



智能蛋白质工程*

王凯悦^{1,3)} 叶盛^{1,2,3)**}⁽¹⁾ 北京航空航天大学生物与医学工程学院, 北京 100191; ⁽²⁾ 北京航空航天大学医学科学与工程学院, 北京 100191;⁽³⁾ 北京航空航天大学大数据精准医疗工业和信息化部重点实验室, 北京 100191)

摘要 蛋白质是生命活动的重要基础物质, 在生物体内扮演着多种关键角色, 包括构建细胞结构、参与代谢和能量转化、调节生理功能、提供免疫保护、传递信号等。蛋白质多样化的功能是通过其特定的氨基酸序列, 以及相应的三维结构来实现的。蛋白质工程通过改变或设计蛋白质的序列与结构来实现特定的功能, 从而扩展对蛋白质的理解, 并为生物医学、生物材料、生物工程及其他领域的研究提供强大的工具和技术支持。近年来, 随着算法的进步、大数据的积累, 以及硬件计算能力的提升, 人工智能技术得到了快速发展, 并逐渐应用于蛋白质工程领域, 形成了智能蛋白质工程。通过运用基因组、蛋白质组、蛋白质结构数据库等生物学大数据, 以及在数据基础上建立各类先进的深度学习模型, 智能蛋白质工程能够实现高效、精准、可预测的蛋白质设计和改造。本文主要侧重于智能蛋白质工程的四个方面, 即结构设计、无骨架的序列设计、基于骨架的序列设计, 以及其他辅助设计方法, 总结了人工智能技术在这些领域取得的最新进展, 并汇总了近年来采用智能蛋白质工程技术取得的实践成果。智能蛋白质工程作为一种新兴的技术和方法, 展示了巨大的潜力和前景, 为未来科学研究和技术创新带来深远的影响, 并为解决全球性挑战提供新的解决方案和工具。

关键词 蛋白质工程, 蛋白质设计, 密码子优化, 蛋白质结构, 深度学习, 深度生成, 人工智能

中图分类号 Q816

DOI: 10.16476/j.pibb.2024.0402

蛋白质是生命体不可或缺的基本组成部分, 其重要性不言而喻。无论是在生命体内还是在病毒等类生命体内, 蛋白质都承担着多种多样的关键功能, 从搭建各类基本结构到催化复杂的生物化学反应, 在生命的各个层面扮演着至关重要的角色^[1]。随着对蛋白质功能的深入研究和实际应用的推进, 传统的天然蛋白质已无法满足人类日益增长的多样化需求。蛋白质工程旨在现有天然蛋白质的基础上, 创造出具有特定功能的新蛋白质分子。但是, 受限于蛋白质序列近乎无限的可能性, 以及尚不能充分了解蛋白质的折叠规律与“结构-功能”关系, 所以很难实现任意功能蛋白质的设计。

近年来, 新一代基于深度学习 (deep learning) 模型的人工智能 (artificial intelligence, AI) 技术有了突破性的发展, 促使研究者将人工智能技术应用于生命科学领域的研究, 令该领域的研究方法和思维范式发生了转变。特别是2024年的诺贝尔化学奖分别授予了David Baker (以表彰其在计算蛋

白质设计方面的贡献), 以及Demis Hassabis和John Jumper (以表彰两人在蛋白质结构预测方面的贡献)。后两者正是基于深度学习的蛋白质结构预测程序AlphaFold的主要开发者, 而David Baker则在近年推出了诸如RFdiffusion等重要的基于人工智能的蛋白质设计算法, 由此大大加速了他的研究团队在蛋白质设计方面的研究进展。

蛋白质科学与人工智能先进算法相融合的发展趋势体现在蛋白质工程领域, 就催生了“智能蛋白质工程” (intelligent protein engineering), 即利用人工智能技术辅助解决蛋白质工程在蛋白质的设计、改造、生产等诸多环节中的挑战^[2]。相比于传统蛋白质工程当中采用的局部饱和和突变或定向进化策略, 智能蛋白质工程通过人工智能算法所能探

* 北京市科技计划课题 (Z221100003522020) 资助项目。

** 通讯联系人。

Tel: 010-82317114, E-mail: yesheng@buaa.edu.cn

收稿日期: 2024-09-12, 接受日期: 2024-11-27

索的序列空间更为广阔也更为准确, 因此更有可能获得具备指定新功能, 或原有功能更加高效、性质更加稳定的定制化蛋白质产品, 以解决医学、药物、生物技术、环境保护、工业生产等领域的产业需求。在智能蛋白质工程的多个组成部分中, 基于人工智能算法的蛋白质设计是关键核心技术之一^[3], 不仅为人类改造和设计蛋白质提供了有力的工具, 也促进了人类对于蛋白质科学的认知^[4]。这一部分将在本文中进行重点综述。此外, 人工智能技术还在密码子优化等其他技术环节上对蛋白质工程做出了贡献, 这些技术也将是本文综述的内容。

1 蛋白质设计方法的分类与问题

蛋白质设计与蛋白质工程的目标有所差异: 前者注重发现新的非天然的蛋白质结构形态, 而后者更关注能够解决科研或生产领域实际问题的蛋白质改造。蛋白质设计的初衷决定其早期研究目标是无参照模板、无约束条件下的蛋白质设计, 因此也称蛋白质的“从头设计”(de novo design)。事实上, 有研究者正是以此作为划分不同蛋白质设计方法的分水岭^[5]。蛋白质从头设计问题的本质是研究蛋

白质的折叠与拓扑, 因此主要关注蛋白质的主链走向。然而, 蛋白质的折叠同时依赖于主链或侧链参与的残基相互作用, 设计主链骨架还需要填充可以实现该折叠的适当氨基酸序列, 因此诞生了依据主链骨架来设计氨基酸序列的算法。近年来, 随着大语言模型的兴起, 又出现了仅把蛋白质考虑为氨基酸序列的处理方法。相应的, 也有研究者主张将蛋白质设计分为固定骨架设计、可变骨架设计和序列结构生成三类问题^[6]。

面对纷繁复杂的蛋白质设计方法, 笔者在本文中主张做以下分类: a. “结构设计”(structure design), 指以蛋白质结构等数据训练深度生成模型, 从而直接生成新的设计蛋白质的结构, 包括仅生成主链原子, 或同时生成主侧链全部原子; b. “无骨架的序列设计”(backbone-free sequence design), 指以蛋白质序列等数据训练深度生成模型, 从而直接生成新的设计蛋白质的氨基酸序列; c. “基于骨架的序列设计”(backbone-based sequence design), 指根据待设计蛋白质已知的主链骨架结构, 设计能够实现该折叠的氨基酸序列。图1体现了这三类蛋白质设计方法之间的关联性与差异性。表1中列出了三个类别中采用了深度学习模型的一些代表性算法。

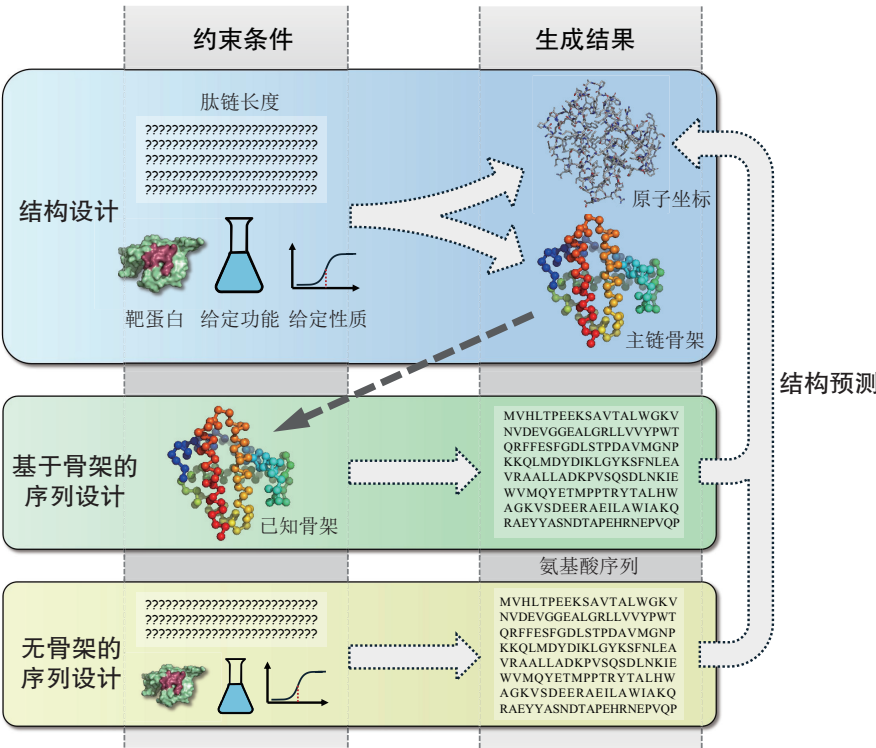


Fig. 1 Relationships between three kinds of protein design approaches
图1 三类蛋白质设计方法之间的联系与差异

Table 1 Classification of protein design methods and representative algorithms using deep learning models

表1 蛋白质设计方法分类及采用了深度学习模型的代表性算法

方法类别	算法名称	开发者	发布时间	技术特点	功能与应用场景
结构设计 (structure design)	RFdiffusion ^[7]	David Baker研究组	2022年	在残基距离矩阵的二维图上应用扩散模型，并利用RosettaFold来构造去噪函数	从用户指定的输入生成图像的网络的方式，能够从简单的分子规范设计出各种对称性的蛋白质，也可以无约束设计蛋白质
	SCUBA-D ^[8]	刘海燕研究组	2022年	使用氨基酸序列语言模型辅助结构扩散，使用非零均值先验的去噪扩散将低质量初始主干转换为高质量主干	可以始终生成高质量的蛋白质骨架结构，再匹配其他算法来填充氨基酸序列
	Ig-VAE ^[9]	Huang等	2022年	提供了一个旋转和平移不变的损失函数，同时还具有扭转感知和直接生成三维坐标的能力	支持高通量构建表位定制的结构引导文库，并可应用于任何结构折叠类别，如酶和核酸结合域，但主要用于抗体的设计
	DiffAb ^[10]	马剑竹研究组	2022年	是最早用于蛋白质结构设计的扩散模型之一，可以设计原子分辨率的且明确针对特定抗原结构的抗体	是一把“瑞士军刀”，能够进行序列结构协同设计、给定主链结构的序列设计和抗体优化
	Chroma ^[11]	Generate Biomedicines公司	2023年	算法的可编程性允许通过设置条件器直接指定和优化蛋白质的高级属性	可以生成具有各种结构和特性的新型多样化蛋白质
无骨架的 序列设计 (backbone-free sequence design)	MSA-VAE ^[12]	Hawkins-Hooker等	2021年	可以更好地捕获反映三维结构影响的长距离依赖关系	使用深度生成模型搜索生成有用变体的可行性，提供了一种补充理性设计的方法
	ProteinGAN ^[13]	Repecka等	2021年	是一种基于自注意力机制的生成对抗网络变体	能够“学习”天然蛋白质序列的多样性并能够生成功能性蛋白质序列
	ProtGPT2 ^[14]	Ferruz等	2022年	具有7.38亿个参数的深度无监督语言模型，探索了蛋白质空间中的新区域	能够以高通量方式生成从头蛋白质序列，能够筛选与天然蛋白质相似的蛋白质
	ProGen ^[15]	Profluent公司	2023年	使用条件语言建模来生成人工蛋白质	基于深度学习的语言模型可以生成跨大型蛋白质家族的功能性人工蛋白质序列
	ProGen2 ^[16]	Profluent公司	2023年	可扩展到64亿个参数，可以更好地捕捉蛋白质序列的分布	在捕获观察到的进化序列的分布、生成新的可行序列和预测蛋白质适应性方面表现出最先进的性能，而无需额外的微调
基于骨架的 序列设计 (backbone-based sequence design)	DARK ^[17]	Moffat等	2022年	可以快速采样具有多样化和有序结构的新蛋白质序列	可为从头设计蛋白质生成高质量的候选者，并有助于开发进一步的新型设计方法
	ProteinMPNN ^[18]	David Baker研究组	2022年	训练数据中添加了高斯噪声，提高了模型的稳定性，同时也增强了模型的泛化能力	适用于几乎所有蛋白质序列设计问题，高实验设计成功率、计算效率和广泛适用性
	LM-Design ^[19]	Zheng等	2023年	能准确处理功能环和暴露于表面等结构非确定性区域	作为一个“封装器”来整合蛋白质序列学习（如pLMs）和结构学习（如几何/图神经网络及蛋白质结构预测）的进展，从而促进未来的蛋白质研究

续表1

方法类别	算法名称	开发者	发布时间	技术特点	功能与应用场景
	LigandMPNN ^[20]	David Baker研究组	2023年	明确地模拟了生物分子系统的所有非蛋白质成分	可以对小分子、核苷酸、金属和其他原子环境进行明确的建模
	ProDesign-LE ^[21]	卜东波研究组	2023年	通过优化残基与局部环境的匹配, 避免了频繁重建全原子结构, 提高了设计效率和准确性	对于设计出差异化的序列有着非常强的实用性
	CarbonDesign ^[22]	张海仓研究组	2024年	核心是Inverseformer, 可以视为Evoformer的逆向架构, 让信息从三维结构流向一维序列	可以在设计序列的同时生成侧链结构, 对于原始序列有着相当高的召回率

事实上, 上述分类方法本身就体现了当前蛋白质设计算法的两个严重问题, 即结构与序列的割裂, 以及主链与侧链的割裂。

结构与序列的割裂问题一直困扰着蛋白质的相关研究领域。其根本原因在于, 人类至今仍旧没有完全掌握蛋白质折叠过程中的全部物理和化学规律, 既不能由氨基酸序列直接计算得到其对应的三维折叠形式, 也不能由给定的三维折叠形式直接计算得到其对应的氨基酸序列。这就导致, 设计方法获得的蛋白质序列无法直接对应于结构, 还要求助于蛋白质结构预测算法来获知设计蛋白可能的结构状态, 添加了额外的一层不确定性。

主链与侧链的割裂在传统的结构生物学研究中是不存在的。这是因为, 结构生物学研究的蛋白质往往已知氨基酸序列, 也就具备了明确的侧链化学信息。结构生物学的实践也表明, 在蛋白质三维结构中, 主链与侧链是一体的, 两者深度耦合。但是在设计蛋白质时, 序列恰恰是未知信息, 而结构也是未知信息, 因此才诞生了先主链折叠, 再侧链装配的设计流程。从方法体系上来讲, 当前缺乏能够一体化解决主侧链设计问题的算法框架。

与上述两个问题有着紧密联系的是蛋白质结构数据的表征方式问题。Wang等^[23]指出, 蛋白质的不同表征方式也与相应深度学习模型的框架选择有着紧密的关系。当前的主链骨架设计模型往往将蛋白质结构表征为所有氨基酸主链原子串联形成的结构, 甚至是简化为 α 碳原子的长链, 因此不具备表征和理解侧链的能力; 全原子生成模型往往将蛋白质理解为所有非氢原子组成的拓扑图; 笔者团队开发的基于电子密度的扩散模型, 将蛋白质表征为三

维空间的电子密度数值矩阵。这些表征方式之间各有优劣, 并且与所采用的深度学习模型的类型有关, 很难直接比较。因此下文的论述中将不再单独就表征方式进行比较讨论。

除此之外, 巨大的蛋白质序列空间也给设计带来了极大的挑战。仅以100个残基的小蛋白质为例, 其肽链上的每个氨基酸都有20种可能, 因此其序列理论上的可能组合数量就达到了 20^{100} , 即 10^{130} 左右, 比可观测宇宙中的原子数量还高了50个数量级。在这样巨大的一个序列空间中, 很难通过随机搜索来找到可以正确折叠的序列, 也就使得生物学和化学研究中经常使用的筛选法在蛋白质设计中难以成立^[24-26]。

在蛋白质工程的角度来看, 蛋白质设计的目标是要达到某种蛋白质功能或令该蛋白质具备某种理化性质, 可记为 T 。若将设计得到的蛋白质结构记为 S , 那么蛋白质设计算法的本质就是求得给定 T 时 S 的后验概率, 即 $P(S|T)$, 而后从中采样。根据贝叶斯定理有

$$P(S|T) = \frac{P(S) \cdot P(T|S)}{P(T)}$$

(1)

其中, 由于目标是设计之初给定的, 所以在每一个特定的设计任务中, T 是定值。 $P(T|S)$ 表示对于给定结构 S , 其具备功能/性质 T 的后验概率(图2中红色子空间)。通常从蛋白质结构到性质是具有确定性的。但是由于在微观层面缺乏相关的科学理论, 因此只有一小部分功能或性质是可以依据公式计算得到的, 而大部分的蛋白质功能或性质都难以准确计算, 令 $P(T|S)$ 的获取变得困难。最后一项 $P(S)$ 代表了所有可能蛋白质结构构型的概率分布

(图2中绿色子空间)。考虑到前述蛋白质序列空间极其巨大的因素,以及目前对蛋白质折叠的原理仍

旧缺乏深刻的理论认知, $P(S)$ 的获取也是非常困难的。

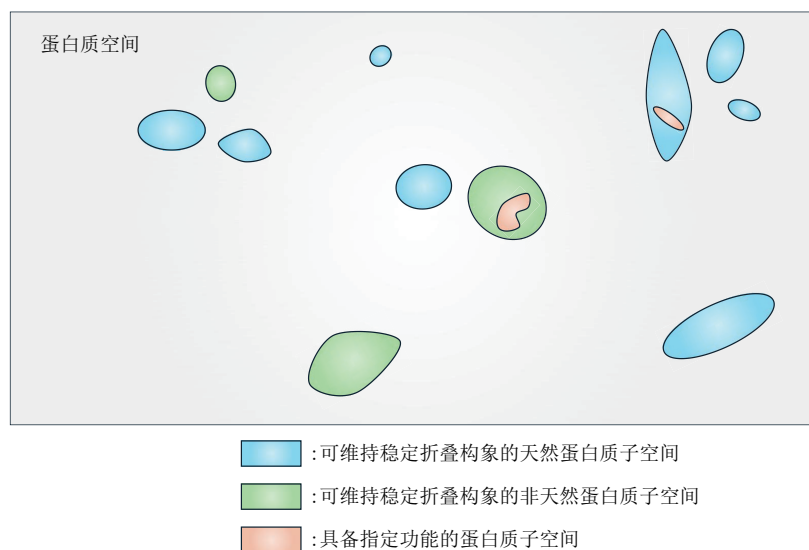


Fig. 2 The composition of protein space

图2 蛋白质空间的构成

2 蛋白质的结构设计

尽管困难重重,从20世纪70年代起,生物学家们仍旧发展了一系列不同的方法,尝试解决蛋白质结构的设计问题。第一个时期是20世纪70年代末到80年代初,主要使用物理模型手动设计蛋白质;第二个时期从20世纪80年代中期持续到21世纪初,主要集中在由基本物理化学原理指导的计算设计,使用数学方程定义骨架构象^[27-29],配合侧链重新装配算法来设计序列^[30-31];第三个时期始于21世纪初,随着晶体结构数据库不断扩大,发展出了基于蛋白质结构片段和生物信息学计算的蛋白质从头设计方法,由已知确定结构的蛋白质片段重新组合以创建新的蛋白质^[32-33]。

总体上,这一时期的蛋白质结构设计强烈依赖于设计人员的专业知识,成功率极低,计算耗时极长,且只能解决个别问题,不具备通用性的方法。直到近年来,人工智能算法和高性能计算的算力都有了突破性的发展,才带来了新的希望。

2.1 从结构预测到结构设计

在基于深度学习模型的人工智能算法介入之前,蛋白质结构设计的主要成果大多出自美国华盛顿大学的David Baker教授团队^[7, 34-35]。他们开发

了一种识别“极低自由能序列-结构对”的通用程序^[36],并在序列优化和结构预测之间迭代,可应用于多种期望目标结构的设计。该程序依赖于Baker团队此前在研究蛋白质结构预测算法Rosetta^[37]的过程中建立的一个经验型的混合能量函数,综合考虑了蛋白质结构涉及的多方面能量项。虽然该函数计算得到的并不是一个可验证的实际能量数值,而只是多个能量项的整体评估,但是确实能够指导蛋白质的折叠预测与结构设计,并得到有益的结果。Baker团队利用该算法从头设计的蛋白质Top7虽然不具备生物学功能,但是其晶体结构证明,该蛋白质的实际结构与给定设计目标一致,且结构高度稳定^[32]。Top7的设计成功是一个重要的里程碑,表明综合运用生物信息学、结构生物学、生物物理学、生物化学等多学科的知识和技术手段,完全有可能成功设计出某些特定的蛋白质结构。

在Top7等同一时期设计蛋白质的流程中,一个核心环节是对于给定氨基酸序列对应的蛋白质三维结构的预测(structure prediction)。广义来讲,预测问题通常可以理解为对一个从输入(氨基酸序列)到输出(三维结构)的转换函数的拟合。如果该函数背后的物理化学或其他科学原理非常复杂,难以计算,但是相关的输入/输出数据对非常丰富,

那么往往可以采用深度学习模型来进行拟合。蛋白质结构领域恰好拥有数据量非常丰富且数据质量相对较高的数据库, 即国际蛋白质结构数据库 (protein data bank, PDB)^[38], 为蛋白质相关深度学习模型的建立奠定了数据基础。目前已经有很多蛋白质结构预测的深度学习模型取得了成功^[39-45], 其中最早取得实质性突破的是由DeepMind公司开发的AlphaFold, 它在2018年的CASP13大赛上取得了优异的成绩^[46]。此后的2020年, 其第二代版本AlphaFold2在CASP14大赛上赢得了冠军, 表现出显著优于传统蛋白质结构预测算法的强大能力^[40]。2024年, DeepMind公司又公布了AlphaFold3, 不仅能够预测单个蛋白质的三维结构, 还能够准确地预测蛋白质-蛋白质、蛋白质-核酸、蛋白质-小分子等不同类型复合物的三维结构^[41]。以AlphaFold系列, 以及RaptorX^[47]、RoseTTAFold^[44]、ESMfold^[48]等算法为代表的深度学习模型, 基本解决了蛋白质结构预测领域的主要问题。虽然在局部准确性、突变敏感性等方面还存在一些挑战, 但这些人工智能算法确实促进了蛋白质相关生命科学问题的解决, 展现了人工智能技术与复杂的蛋白质科学相结合的可能性, 推动了人工智能驱动的蛋白质设计算法的出现。

蛋白质的结构设计常常被误以为是蛋白质结构预测的求逆过程, 然而事实上, 下文将要讨论的基于骨架的序列设计才更像是蛋白质结构预测的逆问题。与直觉相反的是, 从求解未知结构的角度来说, 蛋白质的结构与结构预测是具备极高相似性的问题。两者的差别仅仅在于, 结构设计时可能是无约束的, 也可能给定了某种功能约束 (比如要求与某个给定的靶蛋白能够结合), 但氨基酸序列一定是未知的, 而结构预测时只有一种约束, 即已知氨基酸序列的约束 (图1)。AlphaFold3已经将这种理论构想变成了现实, 采用了深度生成中的扩散模型作为算法框架的核心^[41], 变“预测”为“设计”。不过在AlphaFold等早期的人工智能结构预测算法初步发展起来之后, 研究者们首先想到的是直接利用人工智能算法强大的预测能力来对序列进行迭代修正, 从而实现蛋白质设计。

由Baker团队开发的一种被称为幻想 (hallucination) 设计的蛋白质结构设计方法采用的就是上述思路^[34]。该方法基于的结构预测算法可以是传统的trRosetta (变形限制版的Rosetta), 不仅约束了通常会关注的残基之间的距离关系, 还约

束了从主链肽键到侧链单键的多个转角关系, 能够更准确地预测给定蛋白质序列对应的三维结构^[49], 也可以是基于深度学习的RoseTTAFold或AlphaFold2。以此为基础, 幻想设计使用马尔科夫链蒙特卡洛 (MCMC) 算法生成包含随机突变的蛋白质序列, 并利用trRosetta对这些序列进行结构预测, 获得残基间距离分布。该算法预先针对大量已知蛋白质结构计算了残基间距离的分布作为优化目标, 通过不断优化设计结构的距离分布与目标分布之间的KL散度 (KL-divergence), 使设计结构的约束逐渐逼近真实蛋白质, 从而产生了涵盖广泛序列和预测结构的新蛋白质。结构生物学研究获得了其中三种设计蛋白质的实验结构, 显示与设计结构高度匹配。需要指出的是, 该方法仅设计得到了蛋白质主链的三维结构, 还需要配合其他算法完成序列设计, 因此是典型的结构设计算法。

近日, DeepMind公司宣布开发了一种新的蛋白质设计平台AlphaProteo, 也是基于结构预测算法去完成蛋白质设计^[50]。AlphaProteo首先采用AlphaFold系列算法预测得到了大量蛋白质的结构, 再通过一个过滤器从中筛选得到能够满足设计目标功能的蛋白质。这一过程的逻辑恰好是AlphaFold3的反面, 可谓变“设计”为“预测”。事实上, 从任何一种蛋白质结构预测算法出发, 都可以采用类似的思路来实现蛋白质设计 (图3)。但是很显然, 通过大量的预测计算才能获得期望的设计蛋白质, 一定意味着巨大算力的支持。因此该路径的成本极高, 效率却比较低, 往往不适合科研机构或中小型企业采用。

2.2 基于早期生成模型设计蛋白质结构

生成模型是深度学习模型中发展得较为成功的一个子领域。比如早期的生成对抗网络 (generative adversarial network, GAN) 和变分自编码器 (variational autoencoder, VAE) 都能出色地完成图像的生成任务, 在部分任务领域能够达到以假乱真的生成质量。蛋白质结构的设计问题本质上也是要获得一个符合既有分布的采样, 与生成模型的特性相一致, 天然适合采用生成模型来实现。

Huang团队提出了一种基于GAN的蛋白质设计算法^[51], 根据蛋白质主链上 α 碳原子之间的成对距离来编码蛋白质结构, 这消除了生成模型学习平移和旋转对称性的需要。然后, 引入一个用于对抗噪声的三维结构恢复的凸函数, 以便从生成的成对距离图中折叠蛋白质结构, 并通过交替方向乘子

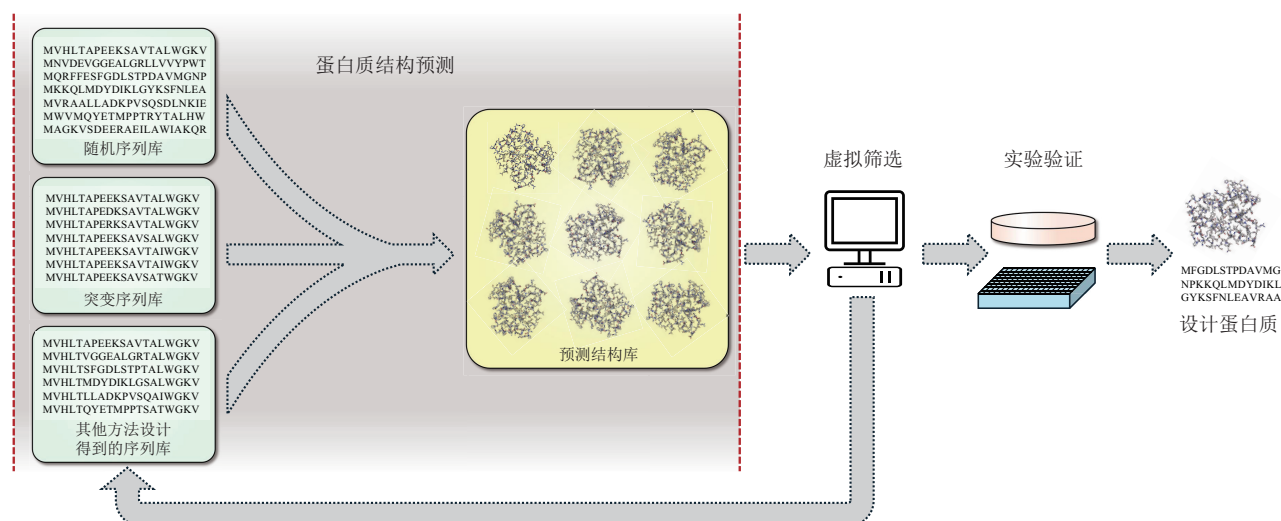


Fig. 3 Protein design based on protein structure prediction

图3 通过蛋白质结构预测来实现蛋白质设计

法 (ADMM) 解决这些问题, 最后通过预测损坏的蛋白质结构来测试模型的有效性。随后, Huang 等进一步优化了该模型, 从真实的蛋白质三维骨架片段分布中进行采样, 用所有骨架原子之间的成对距离来表示蛋白质结构, 并提出一种快速、完全可微分方式直接恢复和细化相应骨架坐标的方法, 提升了其生成结构的性能^[52]。2023年 Huang 等^[9]又提出了一种构建蛋白质骨架的新模型 Ig-VAE, 这是第一个直接生成蛋白质骨架全部原子三维坐标的架构。模型的新颖之处在于它提供了一个旋转和平移不变的损失函数, 同时还具有扭转感知和直接生成三维坐标的能力, 直接生成免疫球蛋白的三维骨架。重要的是, 该模型提出了一种概念上全新的解决蛋白质设计问题的方法, 由于 Ig-VAE 直接生成坐标, 它的所有坐标输出是完全可微分的, 因此就可以通过任何可微分的基于坐标的启发式方法来约束结构生成, 这就让设计者可以指定任何期望的结构特征, 约束模型生成符合该特征的蛋白质分子的全部原子坐标。

2.3 基于扩散模型设计蛋白质结构

在所有深度生成模型当中, 去噪扩散概率模型 (denoising diffusion probabilistic models, DDPM)^[53] 是在图像生成领域表现最优异的人工智能模型, 以之为基础的图像生成程序已经被人们广泛使用。DDPM 首先向有信息的图像中逐步添加噪声, 再通过神经网络来学习逆向的去噪函数, 最终就可以对任一不含特定信息的高斯噪声图像进行

去噪, 从而获得含有信息的图像采样。然而, 传统的扩散模型整体是为二维图像问题设计的, 要用它解决蛋白质结构的生成问题, 就需要从图像的二维空间转换到结构的三维空间。另外, DDPM 只是广义上的扩散模型 (diffusion model) 中的一种, 效率较低。近几年来已经出现了其他一些新的扩散模型, 比如去噪扩散隐式模型 (denoising diffusion implicit models, DDIM) 等, 甚至联合 Transformer 构建的扩散 Transformer (diffusion transformers, DiT) 和可扩展插值 Transformer (scalable interpolant transformers, SiT) 等, 可以在面对蛋白质这类复杂问题时发挥更大的作用。

2022年, 清华大学马剑竹团队开发了一种针对抗体蛋白的互补决定区 (complementarity determining regions, CDR) 同时进行序列和结构设计的扩散概率模型, 也是最早一批应用扩散模型解决蛋白质结构设计问题的算法工具之一^[10]。该模型基于和等变神经网络, 针对给定的抗原表位, 联合建模抗体的序列和结构。它能够完成多种不同类型的任务, 包括序列-结构协同设计、给定主链结构的序列设计, 以及抗体优化。应用该模型针对包括新冠病毒受体结合结构域在内的几类不同抗原进行了抗体设计后, 相关的能量计算、序列比对等评估结果显示该模型的性能优于此前的其他设计算法。

中国科学技术大学刘海燕教授团队^[8] 在此前开发的基于能量计算的蛋白质设计算法 SCUBA^[54]

的基础上, 引入了扩散模型, 形成了蛋白质设计算法 SCUBA-D。在该算法中, 蛋白质的结构骨架被表征为残基间距离矩阵, 与 AlphaFold 所采用的策略一致。因此 SCUBA-D 的扩散和去噪均发生在残基间距离的二维矩阵上, 仍是经典的二维扩散模型。比较有特色的是, SCUBA-D 将整体架构分为顺次衔接的三个模块, 分别是一步去噪的低分辨率骨架生成模块, 融合了语言模型和扩散模型的高分辨率骨架生成模块, 以及类似于 GAN 结构的多重物理判别模块, 用以保证生成结构的合理性。结果表明, 该模型可以始终生成高质量的蛋白质骨架结构, 具有十分广阔的应用前景。

同样在 2022 年, Baker 团队也推出了基于扩散模型的蛋白质设计算法 RFdiffusion^[7]。该算法基于 Baker 团队于 2021 年发布的蛋白质结构预测算法 RoseTTAFold^[44]。与 AlphaFold2 主要依赖于多序列比对 (MSA) 和残基间距离矩阵不同, RoseTTAFold 还引入了其他信息, 包括残基的朝向等。因此, RFdiffusion 几乎直接复用了 RoseTTAFold 已经训练好的模型来构造去噪函数, 也就是在有序列信息的情况下, 采用当前步骤的残基距离矩阵和三维模板来预测结构。该预测结构会被下一步去噪过程视为三维模板使用 (称为“自我条件约束”, self-conditioning), 但不会被直接用于产生下一步的残基距离矩阵, 而是被视为去噪的目标来产生下一步的残基坐标。RFdiffusion 既可以通过指定肽链长度来自由生成蛋白质结构骨架, 也可以通过损失函数的重写和重新训练, 实现特定对称性或特定功能的蛋白质的骨架设计。但是后一种方式对非人工智能领域的研究者来说具有较高的使用门槛。在 RFdiffusion 的基础上, Baker 团队又于 2024 年推出了 RFdiffusion-All-Atom^[55], 能够更好地解决蛋白质与核酸等多种不同类型其他分子的共同设计问题。

DDPM 在处理二维图像时存在对图像分辨率的限制, 即对处理的二维矩阵大小的限制。因此, SCUBA-D 和 RFdiffusion 也都存在大小限制的问题, 仅能直接处理几百个氨基酸残基的蛋白质的设计问题。如果超过这一限制, 就需要通过拼接等方法进行弥补。2023 年, Generate Biomedicines 公司发布了蛋白质设计算法 Chroma^[11], 一种基于图神经网络 (graph neural network, GNN) 的扩散模型, 能够解决无长度限制的蛋白质设计问题。这个 GNN 模块生成的蛋白质骨架仍旧比较粗糙, 还需

要通过扩散模型的低温采样 (low-temperature sampling) 方法来提高骨架质量。不同于 SCUBA-D 和 RFdiffusion 调用其他算法来完成序列设计问题, Chroma 自身带有一个序列和侧链设计模块, 且基于与主模块相似的 GNN 模型, 但是添加了条件约束的序列解码层和侧链解码层。

目前, 扩散模型越来越多地出现在蛋白质结构的设计算法当中。笔者团队也开发了一种基于 DDIM 的蛋白质设计算法, 在电子密度层面上完成蛋白质结构的加噪与去噪过程, 实现了弱监督的蛋白质快速设计 (未发表)。笔者团队的另一项工作中, 采用了流匹配的扩散模型框架, 在生成结合蛋白方面有极高的效率 (未发表)。如前所述, DeepMind 公司近期发表的 AlphaFold3 也是以扩散模型为核心来完成预测结构的生成。这些成果都说明, 扩散模型或许在理解三维空间方面具备其他人工智能模型所不具备的优势, 因此会成为未来蛋白质设计算法的主流模型框架。

3 无骨架的序列设计

蛋白质分子是由一条或多条肽链折叠而成的, 而肽链是一维线性的氨基酸残基链。因此, 蛋白质天然地可以通过其一维氨基酸序列进行表征。考虑到“序列决定结构”的蛋白质折叠理论^[56-58], 可以认为与功能密切相关的三维结构信息就隐含在蛋白质的氨基酸序列当中。因此, 蛋白质设计任务也可以通过人工智能模型对于氨基酸序列数据的学习和生成来完成, 也即“序列设计”。为了区别于已知主链骨架结构时的序列设计, 此处的序列生成方法定义为“无骨架的序列设计”。通常, 在此类算法中基本不考虑结构信息的因素, 因此其算法实现与结构设计算法有着显著的差异。

另外有必要指出的是, 无骨架的序列设计只产生了氨基酸序列信息, 并不直接产生结构信息, 这对于后续的筛选、分析、应用往往是不利的。因此, 后续需要使用 AlphaFold 等结构预测算法对生成的氨基酸序列进行结构预测, 才能获得设计蛋白质的结构信息 (图 1)。一方面, 这种方法引入了一个新的带有不确定性的预测算法, 进一步加大了最终结构信息的不确定性; 另一方面, 受限于当前结构预测算法较慢的运算速度, 无法对快速大量生成的氨基酸序列进行高通量的结构预测, 因而制约了整个技术流程的通量和速度。

3.1 基于早期生成模型设计蛋白质序列

“人类基因组计划”的推进极大地促进了基因组测序技术的发展,令基因信息的数据量急速增长,相应的蛋白质序列数据库也迅速扩大,为人工智能模型学习蛋白质序列规律做好了数据上的准备^[59]。

2021年,Hawkins-Hooker等^[12]在近70 000个与荧光素酶类似的氧化还原酶的氨基酸序列数据集上训练了两种VAE模型,分别处理对齐序列输入(MSA-VAE)和原始序列输入(AR-VAE)。研究表明,虽然两者都能够重现该酶家族使用氨基酸的特征模式,但MSA-VAE能更好地捕获反映三维结构影响的长距离依赖性。他们通过上述模型生成了细菌荧光素酶*luxA*的变体,并通过实验检测到了发光活性,证实了该模型的实用性。

同年,Repecka等^[13]训练了一种GAN模型来生成功能性蛋白质序列,称为ProteinGAN。该模型是一种基于“自注意力”(self-attention)机制的生成对抗网络变体,能够从复杂的多维氨基酸序列空间中学习蛋白质序列的进化关系,并创建具有类似自然物理特性的高度多样化的新序列变体。该团队以苹果酸脱氢酶的氨基酸序列为模板,使用ProteinGAN生成了55条序列,其中13条获得可溶表达,并展现了苹果酸脱氢酶的催化活性。ProteinGAN展示了人工智能在序列空间允许的生物学限制内快速生成高度多样化的功能蛋白质的潜力。

需要指出的是,上述无骨架的序列设计方法有很大的局限性。如果待设计的目标蛋白质并不具备大量的已知同源序列信息,就不足以训练一个高质量的深度学习模型,也就无法实现目标蛋白质序列的有效生成。

3.2 基于Transformer设计蛋白质序列

2022年,由OpenAI公司发布的ChatGPT引发了公众对于新一轮人工智能发展浪潮的广泛关注,同时也展现了以Transformer^[60]为基础的语言模型(language model)作为一种深度生成模型架构的强大能力。应用Transformer解决蛋白质序列的设计问题有着天然的优势:一方面,氨基酸序列可以表征为20种字母组成的字符串,直接就可以纳入语言模型的数据体系;另一方面,Transformer的成功在很大程度上源于所谓的“多头注意力”(multi-head attention)机制,能够在文本的远端之间建立联系,而这与蛋白质折叠后令序列上相距较远的氨

基酸残基之间产生直接相互作用非常类似,使得Transformer有可能学习到此类关联性。

2022年,Moffat等^[17]提出了DARK架构,用于在不断迭代扩展的合成蛋白质序列数据集上有效地训练生成模型,该模型使用了标准的Transformer解码器架构,可生成具有不同有序结构的新序列。随后,Ferruz等^[14]提出了ProtGPT2模型,该模型是一个自回归Transformer模型,拥有7.38亿参数,使用Uniref-50作为训练数据集。训练完成后生成的蛋白质序列,显示出与天然蛋白质序列相似的预测稳定性与动态特性,同时在进化上又与当前的蛋白质序列空间相距甚远。

2023年,Profluent公司提出了一种应用于蛋白质序列数据的条件Transformer语言模型,称为ProGen^[15]。该模型接受了超过19 000个蛋白质家族的2.8亿个蛋白质序列的训练,并通过指定蛋白质特性的控制标签进行了增强。ProGen生成的蛋白质在能量上与天然蛋白质相近,具有理想的生物学功能。而后Profluent公司又提出了ProGen2,一种参数量高达64亿的自回归Transformer模型^[16]。该模型的训练数据集是由从基因组、宏基因组和免疫蛋白数据库中提取的超过10亿种蛋白质序列组成的。为了评估ProGen2生成蛋白质序列的能力,研究者选择在以下三种情境中对模型进行生成,即:预训练后一般序列的生成,微调后的可以折叠成特殊结构的序列生成,以及在抗体序列数据集上进行预训练后的抗体序列生成。结果表明,ProGen2与在之前公布的所有模型相比,在生成合理序列方面有着最佳表现。

虽然使用了Transformer这种当前最强大的人工智能模型架构,但是仅仅针对序列进行设计的方法在一些设计任务上的表现仍旧差强人意。与结构设计算法相比,直接设计序列的优势是训练数据量更充足(数亿与22万)。但是由于未纳入结构信息,而蛋白质的结构又与其功能直接相关,因此等同于加重了模型的学习负担,令纯粹的序列设计模型变得臃肿而低效。

4 基于骨架的序列设计

“基于骨架的序列设计”与前述“无骨架的蛋白质设计”有着不同的内涵。在此场景下,由于待设计蛋白质的骨架结构是已知的,所以设计目标是要获得一条氨基酸序列,使其自然折叠之后的骨架与给定骨架高度吻合。在某种意义上,上述过程恰

恰是蛋白质结构预测的逆问题。在蛋白质设计实践中, 基于骨架的序列设计往往要与 SCUBA-D、RFdiffusion 等蛋白质骨架设计算法配合使用, 为这些算法生成的纯甘氨酸或纯丙氨酸骨架匹配可折叠的序列 (图1)。

中国科学院计算技术研究所的卜东波团队开发了一种精确高效的蛋白质序列设计算法 ProDESIGN-LE^[21]。它采用了一种简洁的残基局部环境表示法, 用残基类型而非具体的原子来描述目标残基的化学环境, 但通过对残基朝向等条件的约束, 能够获得准确的环境关系。在此基础上, ProDESIGN-LE 训练了一个转换器来学习残基局部环境与其氨基酸类型之间的相关性。对于目标骨架结构, ProDESIGN-LE 通过对序列的每个位置进行迭代更新获得最终的设计序列, 而初始序列既可以是纯甘氨酸序列, 也可以是任意的随机序列。对设计序列的结构预测显示, ProDESIGN-LE 设计的序列能够高度还原其设计目标的骨架结构。有趣的是, 如果设计目标为天然蛋白质时, ProDESIGN-LE 并不寻求高度复原目标骨架的原始序列, 因此对于设计出差异化的氨基酸序列有着非常强的实用性。

为了配合其 RFdiffusion 等算法设计的蛋白质骨架, Baker 团队也开发了一种序列设计算法 ProteinMPNN^[18]。该算法参考了 GraphTrans 的架构, 使用具有 3 个编码器层和 3 个解码器层消息传递网络 (message passing neural network, MPNN), 每层宽度为 128 位。该网络使用蛋白质骨架的结构特征作为输入, 以自回归方式从 N 端到 C 端预测蛋白质序列。如果设计目标为天然蛋白质, ProteinMPNN 能够较好地复原目标骨架的原始氨基酸序列, 在 15 760 个蛋白质的测试中复原准确率达到 49.0%。ProteinMPNN 的训练数据中添加了高斯噪声以避免模型学到结构修正过程中引入的误差, 从而提高了模型的稳定性, 同时也增强了模型的泛化能力。除了从 N 端到 C 端逐一预测的方式外, ProteinMPNN 还可以采用与顺序无关的自回归模式, 从而让该算法的应用场景进一步拓宽, 可以在固定一部分序列的情况下设计其他部分的序列, 也可以解决蛋白质-蛋白质复合物、环状蛋白质等复杂构型的序列设计问题。

在 ProteinMPNN 的基础上, Baker 团队又开发了带有原子环境约束的蛋白质序列设计算法 LigandMPNN^[20]。为了表征与蛋白质结合在一起

的其他配体分子, 比如小分子、核酸、金属离子等, 该算法引入了三个图神经网络, 其中配体图用于表征配体分子的所有原子之间的连接关系, 蛋白质图用于表征蛋白质所有残基的主链原子之间的连接关系, 而蛋白质-配体图则用于表征蛋白质所有主链原子之中与配体距离最近的 25 个与配体的连接关系。此外, LigandMPNN 还引入了一个神经网络专门解决侧链装配问题, 因此不仅能给出残基序列, 还能给出每个残基侧链的具体构象。对比测试表明, 在结合有小分子、核酸、金属离子的三个场景下, LigandMPNN 对配体附近的氨基酸序列的复原率都显著超过了 ProteinMPNN (小分子 63.3% 对 50.4%, 核酸 50.5% 对 34.0%, 金属离子 77.5% 对 40.6%)。显然, LigandMPNN 在药物发现、酶的改性设计等领域有较多的应用场景, 能够帮助研究人员快速筛选和设计具有特定活性的化合物或蛋白质。

2024 年, 中国科学院计算技术研究所的张海仓团队开发了一种基于蛋白质语言模型的序列设计算法 CarbonDesign^[22]。该模型采用了端到端的网络循环技术, 并可以在设计序列的同时生成侧链结构, 从而避免了再调用 AlphaFold 等方法来获得预测结构的麻烦。CarbonDesign 的核心是 Inverseformer, 可以视为 AlphaFold2 中的 Evoformer 的逆向架构, 让信息从三维结构流向一维序列。在此基础上, CarbonDesign 使用了马尔可夫随机场来构建氨基酸序列的概率模型, 就可以在生成过程中对氨基酸类型做出选择, 输出设计序列。此外, CarbonDesign 在循环中采用了蛋白质语言模型 ESM2 或 ESM-3B 的进化约束。在 CAMEO 测试集和 CASP15 测试集上的对比测试表明, CarbonDesign 的序列恢复率等关键指标优于 ProteinMPNN 等其他主要的序列设计算法。CarbonDesign 的成功恰恰说明, 蛋白质序列设计和蛋白质三维结构预测是相互关联的逆问题, 可以在类似的架构上逆向实现。

相对而言, 基于骨架的蛋白质序列设计是蛋白质设计当中发展得比较成熟的一个领域, 已经能够得到高精度、可还原的氨基酸序列, 以解决给定主链骨架的折叠问题, 甚至还可以在设计中考考虑小分子、金属离子、核酸等配体的结合问题。造成这一现象的原因或许正如前文所述, 蛋白质序列设计更接近于一种“预测”问题, 是在已知主链骨架的前提下预测对应的氨基酸序列。而蛋白质结构的设计

等问题则没有太多的“已知条件”可用，难度自然也要大得多。

5 其他辅助设计算法

5.1 提升酶活性的深度学习算法

任何深度学习模型的成功建立都有赖于优质、数据量丰富、分布偏差不显著的数据库。如果说PDB中的每一套蛋白质结构数据都呈现了蛋白质折叠的一般性规律，那么酶活性的数据库恰恰难以呈现关于酶催化的一般性规律^[61]。一方面，酶活性数据来自于一些独立的实验研究，实验体系不同，缺乏对数据的评估手段，因此这些实验数据之间很难建立一致性，常常对于同一个酶会出现不同的实验结果。另一方面，酶活性数据的实验通量小，数据集中的全部数据往往只覆盖了有限的若干酶类，难以让模型产生泛化能力。最后，酶活性实验往往是各个研究组针对某个酶的研究中获得的，因此往往带有大量突变体的活性数据，令整个数据集呈现严重的分布偏差。为了解决数据集方面的问题，一个可能的方案就是利用人工智能算法进行学习和整理。浙江大学谢昌谕团队开发了基于多模态交叉注意力机制的深度学习模型EasIFA，融合了来自蛋白质语言模型的隐空间表征和蛋白质三维结构，以及关于酶催化功能的信息，最终实现对任一酶的催化位点注释^[62]。利用这一工具，研究者可以将庞杂的不同来源的酶数据库整合精简为规模更小、信息更准确的酶数据库。

酶的催化能力有多方面的指标，其中催化常数(catalytic constant, 又称周转数, turnover number)是非常重要的一个指标。2022年, Li等^[63]建立了一种基于深度学习的催化常数预测模型DLKcat。该模型在一个重新整合的催化数据集上进行了训练, 其中包括来自851个物种的7 822种酶的催化数据, 底物分子种类多达2 672种, 覆盖面相对较广。DLKcat对于酶的氨基酸突变比较敏感, 能够预测突变对于酶催化常数的影响, 并能够用于寻找对催化常数影响最显著的氨基酸残基位点。在此基础上, 研究者建立了一个基因组规模上的酶约束代谢模型(enzyme-constrained genome scale metabolic model), 能够比传统的同类模型更加准确地预测酵母等微生物的代谢行为。2024年, 上海交通大学的鲁洪中副教授团队进一步开发了一种基于蛋白质结构信息的酶催化常数预测算法DeepEnzyme, 整合了Transformer和图神经网络两

种架构^[64]。相比于此前的同类算法, DeepEnzyme对于与训练数据集序列同源性较低的蛋白质(低至27%)仍展现了较强的预测鲁棒性, 说明模型有更好的泛化能力。利用该算法, 研究者对PafA(phosphate-irrepressible alkaline phosphatase of *Flavobacterium*)和人源的P00558两种酶进行了分析, 成功鉴定到了与催化紧密相关的残基位点。

虽然已经有一些应用于酶学领域的深度学习模型, 但是并不存在一种全自动的算法, 能够直接设计出具备某种给定催化功能的酶。目前来看, 对于酶的设计可能还是要采取“各个击破”的策略^[4]。每一种酶进行单独研究, 甚至为某一类酶专门建立数据库并训练专门的模型^[65]。对于更一般性的情况, 特别是在自然界不存在现有酶类可以当作参考模板的新型反应的催化, 则需要极其复杂的技术方案, 综合运用本文中介绍的不同类型的设计算法。下文关于智能蛋白质工程当前应用现状的部分对于酶的实际设计案例进行了介绍。

5.2 提升热稳定性的深度学习算法

蛋白质的热稳定性问题对于蛋白质产品的实际应用起着重要的作用。在生物工程领域工业化的反应条件下, 发热问题严重, 因此业界更偏好具有高热稳定性的酶分子; 在蛋白质药物领域, 抗体等蛋白质分子的热稳定性较差, 需要全程冷链储运, 大大推高了药物的成本, 且限制了药物在偏远地区等场景下的应用。因此, 在不削弱给定蛋白质功能的前提下提高其热稳定性, 一直都是蛋白质工程领域的强烈需求。

值得注意的是, 通过各种深度学习算法设计的蛋白质分子往往天然就具有较高的热稳定性^[11, 34, 54, 66], 去构温度达到60~90°C, 高于自然界的蛋白质。关于这一现象的成因还缺乏系统性的研究, 但已经有一些理论推测。一种可能的原因来自于深度学习模型本身。由于这些模型关注于强化“折叠”, 因此其生成结果更加追求最优的残基侧链堆叠方式, 从而造成了设计蛋白质分子的整体结构更加紧密, 必然就有着更高的热稳定性。另一种可能的原因则与训练数据有关。一方面, 蛋白质结构的训练数据都来源于PDB, 其中大部分是由X射线晶体学方法获得的, 展现了处于晶体堆积中的受限构象状态, 能量相对更低, 折叠相对更紧密。用这样的数据训练出来的深度学习模型必然更倾向于生成同样堆积紧密的蛋白质结构。另一方面, 几乎所有当前的蛋白质设计深度学习算法都只学习了

PDB 结构文件中的蛋白质分子坐标, 没有学习其中水分子的结构数据。然而天然蛋白质的折叠有时候与水分子形成的氢键或氢桥有关, 所以就具备了一定的柔性, 可以在水分子动态脱离时获得更为自由的构象状态, 从而加大了失去整体构象的机率。蛋白质设计的深度学习模型由于没有学习到水分子的结构特征, 也就不会考虑水分子介导的柔性折叠形式, 因此会比天然蛋白质更为稳定。当然, 上述推测还需要更多数据的验证, 以及更为系统性的深入研究。

关于蛋白质热稳定性的预测算法发展得很快, 从早期应用简单的卷积神经网络, 到近一两年采用复杂的深度学习模型, 都取得了一定的成功。在热稳定性预测算法的构建中, 主要的制约环节是训练数据的不足。一个主要被使用的数据集是 ProThermDB^[67], 收集有 32 000 条关于热稳定性的不同类型数据, 其中去构温度相关的一些子数据集常被用于热稳定性预测模型的训练和横向比较, 比如 S1615 和 S388。华中科技大学的徐莉研究组在此基础上建立了一个更大的子数据集 S3661, 并利用它训练了一个基于卷积神经网络的热稳定性预测算法, 在标准数据集上的预测准确度提升到了 75%^[68]。但是, 由于 ProThermDB 中相当一部分数据是突变体数据, 因此其实际对应的蛋白质数量很有限。比如 S1615 数据集有 1 615 条去构温度数据, 但只对应于 42 种蛋白质和它们的众多单点突变体, 而 S3661 数据集则对应于 178 种不同的蛋白质及其突变体。当蛋白质的数量如此有限时, 训练得到的模型往往不具备很好的泛化能力, 因此对于其他蛋白质的预测表现不佳。

另一个重要的热稳定性数据集出现在 2020 年, Jarzab 等^[69]巧妙地借助蛋白质组的方法建立了一个融化组 (Meltome) 数据集, 包含了来自 13 个物种的 48 000 种蛋白质的去构温度数据。虽然这些蛋白质并没有直接进行去构温度的实验测定, 但是通过对细胞裂解物的整体加温, 并去除不溶物质, 便可通过蛋白质组的质谱方法确定可溶成分中尚存的蛋白质种类。也就是说, 这些存留下来的蛋白质在该温度下仍可保持稳定构象。接下来, 通过梯度抬高温度并重复上述过程, 就可以确定所有蛋白质组分去构温度。2024 年, 北京化工大学的朱玉山研究组从上述融化组中选取了 7 790 个蛋白质的数据建立训练和校验数据集, 在此基础上训练了一个基于图神经网络的热稳定性预测算法 DeepTM^[70]。

该方法在测试数据集上表现优异, 预测去构温度的均方根差为 6.24°C; 在另外两个不同来源的独立数据集上的预测均方根差也均小于 8°C, 相对其他已有预测算法都有 50% 以上的准确度提升。

通过蛋白质热稳定性预测可以有效地帮助设计者预判蛋白质突变体的热稳定性, 从而加快找到更佳突变体的过程, 但仍要配合突变体库的构建和筛选, 有较大的工作量, 并且有着极大的不确定性。更为有效的方法是直接设计能够获得更高热稳定性的突变体, 这就需要采用深度生成模型^[71]。

5.3 密码子及启动子的设计算法

除了针对蛋白质本身的设计与改造之外, 人工智能算法也在其他环节上促进了蛋白质工程的发展, 其中一个经典的问题是蛋白质的密码子设计问题。由于 64 种密码子对应于 20 种基本氨基酸和终止信号, 因此存在同义密码子的现象, 即多个密码子对应于同一种氨基酸。同义密码子不影响从基因到蛋白质的翻译过程的确切性, 但是对于设计蛋白质只有氨基酸序列的情况, 就需要恰当地选择同义密码子来构成对应的核酸序列, 以保证设计蛋白质在细胞内能够高效表达并正确折叠。

就传统的异源重组蛋白而言, 密码子的选择通常被称为密码子优化。就设计蛋白质而言, 不存在其原始的可参考的基因序列, 因此也就不存在“优化”问题, 其密码子需要从头设计。虽然两种技术的名称有所差异, 但是它们的本质是相同的, 都是要为给定的氨基酸序列匹配一条最适应表达细胞体系的密码子序列, 以适应该细胞中的 tRNA 丰度差异等表达调控机制。然而, 由于目前对同义密码子的选用规律仍旧缺乏准确的认知, 因此现有密码子优化算法往往只是选择对应 tRNA 丰度最高的密码子, 同时兼顾核酸内切酶识别序列等一些次要因素。重组蛋白表达的实践证明, 这种追求高丰度和快速表达的策略并不具有必然的有效性, 在相当多的情况下都无法获得目标蛋白的可溶表达。

随着密码子优化问题被广泛研究, 一些研究者将机器学习的算法应用于密码子的设计问题, 例如基于遗传算法的 GASCO^[72], 结合最小二乘回归和遗传算法的建立模型预测密码子与表达量的关系^[73], 针对特定物种 (红球菌) 的密码子优化方法^[74], 利用混合整数线性规划 (MILP) 构建考虑多个优化目标的密码子优化模型等^[75]。2020 年, Fu 等^[76]引入密码子盒的概念, 并利用双向长短期记忆条件随机场训练了大肠杆菌密码子盒的优化模

型。该模型在两种蛋白质的表达实验中表现出了优于传统密码子优化算法的效果,能够显著提升这两种蛋白质在大肠杆菌中的表达量。不过,该模型的训练数据集仅包含4 906个基因,且最终在训练集和测试集上的精度仅达到0.76和0.52,仍有很大的优化空间。Jain等^[77]在2023年利用循环神经网络(recurrent neural network, RNN)实现了密码子优化,并证明RNN通过对序列上下文的深刻理解能够产生与宿主基因组更相似的密码子选择。

笔者团队也开发了一种基于Transformer的密码子设计算法NatCodon。该算法的训练数据中不仅包含了氨基酸序列与核酸序列的对应关系,还引入了蛋白质的结构信息,以帮助模型理解不同丰度tRNA对应的密码子对于蛋白质折叠过程的影响。重组表达实验结果表明,该算法可以有效提高异源蛋白质在大肠杆菌细胞、人细胞、植物细胞等表达体系内的可溶表达效率,甚至达到原始得率的5倍以上(未发表)。

除密码子的选择之外,其他可能影响表达的因素也可以在人工智能算法的辅助下进行优化。清华大学汪小我团队开发了一种结合专家知识与生成对抗网络的算法DeepSEED,可以设计优化待表达基因的启动子,使目标基因获得更高效的转录,进而提高表达效率^[78]。另一种思路是对于mRNA的结构进行优化,因为研究表明,适当的mRNA结构可以提高其翻译效率^[79-80]。日本庆应义塾大学佐藤健吾团队开发了一种RNA二级结构预测算法MXfold2^[81],结合了基于神经网络的折叠评分预测与基于理论的自由能计算,能够比较准确地预测mRNA的二级结构,从而为改善表达效率指导方向。

6 智能蛋白质工程的应用现状

在过去的几十年间,蛋白质工程在医药、农业、工业和环境保护等领域的应用一直在扩大。研究显示,2030年全球蛋白质工程市场规模预计将达到76.2亿美元^[82]。相比之下,智能蛋白质工程的发展才刚刚起步,取得的实践成果仍十分有限,也没有形成显著的经济影响力。但不可否认的是,已经取得的智能蛋白质工程的成果大多突破了人类已有的认知,展现了人工智能算法远远超越传统蛋白质工程技术和人类智慧的设计能力。

酶的设计始终是蛋白质设计领域的一个难点。一方面,对于酶的催化机制与其结构动态性之间的

耦合关系还缺乏深入的认知;另一方面,不同的酶与酶之间往往存在着本质性的差异,很难放在同一个模型的框架下进行学习和采样。前述关于*luxA*细菌荧光素酶^[12]和苹果酸脱氢酶^[13]的设计工作都需要关于同一功能的酶类的大量蛋白质序列信息用于模型训练。但遗憾的是,并非所有酶都具有大量已知的同功酶序列,因此难以推广类似的方法。2023年,Baker团队采用结构设计方法实现了一种新型荧光素酶的从头设计^[83]。首先,研究人员选择了非天然的二苯基特拉嗪(diphenylterrazine, DTZ)作为设计荧光素酶的底物,使用量子化学等计算方法围绕该分子的阴离子构建了构象系统,然后围绕每个构象枚举可能的相互作用氨基酸侧链旋转异构体,并与约4 000个天然蛋白质骨架的中心腔进行对接。他们在对接结果中发现,与荧光素完全无关的一种核转运因子2(nuclear transport factor 2, NTF2)家族蛋白或许是最优选择。在此基础上,他们采用了一种基于深度学习的“family-wide hallucination”方法,集成了前文所述的无约束幻想设计^[34, 84]与Rosetta序列设计方法^[85],最终获得了1 615种骨架蛋白。接下来,研究者使用了传统的蛋白质工程方法进行功能筛选,以及关键残基位点的饱和突变,最终获得了具有高底物选择性和高亮度的设计荧光素酶,可与天然荧光素酶相媲美。Baker团队的这项工作是酶设计的革命性突破,也是智能蛋白质工程的一个里程碑,虽然人工智能算法在其中的贡献仍比较有限,但是它所实现的任务,即以完全不同的蛋白质架构来实现一种酶的功能,在此前的蛋白质工程领域是不可想象的挑战。以此为起点,通过综合运用智能蛋白质工程中的多种技术来实现任意给定催化功能的酶的设计,或许是不远的将来就可以达成的目标。

单克隆抗体药物是近年来重组蛋白质市场快速发展的重要推动力,也是当前蛋白质药物市场的明星,在癌症等多种严重疾病的治疗中取得了传统小分子药物无法替代的优势地位。虽然抗体依靠传统蛋白质工程技术可以进行改良,但是其结构表位特异性、体内稳定性等问题往往不能得到很好的解决,一直存在对于计算生物学方法的需求^[86]。虽然目前抗体库的深度测序与相关功能数据的结合,已经可以用于训练基于机器学习的新型模型,以实现亲和力成熟、人源化和可开发性^[87-88],但是针对给定靶蛋白设计或优化抗体仍旧是一个巨大的挑战,综合运用本文前述的人工智能设计算法也并未

取得显著的成果。这主要是由于两方面的原因:

a. 抗体的数据仍然不足, 特别是序列与功能的对应数据严重不足, 并且存在由大量的新型冠状病毒研究所导致的数据构成偏差, 即来自新冠病毒刺突蛋白受体结合结构域的抗体在数据中占比过高; b. 抗体结合抗原的区域具有高度的结构柔性, 两者的紧密结合往往是诱导变构的结果, 而结构的动态性是当前绝大多数蛋白质结构设计算法都无法纳入考量的因素。毫无疑问, 无论是基于结构还是基于序列的抗体人工智能设计算法, 都将是智能蛋白质工程领域未来的重点发展方向之一。

虽然抗体的设计仍然很困难, 但是其他构型的结合蛋白的设计问题却有望在人工智能技术的推动下得以解决, 特别是大小在 50 个残基左右的“微蛋白”(miniprotein), 或是能够形成单根螺旋的多肽, 以及由多根螺旋反平行构成的螺旋束蛋白。螺旋作为蛋白质中最稳定的一种二级结构单元, 广泛出现在多肽和微蛋白中。在人体内, 许多肽激素就是以 α 螺旋的构型与其受体结合的, 比如司美格鲁肽和替尔泊肽等药物模拟的对象 GLP-1^[89-92]。2023 年, Baker 团队采用前文所述的 RFdiffusion^[7] 实现对柔性靶标的结合物设计, 并通过连续的加噪和去噪来优化输入结构模型, 最终生成了对几种不同螺旋肽都具有皮摩尔量级的极高亲和力和特异性结合蛋白。这类蛋白质将能够作为此类单螺旋肽激素的传感器, 实现这些激素在临床上的高灵敏度检测。在微蛋白的设计方面, Baker 团队最早在 2020 年曾经采用传统的计算设计方法获得了新冠病毒刺突蛋白受体结合结构域的结合微蛋白, 能够取代中和抗体的作用, 显示了微蛋白可能成为蛋白药物的一股新势力^[93]。2024 年, Baker 团队综合采用类似的策略, 实现了 IL-23R 和 IL-17A 的微结合蛋白设计, 并达到了皮摩尔数量级的强结合力^[94]。进一步的实验证明, 这些微蛋白可以通过口服给药在小鼠的血液中达到足够的药物浓度, 从而为治疗 Th17 介导的炎症提供了全新的药物思路。2024 年, Baker 团队还成功设计了能够探测不同小分子的伪环状蛋白质(pseudocycle)^[95]。这种蛋白质由多次重复的结构单元构成, 在不结合小分子时具有一定的结构稳定性, 而小分子的加入能够促使重复单元以小分子为中心围拢成为接近环状的结构状态, 即可作为对应小分子的传感器使用。当前, 结合蛋白的设计在应用模式上还是以抗体为目标的。抗体的一种重要应用是中和蛇等动物产生的毒素蛋白。于是,

Baker 团队利用前述 RFdiffusion 等方法成功设计了三种蛇毒蛋白的结合蛋白, 能够起到对蛇毒的中和作用^[66]。设计过程中通过指定二级结构特征, 实现了结合蛋白与蛇毒蛋白的高特异性结合, 且设计蛋白质自身的热稳定性极佳, 去构温度在 60~90°C, 远远优于中和性抗体, 有着更广泛的应用场景。

7 智能蛋白质工程未来的重大挑战

随着蛋白质类和多肽类药物的兴起, 以及合成生物学的飞速发展, 特定功能蛋白质的设计、改造、生产已经成为了生物医药领域的热点。近年来人工智能的兴起, 让传统的蛋白质工程进入了“智能蛋白质工程”的时代。虽然该领域的重要成果大多出自国际上少数研究团队, 但是已经有越来越多的研究人员在尝试采取类似的方法来解决问题, 甚至开发采取不同思路与架构的全新人工智能方法。通过采用这些创新方法, 有望解决蛋白质设计当中序列与结构相融合的挑战, 以及主链与侧链一体化设计的挑战。除了这两方面的问题, 蛋白质设计中还存在其他一些重大挑战需要去解决。

首先就是除蛋白质之外的其他类型分子能否被纳入设计之中。虽然 RFdiffusion 等算法已经能够处理蛋白质与核酸分子、金属离子或其他小分子的复合物设计问题, 但这还远远不够。一种基本的需要在设计中予以考虑的分子就是水。结构生物学, 特别是 X 射线晶体学几十年来的研究已经表明, 水分子参与到了蛋白质的折叠过程中, 并直接参与了蛋白质之间的相互作用。能够考虑水分子的设计问题, 就能够设计出更加自然的蛋白质相互作用, 有利于获得功能性更好的结合蛋白。除了水分子之外, 蛋白质中的非天然氨基酸, 以及氨基酸上发生的化学修饰, 都意味着氨基酸的化学组成发生了变化, 因而是当前的设计算法无法处理的。比如以氨基酸序列来表征蛋白质的方式, 就无法跳出 20 种基本氨基酸之外来完成设计。但是从蛋白质产业的角度来看, 无论是非天然氨基酸还是蛋白质化学修饰, 都已经出现在蛋白质药物等蛋白质产品中, 是非常重要的蛋白质技术手段。毫无疑问, 未来的蛋白质设计也应该具备这些特殊氨基酸的设计能力。

蛋白质的结构动态性是当前蛋白质设计的另一个重大挑战。众所周知, 绝大多数蛋白质的三维结构都是柔性的, 具备多种不同的构象状态。当前蛋白质设计方法中针对结构进行设计的, 从其训练数

据到其设计时针对的靶蛋白, 往往都是静态的结构, 那么在此基础上设计出来的蛋白质是否能够在相对更自由的溶液环境中展现设计构象呢? 结合是否会令靶蛋白产生严重的诱导构象变化, 从而导致无法实现设计中的紧密结合呢? 这些问题目前都还没有确定的答案。这也很可能是在一部分场景中的蛋白质设计成功率明显较低的原因所在。当前蛋白质结构动态性的研究主要依赖于分子动力学模拟, 非常消耗算力, 难以积累大量的数据用于深度学习模型的训练。其他一些与人工智能技术相结合的方法往往基于复杂的计算体系, 其实用性也有待进一步验证。无论如何, 蛋白质设计未来的发展必然会将结构的动态性纳入到设计场景中, 从而实现能够与实验结果更加匹配的结构设计。

在工程化生产的角度来看, 蛋白质的重组生产等生产环节也需要人工智能技术的介入。当前蛋白质的工程化生产已经具备了一定的自动化程度, 但相关控制算法基本都是基于规则的, 智能程度低, 对人的依赖程度仍然较高。在该领域建立深度学习模型的主要制约条件是数据的不足。大多数生产企业将工程中产生的过程数据视为重要的无形资产, 不会开放给其他企业或机构使用, 因而很难建立像PDB这样大规模的数据集, 也就难以训练得到高质量的人工智能模型。未来在这方面的进一步发展还需要找到打破数据壁垒的渠道或方法。

虽然智能蛋白质工程领域还面临着众多挑战, 但是现有的大量努力和成果, 以及2024年诺贝尔化学奖对蛋白质设计领域的肯定, 都已经证明, 未来的蛋白质工程必然要依赖于先进的人工智能技术才能得到进一步的发展。面对比宇宙还要浩瀚的蛋白质空间, 基于深度学习的蛋白质设计方法将是人类的必然选择。

参 考 文 献

- [1] Woolfson D N. A brief history of *de novo* protein design: minimal, rational, and computational. *J Mol Biol*, 2021, **433**(20): 167160
- [2] Singh R K, Lee J K, Selvaraj C, *et al.* Protein engineering approaches in the post-genomic era. *Curr Protein Pept Sci*, 2018, **19**(1): 5-15
- [3] Chang C, Deringer V L, Katti K S, *et al.* Simulations in the era of exascale computing. *Nat Rev Mater*, 2023, **8**(5): 309-313
- [4] Kouba P, Kohout P, Haddadi F, *et al.* Machine learning-guided protein engineering. *ACS Catal*, 2023, **13**(21): 13863-13895
- [5] Kortemme T. *De novo* protein design-from new structures to programmable functions. *Cell*, 2024, **187**(3): 526-544
- [6] 陈志航, 季梦麟, 戚逸飞. 人工智能蛋白质结构设计算法研究进展合成生物学, 2023, **4**(3): 464-487
- [7] Chen Z H, Ji M L, Qi Y F. *Synthetic Biology Journal*, 2023, **4**(3): 464-487
- [8] Watson J L, Juergens D, Bennett N R, *et al.* *De novo* design of protein structure and function with RFdiffusion. *Nature*, 2023, **620**(7976): 1089-1100
- [9] Liu Y, Chen L, Liu H. *De novo* protein backbone generation based on diffusion with structured priors and adversarial training. *bioRxiv*, 2022. DOI: 10.1101/2022.12.17.520847
- [10] Eguchi R R, Choe C A, Huang P S. Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput Biol*, 2022, **18**(6): e1010271
- [11] Luo S, Su Y, Peng X, *et al.* Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Adv Neural Inform Process Sys*, 2022, **35**: 9754-9767
- [12] Ingraham J B, Baranov M, Costello Z, *et al.* Illuminating protein space with a programmable generative model. *Nature*, 2023, **623**(7989): 1070-1078
- [13] Hawkins-Hooker A, Depardieu F, Baur S, *et al.* Generating functional protein variants with variational autoencoders. *PLoS Comput Biol*, 2021, **17**(2): e1008736
- [14] Repecka D, Jauniskis V, Karpus L, *et al.* Expanding functional protein sequence spaces using generative adversarial networks. *Nat Mach Intell*, 2021, **3**: 324-333
- [15] Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun*, 2022, **13**(1): 4348
- [16] Madani A, Krause B, Greene E R, *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol*, 2023, **41**(8): 1099-1106
- [17] Nijkamp E, Ruffolo J A, Weinstein E N, *et al.* ProGen2: exploring the boundaries of protein language models. *Cell Syst*, 2023, **14**(11): 968-978.e3
- [18] Moffat L, Kandathil S, Jones D T. Design in the DARK: learning deep generative models for *de novo* protein design. *bioRxiv*, 2022. DOI: 10.1101/2022.01.27.478087
- [19] Dauparas J, Anishchenko I, Bennett N, *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 2022, **378**(6615): 49-56
- [20] Zheng Z, Deng Y, Xue D, *et al.* Structure-informed language models are protein designers//Andreas K, Emma B. *Proceedings of the 40th International Conference on Machine Learning*: 202. Honolulu Hawaii: PMLR, 2023: 42317-42338
- [21] Dauparas J, Lee G R, Pecoraro R, *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *bioRxiv*, 2023. DOI: 10.1101/2023.12.22.573103
- [22] Huang B, Fan T, Wang K, *et al.* Accurate and efficient protein sequence design through learning concise local environment of residues. *Bioinformatics*, 2023, **39**(3): btad122
- [23] Ren M, Yu C, Bu D, *et al.* Accurate and robust protein sequence design with CarbonDesign. *Nat Mach Intell*, 2024, **6**: 536-547
- [24] Wang J, Chen C, Yao G, *et al.* Intelligent protein design and molecular characterization techniques: a comprehensive review. *Molecules*, 2023, **28**(23): 7865

- [24] Bryngelson J D, Onuchic J N, Socci N D, *et al.* Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 1995, **21**(3): 167-195
- [25] Korendovych I V, DeGrado W F. *De novo* protein design, a retrospective. *Q Rev Biophys*, 2020, **53**: e3
- [26] Levinthal C. How to fold graciously. *Mossbauer Spectroscopy in Biological Systems*, 1969, **67**: 22-24
- [27] DeGrado W F, Regan L, Ho S P. The design of a four-helix bundle protein. *Cold Spring Harb Symp Quant Biol*, 1987, **52**: 521-526
- [28] Regan L, DeGrado W F. Characterization of a helical protein designed from first principles. *Science*, 1988, **241**(4868): 976-978
- [29] Harbury P B, Tidor B, Kim P S. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc Natl Acad Sci USA*, 1995, **92**(18): 8408-8412
- [30] Desjarlais J R, Handel T M. *De novo* design of the hydrophobic cores of proteins. *Protein Sci*, 1995, **4**(10): 2006-2018
- [31] Dahiyat B, Mayo S L. Protein design automation. *Protein Sci*, 1996, **5**(5): 895-903
- [32] Kuhlman B, Dantas G, Ireton G C, *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science*, 2003, **302**(5649): 1364-1368
- [33] Huang P S, Boyken S E, Baker D. The coming of age of *de novo* protein design. *Nature*, 2016, **537**(7620): 320-327
- [34] Anishchenko I, Pellock S J, Chidyausiku T M, *et al.* *De novo* protein design by deep network hallucination. *Nature*, 2021, **600**(7889): 547-552
- [35] Chevalier A, Silva D A, Rocklin G J, *et al.* Massively parallel *de novo* protein design for targeted therapeutics. *Nature*, 2017, **550**(7674): 74-79
- [36] Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA*, 2000, **97**(19): 10383-10388
- [37] Bonneau R, Tsai J, Ruczinski I, *et al.* Rosetta in CASP4: progress in *ab initio* protein structure prediction. *Proteins*, 2001, **Suppl 5**: 119-126
- [38] Berman H, Westbrook J, Feng Z, *et al.* The protein data bank. *Nucleic Acids Res*, 2000, **28**(1): 235-242
- [39] Wu K E, Yang K K, van den Berg R, *et al.* Protein structure generation via folding diffusion. *Nat Commun*, 2024, **15**(1): 1059
- [40] Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, **596**(7873): 583-589
- [41] Abramson J, Adler J, Dunger J, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024, **630**(8016): 493-500
- [42] Baek M, McHugh R, Anishchenko I, *et al.* Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. *Nat Methods*, 2024, **21**(1): 117-121
- [43] Fang X, Wang F, Liu L, *et al.* A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat Mach Intell*, 2023, **5**: 1087-1096
- [44] Baek M, DiMaio F, Anishchenko I, *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021, **373**(6557): 871-876
- [45] Lin Z, Akin H, Rao R, *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023, **379**(6637): 1123-1130
- [46] Senior A W, Evans R, Jumper J, *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins*, 2019, **87**(12): 1141-1148
- [47] Källberg M, Wang H, Wang S, *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat Protoc*, 2012, **7**(8): 1511-1522
- [48] Lin Z, Akin H, Rao R, *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. DOI: 10.1101/2022.07.20.500902
- [49] Yang J, Anishchenko I, Park H, *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci USA*, 2020, **117**(3): 1496-1503
- [50] Zambaldi V, La D, Chu A E, *et al.* De novo design of high-affinity protein binders with alphaproteo. *arXiv*, 2024. DOI: 10.48550/arXiv.2409.08022
- [51] Anand N, Huang P C. Generative modeling for protein structures// *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Canada: Curran Associates Inc. 2018: 7505-7516
- [52] Anand N, Eguchi R R, Huang P S. Fully differentiable full-atom protein backbone generation. *New Orleans: International Conference on Learning Representations (ICLR)*, 2019 [2024-11-26]. <https://openreview.net/forum?id=SJxnVL8YOV>
- [53] Ho J, Jain A, Abbeel P J a I N I P S. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020, **33**: 6840-6851
- [54] Huang B, Xu Y, Hu X, *et al.* A backbone-centred energy function of neural networks for protein design. *Nature*, 2022, **602**(7897): 523-528
- [55] Krishna R, Wang J, Ahern W, *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 2024, **384**(6693): ead12528
- [56] Anfinsen C B. Principles that govern the folding of protein chains. *Science*, 1973, **181**(4096): 223-230
- [57] Onuchic J, Wolynes P. Theory of protein folding. *Curr Opin Struct Biol*, 2004, **14**(1): 70-75
- [58] Dill K A, MacCallum J L. The protein-folding problem, 50 years on. *Science*, 2012, **338**(6110): 1042-1046
- [59] Wu Z, Johnston K E, Arnold F H, *et al.* Protein sequence design with deep generative models. *Curr Opin Chem Biol*, 2021, **65**: 18-27
- [60] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *arXiv*, 2017. DOI: 10.48550/arXiv.1706.03762
- [61] Mazurenko S, Prokop Z, Damborský J. Machine learning in enzyme engineering. *ACS Catalysis*, 2019, **10**(2): 1210-1223
- [62] Wang X, Yin X, Jiang D, *et al.* Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites. *Nat Commun*, 2024, **15**(1): 7348
- [63] Li F, Yuan L, Lu H, *et al.* Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat*

- Catal, 2022, **5**: 662-672
- [64] Wang T, Xiang G, He S, *et al.* DeepEnzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3D-structures. *Brief Bioinform*, 2024, **25**(5): bbae409
- [65] Zhang F, Naeem M, Yu B, *et al.* Improving the enzymatic activity and stability of N-carbamoyl hydrolase using deep learning approach. *Microb Cell Fact*, 2024, **23**(1): 164
- [66] Torres S V, Valle M B, Mackessy S P, *et al.* *De novo* designed proteins neutralize lethal snake venom toxins. *Res Sq*, 2024: rs.3.rs-rs.4402792
- [67] Nikam R, Kulandaisamy A, Harini K, *et al.* ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res*, 2021, **49**(D1): D420-D424
- [68] Fang X, Huang J, Zhang R, *et al.* Convolution neural network-based prediction of protein thermostability. *J Chem Inf Model*, 2019, **59**(11): 4833-4843
- [69] Jarzab A, Kurzawa N, Hopf T, *et al.* Meltome atlas-thermal proteome stability across the tree of life. *Nat Methods*, 2020, **17**(5): 495-503
- [70] Li M, Wang H, Yang Z, *et al.* DeepTM: a deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences. *Comput Struct Biotechnol J*, 2023, **21**: 5544-5560
- [71] Muellers S N, Allen K N, Whitty A. MENaT: a machine-learning approach for the identification of mutations to increase protein stability. *Proc Natl Acad Sci USA*, 2023, **120**(49): e2309884120
- [72] Sandhu K S, Pandey S, Maiti S, *et al.* GASCO: genetic algorithm simulation for codon optimization. *In Silico Biol*, 2008, **8**(2): 187-192
- [73] Welch M, Govindarajan S, Ness J E, *et al.* Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One*, 2009, **4**(9): e7002
- [74] Saito Y, Kitagawa W, Kumagai T, *et al.* Developing a codon optimization method for improved expression of recombinant proteins in actinobacteria. *Sci Rep*, 2019, **9**(1): 8338
- [75] Şen A, Kargar K, Akgün E, *et al.* Codon optimization: a mathematical programming approach. *Bioinformatics*, 2020, **36**(13): 4012-4020
- [76] Fu H, Liang Y, Zhong X, *et al.* Codon optimization with deep learning to enhance protein expression. *Sci Rep*, 2020, **10**(1): 17617
- [77] Jain R, Jain A, Mauro E, *et al.* ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinformatics*, 2023, **24**(1): 132
- [78] Zhang P, Wang H, Xu H, *et al.* Deep flanking sequence engineering for efficient promoter design using DeepSEED. *Nat Commun*, 2023, **14**(1): 6309
- [79] Leppik K, Byeon G W, Kladwang W, *et al.* Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat Commun*, 2022, **13**(1): 1536
- [80] Wilkinson P, Jackson B, Fermor H, *et al.* A new mRNA structure prediction based approach to identifying improved signal peptides for bone morphogenetic protein 2. *BMC Biotechnol*, 2024, **24**(1): 34
- [81] Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun*, 2021, **12**(1): 941
- [82] Grand View Research. Protein engineering market size & trends. San Francisco: Grand View Research, 2023[2024-11-26]. <https://www.grandviewresearch.com/industry-analysis/protein-engineering-market>
- [83] Yeh A H W, Nom C, Kipnis Y, *et al.* *De novo* design of luciferases using deep learning. *Nature*, 2023, **614**(7949): 774-780
- [84] Wang J, Lisanza S, Juergens D, *et al.* Scaffolding protein functional sites using deep learning. *Science*, 2022, **377**(6604): 387-394
- [85] Nom C, Wicky B I M, Juergens D, *et al.* Protein sequence design by conformational landscape optimization. *Proc Natl Acad Sci USA*, 2021, **118**(11): e2017228118
- [86] Sormanni P, Aprile F A, Vendruscolo M. Third generation antibody discovery methods: *in silico* rational design. *Chem Soc Rev*, 2018, **47**(24): 9137-9157
- [87] Pertseva M, Gao B, Neumeier D, *et al.* Applications of machine and deep learning in adaptive immunity. *Annu Rev Chem Biomol Eng*, 2021, **12**: 39-62
- [88] Carter P J, Rajpal A. Designing antibodies as therapeutics. *Cell*, 2022, **185**(15): 2789-2805
- [89] Pioszak A A, Xu H E. Molecular recognition of parathyroid hormone by its G protein-coupled receptor. *Proc Natl Acad Sci USA*, 2008, **105**(13): 5034-5039
- [90] Park C, Kim J, Ko S B, *et al.* Structural basis of neuropeptide Y signaling through Y1 receptor. *Nat Commun*, 2022, **13**(1): 853
- [91] Sasaki K, Dockerill S, Adamiak D A, *et al.* X-ray analysis of glucagon and its relationship to receptor binding. *Nature*, 1975, **257**(5529): 751-757
- [92] Fukuhara S, Kobayashi K, Kusakizako T, *et al.* Structure of the human secretin receptor coupled to an engineered heterotrimeric G protein. *Biochem Biophys Res Commun*, 2020, **533**(4): 861-866
- [93] Cao L, Goresnik I, Coventry B, *et al.* *De novo* design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science*, 2020, **370**(6515): 426-431
- [94] Berger S, Seeger F, Yu T Y, *et al.* Preclinical proof of principle for orally delivered Th17 antagonist miniproteins. *Cell*, 2024, **187**(16): 4305-4317.e18
- [95] An L, Said M, Tran L, *et al.* Binding and sensing diverse small molecules using shape-complementary pseudocycles. *Science*, 2024, **385**(6706): 276-282

Intelligent Protein Engineering*

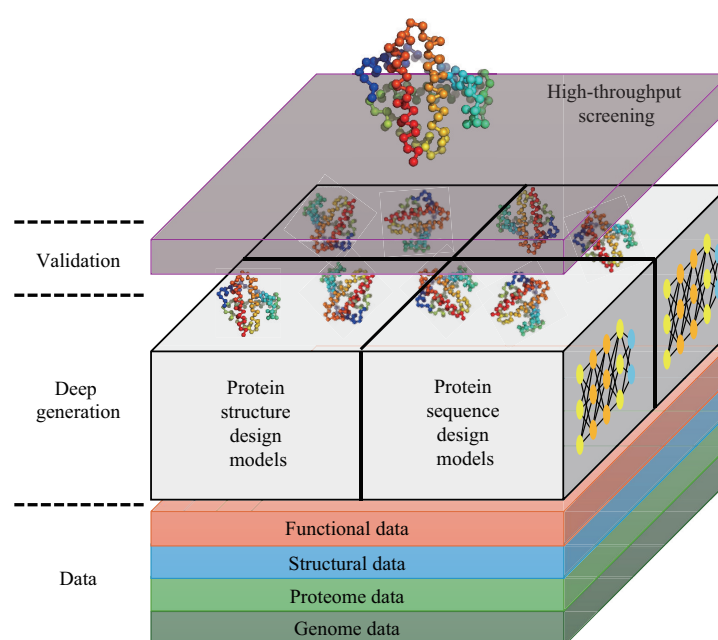
WANG Kai-Yue^{1,3)}, YE Sheng^{1,2,3)**}

⁽¹⁾*School of Biomedical Science and Medicine Engineering, Beihang University, Beijing 100191, China;*

⁽²⁾*School of Engineering Medicine, Beihang University, Beijing 100191, China;*

⁽³⁾*Key Laboratory of Big Data-Based Precision Medicine, Ministry of Industry and Information Technology, Beihang University, Beijing 100191, China)*

Graphical abstract



Abstract Proteins are essential fundamental substance for life processes, performing a variety of key roles in organisms, such as constructing cell structures, participating in metabolism and energy transformation, regulating physiological functions, providing immune protection, and transmitting signals. The diverse functions of proteins are achieved through their unique amino acid sequences and corresponding three-dimensional structures. Protein engineering involves altering or designing protein sequences and structures to achieve specific functions, and these efforts enhance our knowledge of proteins and offer powerful tools and technical support for research in biomedicine, biomaterials, bioengineering, and related fields. In recent years, with the advancements in algorithms, the accumulation of big data, and improvements in hardware computational power, artificial intelligence technology has rapidly developed and gradually been applied in the field of protein engineering, leading to the emergence of intelligent protein engineering. By utilizing biological big data such as genomics,

* This work was supported by a grant from The Project of Beijing Science and Technology Committee (Z221100003522020).

** Corresponding author.

Tel: 86-10-82317114, E-mail: yesheng@buaa.edu.cn

Received: September 12, 2024 Accepted: November 27, 2024

proteomics, protein structure databases, and establishing various advanced deep learning models based on the data, intelligent protein engineering can achieve efficient, precise, and predictable protein design and modification. This article primarily focuses on four aspects of intelligent protein engineering, including structure design, backbone-free sequence design, backbone-based sequence design, and other auxiliary design approaches, summarizing the latest progress in the artificial intelligence technologies employed in these fields, and compiling the practical results achieved in recent years using intelligent protein engineering technology. As an emerging technology and method, intelligent protein engineering demonstrates significant potential and prospects, bringing substantial impacts on future scientific research and technological innovation, and providing new solutions and tools for addressing global challenges.

Key words protein engineering, protein design, codon optimization, protein structure, deep learning, deep generative, artificial intelligence

DOI: 10.16476/j.pibb.2024.0402