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意义深远的影响[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]. 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, 数据驱动的机器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

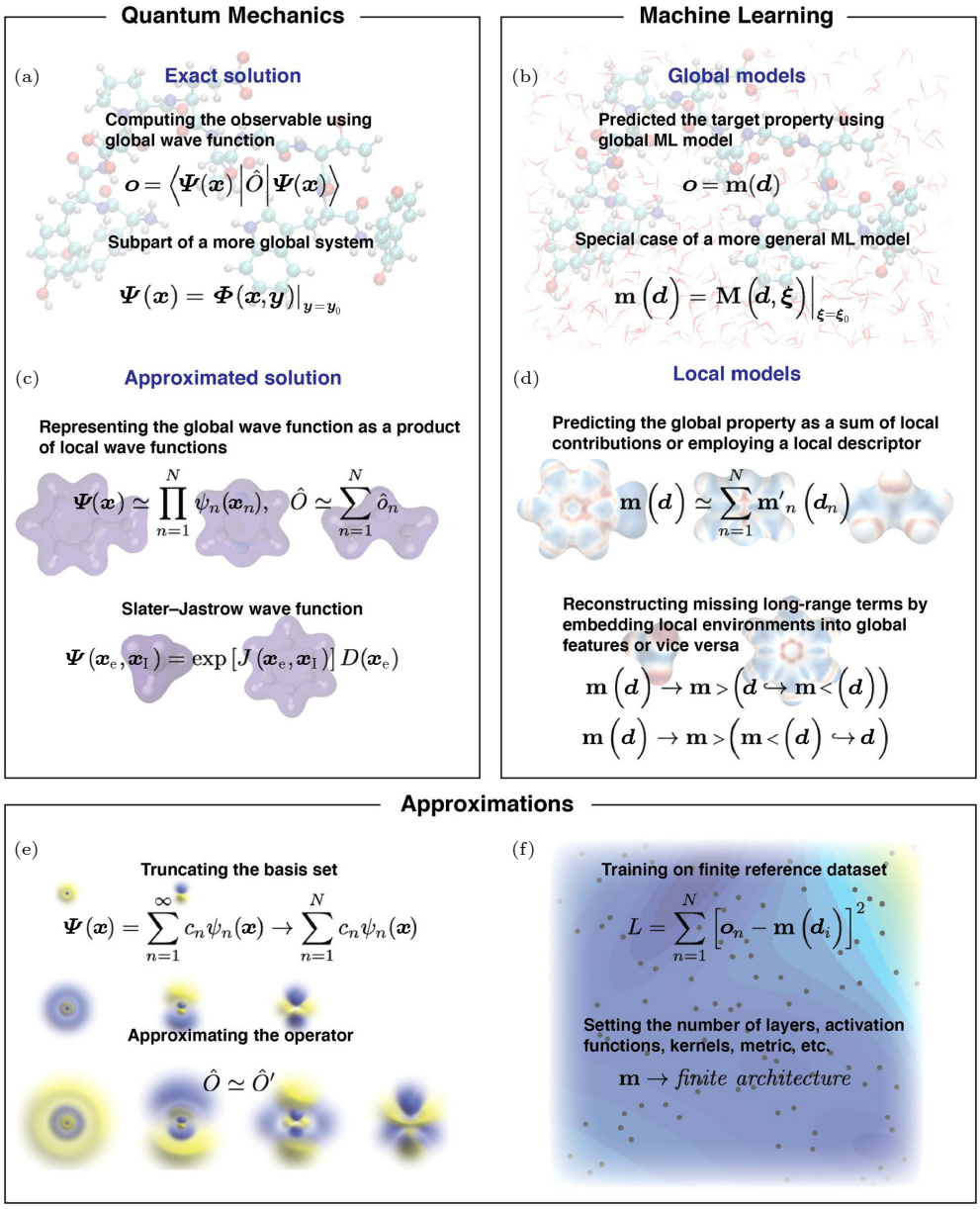

图 1 量子力学与机器学习间的相似性. 从左到右, 从上到下的图片分别是: Chignolin 蛋白质在 (a) 无水环境和 (b) 有水环境下的情况, 使用 SchNet 模型得到的 (c) 可视化电荷密度和 (d) 局部化学势, (e) 氢原子的波函数以及 (f)Müller-Brown 势能. 图片引自文献 [42] (版权属于美国化学会)

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

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

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, 但确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工作流, 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-3β 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的一个潮流[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 neteork, 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, 发现其和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

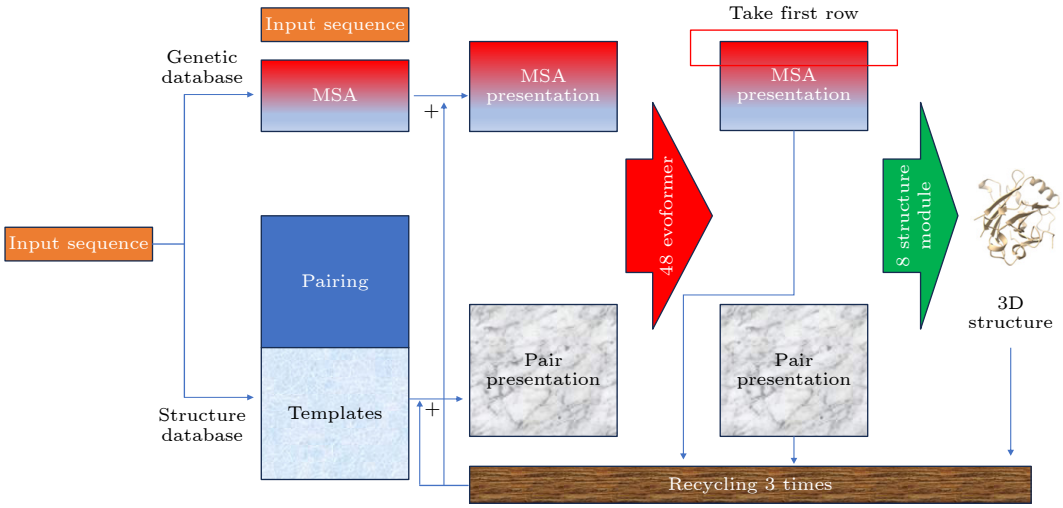

图 2 AlphaFold2 的结构图

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链接, 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.

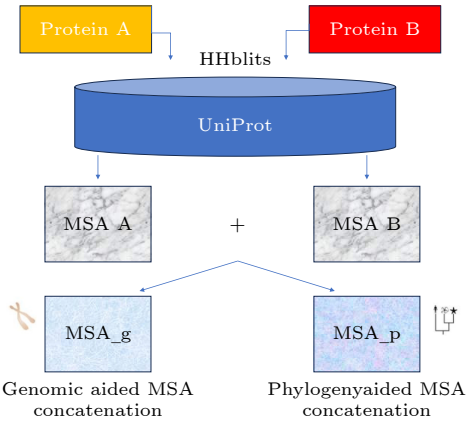

图 3 AlphaFold-Multimer 的多序列比对构建方法

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 (基于线型回归), 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:

$$\begin{aligned}&\text{designed protein}\\&=\text{argmax}_{\text{protein}}P(\text{protein}|\text{condition}).\end{aligned}\tag{2}$$

因为我们把骨架结构设计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:

$$\begin{aligned}&P(\text{protein}|\text{condition})\\&=P(\text{sequence},\text{structure}|\text{condition}).\end{aligned}\tag{3}$$

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:

$$\begin{aligned}&\text{designed structure}\\&=\text{argmax}_{\text{structure}}P(\text{structure}|\text{condition}).\end{aligned}\tag{4}$$

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:

$$\begin{aligned}&\text{designed sequence} = \text{argmax}_{\text{sequence}}\\&P(\text{sequence}|\text{designed structure}, \text{condition}).\end{aligned}\tag{5}$$

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:

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

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

$$P_{\max}^{\text{traditional}} \leqslant P_{\max}^{\text{co-design}},\tag{7}$$

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

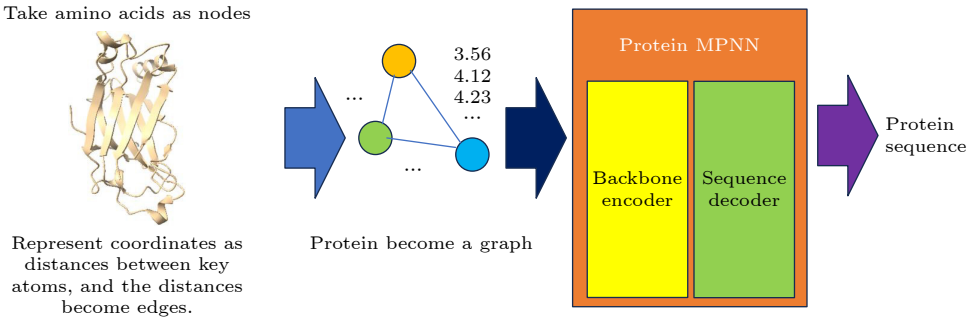

图 4 ProteinMPNN 模型核心思想示意图

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多的信息. 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.

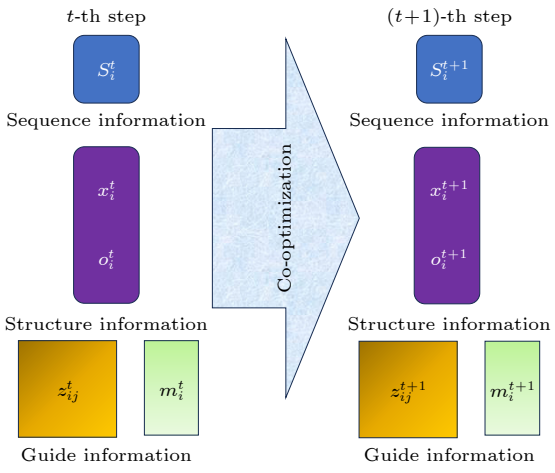

图 5 蛋白质结构序列协同设计的一种机器学习模型示意图

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也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] Markowetz 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, Isralewitz 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] Mirabello 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, Saitô 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 Ł, 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 marching 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 andlooks forward to the application of marching 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

* Project supported by the National Natural Science Foundation of China (Grant No. 22177107).

† Corresponding author. E-mail: jhzhang@ustc.edu.cn

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

蛋白质计算中的机器学习

张嘉晖

Machine learning for *in silico* protein research

Zhang Jia-Hui

引用信息 Citation: Acta Physica Sinica, 73, 069301 (2024) DOI: 10.7498/aps.73.20231618
在线阅读 View online: https://doi.org/10.7498/aps.73.20231618
当期内容 View table of contents: http://wulixb.iphy.ac.cn