·综 述·

人工智能时代下的蛋白质从头设计

刘南[1]，金小程[1]，杨崇周[2]，王梓洋[2]，闵小平[2*]，葛胜祥[1*]

1 厦门大学 公共卫生学院 分子疫苗学与分子诊断国家重点实验室，福建 厦门 361005
2 厦门大学 信息学院 人工智能研究院，福建 厦门 361005

刘南，金小程，杨崇周，王梓洋，闵小平，葛胜祥. 人工智能时代下的蛋白质从头设计[J]. 生物工程学报，2024，40(11): 3912-3929.
LIU Nan, JIN Xiaocheng, YANG Chongzhou, WANG Ziyang, MIN Xiaoping, GE Shengxiang. *De novo* protein design in the age of artificial intelligence[J]. Chinese Journal of Biotechnology, 2024, 40(11): 3912-3929.

摘 要：具有特定功能和特性的蛋白质在生物医药、纳米材料等领域至关重要。蛋白质从头设计能够定制序列以生成具有所需结构、自然界中未存在的蛋白质。近年来，随着人工智能的迅猛发展，深度学习生成模型逐渐成为强大工具，许多功能性蛋白质的设计都达到了原子级别的精度。本文概述了蛋白质从头设计的演进，着重介绍了其最新算法模型，并分析了其存在的问题，如设计成功率低、精度不足以及对实验验证的依赖性，最后探讨了蛋白质设计的未来趋势，旨在为研究者和从业者提供有益参考。

关键词：蛋白质从头设计；人工智能；深度学习；扩散模型

De novo protein design in the age of artificial intelligence

LIU Nan[1], JIN Xiaocheng[1], YANG Chongzhou[2], WANG Ziyang[2], MIN Xiaoping[2*], GE Shengxiang[1*]

1 State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics, School of Public Health, Xiamen University, Xiamen 361005, Fujian, China
2 Institute of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, Fujian, China

Abstract: Proteins with specific functions and characteristics play a crucial role in biomedicine and nanotechnology. *De novo* protein design enables the customization of sequences to produce proteins with desired structures that do not exist in the nature. In recent years, with the rapid development of

资助项目：国家自然科学基金(62272399)；医学科学院医学科学创新基金(2019RU022)；中国福建省重点项目基金(2021J02006)；中央高校基本科研业务费(20720220005, 20720220006)
This work was supported by the National Natural Science Foundation of China (62272399), the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2019RU022), the Key Program Foundation of Fujian Province (2021J02006), and the Fundamental Research Fund for the Central Universities of China (20720220005, 20720220006).
*Corresponding authors. E-mail: GE Shengxiang, sxge@xmu.edu.cn; MIN Xiaoping, mxp@xmu.edu.cn
Received: 2024-01-30; Accepted: 2024-07-19; Published online: 2024-07-22

artificial intelligence (AI), deep learning-based generative models have increasingly become powerful tools, enabling the design of functional proteins with atomic-level precision. This article provides an overview of the evolution of *de novo* protein design, with focus on the latest algorithmic models, and then analyzes existing challenges such as low design success rates, insufficient accuracy, and dependence on experimental validation. Furthermore, this article discusses the future trends in protein design, aiming to provide insights for researchers and practitioners in this field.

Keywords: *de novo* protein design; artificial intelligence; deep learning; diffusion model

在地球生命演化的 40 亿年中形成了功能丰富的蛋白质，这些蛋白质作为功能性分子，在 DNA/RNA 复制、转录、翻译等生命活动中起着至关重要的作用，是执行生命功能的核心元素[1]。根据热力学原理[2]，蛋白质的功能由结构决定，而其三维结构由 20 种氨基酸序列排列所确定。然而，蛋白质序列多样性极高，一个由 200 个氨基酸组成的蛋白质能有 20^{200} 种可能[3]，超过宇宙原子总和[4]。自然演化形成了其中的一部分序列，组成了众多的蛋白家族，但利用传统方法如随机突变设计蛋白质面临巨大挑战[3]。近年来，蛋白质从头设计取得了显著进展，可以不依赖天然模板设计新型蛋白质，这为探索具有特定功能的全新蛋白质、满足人类不断增长的需求打开了大门。

蛋白质从头设计一般包括两个步骤：蛋白质主链三级结构的生成，残基位置及侧链构象设计[5]。这两步的基础是能量函数的评估。1980 年，研究人员提出了蛋白质折叠的最低能量状态假说，认为氨基酸序列决定了其三维结构[2]。因此，传统设计方法主要依据能量最低原理来寻找合适的能量函数。典型的研究方法有 RosettaDesign[6]、CHARMM[7]、ISAMBARD[8]、FoldX[9]等。Rosetta 提供了最常用的一种能量函数，涵盖了物理的能量组成和确定蛋白质结构的实验因素，包括氢键网络[10]、范德华力[11]、极性及疏水性[12]等，并采用蒙特卡罗模拟退火、死码消除算法、遗传算法和优化理论等[13-16]方法来优化设计结果。在过去的 40 年中，基于传统方法的蛋白质从头设计成功率低且依赖于多轮实验验证以确保其可靠性。

近 10 年来，随着计算能力的提升、蛋白质数据的累积以及人工智能(artificial intelligence, AI)技术的应用，蛋白质从头设计发生了根本性的变革。设计方法由基于第一性原理的传统方式转变成了运用深度学习的现代方法，特别是 AlphaFold2[17]及 RoseTTAFold[18]在蛋白质结构预测上的开创性成就。据报道，通过整合最新的计算设计、筛选方法，实验室工作量可大幅减少至万分之一[19]。2023 年 12 月，David Baker 展示了新型扩散模型 RFdiffusion，其在 Bim 和 PTH 体系的应用表明所设计的多肽结合蛋白具有高度亲和力，达到 pmol 量级[20]。虽有成功实例，但蛋白质从头设计依旧面临挑战，包括低成功率、计算模型局限性以及对实验验证的依赖等。另外，未来的研究中需关注的问题还包括缩短设计周期和降低成本。

总体而言，蛋白质从头设计在先进算法的助力下仍具有巨大的发展潜力，该技术在癌症治疗、纳米技术、下一代人工智能疫苗、生物兼容性材料等众多领域均显示出巨大潜能(图 1)。本文回顾蛋白质从头设计的发展历程和最新成就，分析现存挑战，探讨未来的发展方向，旨在为该领域的研究人员提供参考。

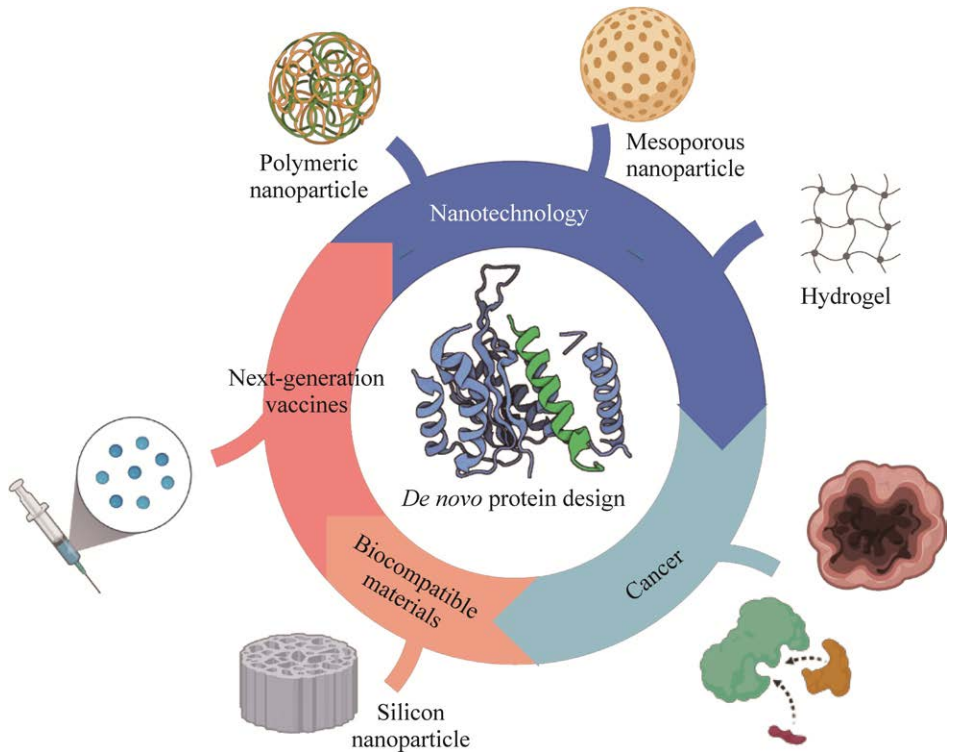

图 1　蛋白质从头设计的应用

Figure 1　The application of *de novo* protein design.

1　蛋白质从头设计概述

1.1　蛋白质从头设计演变

蛋白质从头设计涉及的序列以及构象空间极为复杂，曾被视为不可能完成的任务[21]。自蛋白质从头设计首次被提出以来[22]，该领域面临诸多挑战。蛋白质的从头设计最初是指从头开始设计蛋白质，而不是修改天然存在的蛋白质[23]。随着蛋白质数据库扩充以及计算能力和算法的迭代更新，蛋白质从头设计的概念不断拓展。在该领域的发展历程中，有研究者将其前 20 多年演变过程分成 3 个阶段[24]：依赖物理模型设计、以物理化学原理为导向的计算设计以及结合片段和生物信息学的设计，并回顾了该领域的里程碑进展和成就。同时，有研究者提出了最小化蛋白设计、理性蛋白设计、蛋白质计算设计这 3 种蛋白设计方法[25]，并制定了详尽的时间表来描述从头设计肽及蛋白质的进展。上述工作总结了蛋白质从头设计的发展历程，虽然时间线不同，但是各阶段和定义方法基本一致。

近 10 年间，随着对蛋白质折叠机制理解的不断深入，蛋白质结构预测的准确性显著提高，使得设计的氨基酸序列能够折叠成与晶体结构高度一致的蛋白质构象。特别在 2021 年，深度学习驱动的 AlphaFold2[17]在蛋白质结构预测中实现了突破，彰显了深度学习的巨大优势。同时，蛋白质从头设计领域也产生了包括病毒抑制剂[26-27]、萤光素酶[28]、ProteinMPNN[29]、RFdiffusion[20]等在内的令人振奋的成果及设计工具，开启了新的发展阶段。AlphaFold2 的出现标志着蛋白质设计领域的重大转折点，本文据此将蛋白质从头设计策略分为传统和基于 AI 两大类别(图 2)。

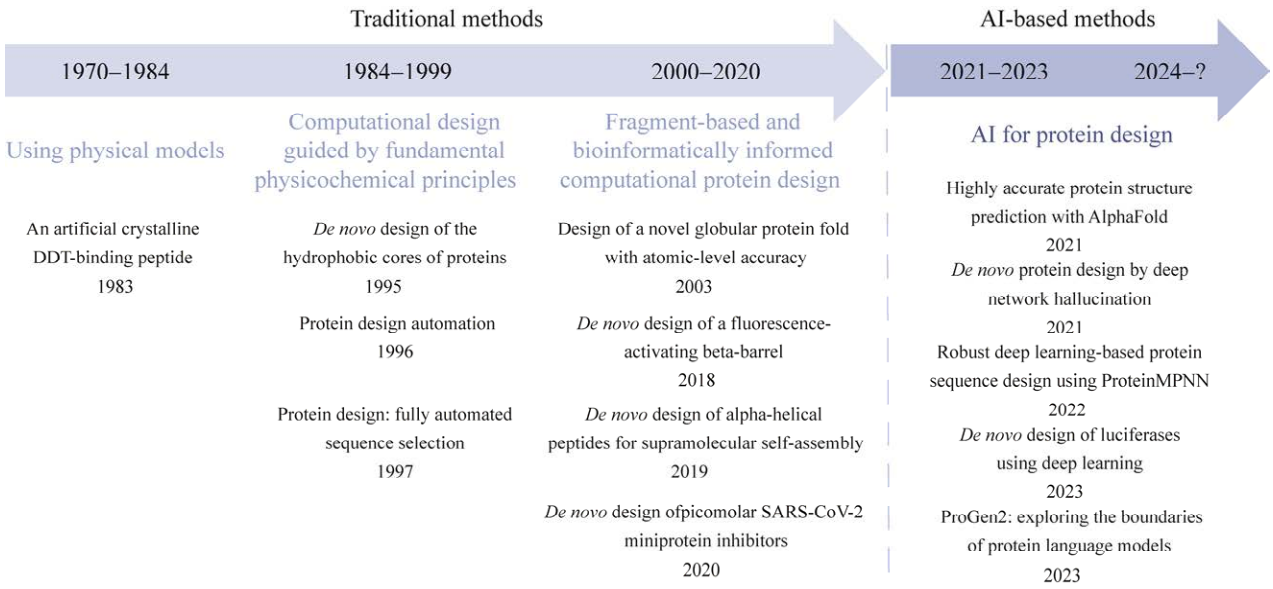

图 2　1970–2024 年的蛋白质从头设计发展史

Figure 2　The development history of protein design from scratch (1970–2024).

1.2　传统蛋白质从头设计基本步骤

传统蛋白质从头设计通常利用能量函数和搜索策略揭示序列、结构和功能之间的关系，该过程分为 4 个阶段[30]：(1) 定义目标蛋白结构；(2) 生成主链骨架；(3) 确定氨基酸序列；(4) 评估设计序列与结构的兼容性(图 3)。

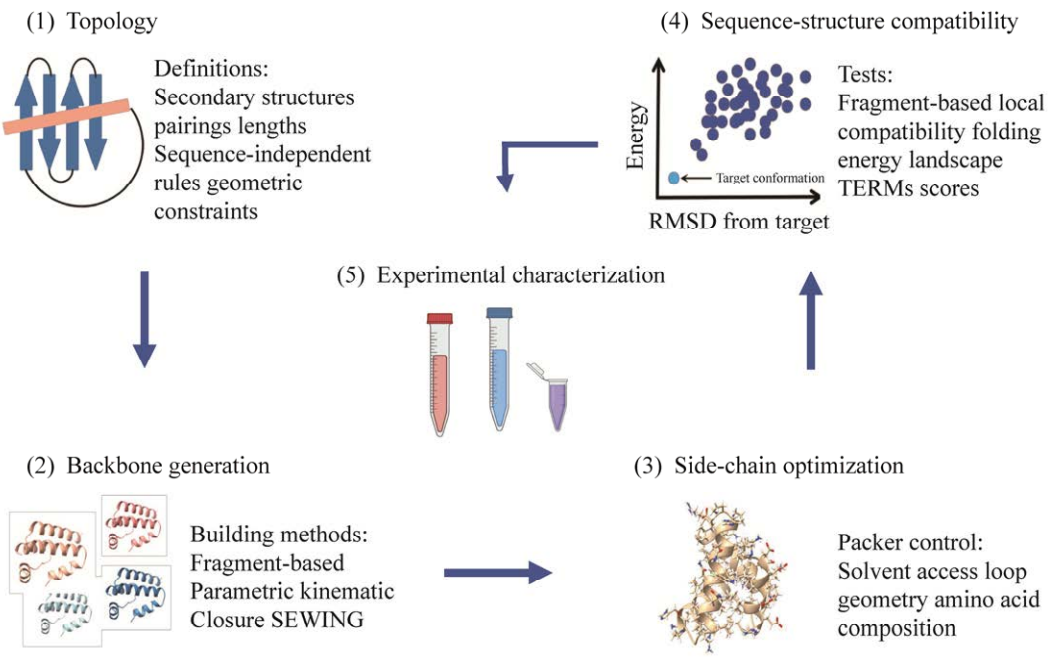

图 3　蛋白质设计流程图

Figure 3　Protein design flowchart.

从头设计从确定目标蛋白拓扑结构入手，考虑诸如二级结构的元素、连接形状和几何形状等因素。接下来，筛选更高度设计性的结构特征，获得优选构型。骨架构建通常基于蛋白质数据库(protein database, PDB)的结构信息，方法包括片段组装[31]、参数化设计[32]、kinematic closure[33]、原生结构图扩展[34]、循环构建[35]等。当前，依靠已有片段组装是最成功的通用方法，例如 Rosetta TopoBuilder[36]可以构建各种蛋白质构型，包括微型蛋白[37]，尽管其难度会随着主干结构元素增加而升高。

为选定骨架结构寻找能量最小化的氨基酸序列需要进行序列填充设计。大部分序列设计算法依赖于能量函数，如 David Baker 课题组的 Rosetta 软件中广泛应用的 Packer[38]以及刘海燕课题组开发的 ABACUS 方法[39-40]。Packer 主要用于模拟退火处理主链的侧链放置，以确定最佳序列顺序。研究人员设计了一系列偏向策略[41-42]，包含层设计[43]、氨基酸组成限制[44]、二硫键搜索[45]、自定义选择关键结构特征的氨基酸[46]等，以偏向于最佳设计方案，减少能量函数和采样算法的局限性。经过 Rosetta 的快速属性评估与排名，这些方法能较大概率地产生结构稳定且符合目标的设计。

尽管得到低能序列有利于折叠成目标结构，但还需进一步评估设计序列与目标结构的兼容性以确保正确折叠。AlphaFold2 出现前，高端的结构预测技术[36]由于不准确、成本高昂等缺点促使人们选择成本较低的序列结构兼容性测试，通常分为局部和非局部验证，但这种方法仅限于少数设计的验证。

1.3 传统方法的局限性

在过去的几十年里，蛋白质设计领域经历了从依赖经验规则到计算驱动方法的显著转变。这一进步不仅将原本难以捉摸的理念转化为现实，而且对精准医疗[27]、疫苗研发[47-48]、纳米材料[49]领域产生了深远影响。虽然传统从头设计能够解决一些挑战性任务，但是人们逐渐认识到其准确性和稳定性有限。传统从头设计受到能量函数描述不精确及搜索策略局限的影响，存在计算量大、很难找到最优解、成功率低、精度不足、能量函数本身立场精度等问题，通常需要通过实验方法的优化来验证设计的可靠性。此外，这些方法通常仅适用于特定生物系统，难以应用于不同生物学问题的迁移设计。

2 蛋白质从头设计算法模型

计算能力的显著提升和高通量技术的应用为蛋白质数据的积累提供了强大支持，推动了蛋白质从头设计从传统方法向数据驱动的计算方法转变[50]。近年来，深度学习技术在计算机视觉、自然语言处理(natural language processing, NLP)等[51]领域的应用已逐步成熟。由于数据结构和处理逻辑的相似性，这些先进技术正对生物科学领域产生深远影响。例如氨基酸序列类似于人类语言，现有的 NLP 模型便能迅速转化为蛋白质序列的有效参数化工具。

在大量蛋白质序列和结构数据的支持下，深度学习在探索序列-结构-功能关系方面取得了重要进展，革新了功能性蛋白的设计方法。基于深度学习的蛋白质从头设计主要分为两大类：第一类基于骨架结构的蛋白设计，包括固定主链和可变骨架的序列优化[52]；第二类不依赖于结构的直接序列生成，用于探索序列空间和蛋白质生成。

2.1 基于骨架结构的蛋白设计

在蛋白质结构预测领域，深度学习从早期的残基接触预测和辅助结构建模[53-57]，到准确预测残基的几何性质和基于几何约束的蛋白质

折叠[58-61]，该领域的多个方面已经彻底革新。蛋白质设计和结构预测相辅相成，设计过程可以看作是预测过程的逆向工程。蛋白质设计可以利用结构特征来指导设计任务，而结构预测技术的进步为设计工作提供了实用的工具，从而推动了蛋白质研究的发展。蛋白质序列中包含了其折叠结构的关键信息，深度学习技术的应用加深了研究人员对序列与结构关系的理解，为基于结构的蛋白质设计奠定了坚实的基础。

在固定骨架的情况下，寻找最大化氨基酸联合概率的序列是设计的关键[62]。传统设计方法倾向于产出与输入骨架紧密匹配的序列，忽略了结构的灵活性与动态性，导致输出限制和序列多样性不足。然而，深度学习能够从数据中学习到人类所不知道的潜在规律，有望解决上述局限性。

2018 年，一种名为 SPIN2 的新模型被提出[63]。最初，SPIN2 模型的平均恢复率为 34.0%，它仅采用了一维属性。随后，采用了基于图像学习框架新方法[64]，在独立的测试集上，序列的恢复率提高到 39.8%。进入 2020 年，进一步开创了 DenseCPD 深度神经网络[65]，这是一种考虑蛋白骨架原子三维密度分布的模型，其在两个测试集上氨基酸回收率分别达到 55.53%和 50.71%，比以前最先进的技术提高了约 10.00%。

TrDesign[66]以 TrRosetta[58]模型为基础进行反向序列设计，利用比对蛋白序列的几何特性，以推测残基间的距离分布图。接着通过蒙特卡罗模拟退火技术，对随机序列进行迭代优化；在这个过程中，可以在任意位置突变氨基酸，当新序列的距离分布满足 Metropolis 标准时接受这次突变，完成所有替换后将创造自然界从来没出现过的序列。然而，模型的反复运行可能会降低计算效率，陷入次优解的方案。刘海燕课题组开发了 ABACUS-R[39-40,67]，使用 Transformer 神经网络，多任务学习策略，消除重构和优化侧链结构的需要，简化序列设计的过程，使得平均恢复率提高到 53%。生物实验验证表明，ABACUS-R 的设计精度和成功率均超过现有最先进的能量函数方法。

2022 年，David Baker 团队提出了一种基于深度学习的蛋白质序列设计方法[29]，称为 ProteinMPNN (图 4)，通过不同采样温度和噪声添加，旨在增强序列多样性及可靠性，该方法在模拟和实验中表现卓越。与基于物理的方法如 Rosetta 不同，ProteinMPNN 无需为每一个特定的设计挑战进行专家级定制，可以广泛应用于多种序列设计问题中。这种广泛适用性源于序列设计构建方式的本质区别，并依赖于通过从 PDB 检索所有蛋白质进行直接训练的深度学习方法，而不是识别最低能量氨基酸序列问题。由于具有较高的实验设计成功率和计算效率，ProteinMPNN 有望成为蛋白质序列设计的标准方法，并可能迅速获得研究人员的广泛认可。此外，它还展现出更高的结晶倾向性，极大地促进了设计蛋白质结构测定。预计 ProteinMPNN 生成的序列将大幅改善天然蛋白质骨架的稳定性和表达能力，对于重组表达的天然蛋白质尤为有益。由于设计方法的最终检验是实验验证，就像翻译的准确性需要人来评估一样，ProteinMPNN 的实用性同样需要通过实验来验证。

2023 年，一种基于深度多层感知机的前馈神经网络 Anand 模型[68]被提出，该模型通过 16 个氨基酸骨架原子定义残基的结构环境。实验证明，该方法能够精准地预测折叠序列，并产生与参考序列差异显著的新氨基酸序列，这为蛋白设计和材料科学提供了更大的自由度。

随着蛋白质结构数据日益丰富和深度学习技术的不断进步，基于结构的序列设计方法层

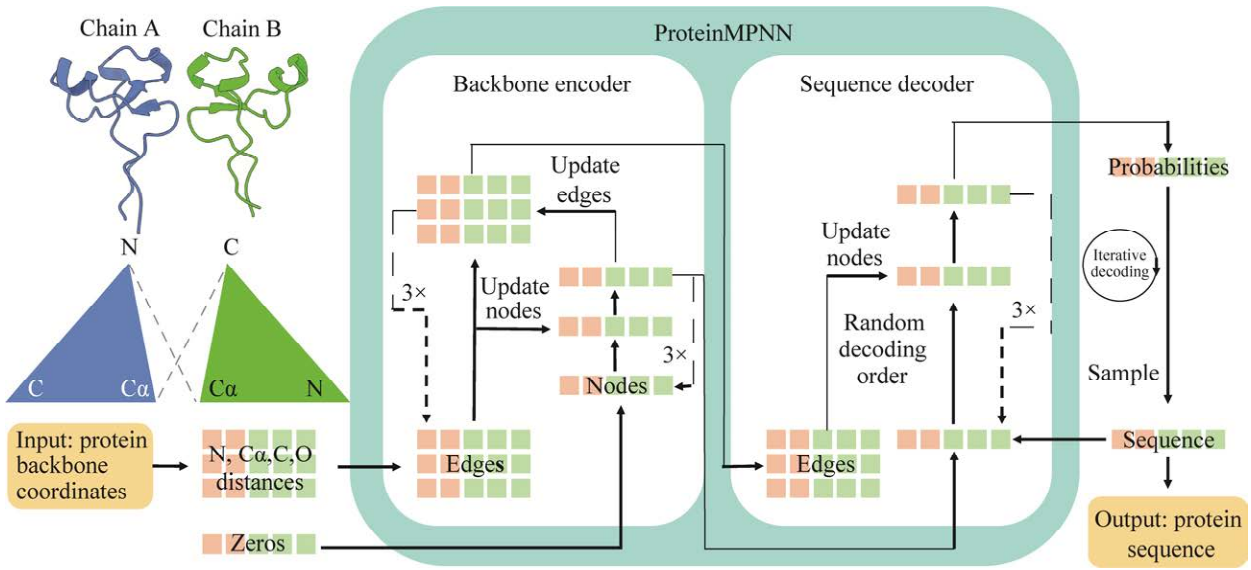

图 4　ProteinMPNN 模型的整体结构

Figure 4　The overall structure of the ProteinMPNN model.

出不穷，这些方法的恢复率和预测精度都在持续提高。不过大部分蛋白设计工作结构类型不确定，因此可变骨架设计在设计蛋白质时需要优化同时序列和结构。深度学习技术通过识别结构的模式，可以极大地提高这些模式的辨识度并输出增强的特征。

2021 年，一个新颖的可变骨架蛋白质设计技术[69]被提出，该技术是训练一个可以迭代优化序列的网络，被称为“幻想(hallucination)设计”。该方法使用 TrRosetta 来预测序列的空间约束，并利用 Kullback-Leibler 散度进行迭代优化，使得预测的结构更加接近真实蛋白的空间分布。在 TrDesign-motif[70]的应用案例中，它结合 TrRosetta 和幻想设计，通过优化目标分布与背景噪声分布的结构和序列，专门设计活性位点和功能 motif。RFDesign 则通过“constrained hallucination”来优化蛋白质设计[71]。“Inpainting”方法使用 RoseTTAFold[18]在补全方面作出了创新，它能够同时生成序列和相应的结构。这些方法已应用于创建结构上有效且具有潜在生物活性的金属蛋白和酶等，且通过实验验证。然而，RFDesign 在使用 RoseTTAFold 时面临挑战，受限于单次的运行预测结构。为了提升蛋白质从头设计的质量和多样性，新工具 AutoFoldFinder[72]采用 CM-Align 进行优化设计，这一方法优化了序列并生成结构，克服了传统 KL 散度方法的局限，大幅提高了新蛋白质结构的比例，同时显著提升了从头设计的质量，产生了与现有结构差异显著的蛋白质。

在近期的研究中，扩散模型在蛋白质设计领域的应用展现出显著的潜力。2022 年，有研究者开发了使用等变去噪扩散概率模型(equivariant denoising diffusion probabilistic models)来生成蛋白质结构和序列的方法[73]。该方法能够从实验数据中学习并生成具有特定三维结构和化学性质的蛋白质，以实现目标功能。紧接着，2023 年，研究人员对 RoseTTAFold 结构预测网络进行微调[20]，将其应用于蛋白质结构去除噪声任务，并开发出一种名为 RFdiffusion 的蛋白质骨架的生成模型；该模型在多个领域表现出色，包括无条件蛋白质单体设计、蛋白质结合剂设计、对称寡聚体设计、酶活性位点和对称基序支架设

计等，特别适用于解决蛋白质设计中的复杂问题，如从头设计结合蛋白、设计具有特定对称性的蛋白质结构以及创建新的蛋白质药物和疫苗；该团队使用了来自蛋白质数据库(protein database, PDB)的结构为基础，并对其施加多至200 步的噪声，以此作为训练输入；RFdiffusion模型利用最小化的预测与未对齐的真实蛋白质结构之间的均方误差损失进行训练，实现噪声过程逐步逆转；在序列设计方面，采用 ProteinMPNN对结构所对应的序列进行编码，通常对每个生成8 个序列变体；尽管 RFdiffusion 在结构和序列设计方面都具有潜力，可是由于其与 ProteinMPNN单独结合已显示出良好的表现，所以并未广泛探索这种潜力。研究者通过实验验证了数百种设计的对称组装体、金属结合蛋白及蛋白质结合剂的结构和功能，证实了 RoseTTAFold 扩散方法的强大功能和广泛适用性[74]。其中，RFdiffusion 作为一种前沿的蛋白质设计技术，在实验层面已成功构建了多样的蛋白质结构，包括具有创新性拓扑特征的蛋白质、对称性寡聚体、酶活性位点的支架以及金属结合位点的支架。该技术能够设计出具有独特 α/β 桶状结构的蛋白质，并能精确地设计出与特定靶标如SARS-CoV-2 病毒的刺突蛋白相结合的作用蛋白。此外，RFdiffusion 还能够设计出具有特定对称性，例如 C3 对称性的三聚体结构，这些设计在电子显微镜下展现出与设计模型高度一致的结构，从而证明了其在创新蛋白质结构生成方面的准确性和潜在应用价值。尽管 RFdiffusion在蛋白质设计领域展现出显著优势，能够创造出具有新颖结构和功能的蛋白质，并且其设计的蛋白质在实验中被证实具有预期的结构和功能，但也存在一些局限性。首先，设计复杂性受到现有计算资源和算法效率的限制，对于大型蛋白质结构的预测可能超出了单序列预测的能力。此外，设计过程可能需要结合其他工具以确保序列的稳定性；对称性设计的限制需要进一步地优化，以确保蛋白质在生物体内的结构稳定性和功能。同时，在实际应用中还需要克服包括外部条件依赖性、生物活性和免疫原性的考量、知识产权和伦理问题、成本和可及性以及软件和算法的局限性等障碍。尽管面临这些挑战，RFdiffusion 在蛋白质设计领域的应用前景仍然非常广阔，尤其是在药物开发和生物技术领域，它为解决复杂的设计挑战提供了新的可能。

2024 年，David Baker 团队在蛋白质设计领域作出了显著贡献，他们提出了一种名为全原子模型(RoseTTAFold All-Atom, RFAA)的先进计算方法[74]。RFAA 通过在去噪任务上的精细调整，进一步发展为 RFdiffusionAA 模型，该模型在多种应用场景中展现出广泛的适用性，包括药物设计、蛋白质工程和生物标志物的识别。RFAA 模型特别擅长设计能够与特定生物标志物特异性结合并调节其功能的蛋白质，以及设计能够靶向并破坏病原体(如病毒或细菌)的蛋白质。例如，该模型已被用于围绕小分子构建蛋白质结构，并通过晶体学和结合测量实验验证了其设计的蛋白质与心脏疾病治疗剂地高辛、酶辅因子血红素和光捕获分子胆红素的结合，显示出高特异性和稳定性。RFAA 的主要优势在于能够处理广泛的生物分子复合体[74]，并在预测蛋白质-小分子复合体结构方面展现出高精度。然而，RFAA 的性能在很大程度上依赖于训练数据集的质量和多样性。如果训练数据集中缺少特定类型的生物分子或蛋白质-小分子相互作用，RFAA 在预测这些特定情况时可能会受到限制。此外，作为一个复杂的深度学习模型，RFAA 的训练和运行需要大量的计算资源，这可能限制了其在资源

有限的环境中的应用。在处理与训练数据集差异较大的新蛋白质或小分子时，RFAA 的预测准确性可能也存在局限。虽然 RFAA 旨在模拟广泛的生物分子组装体，但在特定的应用场景中，可能需要更特化的模型来提高预测精度。将 RFAA 的预测转化为实际的蛋白质设计和实验验证也面临一系列挑战，包括实验条件的优化、蛋白质表达和纯化等。尽管存在这些限制，RFAA 在药物发现和生物分子设计领域仍具有广泛的应用前景，特别是在需要精确模拟和设计蛋白质-小分子相互作用的场景中。

在同期的研究中，该实验室利用经过微调的 RFdiffusion 网络，成功设计了针对特定抗原表位的抗体。这些抗体经过冷冻电镜解析，证实与设计结果高度一致，能够特异性地结合指定的表位[75]。该研究的应用主要集中在快速开发抗体药物和个性化医疗领域。例如，在面对新出现的病毒如 SARS-CoV-2 时，传统的抗体发现方法可能耗时数月甚至数年，而该研究展示的从头设计方法能在几周内设计出与病毒表面特定蛋白(例如受体结合域)结合的抗体。这些抗体设计具有高亲和力和特异性，能有效中和病毒，阻止其感染细胞。研究中设计的抗体与流感血凝素的结合，冷冻电镜结构与设计模型高度一致，显示出原子级的精确度；抗体与流感的结合亲和力达到了 78 nmol，尽管这一亲和力相对较低，但在未经优化的情况下，该结果已经证明了设计方法的有效性；此外，抗体与目标表位的结合显示出高度特异性，如针对 TcdB 细菌毒素的抗体设计在实验中显示出特异性结合，而对结构相似的 TcsL 细菌毒素则没有结合活性[75]。总体而言，该研究提供了一种快速且精确的新型抗体设计方法。其创新之处在于利用 RFdiffusion 网络和 RoseTTAFold2 结构预测网络，实现了从头设计抗体的突破，为抗体药物的快速开发提供了新的可能。然而，目前设计的抗体与目标的亲和力相对较低，设计成功率也有提升空间。该方法适用于需要特定表位结合的抗体设计，特别是在缺乏天然抗体或需要快速开发新型抗体药物的情况下。此外，该方法的通用性可能适用于多种疾病相关抗原，为个性化医疗和精准治疗提供了新的工具，标志着结构引导的大分子抗体设计领域的一个新起点。这不仅为抗体药物的快速和成本效益高的开发开辟了新途径，也为扩散模型在生物工程中的进一步应用奠定了基础。

在国内，刘海燕研究团队提出了一个名为 SCUBA 的全新方法[76]，该方法使用的是一种神经网络形式的能量项统计模型，实现了基于连续采样和优化主链能量的新的蛋白主链骨架设计。SCUBA 模型融合了局部构象倾向性、氢键几何构形以及侧链所需骨架空间等关键因素，利用邻接计数和神经网络方法进行训练，以此生成全新设计的蛋白质主链骨架；并且采用随机动力学和模拟退火算法，结合之前提到的 ABACUS2 技术进行序列优化和骨架设计的迭代过程，实现了可变骨架蛋白的从头设计。研究表明，SCUBA 设计的骨架不仅具有比自然蛋白结构更高的热稳定性，而且序列同一性低，其中约 42%的设计蛋白质能够实现正确的折叠[76]。同期另一项研究介绍了 SCUBA-D[77]，这是一种通过去除噪声扩散的方法，该方法可以从噪声较大的原始骨架中生成高质量骨架。SCUBA-D 的创新特点是引入了基于序列语言模型的扩散辅助和若干生成对抗性网络(generative adversarial networks, GAN)式判别器，这些工具共同增强了产生骨架的物理合理性。SCUBA-D 的效果可以通过在生成的骨架上设计氨基酸序列并利用结构预测进行评估。这些工具还有一定的局限性，例如无法设计大分子蛋白质且成功率

比较低、应用场景较小。

综上，在深度学习领域，通过构建能量函数来捕捉骨架结构与序列之间的关键特征已成为一种有效的策略。与传统的分子模拟方法相比，深度学习在模拟范德华力、氢键、亲疏水性以及其他相互作用关系方面展现出了显著的优势。特别是扩散模型的引入不仅显著提高了预测的准确性，而且为设计具有特定功能的蛋白质提供了一种强大的工具，从而推动了功能蛋白设计领域的进步。此外，深度学习技术的应用促进了蛋白质设计方法的演变，从传统的小蛋白设计逐步过渡到大分子抗体的从头设计，并且在此过程中显著提高了设计的效率。深度学习在基于骨架结构-序列的蛋白质设计方法中扮演着至关重要的角色，为探索已知蛋白质结构空间提供了新的可能。

2.2 不依赖于结构的直接序列生成

如前所述，在蛋白质设计领域，核心目标是找到能稳定展示所需特性并履行其功能的序列。由于蛋白质序列数据较为丰富，且信息流中结构数据过多的转换和中继点可能导致信号偏移，故直接在序列与功能空间建立映射的方法具有潜在优势[78]。相较于在特定骨架上搜索适配度，直接序列设计一旦掌握了序列空间重要分布，便可以指导设计过程，无需依赖预先获得的结构信息。因此深度学习在不依赖结构的直接序列设计方法中展现出强大潜力。

2017 年，一种新颖的策略被引入，该方法使用变分自编码器(variational autoencoder, VAE)嵌入天然蛋白质序列[79]，能够预测蛋白质突变对功能的影响。利用这种无监督学习技术可以捕捉天然蛋白质的变异，并识别出特定位点间的相互作用模式。与传统的不考虑序列间相互作用的基线方法相比，此策略展现了更高的性能，有时超越了利用 Inverse Potts Model[80]的先进技术。作为生成模型，VAE 可以指导蛋白质序列空间的探索，从而增强蛋白质设计的合理性和自动化程度。紧接着，一种改进的生成对抗网络模型被提出[81]，使用生成的对抗性网络(wasserstein generative adversarial networks, W-GAN)用于生成预测具有特定抗生素抗药性的蛋白质序列。W-GAN 模型由于能有效模拟真实数据分布而获得认可，并可生成风格类似于初始训练数据的新数据。然而，上述生成模型在某些方面存在局限性，如生成特定功能的序列时精度不足。2019 年，研究人员进一步开发了蛋白质溶解度生成对抗性网络(protein solubility generation adversarial network, ProGAN)的数据增强算法[82]，此深度神经网络能从序列预测蛋白质溶解度，并借助 ProGAN 提高预测的准确性；该研究结果表明，ProGAN 生成的数据能提升模型的预测性能相比先前同类研究提高了约 10%。

此外，还有研究利用长短期记忆递归神经网络(long short-term memory recurrent neural network, LSTM RNN)单元开发了创新组合和设计多肽序列的新方法[83]。LSTM RNN能捕获顺序数据中的特征，并根据学习到的上下文特征生成新的序列。研究者以氨基酸序列作为输入，针对螺旋抗菌肽的设计模式进行了 LSTM 的训练，并成功从头生成了 82%具有预测抗菌活性的潜在抗菌肽的序列，所生成的序列比随机序列更贴近于训练数据[83]。LSTM RNN作为一种循环神经网络，解决了传统 RNN 训练中的梯度问题，其循环连接和单元状态帮助保持长期的数据关联。这些技术已成功应用于蛋白质序列分析领域，包括二级结构预测、同源性检测和亚细胞区室定位等。对抗菌肽的系统训练的研究揭示了 LSTM RNN 在肽和蛋白质设计中的潜能。

2021 年，有研究者运用 VAE 拟合细菌荧光素酶蛋白在序列水平上的分布，并通过解码邻近的潜在载体生成新的变异靶蛋白[84]。它在近 70 000 个萤光素酶的数据集上经过训练，由条件模型生成的所有 23 个变体都保留了发光功能[84]。这一结果证实了深度生成模型探索蛋白质可能的序列空间，并成功生成新蛋白的可行性，为传统的合理设计和定向进化方法提供了一个补充策略。2023 年，一个新的密集-自动生成的对抗性网络(dense-auto generative adversarial networks, Dense-AutoGAN)模型[85]被提出，该模型融合了注意力机制与 GAN，旨在创造出新的蛋白质序列。此模型结合注意力机制和编码器-解码器框架，不仅提升了蛋白质序列的生成质量，还实现了在保留原始特征的同时进行微调。此外，Dense-AutoGAN 采用密集连接卷积神经网络的结构，在 GAN 的生成器中实现了多层的特征传递，既拓展了训练空间，又增强了序列生成的效率。在蛋白质功能图谱上，该模型成功生成了更为复杂性的蛋白质序列。通过与其他模型进行比较，Dense-AutoGAN 生成的序列在化学和物理性质上显示出了高准确性和效率，证实了该模型的优越性能。

在蛋白质设计领域，VAE 和 GAN 已被证实可以有效预测和生成具有改进的突变序列。经过模型训练，研究人员能够成功预测实验突变扫描中的突变，并通过实验方式验证了这些序列的功能。随着测序技术的迅猛发展，蛋白质序列数据按照指数级别增加，大量积累了未标记序列。这些序列为大语言模型应用于蛋白质设计领域(表 1)奠定了坚实基础。这使得我们能更加深入理解蛋白质序列与功能的关系。

大型语言模型，如 ESM-1b 大型 Transformer 模型[87]，已成功解析数亿条蛋白质序列，深入挖掘了生物特性。这些模型不仅掌握了蛋白质的二级结构，还成功地捕获了蛋白质多层面的空间组织原则，涵盖了从物理化学属性到远程同源性等多个维度。同样地，在 2023 年，一种名为 ProGen 模型[93]被开发出来，这一基于 Transformer 的语言模型能通过标记不同属性的氨基酸序列进行训练，控制生成带有特定属性且天然蛋白相似的蛋白质。同时，ProtTrans 模型[91]在总量为 2 亿的蛋白质序列上进行了自回归和自动编码器的训练，并通过它展示了无监督语言模型在蛋白质生物物理学特性学习方面的有效性。UniRep 模型[86]通过预训练大量未标记的氨基酸序列，有效提取了蛋白质序列的深层特征，并准确预测了天然和从头设计蛋白质

表 1　蛋白质序列大语言模型

Table 1　Large language model for protein sequences

Year	Model	Task
2019	UniRep	Unified rational protein engineering with sequence-based deep representation learning[86]
2021	ESM-1b	Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[87]
2021	ECNet	ECNet is an evolutionary context-integrated deep learning framework for protein engineering[88]
2021	ESM-1v	Language models enable zero-shot prediction of the effects of mutations on protein function[89]
2021	Low-N	Low-N protein engineering with data-efficient deep learning[90]
2022	ProtTrans	ProtTrans: toward understanding the language of life through self-supervised learning[91]
2023	ProtTucker	Protein remote homology detection and structural alignment using deep learning[92]
2023	ProGen	Large language models generate functional protein sequences across diverse families[93]
2023	ProGen2	Exploring the boundaries of protein language models[94]

的稳定性。借助大数据集预训练可以增强模型的迁移学习能力，进而为特定任务进行微调提供可能。

2022 年，ProtGPT2 模型[95]的引入进一步推动了这一领域的发展，这个拥有 7.38 亿参数的自回归模型，经过 Uniref-50 数据集的训练；ProtGPT2 生成的蛋白质中 88%为球形结构，与自然序列相符；对蛋白质数据库的深度搜索结果显示，ProtGPT2 序列与自然蛋白序列存在远程同源性；相似性网络分析也表明，ProtGPT2 有效地探索了蛋白质功能空间中的新区域。AlphaFold2 对 ProtGPT2 序列进行结构预测生成了具有实验价值和大环特征的良好折叠非理想化结构，展示了新的拓扑种类。此外，这些生成的序列在稳定性和动态特性方面与天然蛋白质序列类似，并在进化序列空间中展现出独特性。在 2023 年，ProGen2 模型[94]被推出，该模型是一种规模更大的自回归 Transformer，拥有 6.4 亿参数，并在超过 10 亿种蛋白质序列的多样化数据集上进行训练。经过全面评估，ProGen2 在蛋白质序列生成、经微调的特殊结构序列以及特定于抗体序列生成方面均表现卓越，突显其在合理序列生成上的优势。2024 年，国内之江实验室的科研团队开发了指导蛋白质语言模型(instruct protein language models, InstructPLM)蛋白质设计框架[96]，其利用了大型语言模型的跨模态对齐和指令微调技术来指导蛋白质语言模型生成符合特定结构要求的蛋白质序列。这种方法特别适用于需要设计具有特定功能和结构的蛋白质的场景，例如药物开发、生物技术应用和材料科学。经实验验证，成功从头设计具有所需催化特性的酶、结合特性的蛋白质序列、特定结构和功能的蛋白质基材料等。InstructPLM 提供了一种强大的工具，用于设计具有特定功能和结构的蛋白质，这在药物开发、生物技术和材料科学等领域具有重要的应用价值。

深度学习在特征提取、模式识别和目标生成方面具有先进能力。在蛋白质设计方面，直接从序列出发的方法因其不依赖预先获得的结构及大量序列信息而显示出优势，尤其是基于深度学习的策略，如大语言模型在不依赖结构的序列设计中表现出强大潜力。这些策略与基于结构的蛋白设计不同，它通过学习序列与其结构功能的关系，能够直接探索序列空间，从而带来蛋白质设计范式的创新[97]。

3 总结与展望

近几年，深度学习等人工智能技术已使蛋白设计取得质的飞跃。AI 模型凭借其强大的数据处理能力和模式识别能力，在蛋白质结构与序列数据分析上作出了突破性的贡献。通过复杂的网络架构和算法，AI 能够识别出关键生物学特征，进而构建预测模型以辅助不同的设计场景。例如，某些深度学习模型设计的蛋白质，在实验中验证具备预定的结构和功能。这不仅证明了 AI 在蛋白设计中的实用性，也显著提升了设计过程的效率。

尽管 AI 在这一领域取得了一定进展，但仍存在问题和挑战。首先，蛋白质结构数据库数据规模不足，限制了模型的准确性和泛化能力。这些不足可能通过数据增强、迁移学习等方法得到缓解。其次，对蛋白设计模型性能的现行评估体系，如天然序列恢复率和预测结构之间的差异[52]，未能充分反映蛋白设计的物理化学复杂性，缺乏严格和标准化的基准。另外，蛋白质功能的实现往往涉及到动态过程，现有模型通常只关注单一功能状态，而较少综合考虑表达性、溶解性、稳定性和免疫原性等多样属性。最重要的是，实验验证仍是确认 AI

在蛋白质设计中的关键一环。

随着计算能力增强、算法优化以及新一代高通量实验技术的涌现，研究人员将获得更全面和更深入的生物数据。这将使 AI 模型的训练更充分，加深对生命过程的洞察，推动设计更复杂精准的蛋白质，使得深度学习模型能够更有效处理复杂的生物数据。未来，人工智能将整合表观遗传学、转录组学和蛋白质组学等生物信息，更准确地模拟生物分子交互，设计出具有特定功能的蛋白质。此外，厘清深度学习模型的“黑箱”机制，增强模型的精确性和计算过程的可解释性[97]，将提高药物设计在酶工程、合成生物学以及蛋白工程等领域中的成功率。

最后，随着蛋白质设计的要求越来越个性化和精准化，定制功能蛋白质将成为新常态。人工智能技术将在预测和设计特异性功能、提高设计效率、优化生物合成路径等方面发挥核心作用。这些技术的进步将推动蛋白质设计在个性化医疗、疾病治疗和生物制造等宽广领域的革新，并开启蛋白质从头设计的新时代。

REFERENCES

[1] MIKLOS GLG, MALESZKA R. Protein functions and biological contexts[J]. PROTEOMICS, 2001, 1(2): 169-178.

[2] ANFINSEN CB. Principles that govern the folding of protein chains[J]. Science, 1973, 181(4096): 223-230.

[3] HUANG PS, BOYKEN SE, BAKER D. The coming of age of *de novo* protein design[J]. Nature, 2016, 537: 320-327.

[4] ROMERO PA, ARNOLD FH. Exploring protein fitness landscapes by directed evolution[J]. Nature Reviews Molecular Cell Biology, 2009, 10: 866-876.

[5] MEINEN BA, BAHL CD. Breakthroughs in computational design methods open up new frontiers for *de novo* protein engineering[J]. Protein Engineering, Design and Selection, 2021, 34: gzab007.

[6] LEMAN JK, WEITZNER BD, LEWIS SM, ADOLF-BRYFOGLE J, ALAM N, ALFORD RF, APRAHAMIAN M, BAKER D, BARLOW KA, BARTH P, BASANTA B, BENDER BJ, BLACKLOCK K, BONET J, BOYKEN SE, BRADLEY P, BYSTROFF C, CONWAY P, COOPER S, CORREIA BE, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks[J]. Nature Methods, 2020, 17: 665-680.

[7] SUÁREZ M, TORTOSA P, JARAMILLO A. PROTDES: CHARMM toolbox for computational protein design[J]. Systems and Synthetic Biology, 2008, 2(3): 105-113.

[8] WOOD CW, HEAL JW, THOMSON AR, BARTLETT GJ, IBARRA AÁ, BRADY RL, SESSIONS RB, WOOLFSON DN. ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design[J]. Bioinformatics, 2017, 33(19): 3043-3050.

[9] GUEROIS R, NIELSEN JE, SERRANO L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations[J]. Journal of Molecular Biology, 2002, 320(2): 369-387.

[10] O’MEARA MJ, LEAVER-FAY A, TYKA MD, STEIN A, HOULIHAN K, DiMAIO F, BRADLEY P, KORTEMME T, BAKER D, SNOEYINK J, KUHLMAN B. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta[J]. Journal of Chemical Theory and Computation, 2015, 11(2): 609-622.

[11] ALFORD RF, LEAVER-FAY A, JELIAZKOV JR, O’MEARA MJ, DiMAIO FP, PARK H, SHAPOVALOV MV, RENFREW PD, MULLIGAN VK, KAPPEL K, LABONTE JW, PACELLA MS, BONNEAU R, BRADLEY P, DUNBRACK RL Jr, DAS R, BAKER D, KUHLMAN B, KORTEMME T, GRAY JJ. The Rosetta all-atom energy function for macromolecular modeling and design[J]. Journal of Chemical Theory and Computation, 2017, 13(6): 3031-3048.

[12] CURNOW P. Designing minimalist membrane proteins[J]. Biochemical Society Transactions, 2019, 47(5): 1233-1245.

[13] PARK S, FU STOWELL X, WANG W, YANG X, SAVEN JG. Computational protein design and discovery[J]. Annual Reports Section "C" (Physical Chemistry), 2004, 100(0): 195-236.

[14] YANOVER C, FROMER M, SHIFMAN JM. Dead-end

elimination for multistate protein design[J]. Journal of Computational Chemistry, 2007, 28(13): 2122-2129.

[15] FLOUDAS CA, FUNG HK, McALLISTER SR, MÖNNIGMANN M, RAJGARIA R. Advances in protein structure prediction and *de novo* protein design: a review[J]. Chemical Engineering Science, 2006, 61(3): 966-988.

[16] ROMERO-RIVERA A, GARCIA-BORRÀS M, OSUNA S. Computational tools for the evaluation of laboratory-engineered biocatalysts[J]. Chemical Communications, 2016, 53(2): 284-297.

[17] JUMPER J, EVANS R, PRITZEL A, GREEN T, FIGURNOV M, RONNEBERGER O, TUNYASUVUNAKOOL K, BATES R, ŽÍDEK A, POTAPENKO A, BRIDGLAND A, MEYER C, KOHL SAA, BALLARD AJ, COWIE A, ROMERA-PAREDES B, NIKOLOV S, JAIN R, ADLER J, BACK T, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596: 583-589.

[18] BAEK M, DiMAIO F, ANISHCHENKO I, DAUPARAS J, OVCHINNIKOV S, LEE GR, WANG J, CONG Q, KINCH LN, SCHAEFFER RD, MILLÁN C, PARK H, ADAMS C, GLASSMAN CR, DeGIOVANNI A, PEREIRA JH, RODRIGUES AV, van DIJK AA, EBRECHT AC, OPPERMAN DJ, et al. Accurate prediction of protein structures and interactions using a three-track neural network[J]. Science, 2021, 373(6557): 871-876.

[19] SUN MGF, SEO MH, NIM S, CORBI-VERGE C, KIM PM. Protein engineering by highly parallel screening of computationally designed variants[J]. Science Advances, 2016, 2(7): e1600692.

[20] WATSON JL, JUERGENS D, BENNETT NR, TRIPPE BL, YIM J, EISENACH HE, AHERN W, BORST AJ, RAGOTTE RJ, MILLES LF, WICKY BIM, HANIKEL N, PELLOCK SJ, COURBET A, SHEFFLER W, WANG J, VENKATESH P, SAPPINGTON I, TORRES SV, LAUKO A, et al. *De novo* design of protein structure and function with RFdiffusion[J]. Nature, 2023, 620: 1089-1100.

[21] BRYNGELSON JD, ONUCHIC JN, SOCCI ND, WOLYNES PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis[J]. Proteins, 1995, 21(3): 167-195.

[22] DeGRADO WF, REGAN L, HO SP. The design of a four-helix bundle protein[J]. Cold Spring Harbor Symposia on Quantitative Biology, 1987, 52: 521-526.

[23] REGAN L, DeGRADO WF. Characterization of a helical protein designed from first principles[J]. Science, 1988, 241(4868): 976-978.

KORENDOVYCH IV, DeGRADO WF. *De novo* protein design, a retrospective[J]. Quarterly Reviews of Biophysics, 2020, 53: e3.

[24] WOOLFSON DN. A brief history of *de novo* protein design: minimal, rational, and computational[J]. Journal of Molecular Biology, 2021, 433(20): 167160.

[25] CHEVALIER A, SILVA DA, ROCKLIN GJ, HICKS DR, VERGARA R, MURAPA P, BERNARD SM, ZHANG L, LAM KH, YAO GR, BAHL CD, MIYASHITA SI, GORESHNIK I, FULLER JT, KODAY MT, JENKINS CM, COLVIN T, CARTER L, BOHN A, BRYAN CM, et al. Massively parallel *de novo* protein design for targeted therapeutics[J]. Nature, 2017, 550: 74-79.

[26] CAO LX, GORESHNIK I, COVENTRY B, CASE JB, MILLER L, KOZODOY L, CHEN RE, CARTER L, WALLS L, PARK YJ, STEWART L, DIAMOND M, VEESLER D, BAKER D. *De novo* design of picomolar SARS-CoV-2 miniprotein inhibitors[J]. bioRxiv: the Preprint Server for Biology, 2020: 2020.08.03.234914.

[27] YEH AHW, NORN C, KIPNIS Y, TISCHER D, PELLOCK SJ, EVANS D, MA PC, LEE GR, ZHANG JZ, ANISHCHENKO I, COVENTRY B, CAO LX, DAUPARAS J, HALABIYA S, DeWITT M, CARTER L, HOUK KN, BAKER D. *De novo* design of luciferases using deep learning[J]. Nature, 2023, 614: 774-780.

[28] DAUPARAS J, ANISHCHENKO I, BENNETT N, BAI H, RAGOTTE RJ, MILLES LF, WICKY BIM, COURBET A, de HAAS RJ, BETHEL N, LEUNG PJY, HUDDY TF, PELLOCK S, TISCHER D, CHAN F, KOEPNICK B, NGUYEN H, KANG A, SANKARAN B, BERA AK, et al. Robust deep learning-based protein sequence design using ProteinMPNN[J]. Science, 2022, 378(6615): 49-56.

[29] MARCOS E, SILVA DA. Essentials of *de novo* protein design: methods and applications[J]. WIREs Computational Molecular Science, 2018, 8(6): e1374.

[30] LIN YR, KOGA N, TATSUMI-KOGA R, LIU GH, CLOUSER AF, MONTELIONE GT, BAKER D. Control over overall shape and size in *de novo* designed proteins[J]. Proceedings of the National

Academy of Sciences of the United States of America, 2015, 112(40): E5478-E5485.

[31] SCHAFMEISTER CE, LaPORTE SL, MIERCKE LJW, STROUD RM. A designed four helix bundle protein with native-like structure[J]. Nature Structural Biology, 1997, 4: 1039-1046.

[32] BHARDWAJ G, MULLIGAN VK, BAHL CD, GILMORE JM, HARVEY PJ, CHENEVAL O, BUCHKO GW, PULAVARTI SVSRK, KAAS Q, ELETSKY A, HUANG PS, JOHNSEN WA, GREISEN P Jr, ROCKLIN GJ, SONG YF, LINSKY TW, WATKINS A, RETTIE SA, XU XZ, CARTER LP, et al. Accurate *de novo* design of hyperstable constrained peptides[J]. Nature, 2016, 538: 329-335.

[33] JACOBS TM, WILLIAMS B, WILLIAMS T, XU X, ELETSKY A, FEDERIZON JF, SZYPERSKI T, KUHLMAN B. Design of structurally distinct proteins using strategies inspired by evolution[J]. Science, 2016, 352(6286): 687-690.

[34] TYKA MD, JUNG K, BAKER D. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers[J]. Journal of Computational Chemistry, 2012, 33(31): 2483-2491.

[35] ROHL CA, STRAUSS CEM, MISURA KMS, BAKER D. Protein structure prediction using Rosetta[J]. Methods in Enzymology, 2004, 383: 66-93.

[36] ROCKLIN GJ, CHIDYAUSIKU TM, GORESHNIK I, FORD A, HOULISTON S, LEMAK A, CARTER L, RAVICHANDRAN R, MULLIGAN VK, CHEVALIER A, ARROWSMITH CH, BAKER D. Global analysis of protein folding using massively parallel design, synthesis, and testing[J]. Science, 2017, 357(6347): 168-175.

[37] KUHLMAN B, BAKER D. Native protein sequences are close to optimal for their structures[J]. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(19): 10383-10388.

[38] XIONG P, WANG M, ZHOU XQ, ZHANG TC, ZHANG JH, CHEN Q, LIU HY. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability[J]. Nature Communications, 2014, 5: 5330.

[39] XIONG P, HU XH, HUANG B, ZHANG JH, CHEN Q, LIU HY. Increasing the efficiency and accuracy of the ABACUS protein sequence design method[J]. Bioinformatics, 2020, 36(1): 136-144.

[40] JOHANSSON KE, TIDEMAND JOHANSEN N, CHRISTENSEN S, HOROWITZ S, BARDWELL JCA, OLSEN JG, WILLEMOËS M, LINDORFF-LARSEN K, FERKINGHOFF-BORG J, HAMELRYCK T, WINTHER JR. Computational redesign of thioredoxin is hypersensitive toward minor conformational changes in the backbone template[J]. Journal of Molecular Biology, 2016, 428(21): 4361-4377.

[41] TYKA MD, KEEDY DA, ANDRÉ I, DIMAIO F, SONG YF, RICHARDSON DC, RICHARDSON JS, BAKER D. Alternate states of proteins revealed by detailed energy landscape mapping[J]. Journal of Molecular Biology, 2011, 405(2): 607-618.

[42] LU PL, MIN D, DiMAIO F, WEI KY, VAHEY MD, BOYKEN SE, CHEN ZB, FALLAS JA, UEDA G, SHEFFLER W, MULLIGAN VK, XU WQ, BOWIE JU, BAKER D. Accurate computational design of multipass transmembrane proteins[J]. Science, 2018, 359(6379): 1042-1046.

[43] HOSSEINZADEH P, BHARDWAJ G, MULLIGAN VK, SHORTRIDGE MD, CRAVEN TW, PARDO-AVILA F, RETTIE SA, KIM DE, SILVA DA, IBRAHIM YM, WEBB IK, CORT JR, ADKINS JN, VARANI G, BAKER D. Comprehensive computational design of ordered peptide macrocycles[J]. Science, 2017, 358(6369): 1461-1466.

[44] HUANG PS, FELDMEIER K, PARMEGGIANI F, FERNANDEZ VELASCO DA, HÖCKER B, BAKER D. *De novo* design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy[J]. Nature Chemical Biology, 2016, 12: 29-34.

[45] MARCOS E, BASANTA B, CHIDYAUSIKU TM, TANG YF, OBERDORFER G, LIU GH, SWAPNA GVT, GUAN RJ, SILVA DA, DOU JY, PEREIRA JH, XIAO R, SANKARAN B, ZWART PH, MONTELIONE GT, BAKER D. Principles for designing proteins with cavities formed by curved β sheets[J]. Science, 2017, 355(6321): 201-206.

[46] CORREIA BE, BATES JT, LOOMIS RJ, BANEYX G, CARRICO C, JARDINE JG, RUPERT P, CORRENTI C, KALYUZHNIY O, VITTAL V, CONNELL MJ, STEVENS E, SCHROETER A, CHEN M, MacPHERSON S, SERRA AM, ADACHI Y, HOLMES MA, LI YX, KLEVIT RE, et al. Proof of principle for epitope-focused vaccine design[J]. Nature, 2014, 507:

201-206.

[47] FLEISHMAN SJ, WHITEHEAD TA, EKIERT DC, DREYFUS C, CORN JE, STRAUCH EM, WILSON IA, BAKER D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin[J]. Science, 2011, 332(6031): 816-821.

[48] SHEN H, FALLAS JA, LYNCH E, SHEFFLER W, PARRY B, JANNETTY N, DECARREAU J, WAGENBACH M, VICENTE JJ, CHEN JJ, WANG L, DOWLING Q, OBERDORFER G, STEWART L, WORDEMAN L, de YOREO J, JACOBS-WAGNER C, KOLLMAN J, BAKER D. *De novo* design of self-assembling helical protein filaments[J]. Science, 2018, 362(6415): 705-709.

[49] MADANI A, KRAUSE B, GREENE E, SUBRAMANIAN S, MOHR B, HOLTON J, LUIS-OLMOS J, XIONG C, SUN Z, SOCHER R, FRASER J, NAIK N. Deep neural language modeling enables functional protein generation across families[J]. bioRxiv, 2021. DOI:10.1101/2021.07.18.452833.

[50] LeCUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521: 436-444.

[51] 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[J]. 合成生物学, 2023, 4(3): 464-487.
CHEN ZH, JI ML, QI YF. Research progress of artificial intelligence in desiging protein structures[J]. Synthetic Biology Journal, 2023, 4(3): 464-487 (in Chinese).

[52] WANG S, SUN SQ, LI Z, ZHANG RY, XU JB. Accurate *de novo* prediction of protein contact map by ultra-deep learning model[J]. PLoS Computational Biology, 2017, 13(1): e1005324.

[53] HANSON J, PALIWAL K, LITFIN T, YANG YD, ZHOU YQ. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks[J]. Bioinformatics, 2018, 34(23): 4039-4045.

[54] SHEN T, WU JX, LAN HD, ZHENG LZ, PEI JG, WANG S, LIU W, HUANG JZ. When homologous sequences meet structural decoys: accurate contact prediction by tFold in CASP14-(tFold for CASP14 contact prediction)[J]. Proteins, 2021, 89(12): 1901-1910.

[55] DING WZ, MAO WZ, SHAO D, ZHANG WX, GONG HP. DeepConPred2: an improved method for the prediction of protein residue contacts[J]. Computational and Structural Biotechnology Journal, 2018, 16: 503-510.

[56] LI Y, HU J, ZHANG CX, YU DJ, ZHANG Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks[J]. Bioinformatics, 2019, 35(22): 4647-4655.

[57] YANG JY, ANISHCHENKO I, PARK H, PENG ZL, OVCHINNIKOV S, BAKER D. Improved protein structure prediction using predicted interresidue orientations[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(3): 1496-1503.

[58] XU JB. Distance-based protein folding powered by deep learning[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(34): 16856-16865.

[59] DING WZ, GONG HP. Predicting the real-valued inter-residue distances for proteins[J]. Advanced Science, 2020, 7(19): 2001314.

[60] SENIOR AW, EVANS R, JUMPER J, KIRKPATRICK J, SIFRE L, GREEN T, QIN CL, ŽÍDEK A, NELSON AWR, BRIDGLAND A, PENEDONES H, PETERSEN S, SIMONYAN K, CROSSAN S, KOHLI P, JONES DT, SILVER D, KAVUKCUOGLU K, HASSABIS D. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577: 706-710.

[61] DAHIYAT BI, MAYO SL. *De novo* protein design: fully automated sequence selection[J]. Science, 1997, 278(5335): 82-87.

[62] O'CONNELL J, LI ZX, HANSON J, HEFFERNAN R, LYONS J, PALIWAL K, DEHZANGI A, YANG YD, ZHOU YQ. SPIN2: predicting sequence profiles from protein structures using deep neural networks[J]. Proteins, 2018, 86(6): 629-633.

[63] CHEN S, SUN Z, LIN LH, LIU ZF, LIU X, CHONG YT, LU YT, ZHAO HY, YANG YD. To improve protein sequence profile prediction through image captioning on pairwise residue distance map[J]. Journal of Chemical Information and Modeling, 2020, 60(1): 391-399.

[64] QI YF, ZHANG JZH. DenseCPD: improving the accuracy of neural-network-based computational protein sequence design with DenseNet[J]. Journal of Chemical Information and Modeling, 2020, 60(3): 1245-1252.

[65] NORN C, WICKY BIM, JUERGENS D, LIU SR, KIM

D, TISCHER D, KOEPNICK B, ANISHCHENKO I, PLAYERS F, BAKER D, OVCHINNIKOV S. Protein sequence design by conformational landscape optimization[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(11): e2017228118.

[66] LIU YF, ZHANG L, WANG WL, ZHU M, WANG CC, LI FD, ZHANG JH, LI HQ, CHEN Q, LIU HY. Rotamer-free protein sequence design based on deep learning and self-consistency[J]. Nature Computational Science, 2022, 2: 451-462.

[67] LATEGAN FA, SCHREIBER C, PATTERTON HG. SeqPredNN: a neural network that generates protein sequences that fold into specified tertiary structures[J]. BMC Bioinformatics, 2023, 24(1): 373.

[68] ANISHCHENKO I, PELLOCK SJ, CHIDYAUSIKU TM, RAMELOT TA, OVCHINNIKOV S, HAO JZ, BAFNA K, NORN C, KANG A, BERA AK, DiMAIO F, CARTER L, CHOW CM, MONTELIONE GT, BAKER D. *De novo* protein design by deep network hallucination[J]. Nature, 2021, 600: 547-552.

[69] TISCHER D, LISANZA S, WANG J, DONG RZ, ANISHCHENKO I, MILLES LF, OVCHINNIKOV S, BAKER D. Design of proteins presenting discontinuous functional sites using deep learning[J]. bioRxiv, 2020, DOI: 10.1101/2020.11.29.402743.

[70] WANG J, LISANZA S, JUERGENS D, TISCHER D, WATSON JL, CASTRO KM, RAGOTTE R, SARAGOVI A, MILLES LF, BAEK M, ANISHCHENKO I, YANG W, HICKS DR, EXPÒSIT M, SCHLICHTHAERLE T, CHUN JH, DAUPARAS J, BENNETT N, WICKY BIM, MUENKS A, et al. Scaffolding protein functional sites using deep learning[J]. Science, 2022, 377(6604): 387-394.

[71] ZHANG S. AutoFoldFinder: an automated adaptive optimization toolkit for *de novo* protein fold design[C]. NeurIPS, 2021.

[72] ANAND N, ACHIM T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models[EB/OL]. 2022: arXiv: 2205.15019. http://arxiv.org/abs/2205.15019

[73] KRISHNA R, WANG J, AHERN W, STURMFELS P, VENKATESH P, KALVET I, LEE GR, MOREY-BURROWS FS, ANISHCHENKO I, HUMPHREYS IR, McHUGH R, VAFEADOS D, LI XT, SUTHERLAND GA, HITCHCOCK A, HUNTER CN, KANG A, BRACKENBROUGH E, BERA AK, BAEK M, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom[J]. Science, 2024, 384(6693): eadl2528.

[74] BENNETT NR, WATSON JL, RAGOTTE RJ, BORST AJ, SEE DL, WEIDLE C, BISWAS R, SHROCK EL, LEUNG PJY, HUANG BW, GORESHNIK I, AULT R, CARR KD, SINGER B, CRISWELL C, VAFEADOS D, SANCHEZ MG, KIM HM, TORRES SV, CHAN S, et al. Atomically accurate *de novo* design of single-domain antibodies[J]. bioRxiv: the Preprint Server for Biology, 2024: 2024.03.14.585103.

[75] HUANG B, XU Y, HU XH, LIU YR, LIAO SH, ZHANG JH, HUANG CD, HONG JJ, CHEN Q, LIU HY. A backbone-centred energy function of neural networks for protein design[J]. Nature, 2022, 602: 523-528.

[76] LIU YF, CHEN LH, LIU HY. *De novo* protein backbone generation based on diffusion with structured priors and adversarial training[J]. bioRxiv, 2022, DOI: 10.1101/2022.12.17.520847.

[77] 夏彬彬, 王军. 基于深度学习的蛋白质建模与设计[J]. 生物工程学报, 2021, 37(11): 3863-3879.
XIA BB, WANG J. Protein modeling and design based on deep learning[J]. Chinese Journal of Biotechnology, 2021, 37(11): 3863-3879 (in Chinese).

[78] SINAI S, KELSIC E, CHURCH GM, NOWAK MA. Variational auto-encoding of protein sequences[EB/OL]. 2017: arXiv: 1712.03346. http://arxiv.org/abs/1712.03346.

[79] FUKUNAGA T, IWASAKI W. Inverse Potts model improves accuracy of phylogenetic profiling[J]. Bioinformatics, 2022, 38(7): 1794-1800.

[80] CHHIBBAR P, JOSHI A. Generating protein sequences from antibiotic resistance genes data using Generative Adversarial Networks[J]. CoRR, 2019, abs/1904.13240.

[81] HAN X, ZHANG LH, ZHOU K, WANG XN. ProGAN: protein solubility generative adversarial nets for data augmentation in DNN framework[J]. Computers & Chemical Engineering, 2019, 131: 106533.

[82] MÜLLER AT, HISS JA, SCHNEIDER G. Recurrent neural network model for constructive peptide design[J]. Journal of Chemical Information and Modeling, 2018, 58(2): 472-479.

[83] HAWKINS-HOOKER A, DEPARDIEU F, BAUR S, COUAIRON G, CHEN A, BIKARD D. Generating

functional protein variants with variational autoencoders[J]. PLoS Computational Biology, 2021, 17(2): e1008736.

[84] WANG F, FENG XC, KONG R, CHANG S. Generating new protein sequences by using dense network and attention mechanism[J]. Mathematical Biosciences and Engineering: MBE, 2023, 20(2): 4178-4197.

[85] ALLEY EC, KHIMULYA G, BISWAS S, AlQURAISHI M, CHURCH GM. Unified rational protein engineering with sequence-based deep representation learning[J]. Nature Methods, 2019, 16: 1315-1322.

[86] RIVES A, MEIER J, SERCU T, GOYAL S, LIN ZM, LIU J, GUO DM, OTT M, ZITNICK CL, MA J, FERGUS R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[87] LUO YN, JIANG GD, YU TH, LIU Y, VO L, DING HT, SU YF, QIAN WW, ZHAO HM, PENG J. ECNet is an evolutionary context-integrated deep learning framework for protein engineering[J]. Nature Communications, 2021, 12: 5743.

[88] MEIER J, RAO R, VERKUIL R, LIU J, SERCU T, RIVERS A. Language models enable zero-shot prediction of the effects of mutations on protein function[C]. NeurIPS, 2021: 29287-29303.

[89] BISWAS S, KHIMULYA G, ALLEY EC, ESVELT KM, CHURCH GM. Low-N protein engineering with data-efficient deep learning[J]. Nature Methods, 2021, 18: 389-396.

[90] ELNAGGAR A, HEINZINGER M, DALLAGO C, REHAWI G, WANG Y, JONES L, GIBBS T, FEHER T, ANGERER C, STEINEGGER M, BHOWMIK D, ROST B. ProtTrans: toward understanding the language of life through self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 7112-7127.

[91] HAMAMSY T, MORTON JT, BLACKWELL R, BERENBERG D, CARRIERO N, GLIGORIJEVIC V, STRAUSS CEM, LEMAN JK, CHO K, BONNEAU R. Protein remote homology detection and structural alignment using deep learning[J]. Nature Biotechnology, 2024, 42: 975-985.

[92] MADANI A, KRAUSE B, GREENE ER, SUBRAMANIAN S, MOHR BP, HOLTON JM, OLMOS JL, XIONG CM, SUN ZZ, SOCHER R, FRASER JS, NAIK N. Large language models generate functional protein sequences across diverse families[J]. Nature Biotechnology, 2023, 41: 1099-1106.

[93] NIJKAMP E, RUFFOLO JA, WEINSTEIN EN, NAIK N, MADANI A. ProGen2: exploring the boundaries of protein language models[J]. Cell Systems, 2023, 14(11): 968-978.e3.

[94] FERRUZ N, SCHMIDT S, HÖCKER B. ProtGPT2 is a deep unsupervised language model for protein design[J]. Nature Communications, 2022, 13: 4348.

[95] QIU JZ, XU JD, HU J, CAO HQ, HOU LY, GAO ZJ, ZHOU XY, LI AN, LI XJ, CUI B, YANG F, PENG S, SUN N, WANG FY, PAN AM, TANG J, YE JP, LIN JY, TANG J, HUANG XX, et al. InstructPLM: aligning protein language models to follow protein structure instructions[J]. bioRxiv, 2024, DOI: 10.1101/2024.04.17.589642.

[96] DING WZ, NAKAI KT, GONG HP. Protein design *via* deep learning[J]. Briefings in Bioinformatics, 2022, 23(3): bbac102.

(本文责编 陈宏宇)