领域的应用, 限于作者所熟悉研究工作的范围, 会遗漏一些优秀的研究进展, 在此表示歉意.

本文的内容组织如下, 首先将对 CV、变分和神经网络及自动微分进行简要说明, 其次对目前已有的针对复杂分子体系自由能地貌图的主要变分构造方法加以讨论, 再次对这些基于变分的和其他 CV 相关的神经网络方法进行比较分析, 最后展望将来的发展.

2 集合变量、相关神经网络架构、自动微分和变分简介

对一个在给定温度 T 和势能 $U(\mathbf{R})$ 下的分子体系, 用 \mathbf{R} 表示其 $3N - 3$ 维坐标, 则平衡态玻尔兹曼分布为 $\mu(\mathbf{R}) = e^{-\beta U(\mathbf{R})}/Z$, 其中 $\beta = (k_B T)^{-1}$ 为逆温度, $Z = \int d\mathbf{R} e^{-\beta U(\mathbf{R})}$ 为配分函数, k_B 为玻尔兹曼常数. 在较长时间尺度上, 这个分子体系的动力学一般可以使用比 $3N$ 维度低很多也平滑很多的 d ($d \ll N$) 维自由能面描述, 对应一组由原来坐标 \mathbf{R} 的函数构建的新变量 $\mathbf{s}(\mathbf{R}) = (s_1(\mathbf{R}), s_2(\mathbf{R}), \dots, s_d(\mathbf{R}))$, 分子体系自由能在这个低维空间也可表示为

$$F(\mathbf{s}) = -(1/\beta) \log \int d\mathbf{R} \delta(\mathbf{s} - \mathbf{s}(\mathbf{R})) e^{-\beta U(\mathbf{R})}, \quad (1)$$

人们通常称这组新变量 $\mathbf{s}(\mathbf{R})$ 为集合变量, $\delta(\cdot)$ 表示 δ 函数.

神经网络是目前人工智能技术浪潮的核心理论方法, 简而言之是由多个神经元组成的复合函数网

络. 每个神经元可以接受不同维度的输入, 经过线性组合和非线性激活函数作用后输出. 虽然原则上神经元之间的连接可以是任意的, 但受视神经层分布的启发和随之带来的并行计算方便, 常用的各种神经网络架构都是层状结构. 神经网络最有力的特点是只需要一个隐藏层, 足够多神经元组成的网络就可以无限逼近任意函数映射, 这就是著名的万能逼近理论 (universal approximation theorem) [42–44]. 但这个理论并没有指出如何在有限的神经元数目的情况下有效拟合各种映射, 所以其发现虽然在很大程度上增强了人们使用神经网络拟合各种函数映射的信心, 却并没有迅速推动其在诸多实际问题中的应用. 后来多种神经网络架构 (卷积 [45]、循环 [45]、残差 [46]、注意力机制 transformer [47] 和扩散模型 [48]) 的发展推动了神经网络在多个学科领域应用的爆发. 当然另外一个不可或缺的基础是自动微分的发现 [49] 和在神经网络中的成功应用 [50], 这使得理论上基于任意阶导数的优化方法都能够被有效用来训练神经网络参数, 当然实际应用中由于算力和内存限制, 人们往往限于使用基于一阶和二阶导数的优化方法, 诸多具体实例和相关文献可以参考 PyTorch 中的 Optim 模块. 如下所述, 在众多神经网络架构中, 分子体系自由能地貌图刻画中应用最为广泛的是自编码器 (auto-encoder) 架构 [51] (如图 1 所示), 该架构把高维输入映射到一个低维

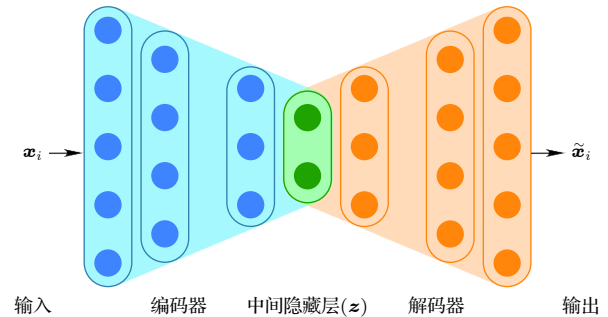


图 1 自编码器神经网络架构示意图, 蓝色部分表示编码器 (encoder) 函数 $f(\cdot)$, 橙色部分表示解码器 (decoder) 函数 $g(\cdot)$, 维度最低的绿色表示中间隐藏层 (z). 对自编码器, 损失函数是输出 ($\tilde{\mathbf{x}}_i$) 与输入 \mathbf{x}_i 的差别的函数 (也可以加正则化项, 如参考文献 [58] (5) 式所示), 每一个输入数据点对应隐藏层空间的一个点

Fig. 1. Schematic representation of an auto-encoder neural network. The blue part on the left represents the encoder, the orange part on the right represents the decoder, and the middle green layer is the hidden layer (z). The loss is always a function of the difference between the input and the output vectors (\mathbf{x}_i and $\tilde{\mathbf{x}}_i$), one may add some form of regularization when necessary (e.g. Eq. (5) in Ref. [58]).

空间的降维部分被称为编码器 (encoder), 而随后从低维逆向映射到高维 (一般与输入同维度以方便训练) 空间的部分则被称为解码器 (decoder). 这显然与人们试图在更低维度空间理解复杂分子体系的目标在形式上较为吻合. 虽然在架构形式上非常相似, 但变分自编码器 (variational auto-encoder, VAE)^[52] 的目标和训练过程却与自编码器显著不同, 其中的隐变量 (\mathbf{z}) 是个概率分布而非特定构型. 如果分别用 ϕ 和 ψ 表示编码器和解码器网络中的参数, $q_\phi(\mathbf{z}|\mathbf{x})$ 和 $p_\theta(\mathbf{x})$ 表示隐变量 (\mathbf{z}) 和 (\mathbf{x}) 的分布, 则似然函数可表述如下:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})} \right] \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \left[\frac{p_\theta(\mathbf{x}, \mathbf{z}) q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x}) p_\theta(\mathbf{z}|\mathbf{x})} \right] \right] \\ &= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \right]}_{=\mathcal{L}_{\theta, \phi}(\mathbf{x}) \text{ (ELBO)}} \\ &\quad + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \left[\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} \right] \right]}_{=D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))}, \end{aligned} \quad (2)$$

其中, $D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) \geq 0$, 所以 $\mathcal{L}_{\theta, \phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})]$ 就是似然函数的下界, 也称为证据下界 (evidence lower bound, ELBO) 或变分下界, 是变分优化的目标, 而非自编码器中解码器输出构型与数据中实际构型差别的函数. 为了对随机隐变量 (\mathbf{z}) 对自动微分, Waterfall 等^[53] 发展了二次参数化技巧 (reparameterization trick).

变分的历史非常悠久, 也是诸多理工科研究生的必修课程内容. 变分在物理、统计和工程领域都已经取得了非常广泛和成功的应用^[49,54], 如量子力学中的 Releigh-Ritz 方法^[55] 也正是本文中要讨论的分子体系变分计算的基础. 另外统计学中的大量应用展示了变分推断方法同采样计算相比高效、收敛性较好和更容易扩展的特点^[56,57]. 在神经网络广泛应用之前, 由于各种未知统计分布的解析和 (或) 参数化构造较为困难, 因此基于平均场的变分成为统计变分分析中最为常用的近似^[56]. 但在分子模拟及其增强采样中的应用却在最近十多年才陆续发生. 原因主要有两点, 其一是和很多统计模型与工程应用不同, 分子体系中的集合变量很难找

到直接的方程或模型解析描述, 其二是传统数值拟合方法 (如最小二乘法^[58]) 中导数计算昂贵且精度不高, 各种优化方法实现困难, 而且在变量较多 (大于 10 个) 时会收敛困难^[53]. 不过最近十多年以来基于自动微分^[49] 的多个人工智能框架 Pytorch^[59], Tensorflow^[60], PaddlePaddle^[61] 迅速发展成熟, 与之伴随的神经网络架构^[62] 也得到了迅猛发展. 这使得在拥有较为充足数据的前提下, 任意函数的稳健拟合成为可能, 因此增强采样和轨迹分析的变分应用也随之发展. 传统上人们探索复杂分子体系自由能地貌图的主要手段是 (加速) 采样, 变分的突出优点是用优化取代采样过程, 从而显著提高效率. 现代神经网络架构的强大拟合能力和基于自动微分的各种优化方法的结合为变分在复杂分子体系中的应用提供了巨大的潜力空间. 这也正是本文想要讨论的话题.

3 分子体系集合变量空间的变分方法

同物理学、工程和统计应用比较, 变分在复杂分子体系自由能地貌图应用相对较少, 主要是近十多年的工作, 不过目前正在迅速增长中. 目前的发展大致可以分为利用转移矩阵算子特征值和特征向量频谱分解分析 (spectral decomposition analysis) 的变分构建^[63-68]; 基于自由能垒跨越概率时间关联函数的变分^[69-71]; 利用偏置势 (bias potential) 的变分构建^[72]; 不受线性假设局限的可汇集性 (lumpability) 与可分解性 (decomposability) 泛函变分构建^[73]; 基于过去-将来信息瓶颈的变分构建^[74,75]; 同时考虑粗粒化、集合变量和增强采样的自适应^[76]; 以及直接利用变分自编码器的分析^[77], 这些方法的简要总结比较见表 1. 具体如下所述.

3.1 频谱分解分析

在严格马尔可夫过程和细致平衡假设下, 针对给定的子态构象空间划分, Perez-Hernandez 等^[63] 发展了利用演化算子 P (propagator) 特征函数自相关构建的变分实现了对最慢动力学过程集合变量 (CV) 的逼近, 分子体系动力学可以被下式表述为演化算子特征函数 ϕ_i ($i = 1, 2, \dots, \infty$) 的叠加:

$$\rho_{t+\tau}(\mathbf{y}) = P(\tau)\rho_t(\mathbf{x}) = \sum_{i=1}^{\infty} e^{-\frac{\tau}{\tau_i}} \langle \psi_i, \rho_t \rangle \phi_i, \quad (3)$$

表 1 复杂分子体系低维隐空间的变分方法简要总结, 表中所述集合空间问题类别是指引言中提到的三类问题
Table 1. A brief summary of variational methods for low-dimensional hidden spaces in complex molecular systems. The category of collective space problems mentioned in the table refers to the three types of problems defined in the introduction.

变分方法		主要目标	关注的集合空间问题类别	特点或主要局限
频谱分解分析	基组线性组合	给定构象子状态空间划分下求解集合变量和子态间转换速率	第1类、第2类	马尔可夫假设与线性基组局限, 需要人工划分构象空间子状态
	神经网络实现	从给定轨迹中直接求解子态划分和对应转换速率	第2类	马尔可夫假设, 没有解析表示的特征函数, 需要人工调整架构测试不同聚类数量
自由能垒跨越概率时间关联函数	基组线性组合	在选定基组空间的线性组合基础上求解状态转换路径和其上的自由能垒跨越概率	第3类	基组线性组合局限, 需要定义始末态
	神经网络实现	在和给定始末态一致的神经网络函数空间求解状态转换路径和其上的自由能垒跨越概率	第3类	需要定义始末态
基于偏置势变分	基组线性组合	利用偏置势增强采样在基组线性组合空间快速求解给定集合变量方向自由能主要能量谷地	第2类	泛函受基组选择限制
	神经网络实现	利用偏置势增强采样在神经网络函数空间快速求解给定集合变量方向自由能主要能量谷地	第2类	泛函导数求解的采样需求导致偏置势(和对应自由能)的精度紧密相关, 收敛受KL散度非对称性限制
Lumpability 和 Decomposability		优化集合变量	第1类	有明确误差控制, 方差取决于隐空间维度, 两种定义的一致性要求可逆过程
信息瓶颈模型		求解信息瓶颈对应集合空间CV表示, 并利用偏置势加速自由能面采样	第2类	线性编码过程假设局限
变分自适应		结合粗粒化信息加速采样求解自由能面	第2类	总体架构较为复杂
变分自编码器		通过集合变量空间加速采样求解自由能面和聚类转化路径	第2类、第3类	特别关注隐空间

其中 t_i 是和第 i 个特征值 $\lambda_i(\tau) = e^{-\tau/t_i}$ 对应的时间尺度, 尖括号代表标量积, $\langle \psi_i, \rho_t \rangle$ 的结果表征概率密度 ρ_t 和 ψ_i 的重叠程度, 体现了第 i 个特征函数对总体动力学的贡献. 因为 $\psi_i = \mu^{-1}(x) \phi_i$, 也可以认为概率密度函数 ρ 是基于特征函数 ϕ_i 展开的. 显然随着 $\tau \rightarrow \infty$, 概率密度会趋于平衡态, (3) 式中只有第 1 项有贡献, 对应于 $\lambda_1 = 1$. 如果人们感兴趣的时间尺度 $\tau \gg t_{d+1}$, 则分子体系的动力学主要取决于对应于 $(\lambda_1, \lambda_2, \dots, \lambda_d)$ 的 d 个特征函数, 也对应于前面 (见方程 (1) 中的 $s(\mathbf{R})$ 定义) 所说的 d 个集合变量. (3) 式可近似为

$$\rho_{t+\tau} = P(\tau) \rho_t \approx \sum_{i=1}^d e^{-\frac{\tau}{t_i}} \langle \psi_i, \rho_t \rangle \phi_i. \quad (4)$$

对于分子体系坐标的任意函数 $f(x)$, 其自相关函数可以表述为

$$\langle f(\mathbf{x}_t) f(\mathbf{x}_{t+\tau}) \rangle_t = \sum_{i=1}^d e^{-\frac{\tau}{t_i}} \langle \phi_i, f \rangle^2. \quad (5)$$

显然如果取 $f = \psi_i(x)$, 则

$$\begin{aligned} \left[\lambda_i^\dagger(\tau) = \langle \psi_i(\mathbf{x}_t) \psi_i(\mathbf{x}_{t+\tau}) \rangle_t = e^{-\frac{\tau}{t_i}} \right], \\ \left[t_i^\dagger = -\tau / \ln |\lambda_i^\dagger(\tau)| = t_i \right]. \end{aligned} \quad (6)$$

对于近似的特征函数 ψ_2^\dagger :

$$\langle \psi_2^\dagger(\mathbf{x}_t) \psi_2^\dagger(\mathbf{x}_{t+\tau}) \rangle_t \leq e^{-\frac{\tau}{t_2}}, \quad t_2^\dagger \leq t_2. \quad (7)$$

因此近似时间尺度 t_2^\dagger 可以作为变分目标. 文中时间尺度最大化结果的实现依靠选定基函数 (具体来说使用了分子体系构型在特定构象子空间的示性函数 (indicator function)) 的线性组合来近似特征函数. 该方法虽然在给定构象聚类的前提下能够给出较好的动力学常数和对应的特征函数 (集合变量) 估计, 但实现构象聚类的过程依然依靠简单的降维方法. 类似地 Tiwary 和 Bern^[78] 通过最大化频谱间距 (spectral gap) 过渡路径熵也展开了给定集合变量空间的线性组合优化, 不过没有使用变分方法. 后来为了拓展该理论的应用范围, Wu 和 Noé^[66] 把该方法推广到不需要细致平衡的一般马尔可夫过程. 尽管多数复杂分子体系的动力学过程一般都是非线性的, 但依据 Koopman 理论^[66,76] 可以构建新的隐空间 $\chi_0(\mathbf{x}) = (\chi_{01}(\mathbf{x}), \chi_{02}(\mathbf{x}), \dots, \chi_{0d}(\mathbf{x}))^T$ 和 $\chi_1(\mathbf{x}) = (\chi_{11}(\mathbf{x}), \chi_{12}(\mathbf{x}), \dots, \chi_{1d}(\mathbf{x}))^T$ 使得原来坐标系中的非线性变换被转化为如下式所示的线性变化:

$$\mathbb{E}[\chi_1(\mathbf{x}_{t+\tau})] \approx \mathbb{K}^T \mathbb{E}[\chi_0(\mathbf{x}_t)]. \quad (8)$$

显然 (8) 式中转化过程 χ_0 和 χ_1 一般是非线性并且未知的. 神经网络的可训练万能逼近能力为其构建

未知转换的提供了可能性. VAMPNets^[67]正是在这种思考下构建的. 对于给定的 χ_0 和 χ_1 变换可以构建 3 个方差矩阵:

$$\begin{aligned} C_{00} &= \mathbb{E}_t[\chi_0(\mathbf{x}_t)\chi_0(\mathbf{x}_t)^\top], \\ C_{01} &= \mathbb{E}_t[\chi_0(\mathbf{x}_t)\chi_1(\mathbf{x}_{t+\tau})^\top], \\ C_{11} &= \mathbb{E}_{t+\tau}[\chi_1(\mathbf{x}_{t+\tau})\chi_1(\mathbf{x}_{t+\tau})^\top]. \end{aligned} \quad (9)$$

这些方差矩阵被用来构建了一个 VAMP-2 打分^[66]:

$$\hat{R}_2[\chi_0, \chi_1] = \left\| C_{00}^{-1/2} C_{01} C_{11}^{-1/2} \right\|_F^2. \quad (10)$$

该分值最大化时对应转化后的低 (d) 维空间分子构象分布被准确复现. 以这个分值作为损失函数的神经网络通过训练就有可能实现从体系原始高维坐标向低维空间较为准确的映射, 实际上起到了低维空间模糊分类器的功能, 消除了前述变分理论^[65]中对人工聚类及以前各个步骤的需求. 具体实现架构如图 2(a) 所示^[67]. 对丙氨酸二肽体系, Mardt 等^[67]利用图 2(b) 的特定架构, 设定低维空间类别数目为 6(也尝试了从 2—8 的其他类别数目), 首先从 250 ns 的分子动力学模拟轨迹中每秒提取一帧得到 250000 个构型, 并通过和第一帧对齐除去分子的整体平移和旋转. 使用十个重原子的三维空间(即长度为 30 的向量)坐标作为神经网络输入, 取延迟时间 $\tau=40$ ps(也尝试了从 4—32 ps 的其他延迟时间), 通过最大化 VAMP 打分, 成功实现了在二维二面角空间 (φ, ψ) 的构象聚类. 他们同时使用 MSM 流程聚类, 当构象类别数目小于 20 时得到 VAMP 打分都低于 VAMPnets 的结果. 此外 Mardt 等^[67]还尝试分析了简单双势阱和 NTL9 蛋白折叠轨迹, 均展示了和原来人工复杂流程可比拟的准确性, 也说明这个思路有望在将来通过逐步发展真正实现自动分析分子模拟轨迹得到动力学特性的可能性. 不过目前该方法还不够成熟, 尚不能用于多系综组合数据^[79–84], 也不能有效集成模拟轨迹与相关实验数据, 另外还缺乏严格清楚定义的误差估算指标^[85–87]. 但该研究结合变分理论和非线性的神经网络拟合, 取代了原来 MSM 方法管线中一系列复杂步骤, 并在简单体系中实现了首次成功应用, 是人工智能用于分析复杂分子体系轨迹的重要进展. VAMPNets 的神经网络架构较为简单, 鉴于图神经网络^[88–91]和注意力机制^[92]在网络型数据中的优异表现, 考虑到复杂分子体系可以被视为由相互作用的单元构成的网络图, Brooks 等^[93]

构建了包含这两种架构要素的 GraphVAMPnet, 该模型实现了更高精度的构象嵌入表示, 也能够通过注意力机制给出蛋白质中对结构聚类起决定性作用的重要氨基酸. 在 20-氨基酸的 Trp-cage 蛋白, 35-氨基酸的 Viliin 蛋白和 NTL9 蛋白轨迹上的成功应用展示了这些神经网络构架改变的好处.

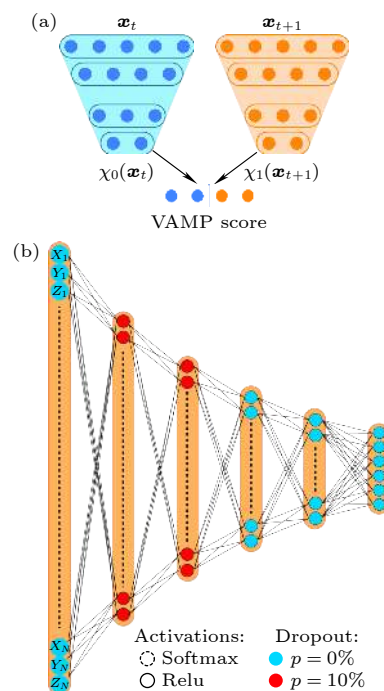


图 2 (a) VAMPnets 构建 VAMP 打分 ((10) 式) 的神经网络总体架构示意图; (b) 丙氨酸二肽轨迹分析实例中的典型神经网络架构, 各层神经元数目为 32-22-16-9-6, 前两层使用 10% 的 dropout, 除最后的 softmax 层外, 其余各层激活函数均使用 ReLU^[67]

Fig. 2. (a) Schematic illustration of VAMP score construction from VAMPnets (see Eq. (10)). (b) A typical neural network architecture for alanine dipeptide analysis, with the number of neurons being 32-22-16-9-6 for five layers. The first two layers utilized a 10% dropout. ReLU was selected as the activation function for all layers except the last softmax layer^[67].

随着人们使用电子显微镜解析生物大分子复合体的能力越来越强, 如何解释这些复合体的动力学过程变成了亟待解决的问题. 为了增进处理较大分子的能力并在将来能够有可能延伸到大复合体, Noé等^[94]结合独立马尔可夫分解方法 (independent Markov decomposition, IMD) 构建了由多个独立的 VAMPNets 构成的 iVAMPNets. 其中不同独立模块的划分由一个可训练的 MASK 实现, 通过竞争训练使每个不同的子网络仅处理不与其他子网络相互重叠的部分. 虽然该方法在 Syna-

ptotagmin-C2A 蛋白质分子中成功应用, 但显然这种处理仅适用于不同子模块间耦合程度较弱的情况, 距离准确描述不同组成分子之间有较强关联的复杂复合体仍然有较大距离. 利用 VAMPNets 输出的子构象空间 (状态) 概率, Kleiman 和 Shukla^[68] 尝试了结合 3 种不同后续处理, 包括最小计数 (least count, LC), 多目标强化学习 (multiagent reinforcement learning-based, MA REAP) 和最大熵 (MaxEnt), 显著促进了构象空间搜索能力. 这 3 种方法的宗旨基本一致, 就是利用前期生成的轨迹对 VAMPNets 进行初步训练后, 在后续的采样中按照上述不同标准聚焦前期采样最少访问的构象空间, 从而实现更进一步的增强采样. 其中最大熵和 VAMPNets 的结合在促进采样的同时消除了聚类步骤.

3.2 自由能垒跨越概率时间关联函数的变分

弦方法^[95-99]和过渡路径理论 (transition path theory, TPT)^[40] 致力于寻找不同亚稳态之间过渡路径及其过渡态的关键细节. 不过这些方法在得到最低自由能过渡路径的同时, 却不能直接给出人们非常感兴趣的路径上任意一点的自由能垒跨越概率. 针对此问题, 文献^[71,100] 基于两个亚稳态之间的净向前反应通量构造了自由能垒跨越概率时间关联函数, 发展了通过变分最小化该函数获得最佳过渡路径并同时给出自由能垒跨越概率的方法. 对两个亚稳态 A 和 B, 集合变量空间从 A 到 B 在时间步长 τ 基于算子 $P_\tau(s'|s)$ 的向前演化可表示为

$$\rho(s'; t + \tau) = \int ds P_\tau(s'|s) \rho(s, t), \quad (11)$$

其中 $\rho(s, t)$ 和 $\rho(s'; t + \tau)$ 分别对应于时刻 $t(t + \tau)$ 在路径位置 $s (s')$ 处的概率密度. 则自由能垒跨越概率 $q(s)$, 即从 s 开始最终到达亚稳态 B 并且在此前从未到达亚稳态 A 的所有过渡路径概率之和, 可定义如下:

$$q(s) = \int ds' q(s') P_\tau(z'|z). \quad (12)$$

则净向前 (从 A 到 B) 反应流为

$$J_{AB} = \frac{1}{2\tau} \int ds \int ds' (q(s') - q(s))^2 P_\tau(s'|s) \rho_{\text{eq}}(s). \quad (13)$$

也可以表达为自由能垒跨越概率的自相关函数:

$$J_{AB} = \frac{1}{2\tau} \left\langle (q(\tau) - q(0))^2 \right\rangle \\ = \frac{1}{\tau} (\langle q(0)q(0) \rangle - \langle q(\tau)q(0) \rangle), \quad (14)$$

其中二次形式可以作为任意给定始末态时尝试自由能垒跨越概率 $q(s')$ 的变分优化目标. 该方法使用基组展开, 通过优化系数来达到变分优化的目标, 在模型双势阱问题中展示了简化子空间 (CV 空间) 中理想一维反应坐标走向沿着自由能垒跨越概率梯度, 与高维空间中的 Kramers-Langer 理论^[101,102] 一致. 文献^[100] 是针对过渡路径变分构建的首次尝试, 并在双势阱问题和丙氨酸二肽中展示了应用. 由于变分函数限于选定基组函数的线性组合空间, 其结果显然会受到基组选择和线性组合的制约. Chipot 等^[69] 将自由能垒跨越概率时间关联函数的变分方法延伸到了神经网络 (variational committor-based neural networks, VCN), 从而可以拟合任意非线性映射. 同基于特征值变分优化的 VAMPNets 相比较, 在双势阱体系 and N-acetyl-N'-methylalanylamine 异构化过程中均得到一致结果. 不过显著不同的是 VCN 需要已知始末态, 针对的目标是一对始末态之间的过渡路径, 而 VAMPNets 则是从轨迹数据开始的无监督学习. 另外一点是有时候人们最感兴趣的慢过程可能不是分子体系中最慢的过程, 这种情况下显然 VCN 更为适合. 这两类方法可以协同使用从而结合其各自优势, 当然也有可能在未来集成到更复杂的神经网络架构中.

3.3 基于偏置势的变分

在给定 CV 的前提下, Valsson 和 Parrinello^[103] 构建了一个基于 CV 空间偏置势 $V(s)$ 的泛函:

$$\Omega[V] = \frac{1}{\beta} \log \frac{\int ds e^{-\beta[F(s)+V(s)]}}{\int ds e^{-\beta F(s)}} + \int ds p(s) V(s), \quad (15)$$

其中 $p(s)$ 是一个自由选择的目标分布, 这赋予人们使用该泛函的灵活性 (当然也伴随着选择的挑战). 该泛函是一个凸函数并且不随偏置势任意给定的有限常数的改变而变化. 用 $F(s)$ 表示体系自由能, 则当 $V(s) = -F(s) - (1/\beta) \log p(s)$ 时, 泛函 $\Omega[V]$ 取极小值, 因此在选定 $p(s)$ 的前提下通过参数化的 $V(s)$, 以 $\Omega[V]$ 极小值为目标的变分优化即

可求解自由能地貌图. 该方法使用线形基组组合在丙氨酸三肽分子中成功应用. 另外, 该泛函同 Kullback-Leibler (KL) 散度 (D_{KL}) 的关系如下所示 [46,104,105]:

$$\beta\Omega[V] = D_{\text{KL}}(p||P_V) - D_{\text{KL}}(p||P_0), \quad (16)$$

其中 P_V 和 P_0 分别是偏势为 V 和 0 时体系的概率密度分布. 由于凸函数特性, 使得偏置势与自由能面有确定关系的驻点也是其极值点. 因此通过参数化偏置势, 就可以对参数实施变分优化从而求解自由能面. 这在原理上比元动力学采样方法要高效很多, 不过, 其表现受限于所选 CV 在较长时间尺度上描述自由能面的能力. 为了克服对该泛函线性展开可能出现的一些麻烦 (比如自由能变化剧烈的区域需要很多项才能实现较好拟合, 集合变量增大时需要变分优化的参数空间指数增长), Bonati 等 [72] 用神经网络表示偏置势泛函, 在给定的集合变量定义下通过优化神经网络参数实现, 如下所示:

$$\frac{\partial\Omega(\alpha)}{\partial\alpha_i} = -\left\langle \frac{\partial V(\mathbf{s}; \alpha)}{\partial\alpha_i} \right\rangle_{V(\alpha)} + \left\langle \frac{\partial V(\mathbf{s}; \alpha)}{\partial\alpha_i} \right\rangle_p. \quad (17)$$

泛函数值微分需要统计平均 ((17) 式中的尖括号表示系综平均), 因此需要采样获取. 直接高精度确定最低点较为困难, 因此 Bonati 等 [72] 在实现过程选择获得达到一定近似程度的偏置势, 评判的标准选用了 $p_V(\mathbf{s})$ 和 $p(\mathbf{s})$ 在迭代次数 n 时的 KL 散度距离:

$$D_{\text{KL}}^{(n)}(p_V || p) = \sum_{\mathbf{s}} p_V^{(n)}(\mathbf{s}) \log \frac{p_V^{(n)}(\mathbf{s})}{p^{(t)}(\mathbf{s})}. \quad (18)$$

显然, 此过程在数值实现中需要选定两个参数, 一个是选定每次迭代计算 KL 散度之间的模拟更新次数, 另一个是每次更新时学习率调整的幅度. 为了集成 CV 构建和偏置势优化, Bonati 等 [106] 利用 VAMPNets 的 VAMP 打分作为损失函数, 利用深度神经网络和 TICA (time-structure based independent component analysis) 结合生成 CV, 随后在更新的 CV 空间采用 OPES [100] 增强采样思路, 实现了 CV 优化和自由能地貌图收敛的迭代. 他们在丙氨酸二肽、chignolin 蛋白折叠和材料结晶过程的成功展示了该方法的应用 [106].

3.4 基于可汇集性 (lumpability) 和可分解性 (decomposability) 的非线性变分描述

由于马尔可夫假设和特征函数构建中的线性

假设, 基于频谱分解分析的变分优化无法正确处理非马尔科夫过程 [40] 和线性无关特征函数之间的非线性关联, 这些根本上的局限无法在后期变分优化中被消除. 针对这个问题, Bitttracher 等 [73] 通过延伸过渡流形理论 (transition manifold theory) 发展了不包含任何线性假设, 只关注于长时间尺度分子体系行为, 显式包含误差量且在可逆体系中互相等价的条件, lumpability 和 decomposability (详见文献 [73] 的 definition 3.2, 3.4), 这两个条件都可以作为损失函数变分. 此外该变分在近似损失函数时只要求在集合变量空间的稀疏采样, 而且损失函数的蒙特卡罗积分误差取决于集合变量空间而非原高维空间的方差, 这会带来巨大的算力节省. 该理论和过渡路径理论连接仍然有待阐明. 另外这些理论上的优势在百万级甚至更大的复杂分子体系如何得以实现也有待于进一步探索.

3.5 过去-将来信息瓶颈模型

Wang 等 [74] 将分子体系中的集合变量空间视为其演化过程中的过去-将来信息瓶颈 (past-future information bottleneck, PIB [107,108]), 对给定分子体系任意时刻坐标 \mathbf{X} 和下一时刻坐标 $\mathbf{X}_{\Delta t}$, 通过瓶颈变量 χ (与集合变量类似的分子体系低维空间描述) 分别和编码器 $P(\chi|\mathbf{X})$ 与解码器 $P(\mathbf{X}_{\Delta t}|\chi)$ 联系 (注意文献 [74] 中结果部分第 1 段把坐标 \mathbf{X} 误解释为 N 个粒子体系中的 d 维 ($1 \ll d \ll N$) 表示, 容易引起混乱). PIB 的目标是瓶颈变量 χ 相对于过去应该尽量简单但对于将来则应该有尽可能好的预测力, Wang 等 [74] 据此构建了如下优化目标:

$$\mathcal{L} \equiv I(\chi, \mathbf{X}_{\Delta t}) - \gamma I(\mathbf{X}, \chi), \quad (19)$$

其中 $I(\chi, \mathbf{X}_{\Delta t})$ 和 $I(\mathbf{X}, \chi)$ 分别表示瓶颈变量与 $\mathbf{X}_{\Delta t}$ 和 \mathbf{X} 的互信息, 常数 $\gamma \in [0, \infty)$ 用来平衡瓶颈变量 χ 的复杂程度和预测力. 进一步通过选择确定性的线性编码器, 则第 2 项可以忽略. 他们然后利用 Gibbs 不等式构建了可变分优化的 PIB 下限近似:

$$\begin{aligned} I(\chi, \mathbf{X}_{\Delta t}) &= H(P_\theta(\mathbf{X}_{\Delta t})) - H(P_\theta(\mathbf{X}_{\Delta t}|\chi)) \\ &\geq H(P_\theta(\mathbf{X}_{\Delta t})) - C(P_\theta(\mathbf{X}_{\Delta t}|\chi)|Q_\phi(\mathbf{X}_{\Delta t}|\chi)), \end{aligned} \quad (20)$$

其中 H 和 C 分别表示香农和交叉熵, Q_ϕ 为随机深度神经网络构建的解码器. 由于选择 P_θ 为确定性

线性编码器, 香农熵项退出优化目标, 可得更新变分下界:

$$\mathcal{L} \geq \mathcal{L}' = -C(P_\theta(\mathbf{X}_{\Delta t}|\chi) | Q_\phi(\mathbf{X}_{\Delta t}|\chi)), \quad (21)$$

其中 ϕ 为随机神经网络中的变分优化参数. 对平衡态轨迹 $\{\mathbf{X}^1, \dots, \mathbf{X}^{M+k}\}$ (\mathbf{X}^n 和 \mathbf{X}^{n+k} 之间的时间间隔为 Δt), 方程 (20) 可被离散为

$$\mathcal{L}' = \frac{1}{M} \sum_{n=1}^M \log Q(\mathbf{X}^{n+k}|\mathbf{X}^n), \quad (22)$$

其中 χ^n 从 $P(\chi^n|\mathbf{X}^n)$ 中采样得到. 对于有对应偏置势 $\{V^1, V^2, \dots, V^{M+k}\}$ 下模拟的轨迹则可在假设偏置势不改变解码器的情况下近似表述为

$$\mathcal{L}' = \left\{ \sum_{n=1}^M e^{\beta V^n} \right\}^{-1} \sum_{n=1}^M e^{\beta V^n} \log Q(\mathbf{X}^{n+k}|\mathbf{X}^n). \quad (23)$$

实际计算中 Wang 等 [74] 选择用坐标的线性基组组合得到 CV, 首先对平衡态轨迹通过逐步增加 Δt 观察基组各项的权重变化, 并取其趋于稳定后最小的 Δt . 随后则按照方程 (24) 和 (25) 计算偏置势并重新估算机组系数, 反复迭代:

$$V_{\text{bias}}(\chi) = k_B T \log P^u(\chi), \quad (24)$$

$$P^u(\chi) \propto \frac{\langle w^\delta(\chi - \chi(t)) \rangle_b}{\langle w \rangle_b} \equiv e^{-\beta F(x)}, \quad (25)$$

其中 $w = e^{\beta V_{\text{bias}}}$, $P^u(\chi)$ 是没有偏置势的情况下 χ 的平衡态分布. 简单的确定性线性编码器在带来方便的同时也在一定程度上限制了该方法的灵活性, 但 PIB 的优点之一是原则上没有其他线性假设, 不过在 PIB 思路下 (见 (19) 式) 使用非线性编码器后的变分优化方法仍有待发展. 该方法在苯-溶菌酶复合体模拟中获得了成功, 在几百纳秒的加速模拟中观察到了几百毫秒常规模拟所观察到的解离过程. Beyerle 等 [75] 后来使用该方法成功描述了双势阱模型和苯甲酸在双分子层膜中扩散这两个分别由能量和熵主导的过渡路径, 进一步展示了该方法的稳健性.

3.6 变分自适应

与前述变分方法主要关注集合变量和偏置势不同, 对有明确集合变量的情况, Zhang 等 [76] 结合生成式深度学习和基于能量模型 [109] (energy based models, EBM) 发展了变分对抗密度估计变分, 直接计算自由能地貌图中的概率密度分布. 将平衡态真实

自由能对应的概率分布记为 p , 在集合变量空间的参数化自由能地貌图和对应的分布分别记为 $F_\theta(s)$ 和 $p_\theta(s)$, 则 KL 散度 $D_{\text{KL}}(p||p_\theta)$ 对 θ 的导数可表示为

$$\nabla_\theta D_{\text{KL}}(p||p_\theta) = \langle \beta \nabla_\theta F_\theta(s) \rangle_{p(s)} - \langle \beta \nabla_\theta F_\theta(s) \rangle_{p_\theta(s)}, \quad (26)$$

其中 $\langle f(x) \rangle_{p(x)}$ 表示函数 $f(x)$ 在分布 $p(x)$ 下的期望值. (26) 式和对抗神经网络 [76,110] 高度相似, 因此在原文中被称为变分对抗密度估计 (variational adversarial density estimation, VADE). 在实际操作中可以用粗粒化实验数据 $P_{\text{FG}}(s)$ 取代真是分布 p . 再通过粗粒化模拟计算 $\langle \beta \nabla_\theta F_\theta(s) \rangle_{p_\theta(s)}$. 对于集合变量维度较高的情况, 由于直接采样计算代价过于昂贵, Zhang 等 [76] 通过加入可训练生成神经网络模块作为神经采样器 (neural sampler) q_ψ , 采用下式实现变分训练:

$$D_{\text{KL}}(q_\psi||p_\theta) = \langle \log q_\psi(s) + \beta F_\theta(s) \rangle_{q_\psi(s)} + \log Z_\theta. \quad (27)$$

加速分子体系自由能地貌图统计概率分布参数的训练. 对于没有集合变量函数的更一般情况, 通过加入了强化学习模式, 较好地解决了固定偏置势在动态采样中尴尬的同时, 实现了粗细两个不同粒度的有效采样补充. 这些方法都被集成在 SPONGE [111] 平台上.

3.7 变分自编码器的直接应用

上述构建显式变分优化目标函数的做法能够给出更有效的物理图像, 神经网络主要用于拟合其中未知非线性映射过程. 不过即使没有直接显式变分目标函数的构建, 变分的思想依然可以被利用. 最简单的做法就是直接使用变分自编码器 VAE 架构 [77] 对自己感兴趣的目标数值分布进行优化, 同时在生成的隐空间 (对应于分子体系集合变量空间) 展开一系列增强采样的操作, 必要时再引入迭代机制.

Ribeiro 等 [112] 发展了重配权变分贝叶斯增强采样 (reweighted auto encoded variational Bayes for enhanced sampling, RAVE) 方法, 通过隐空间分布和模拟的 KL 散度优化自编码器, 更新偏置势模拟后迭代优化直至收敛, 实现了独立于传统方法的隐空间增强采样. 针对在 MSM 模型中使用过渡路径理论方法时会得到大量量子状态之间的路径, 从

而使结果难以理解的困境, Qiu 等^[113,114] 利用 VAE 的数值分布变分优化, 在隐空间实现了类似过渡路径的合并. 该方法被成功应用在两个不同的简单体系, 分别是一对疏水粒子在水溶液中的聚集和 Fip35WW 结构域折叠路径的分析中. 利用 VAE 能够有效预测编码空间、隐空间和解码空间概率密度的特性, Monroe 和 Shen^[115] 发展了基于隐空间的蒙特卡罗移动建议方法, 再通过编码和解码, 从而在真实高维空间有效且高接受率的移动. 该方法的突出优点是直接满足细致平衡要求, 不需要一般偏置势加速采样生成轨迹后的权重调整, 从而避免了与之伴随的所有潜在问题和困难. 这个思路和通过粗粒化模拟促进 (细粒度) 全原子模拟^[76], 以及把低维子空间视为信息瓶颈^[74,75] 的具体方式虽然差别较大, 但总体基本思路一致. 不同粒度之间更加高效准确的构型映射和信息传递还有很大的方法学发展空间, 这方面的新发展也大概率会显著促进复杂分子体系高精度多尺度模型的构建.

4 其他神经网络方法在自由能地貌图相关研究中的应用

神经网络网络的万能逼近能力使得其在自由能面探索中从多个角度被加以应用. 其中很多工作都致力于获得更好的集合变量以改善复杂体系的增强采样. 早在 2005 年, Ma 和 Dinner^[116] 就开始使用神经网络用来寻找复杂体系的反应坐标. 针对各种传统降维方法不能直接把结果中的低维空间 (集合) 变量表达为原空间坐标的问题, Chen 和 Ferguson^[117] 利用自编码器可以实现从高维输入空间到低维隐空间之间的可训练映射, 把通过已有轨迹数据训练生成的隐空间自由度作为集合变量, 从而实现了集合变量偏置势通过自编码器对高维空间坐标的直接微分计算偏向受力, 集成了集合变量的神经网络构建和在加速采样中的直接应用, 该方法在丙氨酸二肽和 TrpCage 蛋白体系中被成功使用. 与此类似, Chen 等^[118] 也采用自编码器进行降维训练获得 CV, 然后通过自动微分把施加于 CV 上的偏置势传递到分子体系中去实现模拟采样和自由能计算.

与使用变分优化特征函数不同的另外一种思路是回归方法. Wehmeyer 和 Noé^[119] 尝试了选择对 N 个连续时间坐标序列 ($\mathbf{X}_t, \mathbf{X}_{t+\tau}, t = 1, 2, \dots, N$) 最小化回归误差^[120-122]:

$$\min_{D,E} \sum_t |\mathbf{X}_{t+\tau} - D(E(\mathbf{X}_t))|^2, \quad (28)$$

其中 D 和 E 分别为编码器和解码器. 在对已有轨迹数据的时间序列坐标构型按照下式进行均值归零 ((28) 式和 (29) 式) 和白化 ((30) 式和 (31) 式):

$$\mathbf{x}_t^{mf} = \mathbf{X}_t - \frac{1}{T-\tau} \sum_{s=1}^{T-\tau} \mathbf{X}_s, \quad (29)$$

$$\mathbf{y}_t^{mf} = \mathbf{x}_{t+\tau} - \frac{1}{T-\tau} \sum_{s=1}^{T-\tau} \mathbf{X}_{s+\tau}, \quad (30)$$

$$\tilde{\mathbf{x}}_t = C_{00}^{-\frac{1}{2}} \mathbf{x}_t^{mf}, \quad C_{00} = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathbf{x}_t^{mf} \mathbf{x}_t^{mf\top}, \quad (31)$$

$$\tilde{\mathbf{y}}_t = C_{\tau\tau}^{-\frac{1}{2}} \mathbf{y}_t^{mf}, \quad C_{\tau\tau} = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathbf{y}_t^{mf} \mathbf{y}_t^{mf\top}. \quad (32)$$

然后对处理后的坐标优化训练, 实现编码器降维和解码器对原空间的映射:

$$\min_{E,D} \sum_{t=1}^{T-\tau} \|\tilde{\mathbf{y}}_t - D(E(\tilde{\mathbf{x}}_t))\|_2^2. \quad (33)$$

通过训练过程中在输出端使用相对输入端 t 时刻的延后 $t + \tau$ 时刻坐标, 也实现了演化的预测. 对于在构象空间中线性可分的不同亚稳态, 该方法被证明同 Koopman 模型^[49,68] 等价. 但对非线性可分的体系, 与 PCA 和 TICA 及人工构造特征空间相比, 文献^[119] 的丙氨酸二肽体系显示通过编码器和解码器的深度学习拟合则可以更好地处理.

Zhang 和 Chen^[123] 针对不恰当的 CV 会在其正交空间出现亚稳态简并 (degeneracy) 从而导致对应方向不能加速采样的问题, 发展了利用随机动力学嵌入 (stochastic kinetic embedding, StKE) 的半监督学习方法增加对当前信息最匮乏区域 (current least informative regions, CLIRs) 的主动学习采样 (active enhanced sampling, AES), 这与 Kleiman 和 Shukla^[68] 在 VAMPNets 输出构象类型采样最少的部分增加后续采样的思路类似. 该方法成功在丙氨酸二肽和五肽 met-enkephalin 体系中从随意给定的无效 CV 开始, 以较短时间实现了对自由能地貌图的可靠采样. Rydzewski 和 Valsoson^[124] 提出的多尺度重配权重随机嵌入 (multi-scale reweighted stochastic embedding, MRSE) 则在此基础上更进一步, 通过高斯混合模型描述高维特征空间和重配权重, 实现对平衡态和偏置势采

样数据在训练中的有效使用. 该方法被 Rydzewski 和 Valsson 应用到 Müller-Brown Potential 以及丙氨酸二肽和四肽体系, 也已被整合到开源的 PLUMMD 软件包 (<https://www.plumed-nest.org/eggs/21/023/>). 类似地, Belkacemi 等^[125] 发展了利用自编码器的自由能偏置势迭代学习 (free energy biasing and iterative learning with auto encoders, FEBILAE), 该方法可以对在平衡态或者偏置势下采样的轨迹重配权重后作为自编码器的输入 (既可以是原来构象空间的, 也可以是某种转换之后的构型). 其中自编码器的瓶颈层确定了 CV 的维度, 但显然需要自行选择, 他们也给出了探索的建议. 可能的问题是迭代收敛的 CV 并不能保证自由能地貌图的全局充分采样. 和大多数类似研究一样, 这类编码过程不具备直接可解释性, 人们无从知道输入构型中不同的参数对 CV 的贡献. 虽然原则上可以间接从计算过程中的自动微分步骤获取一定信息, 但目前所有的方法中没有提供这种分析. 针对这个问题, Kikutsuji 等^[126] 利用模型无关的局部解释 (local interpretable model agnostic explanation, LIME) 和沙普利加和解释 (shapley additive explanations, SHAP) 框架, 给出各个输入量对 RC 的贡献, 能够在一定程度上增进我们对体系的直观物理认知.

Sun 等^[127] 发展了由一个降维编码器, 构象分类器和势能预测器组成的多任务 CV 学习构架, 在几个简单测试系统 (包括 5D Müller Brown model、丙氨酸二肽和金 (110) 晶面重建单元反应体系) 与单目标训练优化相比较展示了一定优势. 与很多应用中系统演化过程在原有高维空间进行不同, 隐空间模拟器^[128] (latent space simulator, LSS) 在训练产生编码器和解码器后, 在 CV 空间快速展开系统演化, 然后通过解码器生成原有高维空间的细粒度轨迹. 这些在隐空间或者集合变量空间进行操作的思路是很多工作中利用自编码器的重要方式. 比大多数方法在诸如丙氨酸二肽或类似模型体系中展示更进一步的是该方法在两个较大体系 (264 残基的 PROTAC 蛋白和 DNA 序列 5'-GCCGTTTCCGC-3' 对应的双螺旋结构) 获得了较为成功的应用.

Jung 等^[129] 以水溶液中离子的聚集和聚合物折叠为例, 集成了深度学习和过渡路径理论实现了复杂分子体系自组织模型的构建、验证和更新, 并

在此基础上通过符号回归总结出更容易理解的可观测量连接, 是分子复杂体系的深度学习和可解释性方面有意义的尝试和进步. 比变分求解自由能上界更进一步, Zhao 和 Wang^[130] 用流匹配 (flow matching) 同时求解上下界, 从而提供更好地逼近目标体系自由能的可能途径.

鉴于生成式模型在语言图像绘画等方面的巨大成功^[131], Janson 等^[132] 基于生成对抗模型和 transformer 架构训练的构象系综生成神经网络成功产生了训练数据集中没有的内秉无序蛋白 (IDP) 构象, 该过程与分子模拟直接采样相比所用计算代价非常小, 不过正确性依然有待进一步在更多体系中验证.

5 结 论

综上所述, 变分方法处理分子体系自由能地貌图目前已经有了较多不同视角的尝试, 但还都限于在较为简单的体系探索, 和其他理论上不甚严格的刻画分子体系自由能地貌图的神经网络方法相比较也还没有展示出明显的系统优势. 比如使用变分的 VAMPNets^[67] 和使用回归^[119] 两种方法在丙氨酸二肽体系中就没有明显的表现差异. 不过变分更严格的理论基础有可能会让误差控制更加容易, 也很可能会在将来较大分子体系的应用和进一步发展中体现出更多的优势. 从理论方法的角度, 现有的这些不同变分目标函数都是为了更好地逼近分子体系自由能地貌图的准确描述, 如何将它们集成并能够依据应用需求灵活选择关注视角显然是个有价值的任务. 当前的变分和自编码器模型中还有很多需要人工调节和尝试的环节, 最为突出的就是目前的所有方法都不能通过自主学习优化获得自编码器中间低维隐空间的适当维度. 另外变分计算本身原则上也可以在神经网络中数值实现, 从而有可能增加灵活性和可泛化能力, 不过目前尚没有见到这类尝试, 有可能是个有价值的发展方向.

从应用的角度, 目前最迫切需要解决的问题可能是将这些变分构建向更大更复杂分子体系的延伸. 从自由能地貌图构象空间的层次来看, 超过两个时空间尺度的体系显然会带来更多挑战, 在同一个自由能地貌图时空间尺度层次上, 多个亚稳态之间过渡路径交汇的可能性和准确处理也有待解决. 这些问题的可靠处理在较大的复合体分子机器的

理解中很有必要.

目前大模型的应用如火如荼^[133], 不过在 AI 的科学应用领域尚没有发力. 主要原因之一是作为通用大模型训练素材的语音图像材料非常丰富, 而特定科学领域的数据一般都不够丰富或者很多都难以理解. 不过这些模型集成多模态的能力显然对 AI 在广泛科学应用中和特定的复杂分子体系中都有参考价值. 已有的这些变分构建方法, 还有未来可能出现的其他新颖构建, 很可能在将来被统一到一个多目标大模型中.

参考文献

- [1] Thomas C, Tampe R 2020 *Annu. Rev. Biochem.* **89** 605
- [2] Jiang F, Doudna J A 2017 *Annu. Rev. Biophys.* **46** 505
- [3] Latorraca N R, Venkatakrishnan A J, Dror R O 2017 *Chem. Rev.* **117** 139
- [4] Wei G, Xi W, Nussinov R, Ma B 2016 *Chem. Rev.* **116** 6516
- [5] Dignon G L, Best R B, Mittal J 2020 *Annu. Rev. Phys. Chem.* **71** 53
- [6] Choi J M, Holehouse A S, Pappu R V 2020 *Annu. Rev. Biophys.* **49** 107
- [7] Spomer J, Bussi G, Krepl M, et al. 2018 *Chem. Rev.* **118** 4177
- [8] Bussi G, Laio A 2020 *Nat. Rev. Phys.* **2** 200
- [9] Mobley D L, Gilson M K 2017 *Annu. Rev. Biophys.* **46** 531
- [10] Rodnina M V, Beringer M, Wintermeyer W 2007 *Trends Biochem. Sci.* **32** 20
- [11] Bernardi R C, Melo M C R, Schulten K 2015 *Biochim. Biophys. Acta* **1850** 872
- [12] Sugita Y, Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141
- [13] Faraldo-Gomez J D, Roux B 2007 *J. Comput. Chem.* **28** 1634
- [14] Laio A, Parrinello M 2002 *Proc. Natl. Acad. Sci. U. S. A.* **99** 12562
- [15] Barducci A, Bussi G, Parrinello M 2008 *Phys. Rev. Lett.* **100** 020603
- [16] Maragliano L, Vanden-Eijnden E 2006 *Chem. Phys. Lett.* **426** 168
- [17] Abrams J B, Tuckerman M E 2008 *J. Phys. Chem. B* **112** 15742
- [18] Darve E, Rodriguez-Gomez D, Pohorille A 2008 *J. Chem. Phys.* **128** 144120
- [19] Torrie G M, Valleau J P 1977 *J. Comput. Phys.* **23** 187
- [20] Carter E A, Ciccotti G, Hynes J T, Kapral R 1989 *Chem. Phys. Lett.* **156** 472
- [21] Sprik M, Ciccotti G 1998 *J. Chem. Phys.* **109** 7737
- [22] Zwanzig R W 1954 *J. Chem. Phys.* **22** 1420
- [23] Kirkwood J G 1935 *J. Chem. Phys.* **3** 300
- [24] Oberhofer H, Dellago C, Geissler P L 2005 *J. Phys. Chem. B* **109** 6902
- [25] Chen M, Cuendet M A, Tuckerman M E 2012 *J. Chem. Phys.* **137** 024102
- [26] Lesage A, Lelievre T, Stoltz G, Henin J 2017 *J. Phys. Chem. B* **121** 3676
- [27] Tribello G A, Gasparotto P 2019 *Front. Mol. Biosci.* **6** 46
- [28] Comer J, Gumbart J C, Henin J, Lelievre T, Pohorille A, Chipot C 2015 *J. Phys. Chem. B* **119** 1129
- [29] Darve E, Pohorille A 2001 *J. Chem. Phys.* **115** 9169
- [30] Huber T, Torda A E, van Gunsteren W F 1994 *J. Comput. Aided. Mol. Des.* **8** 695
- [31] Wang F, Landau D P 2001 *Phys. Rev. Lett.* **86** 2050
- [32] Valsson O, Tiwary P, Parrinello M 2016 *Annu. Rev. Phys. Chem.* **67** 159
- [33] Husic B E, Pande V S 2018 *J. Am. Chem. Soc.* **140** 2386
- [34] Dellago C, Bolhuis P G, Csajka F S, Chandler D 1998 *J. Chem. Phys.* **108** 1964
- [35] Bolhuis P G, Chandler D, Dellago C, Geissler P L 2002 *Annu. Rev. Phys. Chem.* **53** 291
- [36] van Erp T S, Moroni D, Bolhuis P G 2003 *J. Chem. Phys.* **118** 7762
- [37] Moroni D, Bolhuis P G, van Erp T S 2004 *J. Chem. Phys.* **120** 4055
- [38] Hummer G 2004 *J. Chem. Phys.* **120** 516
- [39] Bolhuis P G, Swenson D W H 2021 *Front. Data Comput.* **4** 2000237
- [40] E W, Vanden-Eijnden E 2010 *Annu. Rev. Phys. Chem.* **61** 391
- [41] Sarich M, Banisch R, Hartmann C, Schütte C 2013 *Entropy* **16** 258
- [42] Cybenko G 1989 *Math. Control Signal Syst.* **2** 303
- [43] Leshno M, Lin V Y, Pinkus A, Schocken S 1993 *Neural Netw.* **6** 861
- [44] Zhou D X 2020 *Appl. Comput. Harmon. Anal.* **48** 787
- [45] Alzubaidi L, Zhang J, Humaidi A J, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaria J, Fadhel M A, Al-Amidie M, Farhan L 2021 *J. Big Data* **8** 53
- [46] He K, Zhang X, Ren S, Sun J 2016 *IEEE Conference on Computer Vision and Pattern Recognition Las Vegas, USA, 27–30 June, 2016* pp770–778
- [47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I 2017 *Advances in Neural Information Processing Systems Long Beach, USA, December 4–9, 2017*
- [48] Ho J, Jain A, Abbeel P 2020 *Advances in Neural Information Processing Systems Virtual* pp6840–6851
- [49] Baydin A G, Pearlmutter B A, Radul A A, Siskind J M 2018 *J. Mach. Learn. Res.* **18** 1
- [50] Rumelhart D, Hinton G, Williams R 1986 *Nature* **323** 533
- [51] Michelucci U 2022 arXiv: 1312.6114 [stat. ML]
- [52] Kingma D P, Welling M 2019 arXiv: 1906.02691 [cs. LG]
- [53] Waterfall J J, Casey F P, Gutenkunst R N, Brown K S, Myers C R, Brouwer P W, Elser V, Sethna J P 2006 *Phys. Rev. Lett.* **97** 150601
- [54] Rumelhart D E, Hinton G E, Williams R J (Anderson J A, Rosenfeld E, ed) 1988 *Neurocomputing* (Vol. 1) (Cambridge: The MIT Press) pp696–700
- [55] Arfken G B, Weber H J, Harris F E 2011 *Mathematical Methods for Physicists: A Comprehensive Guide* (Cambridge: Academic Press)
- [56] Blei D M, Kucukelbir A, McAuliffe J D 2017 *J. Am. Stat. Assoc.* **112** 859
- [57] Ganguly A, Earp S W 2021 arXiv: 2108.13083 [cs. LG]
- [58] Marquardt D W 1963 *J. Soc. Ind. Appl. Math.* **11** 431
- [59] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L 2019 *Advances in Neural Information Processing Systems* pp8026–8037
- [60] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M 2016 *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* Savannah, GA, USA, November 2–4, 2016 pp265–283
- [61] Ma Y, Yu D, Wu T, Wang H 2019 *Front. Data Comput.* **1** 105

- [62] Hadji I, Wildes R P 2018 arXiv: 1803.08834 [cs. CV]
- [63] Ghorbani M, Prasad S, Klauda J B, Brooks B R 2022 *J. Chem. Phys.* **156** 184103
- [64] Mardt A, Hempel T, Clementi C, Noe F 2022 *Nat. Commun.* **13** 7101
- [65] Perez-Hernandez G, Paul F, Giorgino T, De Fabritiis G, Noe F 2013 *J. Chem. Phys.* **139** 015102
- [66] Wu H, Noé F 2019 *J. Nonlinear Sci.* **30** 23
- [67] Mardt A, Pasquali L, Wu H, Noe F 2018 *Nat. Commun.* **9** 5
- [68] Kleiman D E, Shukla D 2023 *J. Chem. Theory Comput.* **19** 4377
- [69] Chen H, Roux B, Chipot C 2023 *J. Chem. Theory Comput.* **19** 4414
- [70] Schütte C, Fischer A, Huisinga W, Deuffhard P 1999 *J. Comput. Phys.* **151** 146
- [71] He Z, Chipot C, Roux B 2022 *J. Phys. Chem. Lett.* **13** 9263
- [72] Bonati L, Zhang Y Y, Parrinello M 2019 *Proc. Natl. Acad. Sci. U. S. A.* **116** 17641
- [73] Bittracher A, Mollenhauer M, Koltai P, Schütte C 2023 *Multiscale Model. Simul.* **21** 449
- [74] Wang Y, Ribeiro J M L, Tiwary P 2019 *Nat. Commun.* **10** 3573
- [75] Beyerle E R, Mehdi S, Tiwary P 2022 *J. Phys. Chem. B* **126** 3950
- [76] Zhang J, Lei Y K, Yang Y I, Gao Y Q 2020 *J. Chem. Phys.* **153** 174115
- [77] Kingma D P, Welling M 2013 arXiv: 1312.6114 [stat. ML]
- [78] Tiwary P, Berne B J 2016 *Proc. Natl. Acad. Sci. U. S. A.* **113** 2839
- [79] Wu H, Paul F, Wehmeyer C, Noe F 2016 *Proc. Natl. Acad. Sci. U. S. A.* **113** E3221
- [80] Wu H, Mey A S, Rosta E, Noé F 2014 *J. Chem. Phys.* **141** 214106
- [81] Chodera J D, Swope W C, Noé F, Prinz J H, Shirts M R, Pande V S 2011 *J. Chem. Phys.* **134** 244107
- [82] Prinz J H, Chodera J D, Pande V S, Swope W C, Smith J C, Noe F 2011 *J. Chem. Phys.* **134** 244108
- [83] Rosta E, Hummer G 2015 *J. Chem. Theory Comput.* **11** 276
- [84] Mey A S, Wu H, Noé F 2014 *Phys. Rev. X* **4** 041018
- [85] Hinrichs N S, Pande V S 2007 *J. Chem. Phys.* **126** 244101
- [86] Noe F 2008 *J. Chem. Phys.* **128** 244103
- [87] Chodera J D, Noé F 2010 *J. Chem. Phys.* **133** 265
- [88] Schütt K, Kindermans P J, Saucedo Felix H E, Chmiela S, Tkatchenko A, Müller K R 2017 *Advances in Neural Information Processing Systems* Long Beach, ACM, USA, 2017 pp991–1001
- [89] Husic B E, Charron N E, Lemm D, Wang J, Perez A, Majewski M, Kramer A, Chen Y, Olsson S, de Fabritiis G, Noe F, Clementi C 2020 *J. Chem. Phys.* **153** 194101
- [90] Battaglia P W, Hamrick J B, Bapst V, et al. 2018 arXiv: 1806.01261 [stat. ML]
- [91] Kipf T N, Welling M 2016 arXiv: 1609.02907 [cs. LG]
- [92] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y 2017 arXiv: 1710.10903 [stat. ML]
- [93] Ghorbani M, Prasad S, Klauda J B, Brooks B R 2022 arXiv:2201.04609 [physics.comp-ph]
- [94] Hempel T, Del Razo M J, Lee C T, Taylor B C, Amaro R E, Noe F 2021 *Proc. Natl. Acad. Sci. U. S. A.* **118** e2105230118
- [95] Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G 2006 *J. Chem. Phys.* **125** 24106
- [96] Pan A C, Sezer D, Roux B 2008 *J. Phys. Chem. B* **112** 3432
- [97] Weinan E, Ren W, Vanden-Eijnden E 2005 *Chem. Phys. Lett.* **413** 242
- [98] Branduardi D, Gervasio F L, Parrinello M 2007 *J. Chem. Phys.* **126** 054103
- [99] Leines G D, Ensing B 2012 *Phys. Rev. Lett.* **109** 020601
- [100] Invernizzi M, Parrinello M 2020 *J. Phys. Chem. Lett.* **11** 2731
- [101] Berezhkovskii A, Szabo A 2005 *J. Chem. Phys.* **122** 14503
- [102] Langer J S 1969 *Ann. Phys.* **54** 258
- [103] Valsson O, Parrinello M 2014 *Phys. Rev. Lett.* **113** 090601
- [104] Bilonis I, Koutsourelakis P S 2012 *J. Comput. Phys.* **231** 3849
- [105] Dempster A P, Laird N M, Rubin D B 2018 *J. R. Stat. Soc. B* **39** 1
- [106] Bonati L, Piccini G, Parrinello M 2021 *Proc. Natl. Acad. Sci. U.S.A.* **118** e2113533118
- [107] Tishby N, Pereira F C, Bialek W 2000 arXiv: physics/0004057 [physics.data-an]
- [108] Still S 2014 *Entropy* **16** 968
- [109] Song Y, Kingma D P 2021 arXiv: 2101.03288 [cs. LG]
- [110] Arjovsky M, Chintala S, Bottou L 2017 *International Conference on Machine Learning* Sydney pp214–223
- [111] Huang Y P, Xia Y, Yang L, Wei J, Yang Y I, Gao Y Q 2021 *Chin. J. Chem.* **40** 160
- [112] Ribeiro J M L, Bravo P, Wang Y, Tiwary P 2018 *J. Chem. Phys.* **149** 072301
- [113] Chen M 2021 *Eur. Phys. J. B* **94** 211
- [114] Qiu Y, O'Connor M S, Xue M, Liu B, Huang X 2023 *J. Chem. Theory Comput.* **19** 4728
- [115] Monroe J I, Shen V K 2022 *J. Chem. Theory Comput.* **18** 3622
- [116] Ma A, Dinner A R 2005 *J. Phys. Chem. B* **109** 6769
- [117] Chen W, Ferguson A L 2018 *J. Comput. Chem.* **39** 2079
- [118] Chen H, Liu H, Feng H, Fu H, Cai W, Shao X, Chipot C 2022 *J. Chem. Inf. Model.* **62** 1
- [119] Wehmeyer C, Noe F 2018 *J. Chem. Phys.* **148** 241703
- [120] Williams M O, Kevrekidis I G, Rowley C W 2015 *J. Nonlinear Sci.* **25** 1307
- [121] Mezić I 2005 *Nonlinear Dyn.* **41** 309
- [122] H. Tu J, W. Rowley C, M. Luchtenburg D, L. Brunton S, Nathan Kutz J 2014 *J. Comput. Dynam.* **1** 391
- [123] Zhang J, Chen M 2018 *Phys. Rev. Lett.* **121** 010601
- [124] Rydzewski J, Valsson O 2021 *J. Phys. Chem. A* **125** 6286
- [125] Belkacemi Z, Gkeka P, Lelievre T, Stoltz G 2022 *J. Chem. Theory Comput.* **18** 59
- [126] Kikutsuji T, Mori Y, Okazaki K I, Mori T, Kim K, Matubayasi N 2022 *J. Chem. Phys.* **156** 154108
- [127] Sun L, Vandermause J, Batzner S, Xie Y, Clark D, Chen W, Kozinsky B 2022 *J. Chem. Theory Comput.* **18** 2341
- [128] Wang Y, Lamim Ribeiro J M, Tiwary P 2020 *Curr. Opin. Struct. Biol.* **61** 139
- [129] Jung H, Covino R, Arjun A, Leitold C, Dellago C, Bolhuis P G, Hummer G 2023 *Nat. Comput. Sci.* **3** 334
- [130] Zhao L, Wang L 2023 *Chin. Phys. Lett.* **40** 120201
- [131] Wu T, He S, Liu J, Sun S, Liu K, Han Q L, Tang Y 2023 *IEEE-CAA J. Automatica Sin.* **10** 1122
- [132] Janson G, Valdes-Garcia G, Heo L, Feig M 2023 *Nat. Commun.* **14** 774
- [133] Naveed H, Ullah Khan A, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N, Mian A 2023 arXiv: 2307.06435 [cs. CL]

SPECIAL TOPIC—Machine learning in biomolecular simulations

Variational analysis and AI algorithm implementation of free energy landscapes of molecular system*

Du Bo-Chuan¹⁾ Tian Pu^{1)2)†}¹⁾ (*School of Life Sciences, Jilin University, Changchun 130012, China*)²⁾ (*School of Artificial Intelligence, Jilin University, Changchun 130012, China*)

(Received 14 November 2023; revised manuscript received 18 January 2024)

Abstract

Accurate description of the free energy landscape (FES) is the basis for understanding complex molecular systems, and for further realizing molecular design, manufacture and industrialization. Major challenges include multiple metastable states, which usually are separated by high potential barriers and are not linearly separable, and may exist at multiple levels of time and spatial scales. Consequently FES is not suitable for analytical analysis and brute force simulation. To address these challenges, many enhanced sampling methods have been developed. However, utility of them usually involves many empirical choices, which hinders research advancement, and also makes error control very unimportant. Although variational calculus has been widely applied and achieved great success in physics, engineering and statistics, its application in complex molecular systems has just begun with the development of neural networks. This brief review is to summarize the background, major developments, current limitations, and prospects of applying variation in this field. It is hoped to facilitate the AI algorithm development for complex molecular systems in general, and to promote the further methodological development in this line of research in particular.

Keywords: variation, neural networks, complex molecular system, free energy landscape

PACS: 87.80.-y, 87.15.A-

DOI: [10.7498/aps.73.20231800](https://doi.org/10.7498/aps.73.20231800)

* Project supported by the Interdisciplinary Integration and Innovation Project of JLU, China (Grant No. JLUXKJC2021ZZ05).

† Corresponding author. E-mail: tianpu@jlu.edu.cn

专题: 生物分子模拟中的机器学习

融合结构知识的蛋白质预训练模型进展*

汤天一¹⁾ 熊翊名¹⁾ 张睿格¹⁾ 张建^{1)2)†}李文飞¹⁾²⁾ 王骏¹⁾²⁾ 王炜^{1)2)‡}

1) (南京大学物理学院, 南京 210093)

2) (南京大学脑科学研究院, 南京 210093)

(2024年6月7日收到; 2024年7月12日收到修改稿)

自然语言和图像处理领域引发的人工智能革命给蛋白质计算领域带来了新的思路和研究范式. 其中一个重大的进展是从海量蛋白质序列通过自监督学习得到预训练的蛋白质语言模型. 这类预训练模型编码了蛋白质的序列、进化、结构乃至功能等多种信息, 可方便地迁移至多种下游任务, 并展现了强大的泛化能力. 在此基础上, 人们正进一步发展融合更多种类数据的多模态预训练模型. 考虑到蛋白质结构是决定其功能的主要因素, 融合了结构信息的蛋白质预训练模型可更好地支持下游多种任务, 本文对这一方向的研究工作进行了介绍和总结. 此外, 还简介了融合先验知识的蛋白质预训练模型、RNA语言模型、蛋白质设计等方面的工作, 讨论了这些领域目前的现状、困难及可能的解决方案.

关键词: 蛋白质基础模型, 蛋白质多模态模型, 蛋白质结构, 机器学习**PACS:** 87.10.Vg, 87.16.A-, 87.14.E-, 87.15.A-**DOI:** 10.7498/aps.73.20240811

1 引言

随着2019年AlphaFold以及后来的AlphaFold2在蛋白质结构预测领域取得巨大成功^[1,2], 深度学习在各个科学研究领域攻城略地, 颠覆了诸多传统的研究方法, 催生出一批令人兴奋的成果. 2023年初, OpenAI公司推出了ChatGPT^[3-6], 更是在全球范围掀起了一股人工智能热潮. ChatGPT背后的技术支持来自于语言大模型, 它通过海量的数据训练极大规模的模型. 事实上, 在ChatGPT之前, 科学界和工业界已经开始重点关注语言大模型, 包括Google的Bert和T5、DeepMind的Gopher、阿里的八卦炉、清华的GLM、华为的盘古、百度的文心、浪潮的源1.0等^[7-10], 不一而足. 其中阿里

的八卦炉是第一个参数量达到了 10^{14} 规模的模型^[8], 这和人脑中的突触数量处于同一数量级. 在此类大模型的支持下, 人们可能不再需要为特定任务搭建特定的数据集和模型, 如翻译、情感分析、阅读理解等, 而是直接训练一个超大的通用模型, 其他任务只需要在此模型基础上微调即可. 这颠覆了传统的工作模式, 且为通用人工智能 (artificial general intelligence, AGI) 提供了一条可能的道路. 更重要的是, 随着规模的增大, 这类大模型会突然在某个方面展现出出乎意料的智能, 类似物理复杂系统的“涌现”效应, 催生意外的能力^[11]. 这也已经在ChatGPT中被观察到. 自ChatGPT推出以来, 大模型进化之路正飞速向多模态前进, 以融合从文本到图像、语音、视频等多种模态的更海量的数据, 典型模型如Flamingo^[12], GPT-4^[13], PaLM-E^[14], LLaMA^[15],

* 科技部科技创新项目 (批准号: 2030-2021ZD0201300) 和国家自然科学基金 (批准号: 11934008) 资助的课题.

† 通信作者. E-mail: jzhang@nju.edu.cn

‡ 通信作者. E-mail: wangwei@nju.edu.cn

Gemini^[16], X-LLM^[17], VideoChat^[18] 等. 这里, 多模态模型指一种能够处理来自不同模态 (如图像、语音、文本等) 的多种信息的机器学习模型. 多模态技术可以将这些不同形式的信息整合起来, 实现对数据更加全面和准确的分析与理解. 在生物计算领域, 多模态模型指在序列信息之外, 还将如结构、功能、动力学等其他模态信息融入模型.

深度学习在自然语言处理 (natural language processing, NLP) 大模型技术方向的突破给了其他领域的工作者极大启发. 在蛋白质计算领域, 上述技术被移植过来用于从海量蛋白质序列信息学习其内在数据分布. 通过设计合适的深度神经网络和进行相应的训练, 网络把输入数据映射到其对应的特征表示空间 (representation space), 或称潜在空间 (latent space), 或称嵌入 (embedding), 得到数据的表示 (representation) 或称编码 (encoding). 一般认为此表示编码了蛋白质的序列、结构、进化、乃至功能等信息, 可加速下游多种任务的开发. 这类从海量序列数据出发, 并借鉴 NLP 技术进行预训练得到的模型通常被称为蛋白质语言模型 (protein language model, PLM), 它是一种蛋白质基础模型 (protein foundation model) 或蛋白质预训练模型 (pre-trained model, PTM).

基于蛋白质基础模型或预训练模型的方案相对于传统建模方法有诸多显著优势. 首先, 预训练模型从海量数据进行学习, 能自动挖掘和捕捉其中的深层次特征, 从而更好地编码蛋白质的序列、进化、结构、功能等多种信息, 在预测蛋白质结构和功能方面常表现出更高的准确性. 其次, 预训练模

型通常采用自监督学习方式 (self-supervised learning), 不依赖特定的标签 (labels) 或标注 (annotation) 数据, 使模型在数据稀疏或标签不足的情况下仍然能够进行有效的学习, 降低了学习成本, 加速了开发过程. 再次, 海量数据和大型算力赋予了预训练模型强大的泛化能力, 通常无需对每个下游任务进行额外训练. 只需用预训练模型的特征表示作为输入, 通过采用少量样本微调 (fine-tuning)、零样本学习 (zero-shot learning)、或提示学习 (prompt learning) 等方式即可迅速开发出适应下游任务的模型. 这同时也有利于解决特定下游任务标签数据稀少的问题.

2019 年以来, 各种蛋白质预训练模型如雨后春笋般发展起来. 知名的工作如 BB-model^[19], SeqVec^[20], UniRep^[21], ESM 系列^[22-26], Progen^[27,28], PMLM^[29], ProtTrans^[30], xTrimoPGLM^[31], Evo^[32] 等. 在这些预训练模型的加持下, 人们测试了大量的下游任务并展示了预训练模型的强大. 这些任务包括但不限于二级三级结构预测、折叠类型分类、蛋白质相互作用、蛋白-药物相互作用、配体亲和性预测、蛋白质功能预测、细胞内定位、突变功能预测、适应性预测等. 由于此类工作数量众多, 不一而足, 详细的介绍可参考相关综述^[33-40].

近三年以来, 几乎与自然语言处理领域齐头并进, 蛋白质预训练模型也由单纯从序列进行学习, 进化到同时学习序列、结构、功能、动力学信息等多种模态数据, 涌现出了一批多模态预训练模型. 如图 1 所示. 假如把蛋白质一级序列信息类比于人类语言的话, 那么三维结构就可类比为图片, 而三

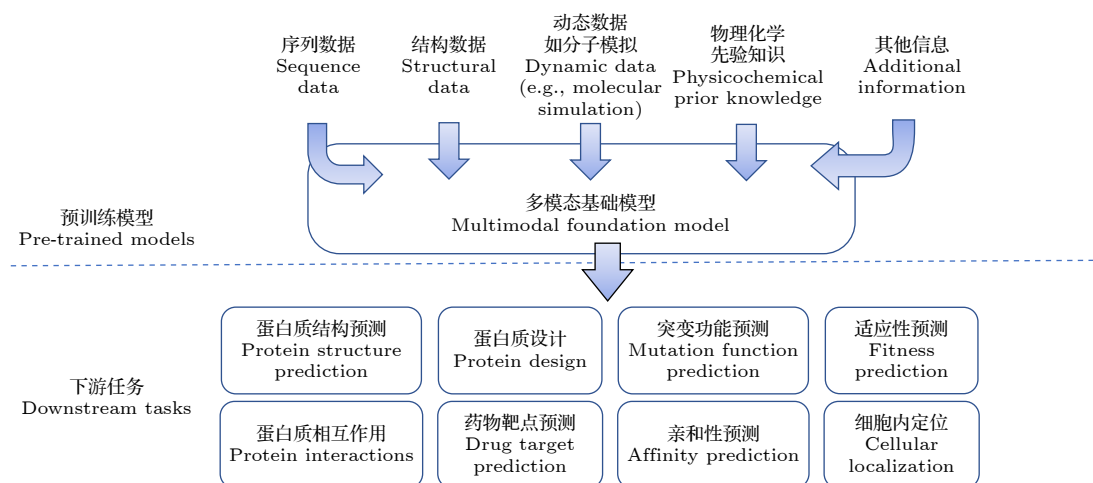


图 1 蛋白质多模态基础 (预训练) 模型及其应用 (只示意性列出若干下游任务)

Fig. 1. Protein multi-modal foundation (pre-trained) models and the downstream tasks.

维动态结构则可被类比为视频. 将更多模态的数据和知识融合在一个大模型内, 可显著地提高模型的智能, 这已在自然语言处理领域有清楚的展现^[41].

除序列信息之外, 在蛋白质预训练模型中融合结构信息尤其重要, 这是因为存在序列类似而结构全然不同的蛋白质, 同时也存在序列相似度极低但空间结构相似的例子. 另外, 由于 AlphaFold 系列的革命性突破, 包括 AlphaFold 预测的结构在内, 可用蛋白质三级结构已到数亿量级, 也为训练对应模型提供了大量数据^[2].

基于上述原因, 本文重点关注融合了结构信息的蛋白质多模态预训练模型. 此外, 如 AlphaFold2 中的 EvoFormer 模块等为特殊目的而优化的网络^[2], 虽并非为通用目的而设计的预训练模型, 也是本文关注的对象. 这是考虑到此类模型具有高度优化的编码器, 性能优异, 且容易从网络中分离出优化的蛋白质特征表示, 对一些特定下游任务, 亦可作为预训练模型使用. 特别需要声明的是, 由于作者学识所限, 并且由于这个领域发展极为迅速, 可能会遗漏部分优秀的工作, 在此致歉.

2 融合了结构信息的通用蛋白质预训练模型

Bepler 和 Berger^[19,42] 开创性地提出把蛋白质语言模型通过迁移学习用于下游各种任务的方案. 在他们的方案中, 在通过自监督学习从大量蛋白质序列中学习其语义表示的基础上, 进一步利用监督学习, 把蛋白质三维结构信息也进行编码, 获取同时编码了蛋白质序列和结构信息的表示, 用于支持下游任务. 具体来说, 他们采用多任务 (multi-task) 方式训练一个双向三层 LSTM 网络, 任务包括基于自监督学习的掩码语言建模 (masked language modeling, MLM) 和基于监督学习的残基间接触预测与结构相似性预测. 其中结构相似性根据 SCOP 数据库中的分类结果定义. 所使用的训练集包括来自 UniRef 数据库的 76M 条蛋白质序列和来自 SCOP 数据库的 28K 个蛋白质结构. 网络把输入蛋白质序列映射到一个低维语义空间, 得到一个和输入序列同长度的表示 (MT-LSTM), 并可通过如池化操作 (pooling) 得到对整个蛋白质的表示. 上述表示编码了蛋白质进化、结构和功能的信息, 可用于多种下游任务. 他们首先测试了基于此

表示的模型在区分蛋白质类别 (class)、折叠类型 (fold)、家族 (family) 方面的能力, 这可以通过简单地比较蛋白质在表示空间的矢量得到. 结果表明, 基于多任务 MT-LSTM 训练得到的表示模型优于只基于序列训练得到的表示模型 (DLM-LSTM), 也优于传统的序列比对或结构比对模型. 在预测蛋白质跨膜区域的任务上, 与其他模型相比, MT-LSTM 同样具有更优的表现. 通过结合 MT-LSTM 表示和高斯过程回归方法, 他们还在预测序列突变表现型的任务上取得了领先的结果. 关于模型的简要信息和模型对比见表 1.

Guo 等^[43] 发展了一个通过自监督方案直接从蛋白质三维结构进行学习的预训练模型. 这一方案没有使用大量序列数据. 训练所用结构数据来自 PDB 数据库, 经过处理后得到约 7 万个蛋白质三维结构. Guo 等在蛋白质 C α 原子三维结构坐标上添加高斯噪声, 把扰动后的残基距离矩阵输入网络. 网络的训练目标是估计扰动后的距离矩阵的梯度. 通过这种自监督学习方式, 网络可以获取蛋白质结构的三个层级的表示: 残基层次、残基对层次和蛋白质层次. 通过对两个下游任务进行测试, 包括蛋白质结构质量评估和蛋白-蛋白互作用位点预测, 他们发现与不使用预训练模型和使用基于纯序列的预训练模型相比, 这一新方案具有明显的优势. 另外还指出, 虽然蛋白质结构数量显著小于序列数量, 但由于结构包含更多的信息, 基于结构进行的预训练模型很有效.

自监督学习的另一个常用方案是对比学习. Hermosilla 和 Ropinski^[44] 发展了 New IEConv 模型用于从蛋白质三维结构中学习其表示. 具体地, 他们将蛋白质的三维结构转化为一张图, 从中随机采样两个子片段, 经编码器映射到表示空间后得到两个矢量, 然后计算两个矢量之间的余弦距离. 网络训练的目标是最小化来自同一个蛋白质的两个片段在表示空间的距离, 同时最大化来自不同蛋白质的片段在表示空间的距离. 网络的训练使用了来自 PDB 数据库的所有长度大于 25 残基的蛋白质结构. 经过实验, 他们发现训练中所采用的子片段的最优长度为蛋白质总长度的 40%—60%. 他们在多个下游任务测试了这一蛋白质结构表示的有效性, 包括基于 SCOP 分类的蛋白质结构相似性、折叠类型分类、蛋白质功能预测、酶催化反映类型预测、蛋白质-配体亲和性预测. 与不经预训练的模

表 1 多模态蛋白质预训练模型
Table 1. Multimodal protein pre-trained models.

模型名	时间	模型	数据模态	预训练方法	训练集	参数量	算力要求	下游任务	文献
融合了结构信息的通用蛋白质预训练模型									
Beppler & Berger	2019	Bi-LSTM	Sequence, structure	MLM for sequences, supervised learning for 3D structures	76M sequences, 28K structures	—	1X 32G-V100, 13 to 51 days	Fold classification transmembrane region prediction	[19,42]
Guo model	2022	CNN	Structure	Self-supervised pre-training on noised pair-distance	73K structures	—	—	QA, PPI	[43]
New IECConv	2022	GCN	Sequence, structure	Contrastive learning between randomly sampled 3D substructures	476K chains	30M	—	protein function prediction, protein fold classification, structural similarity prediction, protein-ligand binding affinity prediction	[44]
GearNet	2023	ESM-1b, GearNet	Sequence, structure	PLM, contrastive learning	805K structures from AlphaFoldDB	—	4X A100	Fold classification, EC, GO	
STEPS	2023	BERT, GCN	Sequence, structure	PLM, supervised learning from 3D structures	40K structures	—	—	Membrane protein classification, cellular location prediction, EC	
UNI-MOL	2023	Transformer	Sequence, structure	Atom 3D position denoise, masked atom type prediction	209M molecule conformations, 3.2M protein pockets structure	—	8X 32G-V100, 3 days	molecular property prediction, molecular conformation generation, pocket property prediction, protein-ligand binding pose prediction	
SaProt	2023	BERT	Sequence, structure	Convert structures to structure-aware vocabulary, MLM	40M sequences and structures from PDB/AlphaFoldDB	650M	64X 80G-A100, 3 months	Thermostability, HumanPPI, Metal Ion Binding, EC, GO, DeepLoc, contact prediction	[51]
融合了结构信息的非通用蛋白质预训练模型									
Evoformer	2021	Evoformer	Sequence, structure	MLM, Supervised learning	BPD+Unichst30, PDB	—	128TPU-v3, 11 days	Structure prediction	[2]
DeepFRI	2021	LSTM+GCN	Sequence, structure	PLM(pretrained, frozen), supervised learning for 3D structures	10M sequences for pre-training	—	—	GO, EC, PPI interaction sites	[47]
LM-GVP	2022	Transformer +GVP	Sequence, structure	PLM(changeable), supervised learning for 3D structures	—	—	8X 32G-V100	fluorescence, protease stability, GO, mutational effects	[48]
ProNet	2023	GCN	Sequence, structure	Supervised learning	—	—	—	Fold classification, reaction classification, binding affinity, PI	
HoloProt	2022	MPN	Sequence, structure surface	Supervised learning	—	1.8M	1X 1080Ti, 1 day	Ligand binding affinity, EC	[56]

表 1 (续) 多模态蛋白质预训练模型

Table 1 (continued). Multimodal protein pre-trained models.

模型名	时间	模型	数据模态	预训练方法	训练集	参数量	算力要求	下游任务	文献
编码动态三维结构信息的预训练模型									
ProtMD	2022	E(3)-	Sequence,	Self-supervised learning,	62.8K snapshots from MD for 64 protein- ligand pairs	5.2M	4X V100	Binding affinity prediction, binary classification of ligand efficacy	[58]
		Equivariant	structure	atom-level prompt-based denoising generative task,					
		Graph	trajectory	conformation-level snapshot ordering task					
融合了知识的蛋白质预训练模型									
OntoProtein	2022	ProtBert, Gu- model	Sequence, knowledge	MLM, contrastive learning	ProteinKG25 with 5M knowledge triples	—	V100	TAPE, PPI, Protein function prediction	[60]
KeAP	2023	ProtBert, Gu- model	Sequence, knowledge	MLM	ProteinKG25	—	—	TAPE, PPI, Protein function prediction	[62]
ProtST	2023	ProtBert, ESM-1b, ESM-2, PubMedBert	Sequence, knowledge	MLM, Multimodal Representation Alignment, Multimodal Mask Prediction	ProtDescribe with 553K sequence-property pairs	—	4X V100	Protein localization prediction, Fitness landscape prediction, Protein function annotation	[63]
RNA语言模型									
RNA-FM	2022.8	BERT	Sequence	MLM	RNAcentral, 23.7M ncRNA sequences	—	8X A100 80G, 1 month	SS prediction, 3D contact/distance map, 3D reconstruction, evolution study, RNA- protein interaction, MRL prediction	[78]
RNAbert	2022	BERT	Sequence	MLM	RNAcentral (762K) & Rfam 14.3 dataset	—	V100	structural alignment, clustering	[86]
SpliceBERT	2023	BERT	Sequence	MLM	Pre-mRNA of 72 vertebrates, 2M sequences, 64B nucleotides	19.4M	8X V100, 1 week	multi-species splice site prediction, human branch point prediction	[79]
RNA-MSM	2023	MSA- transformer	Sequence	MLM	4069 RNA families from Rfam 14.7	—	8X V100 32G	SS prediction, solvent accessibility prediction	[83]
Uni-RNA	2023	BERT	Sequence	MLM	RNAcentral & nt & GWH (1billion sequences)	25—400M	128X A100	SS prediction, 3D structure prediction, MRL, Isoform percentage prediction on 3' UTR, splice site prediction, classification of ncRNA functional families, modification site prediction	[84]
RNAErnie	2024	ERNIE	Sequence, motif information	MLM at base/subsequence/motif level masking	RNAcentral, 23M ncRNA sequences	105M	4X V100 32G, 250 hours	sequence classification, RNA-RNA interaction, SS prediction	[85]

*PLM, protein language model; MLM, masked language model; GCN, graph convolutional network; GVP, geometric vector perceptrons; EC, enzyme commission number prediction; GO, gene ontology term prediction; PPI, protein-protein interaction; TAPE, the tasks assessing protein embeddings database; QA, quality assessment of structures; SS, secondary structure; MRL, mean ribosome load prediction in mRNA.

型、基于监督学习的模型和基于纯序列预训练的模型如 EMS-1b 等相比, 均具有更好的性能。

GearNet 借鉴 SimCLR 的多视角对比学习方案以编码蛋白质结构信息^[45,46]。模型把蛋白质结构转化为一张图, 从中随机抽取两个子图, 使用不同的加噪方案以获取不同视角 (view), 然后通过网络计算它们相应的表示。网络优化的目标是根据两个子图是来自于同一蛋白或不同蛋白, 分别增加或减少两个视角在表示空间的相似度。预训练使用来自 PDB 数据库和 AlphaFold2 预测的约 805K 个蛋白质结构。文献中在 4 个下游任务测试了 GearNet 表示的有效性, 包括酶 EC 编号预测、基因 Ontology(GO) 条目预测、折叠类型分类、酶催化反应类型预测。通过和基于序列的预训练得到的蛋白质表示 (ProtTrans, ESM-1b)、基于序列和结构结合的表示 (DeepFRI^[47], LM-GVP^[48]), 以及基于结构的表示 (NewIEConv)^[44] 进行对比实验, 发现 GearNet 在 8 个测试集的 7 个中给出了最好的结果。另外, 考虑到 GearNet 用较少数量的蛋白质结构 (805K) 进行训练, 性能却优于基于大量序列预训练的编码器 (ESM-1b: 250M 序列, ProtTrans: 2.1B 序列), 证明结构比序列中蕴含了更多的信息, 能导致更好的表示。另外, GearNet 还测试了使用 PDB 数据库和使用 AlphaFold2 预测结构数据库进行预训练的差异, 结果显示不同的数据库选择对模型性能影响很小, 模型有很好的健壮性。

Chen 等^[49] 发展了 STEPS 方法, 以融合从序列和结构得到的两个特征表示。对于序列, 使用蛋白质语言模型得到其表示 h^s 。对于结构, 将其转化为一张图 G , 计算其隐含层表示 h_G 。为优化 h^s 和 h_G 之间的关系, Chen 等设计了两个自监督学习代理任务。第一个为残基间距离预测, 第二个为残基二面角掩码预测 (对特定蛋白, 遮蔽其中 15% 的二面角信息)。注意在自监督学习过程中, 语言模型的输出 h^s 被冻结保持不变。预训练使用了包含 AlphaFold 预测结构在内的约 4 万蛋白质三维结构。他们在三个下游任务对模型进行了微调 and 测试, 包括判定是否膜蛋白、蛋白质细胞内定位、酶催化反应分类。与蛋白语言模型 (基于 BERT)、DeepFRI 等相比, STEPS 均具有较大优势。另外, 消融实验证明, 蛋白质结构中的残基对距离信息、对获取更好的表示具有决定性的贡献。

UNI-MOL 是一个为蛋白质和小分子结合而

设计的预训练模型^[50]。Zhou 等^[50] 收集了 209M 小分子构象以及 3.2M 个蛋白质结合口袋的三维模型, 在原子层面上, 设计了两个代理任务来对模型进行预训练。第一个任务为给原子位置加入噪声, 然后训练网络预测其正确的位置。第二个任务为遮蔽原子类型, 训练网络对其类型进行预测。并在多个下游任务对预训练模型的性能进行了测试, 包括分子属性预测、分子构象生成任务、蛋白质结合口袋性质、配体结合构象预测。发现 UNI-MOL 在大部分任务中优于其他模型。尤其是当下游任务只有很少的标签数据情况下, 如蛋白质结合口袋性质预测, 相比其他模型更是有显著的提高。他们将其归因为预训练模型编码了蛋白质的三维结构信息。

Su 等^[51,52] 提出了一个统一处理序列与结构信息的方案 SaProt, 其创新之处在于把蛋白质三级结构通过 Foldseek 工具编码成与原序列等长的含有结构信息的 token 序列。Foldseek 的输入是指定氨基酸临近区域的三维 (3D) 构象, 它通过一个离散化变分自编码器 (VQ-VAE) 网络, 把构象转化为 20 个离散矢量中的一个, 称为 3D token。相比于通过图来表示蛋白质结构, 这一方案的优势在于把蛋白质序列和三维结构都转化为一个语句, 可无缝地使用 NLP 领域的各种大模型架构。SaProt 模型采用了 ESM^[22,24] 的训练框架, 即掩码语言模型, 在一个包含 40M 蛋白质序列和结构的数据集上对网络进行预训练, 得到一个大范围的具有通用性的蛋白质表示 SaProt。在 10 个下游任务的测试表明, 此表示方案具有优异的性能和广泛的适用性。

3 融合了结构信息的非通用蛋白质模型

如引言所述, EvoFormer 等为特殊目的而优化的网络模型, 虽非通用预训练模型, 亦在本文讨论之列。另外, 由于此类模型众多, 只选其中的一部分予以介绍。

Evoformer 是著名的 AlphaFold2 的编码器部分, 即去掉后部生成模块之后余下的部分^[2]。它接受 MSA 和残基对信息作为输入、输出对应的表示。由于 Evoformer 同时使用大量蛋白质序列和结构进行监督训练, 它输出的表示融合了序列和结构的信息。Hu 等^[53] 在结构预测、功能预测、适应度 (fitness) 预测三类共 7 个任务上, 详细测试了 Evoformer 的表征能力, 并与 ESM-1b 和 MSA-Transformer 进

行了对比. 他们发现: 1) 经 AlphaFold2 训练的 Evoformer 参数是通用的, 可被用于各种结构和功能预测任务. 2) AlphaFold 在结构预测任务 (包括二级结构预测和接触图预测) 和小蛋白稳定性预测中, 比 ESM-1b 和 MSA-Transformer 具有更优的性能. 但在蛋白质功能预测任务上不如后两者, 在零样本适应度预测上表现不好. 3) Evoformer 对输入 MSA 信息的依赖很强, 另外, 如使用从 ESM-1b 转化来的 MSA 信息替代原 Evoformer 的输入, 几乎没有性能损失.

另外, 我们注意到在刚刚发布的 AlphaFold-3 中^[54], Evoformer 被一个更简单的 Pairformer 代替, 它简化了对 MSA 信息处理的过程. Pairformer 只对单个氨基酸表示 (single representation) 和氨基酸对表示 (pair representation) 进行处理, MSA 表示不再传递给下游模块. 虽然 AlphaFold3 具有更强大的性能, 尤其是在复合体结构预测上, 但 Pairformer 模块本身对蛋白质信息的表征能力尚未被系统地测试.

DeepFRI 是一个两阶段蛋白质功能预测模型^[47]. 第一阶段在一个大小为 10M 的蛋白质序列数据集上训练一个基于 LSTM 的语言模型, 从中抽取残基分辨率的序列特征. 具体训练方法借鉴了 Bepler 和 Berger^[19] 采用的掩码语言建模 (MLM) 方法. 在第二阶段, 上述序列特征和残基接触图以及用于表示三维结构的图网络向量一起, 被输入到下游的图卷积层, 得到一个融合了序列和结构信息的表示层, 再经两个全连接层后, 输出蛋白质的功能信息. 网络第二阶段利用具体任务对应的标签数据进行监督学习, 并冻结第一阶段获得的语言模型参数. DeepFRI 这一融合了序列和结构信息的模型具有良好的抗噪声特性, 即使在模型中以预测的蛋白质结构代替实验结构, 预测准确度也只有可忽略的下降.

LM-GVP 模型结合了蛋白质语言模型 (PLM) 和一个对三维空间平移和旋转具有不变性的网络模块 (geometric vector perceptrons, GVP), 在序列数据、结构数据以及若干下游数据集上进行训练, 用于预测蛋白质特性^[48]. 其中 PLM 模块基于 Transformer 架构并且是预训练的, 它的输出向量与由蛋白质结构转化来的图向量结合, 被输入给下游 GVP 模块. 与 DeepFRI 模型不同, LM-GVP 在使用下游数据集进行监督学习时, 允许梯度回传

至 PLM 模块, 因此模型给出的特征表示融合了蛋白质的结构信息. 然而, 由于这些结构信息不是经由下游任务无关的方式融入, 因此 LM-GVP 给出的不是一个通用表示, 可能只在特定任务具有良好性能.

ProNet 在三个不同的层级学习蛋白质的三维结构, 包括氨基酸级 (C^α 原子)、主链级 (主链原子) 和全原子级^[55]. 这种分级方案的优势是: 1) 用不同层级的表示适配不同的下游任务, 如蛋白质功能预测只需要氨基酸层次的表示即可, 而亲和能预测可能需要原子级的表示. 2) 训练和推理的速度大大增加. Wang 等^[55] 在蛋白质折叠类型分类、酶反应类型分类、配体亲和能预测共三个任务上测试了这一模型, 结果显示比其他同类模型具有持平或略优的性能, 并且运算速度最高有 6 倍的提升.

HoloProt 从多个尺度对蛋白质结构进行表征, 包括序列、二级结构、三级和四级结构, 以及蛋白质表面形貌^[56]. 其中前四个层次被统称为结构信息, 并被转化为一张图. 图的节点为残基, 空间距离小于某个阈值的两个残基之间用一条边相连. 对于蛋白质表面形貌, 也在三角剖分后被转化为一张图. 与 MaSIF 模型类似^[57], 每个图节点的特征包括氨基酸标识、电荷、疏水性和局域曲率等信息, 如果两个节点同属于一个三角形, 则它们之间有一条边相连. 此外, 为了在两张图之间传递信息, HoloProt 模型还在分属于两张图的节点之间引入了边 (如果这两个节点属于同一个残基). HoloProt 模型在两个下游任务, 包括配体结合亲和性预测和酶催化反应分类任务上进行了训练和测试, 发现与之前的多个基于序列的模型和基于结构的模型相比, 多尺度的 HoloProt 模型具有更优的性能. 此外, 消融实验指出, 对于配体结合亲和性预测, 只基于蛋白质表面形貌的模型已经工作得很好. 对于酶催化反应分类任务, 只考虑蛋白质表面形貌会导致预测性能大幅下降, 因此对于这一任务, 结构信息非常重要.

4 编码动态三维结构信息的预训练模型

蛋白质结构的动态性对其生物功能至关重要, 尤其是可变构蛋白和天然无序蛋白. 在蛋白质相互作用和蛋白质-药物相互作用中, 结合口袋的构象

动力学对亲和性有重要影响. 然而大部分蛋白质表示模型仅从静态结构进行学习, 未考虑蛋白质的动态性. 可以预期, 如能在预训练模型中融入蛋白质的动态信息, 将有力地促进诸如蛋白-蛋白相互作用、蛋白-药物相互作用等下游任务的进行.

基于类似考虑, Wu 等^[58]发展了 ProtMD 方法, 从蛋白质-配体相互作用的动态结构中学习其特征表示. 他们首先对 64 个蛋白质-配体复合体进行分子动力学模拟, 得到共约 62.8 K 构象. 在自监督学习过程中, 第一个代理任务采用基于提示的去噪生成, 从 t 时刻加了噪声的蛋白质结构预测 $t+i$ 时刻的结构, 并与模拟的结果进行对比来计算损失函数. 这一任务被用于学习原子级别的、局域的时空相关信息. 第二个代理任务为构象重排序任务, 即把模拟中的若干构象顺序打乱, 迫使网络学习其正确顺序. 此任务被用于学习构象级别的时间域的上下文关系.

Wu 等^[58]对两个下游任务采用线性探测 (linear-probing) 和微调 (fine-tuning) 两种模式测试了模型性能, 并与之前的基于监督学习的多种基线模型进行了对比. 这些基线模型包含四类: 基于序列的、基于表面形状的、基于结构的以及多尺度方法. 与基线模型相比, 在基于 PDBbind 数据集的配体亲和性预测任务中, 线性探测模式的 ProtMD 具有良好的性能, 而经过任务微调的 ProtMD 版本性能更优, 具有最小的误差 (RMSE) 和最高的相关系数. 在配体效力预测中 (预测一个配体分子的结合是否能激活蛋白质的功能), 经过微调的 ProtMD 模型预测准确度高于所有基线模型.

Wu 等^[58]还研究了预训练数据集的大小对模型性能的影响. 发现线性探测模式显著地依赖于样本量大小, 当蛋白质-复合体数目超过 50 对时, 模型性能达到最高. 与之相比, 微调模式对预训练样本量依赖程度较低, 较小样本量即可得到好的效果. 他们最后选用了 64 对蛋白质-配体复合物进行预训练, 所得模型对于一个大小为 3K 的测试集依然表现良好, 说明复合物三维结构中蕴含了足够多的相关信息, 从中学习的模型具有优异的泛化性能.

到目前为止, 从分子动力学轨迹学习蛋白质动态性质的预训练模型仍然很少, 大规模的为通用目的而设计的预训练模型还未见报道. 一个于 2010 年开始建立的大型分子模拟轨迹数据库对此类任务可能有帮助^[59].

5 融合了知识的蛋白质预训练模型

蛋白质多模态模型另一个重要发展方向是在语言模型的基础上融合基于描述的知识. OntoProtein 是第一个把蛋白质功能知识 (gene ontology) 融合到蛋白质表示中的多模态预训练模型^[60]. Zhang 等^[60]整理了一个大型的蛋白质知识数据库 (ProteinKG25), 包含约 5M 数据条目, 其形式为三元组 (蛋白质-关系-属性). OntoProtein 的蛋白质编码器采用预训练的 ProtBert^[30], 知识编码器使用微软开发的一个针对生物学语言开发的预训练模型 PubMedBERT^[61]. 序列输入和知识三元组分别被两个编码器编码, 并映射到同一个表示空间. 对于序列数据, 预训练采用代理任务为遮蔽率为 15% 的 MLM 方案, 损失函数为真实值与预测值之间的交叉熵. 而对于三元组形式的功能数据, 则利用对比学习技术设计和计算损失函数. 模型同时优化上述两个损失函数, 以获取融合了序列和知识的蛋白质表示. 他们在 TAPE 数据集的三类任务、蛋白-蛋白相互作用和蛋白质功能预测等多方面测试了模型性能. 相比之前在大型语料数据集上训练的蛋白质语言模型, OntoProtein 性能稍有提高. Zhang 等^[60]将其归结为目前的功能知识条目偏少, 只能覆盖少部分蛋白质空间.

与 OntoProtein 模型类似, KeAP 致力于在一个更精细的令牌层次对蛋白质和知识进行融合^[62]. 具体来说, 对于一个输入的知识三元组 (蛋白质-关系-属性), 蛋白质序列被遮蔽一部分 (约 20%) 后被一个 BERT 型编码器编码, 关系和属性则通过另一个自然语言编码器 PubMedBERT 得到其表示, 然后利用跨模态注意力机制先后从关系数据和属性数据查询与预测被遮蔽氨基酸相关的信息, 并对其预测. 与 OntoProtein 相比, KeAP 简化了代理任务, 只使用了 MLM 技术对网络进行预训练. 通过使用和 OntoProtein 类似的微调技术, KeAP 在残基接触预测, 同源探测、稳定性预测、蛋白相互作用、亲和能预测、语义相似性推理等多个下游任务对预训练模型进行了测试, 发现相比于 EMS-1b, ProtBert, OntoProtein 等模型, 预测准确度有显著的提高.

ProtST 也是一个融合蛋白质序列信息与功能信息的多模态预训练模型^[63]. 它使用预训练的蛋

白质语言模型 (包括 ProtBert, ESM-1b, ESM-2) 来初始化序列编码器, 用 PubMedBERT 对功能知识进行编码并在后续训练中保持网络权重不变. 模型使用三个代理任务进行预训练, 目的是把序列的表示和知识的表示在语义空间进行对齐. 第一个任务为单模态 MLM 任务, 随机遮蔽 15% 的残基并利用上下文预测这一遮蔽信息. 第二个为多模态对齐任务, 在蛋白质序列和文本描述之间进行对比学习, 以拉近成对的信息在表示空间的距离. 第三个为多模态掩码预测任务. 这一任务随机的遮蔽 15% 的蛋白质序列以及 15% 的文本, 经过一个具有自注意力和交叉注意力的融合网络后, 输出对遮蔽信息的预测. 预训练所用知识数据库 ProtDescribe 包含约 553K 蛋白质序列-属性对. Xu 等^[63] 在三类下游任务, 包括蛋白质定位预测、蛋白质突变适应度预测、蛋白质功能预测上对 ProtST 模型进行了微调和测试, 发现它显著优于 CNN 等基线模型, 也优于 ProtBert, OntoProtein, ESM-1b, ESM-2 等模型. 此外, 在亚细胞定位预测、反应类型预测、文本到蛋白搜索几个零样本实验中, 模型也表现出了较好的泛化能力.

6 RNA 预训练模型

RNA 结构和功能预测问题和蛋白质相关问题具有很高的相似性, 相当一部分针对蛋白质发展的计算方法稍加修改即可用于 RNA 领域. 人们很早就开始使用机器学习方法进行 RNA 结构预测, 如 SPOT-RNA 系列^[64,65], 3DRNA 系列^[66-68], FebRNA^[69-71], RNA3DCNN^[72], UFold^[73], DeepFoldRNA^[74], RoseTTAFoldNA^[75] 等. 这方面工作完整的介绍可参考最新的综述文献^[76, 77]. 本文只针对 RNA 预训练模型进行介绍.

和蛋白质相比, 针对 RNA 的预训练模型还相对较少. 这可能是因为相对于 20 字符编码的蛋白质序列, 核酸序列是一种四字符编码语言, 相同长度的序列信息量远小于前者, 且 RNA 序列保守性也相对较低. 另外, 相对于 DNA, RNA 具有较多的修饰及高级结构, 更为复杂.

RNA-FM 是一个为通用目的而设计的大型 RNA 预训练语言模型^[78]. 模型采用 BERT 架构, 在一个超过 2 千万非编码 RNA 序列数据集上通过自监督学习进行预训练, 训练过程中 15% 的核

苷酸被随机遮盖并被模型预测. Chen 等^[78] 在多个任务上对这一预训练模型得到的 RNA 表示进行了测试. 在二级结构预测任务上, RNA-FM 相较于之前的如 LinearFold 和 SPOT-RNA 有大幅度的提高. 在三维 contact map 预测任务上, 基于 RNA-FM 的 ResNet 模型大幅度领先于一个基于 100 个子模型的集成学习方案. 把 RNA-FM 预测的二级结构和 3dRNA 相结合, 可用于预测 RNA 三级结构. 这一方案在 RNApuzzle 测试集上的平均 RMSD 为 4Å. 基于 RNA-FM 预训练模型, 他们还预测了 SARS-CoV-2 基因主要调控区域的二级结构, 并研究了这一病毒的演化路径, 所得结果均与 ground truth 高度符合. RNA-FM 还被用于协助预测 RNA-蛋白质相互作用. Chen 等^[78] 把 RNA-FM 预测的二级结构代替 icSHAPE 实验结构, 使用 Prism Net 预测了海拉细胞中的 RNA-蛋白质相互作用, 发现预测结果全部优于基线模型. 和使用实验结果作为输入的 PrismNet 相比, 使用 RNA-FM 预测值作为输入的模型在 7 种蛋白情况下更优 (共 17 种). 最后, 虽然 RNA-FM 使用非编码 RNA 序列进行训练得到, 它在 mRNA 的 5' 非翻译区核糖体载量预测任务上也展现了良好的性能.

SpliceBERT 是一个在 pre-mRNA 序列数据集上训练的 RNA 语言模型, 主要用于预测 RNA 剪切位点^[79]. 训练数据集包含来自 72 种脊椎动物的约 200 万 pre-mRNA, 碱基数目达到了 650 亿. 训练采用 BERT 架构, 单个核苷酸对应一个令牌, 随机遮盖 15% 的核苷酸并对其进行预测, 以强迫网络学习不同位点间的相互关系. 在多物种剪切位点预测和人类分支点预测两个下游任务进行的微调和测试表明, 基于 SpliceBERT 的模型优于传统的基线模型、DNABERT 和只在人类数据上预训练的 SpliceBERT-human. 这显示了在多物种数据集上进行预训练的有效性. 与 SpliceBert 类似, 针对 mRNA 发展的语言模型还有 CodonBERT, UTR-LM, 3UTR-BERT 等^[80-82], 篇幅关系不能一一详述.

考虑到 RNA 的序列保守性低于蛋白质, Zhang 等^[83] 发展了 RNACmap 方法, 它可提供比 Rfam 数据集更多的同源序列. 在此基础上, 他们采用 MSA Transformer 结构和 BERT 目标函数训练得到了一个 RNA 语言模型 RNA-MSM, 在其输出的二维注意力图和一维嵌入中编码了序列和结构信息. 针对下游任务微调后, 模型在二维碱基对概率预

测和一维溶液可及表面预测任务上, 优于目前的 SOTA 方法如 SPOT-RNA2 和 RNA snap2, 也优于基于之前的语言模型 RNA-FM.

Uni-RNA 是一个利用约 10 亿条 RNA 序列进行大规模训练的 RNA 语言模型, 充分挖掘了 RNA 序列的潜在信息^[84]. 预训练采用经过效率优化的 BERT 模型. 与 RNA-FM, SPOT-RNA 等方法相比, 基于 Uni-RNA 微调的模型在 RNA 二级结构预测、contact map 预测、mRNA 5'UTR 核糖体载量预测, 3'UTR 亚型占比预测、ncRNA 功能聚类, 剪切位点预测, RNA 修饰位点预测七个任务中均取得了优秀的结果.

RNAErnie 也是一个 RNA 语言模型^[85]. 它使用了来自 RNACentral 数据库的约 2 千万序列进行训练. 训练使用支持连续学习的 Ernie Transformer 架构. 与之前语言模型不同, RNAErnie 进一步把 RNA 片段 (motif) 信息作为先验引入模型. 具体来说, 在自监督预训练阶段, 除在碱基水平的随机遮盖、4—8 碱基长度的子序列随机遮盖之外, 模型还加入了一个片段水平的随机掩码任务, 并将 RNA 类型, 如 miRNA, mRNA, lncRNA 等, 以一个停止词的方式加入到序列尾部, 鼓励模型把不同类型的序列映射到 latent 空间的不同位置, 以更好地支持下游类型引导的微调任务. 在多个下游任务, 包括序列分类、RNA-RNA 互作用预测、和 RNA 二级结构预测, RNAErnie 的性能均大幅优于传统的方法以及之前的语言模型如 RNABert^[86], RNA-FM^[78] 等.

到目前为止, 据我们所知, RNA 预训练模型均基于序列数据, 尚未见到整合结构信息的模型. 只有 RNAErnie 通过遮盖 RNA 片段序列, 部分地引入了结构信息. 这可能是由于实验解出的 RNA 结构数量远少于蛋白质, 且虽有很多优秀的结构预测模型^[65,68,75,87,88], 但尚未见到如 AlphaFold 的革命性突破, 这显示了 RNA 结构预测的难度, 同时说明这是一个大有可为的领域.

7 蛋白质预训练模型与蛋白质设计

蛋白质设计是蛋白质计算领域的一个重要方向. 这方面已经有大量优秀的工作^[89-96]和综述性报告^[97-103].

和结构相关的蛋白质设计中, ProteinMPNN

是一个典型的从结构到序列的 Inverse-folding 模型. 它包括编码器和解码器两部分, 其中编码器学习一个和序列无关的蛋白质结构表示, 解码器则通过自回归的方式预测相应的序列^[104,105]. ESM-IF1 模型也采用了类似的架构^[106].

Baker 组^[92,93]发展了 hallucination 方法. 它首先在序列空间进行蒙特卡罗采样, 并使用 trRosetta 预测结构. 他们还使用类似的框架发展了 Protein Generator, 但把蒙特卡罗采样替换为序列空间的去噪扩散概率模型 (DDPM)^[107,108]. 这类模型的特色是把序列空间的优化采样算法和成熟的结构预测模块相结合.

Baker 组还发展了 RFdiffusion 模型进行蛋白质从头设计 (de novo design). 这一模型使用去噪扩散概率模型直接在三维空间从初始噪声生成蛋白质结构, 并利用 ProteinMPNN 设计相匹配的蛋白质序列^[109]. RFdiffusion 支持无条件 and 条件生成, 可进行蛋白质单体、高阶对称寡聚体、功能片段框架、结合蛋白设计等多种任务. 由于直接在结构空间进行去噪扩散生成, 模型生成的结构具有更好的多样性.

与 RFdiffusion 不同, ProteinSGM 在残基间 6 维坐标空间进行去噪扩散以生成结构. 它采用了一个基于随机微分方程的评分生成模型框架, 实现了一个连续的噪声注入和移除策略^[110], 并使用 Rosetta 对主链结构进行能量最小化^[111]. 这一方案还通过条件生成支持准确和模块化的设计, 可获得和天然蛋白相近的新型蛋白质结构.

上述去噪扩散模型倾向于生成刚性的蛋白质结构, 含有较多的螺旋和较短的 loop 区, 而较少生成对蛋白质功能更重要的柔性和动态结构. PVDQ (protein vector quantization and diffusion) 针对这一问题进行了改进^[112]. 这一模型把蛋白质主链结构映射到潜在空间, 并使用一个离散自编码器学习对应的离散表示. 这些离散的表示构成一个代码本 (code book). 通过这种方式, 一个蛋白质主链被映射为一个离散表示序列. PVDQ 在这一潜在空间通过去噪扩散模型进行结构生成, 这一设计允许更高效的采样效率和更平滑的数据分布. 去噪扩散生成的离散表示序列被一个解码器翻译为三维结构, 另一个辅助解码器被用来生成对应的氨基酸序列. 与之前直接在结构空间进行去噪生成的模型不同, PVDQ 模型展现出了更强的生成 β 片和长 loop

区的能力, 这些结构具有较小的刚性和更好的动态性. PVQD 模型也支持条件概率生成.

蛋白质设计模型通常仅利用具有实验或预测结构的序列进行训练, 无法利用海量的结构未知的序列. 本文介绍的融合蛋白质结构的预训练模型可用于解决这一问题. 正如 LM-DESIGN 工作所指出的, 融合结构信息的语言模型是一个蛋白质设计器^[113]. 这一模型把结构编码器 (如 GNN) 的输出和语言模型 (如 ESM 系列) 相结合, 利用语言模型的生成能力进行序列解码, 并通过反复迭代的方法对序列进行优化. 又如 MIF-ST 模型把一个预训练的蛋白质语言模型 (CARP-640M) 和一个表征蛋白质结构的图网络结合起来, 并使用 MLM 方案进行预训练^[114]. 这些模型在核心架构上和前文介绍的 STEPS^[49] 和 LM-GVP^[48] 等模型非常类似, 显示了蛋白质预训练模型和蛋白质设计等不同任务在架构设计上逐渐合流的趋势^[102].

8 讨论

深度学习技术的成功, 在多个科学领域催生了新的思路和研究范式. 其中最具有代表性的是 AlphaFold 系列. 2023 年来, 以 ChatGPT 为代表的自然语言大模型取得了空前的成功, 并且快速朝着多模态大模型发展, 以融合更多的数据, 训练更大的模型. 在这个领域, 模型的大小和算力是推动性能提升的主要力量^[115]. 自然语言处理领域的若干关键技术, 如模型预训练、自监督学习范式、被迅速借鉴到生物学领域. 自 2019 年开始, 尤其是近三年来, 人们发展了多种蛋白质预训练模型并应用于各种下游任务. 这一新的研究范式, 不仅可以充分利用海量无标注数据以提供强大且通用的表征能力, 为多种下游任务提供统一的框架并便于快速部署, 且特别有利于某些缺乏标注数据的下游任务, 可在相当程度上解决某些领域标注数据严重不足的问题.

到目前为止, 针对蛋白质序列的预训练模型已基本成熟. 考虑到蛋白质结构承载了更大的信息量, 且蛋白质功能主要和结构相关, 越来越多的工作开始关注如何把结构信息更好地融入蛋白质的表示空间. 从学科趋势上看, 蛋白质预训练模型明显地朝着更多模态发展, 以融合空间结构、物理化学知识、功能数据、甚至动态结构等信息, 以期多

种数据的交叉融合能够催生出更强大的模型. 本文对这一方向的进展进行了回顾和总结.

本文所介绍的模型各有其优缺点和特色. Evoformer 和 LM-GVP 代表了同时融合序列和结构信息的、为特定目的而设计的蛋白质模型, 其中 Evoformer 针对蛋白质结构预测、而 LM-GVP 针对功能预测. 虽然它们都是为特定任务而设计, 但它们给出的特征表示均具有一定的通用性. 尤其是 Evoformer, 已被实验证实可泛化到比如功能预测任务, 虽然性能上相比 ESM 系列略差. 在为通用目的而设计的蛋白质预训练模型中, BB-model^[19,42] 具有开创性且富有特色, 这一模型利用多任务学习框架同时在序列和结构上对模型进行训练, 且在下游任务只使用序列进行推理 (经过微调). 相比较而言, Guo-model^[43], New IECConv^[44], GearNet^[46] 等模型, 在下游任务进行推理时必须提供蛋白质结构, 虽然这提高了模型的准确度, 但也同时限制了其应用范围. 从训练方法看, 大部分预训练模型借鉴了自然语言处理中的 MLM 方法或图像处理中的 Masked AutoEncoder(MAE) 方法^[116], 也有部分采用对比学习方案^[117], 不同训练方案在分子结构领域的有效性目前尚无定论. 另外, SaProt 提出了一个创新性的训练方案, 它把蛋白质三级结构信息通过 Foldseek 工具编码成与氨基酸序列等长的 token 序列, 和氨基酸对应的 token 结合, 将输入序列和结构转化成一句. 这样做的好处是可以无缝地使用 NSP 领域成熟的语言模型, 且可用于处理大规模数据^[51,52]. 从数据模态角度看, 主流模型重点关注如何融合序列和结构信息, 如 BB-model^[19,42], Guo-model^[43], GearNet, STEPS, SaProt 等, 而 HoloProt 则引入了蛋白质表面形貌信息, ProtMD 模型从分子模拟数据中进行学习以建模蛋白质的动态特征, OntoProtein 和 ProtST 等模型则侧重于融合序列和功能信息. 另外, 生物计算领域还有相当数量的工作致力于集成 DNA、RNA、功能等多来源、多模态数据, 如 xTrimmo^[31]、Evo^[32] 等. 一个全面的总结见文献^[35].

多模态模型的预训练需要大量配对数据, 如匹配蛋白质序列和功能描述. 然而配对数据通常很稀缺, 导致多模态模型训练困难. Biobridge 方案尝试解决这一问题^[118]. 它不试图训练一个多模态模型, 而是使用知识图谱训练一个对齐模型, 把多个单模态的表示空间进行对齐, 把它们连接起来以解决多

模态任务. 这一方案同时解决了多模态模型计算量过大的问题, 是一个有益的探索. 最后, 通过采用主动学习 (active learning) 的方式, 有目的地选择配对数据样本, 亦可降低对数据量的要求.

和海量蛋白质序列相比, 三维结构数据的数量偏少可能并不是一个严重的问题. 这是考虑到与序列信息相比, 空间结构信息可能更类似于自然语言, 具有较高的信息密度. 首先, 和自然语言中的句子不同, 单一蛋白质序列并没有明显的语义特征, 难以找出相当于词的单位以及它们之间的相互关系. 反观空间结构, 由于共价键的刚性, 原子团具有明显的化学意义且种类并不太多, 可被视为基本结构单元, 并对应于自然语言中的词. 原子团之间的相互作用相当于句子中词的相互作用. 此外, 由于物理化学上的限制, 原子团之间的堆积模式可能并不太多, 无需海量实验结构即可覆盖大部分可能的相空间. 当然, 由于 AlphaFold 系列的成功, 目前可用的蛋白质三维结构被大大扩充了. 通过对目前融合了结构的多模态模型进行分析 (见表 1), 我们预测, 与蛋白质语言模型对序列数量的需求相比, 融合结构信息的多模态模型可能并不需要海量的三维结构.

将先验知识引入预训练模型也是一个重要研究方向. 这不仅可以利用现有的知识, 且可以丰富训练数据、增强模型泛化能力、提高模型的可解释性等. 在蛋白质计算领域, 目前常见工作是把功能相关的描述以自然语言编码器编码, 或以知识图谱形式通过对比学习融入模型. 然而, 先验知识的形式多种多样, 如逻辑规则、知识图谱、数学物理方程、人类反馈等^[119]. 对于生物大分子结构来说, 如何把诸如长程静电相互作用等物理化学知识直接引入预训练模型, 是一个值得探索的方向. 这对于 RNA 结构尤其重要, 因为长程的静电相互作用是其结构稳定性的决定性因素之一^[120-122]. 而目前常见的预训练模型中, 无论是遮蔽重建还是对比学习方案, 均局限于短程相互作用.

训练大模型通常需要庞大的算力. 如 ESM 系列, xTrimo 系列, 均需要大量的 GPU 进行训练. 然而, 纵观本文提到的多模态模型, 大部分并不需要十分强大的算力. 这一方面是因为多模态如结构数据、蛋白质功能数据并不十分庞大, 另一方面是因为这些模型利用了已预训练的单模态模型. 如 ProtST 模型分别利用 ESM 系列和 PubMedBERT

编码蛋白质序列信息和功能描述, 并冻结 PubMedBERT 的模型权重, 通过对比学习把蛋白质序列的表示和功能的表示进行对齐, 极大地降低了训练所需算力. 另外, 对于训练多模态模型, Biobridge 方案也可降低对算力的需求.

虽然蛋白质预训练模型领域已经取得了很多进展, 但仍面临诸多挑战. 最显著的问题是缺乏统一 benchmark, 难以判断各模型优劣. 另外, 由于使用大量数据训练模型, 测试数据的信息泄露到训练集中也是常见问题. 蛋白质结构的动态性也是目前大部分模型未考虑的问题. 然而蛋白质这一特性对其生物功能至关重要, 尤其是对于可变构蛋白和天然无序蛋白, 以及蛋白质-药物的非刚性结合, 蛋白质-RNA 相互作用等问题. 虽然 ProtMD 方法从 64 个蛋白质-配体复合体的分子动力学模拟轨迹中学习了结合界面的动态特性, 但由于训练数据集偏小 (62.8 K 构象), 模型的通用性和泛化能力尚未可知.

总之, 近三年来, 融合了蛋白质结构信息的预训练模型, 以及融合了更多模态信息的预训练模型如雨后春笋般出现. 这是一个令人兴奋的、新兴的交叉学科. 然而, 由于其多学科交叉特性、可用数据及算力的限制, 这一领域还处于发展早期, 仍面临诸多困难和挑战, 有大量工作可做. 本文希望能为刚进入这一领域的研究者提供一些指引和帮助.

参考文献

- [1] Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Židek A, Nelson A W, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones D T, Silver D, Kavukcuoglu K, Hassabis D 2020 *Nature* **577** 706
- [2] 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 S A A, Ballard A J, 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 A W, Kavukcuoglu K, Kohli P, Hassabis D 2021 *Nature* **596** 583
- [3] Radford A, Narasimhan K, Salimans T, Sutskever I 2018 *Improving Language Understanding by Generative Pre-Training* [2024-6-9]
- [4] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I 2019 *Language Models are Unsupervised Multitask Learners* [2024-6-9]
- [5] Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen

- M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodeis D 2020 arXiv: 2005.14165[cs.CV]
- [6] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Low R 2022 arXiv: 2203.02155[cs.CV]
- [7] Devlin J, Chang M W, Lee K, Toutanova K 2018 arXiv: 1810.04805[cs.CV]
- [8] Ma Z, He J, Qiu J, Cao H, Wang Y, Sun Z, Zheng L, Wang H, Tang S, Zheng T, Lin J, Feng G, Huang Z, Gao J, Zeng A, Zhang J, Zhong R, Shi T, Liu S, Zheng W, Tang J, Yang H, Liu X, Zhai J, Chen W 2022 *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* Seoul, Republic of Korea, April 2–6, 2022 p192
- [9] Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, Qiu J, Yao Y, Zhang A, Zhang L, Han W, Huang M, Jin Q, Lan Y, Liu Y, Liu Z, Lu Z, Qiu X, Song R, Tang J, Wen J R, Yuan J, Zhao W X, Zhu J 2021 arXiv: 2106.07139[AI]
- [10] Yuan S, Zhao H, Zhao S, et al. 2022 arXiv: 2203.14101 [cs.LG]
- [11] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi E H, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W 2022 arXiv: 2206.07682[cs.CV]
- [12] Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T, Gong Z, Samangooei S, Monteiro M, Menick J, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Binkowski M, Barreira R, Vinyals O, Zisserman A, Simonyan K 2022 arXiv: 2204.14198[cs.CV]
- [13] OpenAI, Achiam J, Adler S, et al. 2024 arXiv: 2303.08774 [cs.CV]
- [14] Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I, Florence P 2023 arXiv: 2303.03378[cs.LG]
- [15] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G 2023 arXiv: 2302.13971[cs.CV]
- [16] Gemini Team Google, Anil R, Borgeaud S, et al. 2024 arXiv: 2312.11805[cs.CV]
- [17] Chen F, Han M, Zhao H, Zhang Q, Shi J, Xu S, Xu B 2023 arXiv: 2305.04160[cs.CV]
- [18] Li K, He Y, Wang Y, Li Y, Wang W, Luo P, Wang Y, Wang L, Qiao Y 2023 arXiv: 2305.06355[cs.CV]
- [19] Bepler T, Berger B 2019 arXiv: 1902.08661[cs.LG]
- [20] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B 2019 bioRxiv: 614313[Bioinformatics]
- [21] Alley E C, Khimulya G, Biswas S, Alquraishi M, Church G M 2019 *Nat. Methods* **16** 1315
- [22] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick C L, Ma J, Fergus R 2021 *Proc. Natl. Acad. Sci.* **118** e2016239118
- [23] Rao R, Liu J, Verkuil R, et al. 2021 bioRxiv: 2021.02.12. 430858 [Synthetic Biology]
- [24] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A 2021 *Advances in Neural Information Processing Systems* **34** 29287
- [25] 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 2023 *Science* **379** 1123
- [26] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Santos Costa A d, Fazel-Zarandi M, Sercu T, Candido S, Rives A 2022 bioRxiv: 2022.07.20.500902[Synthetic Biology]
- [27] Madani A, McCann B, Naik N, Keskar N S, Anand N, Eguchi R R, Huang P S, Socher R 2020 arXiv: 2004.03497[q-bio.QM]
- [28] Madani A, Krause B, Greene E R, Subramanian S, Mohr B P, Holton J M, Olmos J L, Xiong C, Sun Z Z, Socher R, Fraser J S, Naik N 2023 *Nat. Biotechnol.* **41** 1099
- [29] He L, Zhang S, Wu L, Xia H, Ju F, Zhang H, Liu S, Xia Y, Zhu J, Deng P, Shao B, Qin T, Liu T Y 2021 arXiv: 2110.15527[cs.CV]
- [30] Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B 2021 arXiv: 2007.06225[cs.LG]
- [31] Chen B, Cheng X, Li P, Geng Y, Gong J, Li S, Bei Z, Tan X, Wang B, Zeng X, Liu C, Zeng A, Dong Y, Tang J, Song L 2024 arXiv: 2401.06199[q-bio.QM]
- [32] Nguyen E, Poli M, Durrant M G, Thomas A W, Kang B, Sullivan J, Ng M Y, Lewis A, Patel A, Lou A, Ermon S, Baccus S A, Hernandez-Boussard T, Ré C, Hsu P D, Hie B L 2024 bioRxiv: 2024.02.27.582234[Synthetic Biology]
- [33] Gao W, Mahajan S P, Sulam J, Gray J J 2020 *Patterns* **1** 100142
- [34] Unsal S, Atas H, Albayrak M, Turhan K, Acar A C, Doğan T 2022 *Nature Machine Intelligence* **4** 227
- [35] Zhang Q, Ding K, Lyv T, Wang X, Yin Q, Zhang Y, Yu J, Wang Y, Li X, Xiang Z, Feng K, Zhuang X, Wang Z, Qin M, Zhang M, Zhang J, Cui J, Huang T, Yan P, Xu R, Chen H, Li X, Fan X, Xing H, Chen H 2024 arXiv: 2401.14656[cs.CV]
- [36] Guan X Y, Huang H Y, Peng H Q, Liu Y H, Li W F, Wang W 2023 *Acta Phys. Sin.* **72** 248708 (in Chinese) [管星悦, 黄恒焱, 彭华祺, 刘彦航, 李文飞, 王炜 2023 物理学报 **72** 248708]
- [37] Chen G L, Zhang Z Y 2023 *Acta Phys. Sin.* **72** 248705 (in Chinese) [陈光临, 张志勇 2023 物理学报 **72** 248705]
- [38] Zhang J H 2024 *Acta Phys. Sin.* **73** 069301 (in Chinese) [张嘉晖 2024 物理学报 **73** 069301]
- [39] Zeng C, Jian Y, Vosoughi S, Zeng C, Zhao Y 2023 *Nat. Commun.* **14** 1060
- [40] Zeng C, Zhao Y 2023 *Scientia Sinica Physica, Mechanica & Astronomica* **53** 290018
- [41] Huh M, Cheung B, Wang T, Isola P 2024 arXiv: 2405.07987 [cs.LG]
- [42] Bepler T, Berger B 2021 *Cell Systems* **12** 654
- [43] Guo Y, Wu J, Ma H, Huang J 2022 *Proceedings of the AAAI Conference on Artificial Intelligence* **36** 6801
- [44] Hermosilla P, Ropinski T 2022 arXiv: 2205.15675[q-bio.BM]
- [45] Zhang Z, Xu M, Jamasb A, Chenthamarakshan V, Lozano A, Das P, Tang J 2022 arXiv: 2203.06125[cs.LG]
- [46] Zhang Z, Xu M, Lozano A, Chenthamarakshan V, Das P, Tang J 2023 arXiv: 2303.06275[q-bio.QM]
- [47] Gligorijević V, Renfrew P D, Kosciolk T, Leman J K, Berenberg D, Vatanen T, Chandler C, Taylor B C, Fisk I M, Vlamakis H, Xavier R J, Knight R, Cho K, Bonneau R 2021 *Nat. Commun.* **12** 3168
- [48] Wang Z, Combs S A, Brand R, Calvo M R, Xu P, Price G, Golovach N, Salawu E O, Wise C J, Ponnappalli S P, Clark P M 2022 *Sci. Rep.* **12** 6832
- [49] Chen C, Zhou J, Wang F, Liu X, Dou D 2023 arXiv: 2204.04213[cs.LG]
- [50] Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, Zhang L, Ke G 2022 DOI: 10.26434/chemrxiv-2022-jjm0j-v4
- [51] Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F 2023

- bioRxiv: 2023.10.01.560349[Bioinformatics]
- [52] Su J, Li Z, Han C, Zhou Y, Shan J, Zhou X, Ma D, OPMC T, Ovchinnikov S, Yuan F 2024 bioRxiv: 2024.05.24.595648 [Bioinformatics]
- [53] Hu M Y, Yuan F J, Yang K K, Ju F S, Su J, Wang H, Yang F, Ding Q Y 2022 arXiv:2206.06583 [q-bio.QM]
- [54] Abramson J, Adler J, Dunger J, et al. 2024 *Nature* **630** 493
- [55] Wang L, Liu H, Liu Y, Kurtin J, Ji S 2022 arXiv: 2207.12600[cs.LG]
- [56] Somnath V R, Bunne C, Krause A 2021 arXiv: 2204.02337[cs.LG]
- [57] Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein M M, Correia B E 2020 *Nat. Methods* **17** 184
- [58] Wu F, Jin S, Jiang Y, Jin X, Tang B, Niu Z, Liu X, Zhang Q, Zeng X, Li S Z 2022 arXiv: 2204.08663[CE]
- [59] Meyer T, D'Abramo M, Rueda M, Ferrer-Costa C, Pérez A, Carrillo O, Camps J, Fenollosa C, Repchevsky D, Gelpi J L, Orozco M 2010 *Structure* **18** 1399
- [60] Zhang N, Bi Z, Liang X, Cheng S, Hong H, Deng S, Lian J, Zhang Q, Chen H 2022 arXiv: 2201.11147[q-bio.BM]
- [61] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H 2021 arXiv: 2007.15779[cs.CV]
- [62] Zhou H Y, Fu Y, Zhang Z, Bian C, Yu Y 2023 arXiv: 2301.13154[cs.LG]
- [63] Xu M, Yuan X, Miret S, Tang J 2023 arXiv: 2301.12040 [q-bio.BM]
- [64] Singh J, Hanson J, Paliwal K, Zhou Y 2019 *Nat. Commun.* **10** 5407
- [65] Singh J, Paliwal K, Zhang T, Singh J, Litfin T, Zhou Y 2021 *Bioinformatics* **37** 2589
- [66] Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y, Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
- [67] Wang J, Xiao Y 2017 *Current Protocols in Bioinformatics* **57** 5
- [68] Wang J, Wang J, Huang Y, Xiao Y 2019 *Int. J. Mol. Sci.* **20** 4116
- [69] Tan Y L, Wang X, Shi Y Z, Zhang W, Tan Z J 2022 *Biophys. J.* **121** 142
- [70] Zhou L, Wang X, Yu S, Tan Y L, Tan Z J 2022 *Biophys. J.* **121** 3381
- [71] Wang X, Tan Y L, Yu S, Shi Y Z, Tan Z J 2023 *Biophys. J.* **122** 1503
- [72] Li J, Zhu W, Wang J, Li W, Gong S, Zhang J, Wang W 2018 *PLoS Comput. Biol.* **14** e1006514
- [73] Fu L, Cao Y, Wu J, Peng Q, Nie Q, Xie X 2022 *Nucleic Acids Res.* **50** e14
- [74] Pearce R, Omenn G S, Zhang Y 2022 bioRxiv: 2022.05.15.491755[Bioinformatics]
- [75] Baek M, McHugh R, Anishchenko I, Baker D, DiMaio F 2022 bioRxiv: 2022.09.09.507333[Bioinformatics]
- [76] Zhang J, Lang M, Zhou Y, Zhang Y 2024 *Trends in Genetics* **40** 94
- [77] Li J, Zhou Y, Chen S J 2024 *Curr. Opin. Struct. Biol.* **87** 102847
- [78] Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, Zong L, Hong L, Xiao J, Shen T, King I, Li Y 2022 arXiv: 2204.00300[q-bio.QM]
- [79] Chen K, Zhou Y, Ding M, Wang Y, Ren Z, Yang Y 2023 bioRxiv: 2023.01.31.526427[Bioinformatics]
- [80] Babjac A N, Lu Z, Emrich S J 2023 *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* New York, United States, September 3–6, 2023 p1
- [81] Chu Y, Yu D, Li Y, Huang K, Shen Y, Cong L, Zhang J, Wang M 2024 *Nature Machine Intelligence* **6** 449
- [82] Yang Y, Li G, Pang K, Cao W, Li X, Zhang Z 2023 bioRxiv: 2023.09.08.556883[Bioinformatics]
- [83] Zhang Y, Lang M, Jiang J, Gao Z, Xu F, Litfin T, Chen K, Singh J, Huang X, Song G, Tian Y, Zhan J, Chen J, Zhou Y 2024 *Nucleic Acids Res.* **52** e3
- [84] Wang X, Gu R, Chen Z, Li Y, Ji X, Ke G, Wen H 2023 bioRxiv: 2023.07.11.548588[Bioinformatics]
- [85] Wang N, Bian J, Li Y, Li X, Mumtaz S, Kong L, Xiong H 2024 *Nature Machine Intelligence* **6** 548
- [86] Akiyama M, Sakakibara Y 2022 *NAR Genomics and Bioinformatics* **4** lqac012
- [87] Shen T, Hu Z, Peng Z, Chen J, Xiong P, Hong L, Zheng L, Wang Y, King I, Wang S, Siqi S, Yu L 2022 arXiv: 2207.01586[q-bio.QM]
- [88] Li Y, Zhang C, Feng C, Pearce R, Lydia Freddolino P, Zhang Y 2023 *Nat. Commun.* **14** 5745
- [89] Ferruz N, Schmidt S, Höcker B 2022 *Nat. Commun.* **13** 4348
- [90] Wang J, Lisanza S, Juergens D, Tischler D, Watson J L, Castro K M, Ragotte R, Saragovi A, Milles L F, Baek M, Anishchenko I, Yang W, Hicks D R, Expòsit M, Schlichthaerle T, Chun J H, Dauparas J, Bennett N, Wicky B I M, Muenks A, DiMaio F, Correia B, Ovchinnikov S, Baker D 2022 *Science* **377** 387
- [91] Trippe B L, Yim J, Tischler D, Baker D, Broderick T, Barzilay R, Jaakkola T 2022 arXiv: 2206.04119[q-bio.BM]
- [92] Anishchenko I, Pellock S J, Chidyausiku T M, Ramelot T A, Ovchinnikov S, Hao J, Bafna K, Norr C, Kang A, Bera A K, DiMaio F, Carter L, Chow C M, Montelione G T, Baker D 2021 *Nature* **600** 547
- [93] Wicky B I M, Milles L F, Courbet A, Ragotte R J, Dauparas J, Kinfu E, Tipps S, Kibler R D, Baek M, DiMaio F, Li X, Carter L, Kang A, Nguyen H, Bera A K, Baker D 2022 *Science* **378** 56
- [94] Anand N, Achim T 2022 arXiv: 2205.15019[q-bio.QM]
- [95] Luo S, Su Y, Peng X, Wang S, Peng J, Ma J 2022 *Advances in Neural Information Processing Systems* **35** 9754
- [96] Cao L, Coventry B, Goreshnik I, et al 2022 *Nature* **605** 551
- [97] Kuhlman B, Bradley P 2019 *Nat. Rev. Mol. Cell Biol.* **20** 681
- [98] Pan X, Kortemme T 2021 *J. Biol. Chem.* **296** 100558
- [99] Khakzad H, Igashov I, Schneuing A, Goverde C, Bronstein M, Correia B 2023 *Cell Systems* **14** 925
- [100] Malbrancke C, Bikard D, Cocco S, Monasson R, Tubiana J 2023 *Curr. Opin. Struct. Biol.* **80** 102571
- [101] Kortemme T 2024 *Cell* **187** 526
- [102] Notin P, Rollins N, Gal Y, Sander C, Marks D 2024 *Nat. Biotechnol.* **42** 216
- [103] Listov D, Goverde C A, Correia B E, Fleishman S J 2024 *Nat. Rev. Mol. Cell Biol.* **25** 639
- [104] Ingraham J, Garg V K, Barzilay R, Jaakkola T 2019 *Proceedings of the 33rd International Conference on Neural Information Processing Systems* Vancouver, BC, Canada, December 8–14, 2019 p15820
- [105] Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte R J, Milles L F, Wicky B I M, Courbet A, de Haas R J, Bethel N, Leung P J Y, Huddy T F, Pellock S, Tischler D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera A K, King N P, Baker D 2022 *Science* **378** 49
- [106] Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, Lerer A, Rives A 2022 bioRxiv: 2022.04.10.487779[Systems Biology]
- [107] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, Ganguli S 2015 arXiv: 1503.03585[cs.LG]
- [108] Ho J, Jain A, Abbeel P 2020 *Advances in Neural Information Processing Systems* **33** 6840
- [109] Watson J L, Juergens D, Bennett N R, et al 2023 *Nature* **620** 1089
- [110] Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S, Poole B 2020 arXiv: 2011.13456[cs.LG]

- [111] Lee J S, Kim J, Kim P M 2023 *Nature Computational Science* **3** 382
- [112] Liu Y, Chen L, Liu H 2023 *bioRxiv*: 2023.11.18.567666 [Bioinformatics]
- [113] Zheng Z, Deng Y, Xue D, Zhou Y, YE F, Gu Q 2023 *arXiv*: 2302.01649[cs.LG]
- [114] Yang K K, Zanichelli N, Yeh H 2023 *Protein Eng. Des. Sel.* **36** gzad015
- [115] Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D 2020 *arXiv*: 2001.08361[cs.LG]
- [116] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R 2021 *arXiv*: 2111.06377[cs.CV]
- [117] Chen T, Kornblith S, Norouzi M, Hinton G 2020 *arXiv*: 2002.05709[cs.LG]
- [118] Wang Z, Wang Z, Srinivasan B, Ioannidis V N, Rangwala H, Anubhai R 2023 *arXiv*: 2310.03320[cs.LG]
- [119] Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Walczak M, Pfrommer J, Pick A, Ramamurthy R, Garcke J, Bauckhage C, Schuecker J 2021 *IEEE Trans. Knowl. Data Eng.* **35** 614
- [120] Bao L, Zhang X, Jin L, Tan Z J 2015 *Chin. Phys. B* **25** 018703
- [121] Qiang X W, Zhang C, Dong H L, Tian F J, Fu H, Yang Y J, Dai L, Zhang X H, Tan Z J 2022 *Phys. Rev. Lett.* **128** 108103
- [122] Dong H L, Zhang C, Dai L, Zhang Y, Zhang X H, Tan Z J 2024 *Nucleic Acids Res.* **52** 2519

SPECIAL TOPIC—Machine learning in biomolecular modelling

Progress in protein pre-training models integrating structural knowledge*

Tang Tian-Yi¹⁾ Xiong Yi-Ming¹⁾ Zhang Rui-Ge¹⁾ Zhang Jian^{1)2)†}

Li Wen-Fei¹⁾²⁾ Wang Jun¹⁾²⁾ Wang Wei^{1)2)‡}

1) (School of Physics, Nanjing University, Nanjing 210093, China)

2) (Institute of Brain Science, Nanjing University, Nanjing 210093, China)

(Received 7 June 2024; revised manuscript received 12 July 2024)

Abstract

The AI revolution, sparked by natural language and image processing, has brought new ideas and research paradigms to the field of protein computing. One significant advancement is the development of pre-training protein language models through self-supervised learning from massive protein sequences. These pre-trained models encode various information about protein sequences, evolution, structures, and even functions, which can be easily transferred to various downstream tasks and demonstrate robust generalization capabilities. Recently, researchers have further developed multimodal pre-trained models that integrate more diverse types of data. The recent studies in this direction are summarized and reviewed from the following aspects in this paper. Firstly, the protein pre-training models that integrate protein structures into language models are reviewed: this is particularly important, for protein structure is the primary determinant of its function. Secondly, the pre-trained models that integrate protein dynamic information are introduced. These models may benefit downstream tasks such as protein-protein interactions, soft docking of ligands, and interactions involving allosteric proteins and intrinsic disordered proteins. Thirdly, the pre-trained models that integrate knowledge such as gene ontology are described. Fourthly, we briefly introduce pre-trained models in RNA fields. Finally, we introduce the most recent developments in protein designs and discuss the relationship of these models with the aforementioned pre-trained models that integrate protein structure information.

Keywords: protein foundation model, protein multi-modal model, protein structure, machine learning

PACS: 87.10.Vg, 87.16.A–, 87.14.E–, 87.15.A–

DOI: 10.7498/aps.73.20240811

* Project supported by the Science and Technology Innovation Project of the Ministry of Science and Technology (Grant No. 2030-2021ZD0201300) and the National Natural Science Foundation of China (Grant No. 11934008).

† Corresponding author. E-mail: jzhang@nju.edu.cn

‡ Corresponding author. E-mail: wangwei@nju.edu.cn