·生物元器件智能设计合成·

廖小平 博士，中国科学院天津工业生物技术研究所研究员。主要围绕工业生物大数据智能分析展开研究，开发核心的数据库、算法和工具。构建了糖基转移酶数据库 pUGTdb、大肠杆菌代谢调控图谱 ERMer 等系列数据库；开发了编辑序列设计平台 AutoESD、途径设计平台 CAVE 等系列工具；发展了新一代蛋白功能预测算法 HDMLF、酶挖掘与评估工具 REME 等系列 AI 算法。近年来在 *Nucleic Acids Research*、*Science Advances*、*Molecular Plant*、*Research* 等国内外期刊发表文章 50 余篇，被引用 2 000 余次。主持基金委交叉重点专项、中国科学院战略性先导专项课题等多项国家级、省部级项目。

蛋白元件的智能挖掘、改造和从头设计

刘翠 [1,2]，史振坤 [1,2]，马红武 [1,2]，廖小平 [1,2*]

1 中国科学院天津工业生物技术研究所，天津 300308
2 国家合成生物技术创新中心，天津 300308

刘翠, 史振坤, 马红武, 廖小平. 蛋白元件的智能挖掘、改造和从头设计[J]. 生物工程学报, 2025, 41(3): 993-1010.
LIU Cui, SHI Zhenkun, MA Hongwu, LIAO Xiaoping. Intelligent mining, engineering, and *de novo* design of proteins[J]. Chinese Journal of Biotechnology, 2025, 41(3): 993-1010.

摘 要：天然元件服务于细胞长期进化获得的生存本能，难以满足工程细胞在工业等特殊环境下高效执行生物功能的需求。酶作为生物催化剂，在生物合成途径中发挥着关键作用，它们能够显著提高生化反应的速率和选择性。然而，天然酶的催化效率、稳定性、底物特异性和耐受性等方面往往不能满足工业生产的需求。因此，挖掘、设计和改造酶以适应特定的生物制造过程至关重要。近年来，人工智能(artificial intelligence, AI)技术在蛋白的挖掘、评估、改造和从头设计中发挥着越来越重要的作用。AI 技术可以通过机器学习和深度学习算法，分析大量的生物信息学数据，预测蛋白的功能和特性，从而加速蛋白的发现和优化过程。此外，AI 还可以辅助科研人员从头设计新的蛋白结构，通过模拟和预测其在不同条件下的性能，为蛋白的设计提供指导。本文综述了面向生物制造的蛋白元件挖掘、评估、改造以及从头设计的最新研究进展，探讨了该领域的热点问题、难点以及新兴技术方法，旨在为相关领域的科研工作提供指导。

关键词：生物制造；酶挖掘；评估；蛋白改造；从头设计

资助项目：中国科学院战略性先导科技专项(XDC0110203)；国家自然科学基金(12326611)
This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDC0110203) and the National Natural Science Foundation of China (12326611).
*Corresponding author. E-mail: liao_xp@tib.cas.cn
Received: 2024-08-02; Accepted: 2024-10-31; Published online: 2024-11-01

Intelligent mining, engineering, and *de novo* design of proteins

LIU Cui[1,2], SHI Zhenkun[1,2], MA Hongwu[1,2], LIAO Xiaoping[1,2*]

1 Biodesign Center, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
2 National Center of Technology Innovation for Synthetic Biology, Tianjin 300308, China

Abstract: Natural components serve the survival instincts of cells that are obtained through long-term evolution, while they often fail to meet the demands of engineered cells for efficiently performing biological functions in special industrial environments. Enzymes, as biological catalysts, play a key role in biosynthetic pathways, significantly enhancing the rate and selectivity of biochemical reactions. However, the catalytic efficiency, stability, substrate specificity, and tolerance of natural enzymes often fall short of industrial production requirements. Therefore, exploring and modifying enzymes to suit specific biomanufacturing processes has become crucial. In recent years, artificial intelligence (AI) has played an increasingly important role in the discovery, evaluation, engineering, and *de novo* design of proteins. AI can accelerate the discovery and optimization of proteins by analyzing large amounts of bioinformatics data and predicting protein functions and characteristics by machine learning and deep learning algorithms. Moreover, AI can assist researchers in designing new protein structures by simulating and predicting their performance under different conditions, providing guidance for protein design. This paper reviews the latest research advances in protein discovery, evaluation, engineering, and *de novo* design for biomanufacturing and explores the hot topics, challenges, and emerging technical methods in this field, aiming to provide guidance and inspiration for researchers in related fields.

Keywords: biomanufacturing; enzyme mining; evaluation; protein engineering; *de novo* design

蛋白质是生命的基石，在生物体内发挥着关键作用，具有催化生化反应、构建细胞结构及调节生物体的复杂功能。在工业应用中，例如在生物制药、绿色农业、生物化工等方面[1]，蛋白质同样发挥着不可或缺的作用。通过改造和优化蛋白质，科学家们能够开发出更高效的生物催化剂，提高工业生产过程的效率和可持续性。然而，蛋白质的复杂性和多样性使得其设计和改造充满挑战。传统的蛋白质设计方法，如定向进化和理性设计，虽然取得了一定的进展，但往往受到实验条件和理论模型的限制。

传统的蛋白质设计方法主要依赖于实验和理论计算。定向进化是一种通过随机突变和高通量筛选来优化蛋白质特性的方法，虽然能够产生一些有益的突变，但过程繁琐且效率低下。理性设计则基于蛋白质的结构和功能信息，通过预测和计算来设计蛋白质的突变，但这种方法往往受限于对蛋白质结构和功能关系的理解[2]。此外，这些方法通常需要大量的实验验证，成本高昂且周期长。随着生物信息学和计算生物学的发展，虽然在一定程度上提高了蛋白质设计的能力，但仍然面临着预测精度和实验验证的双重挑战。因此，探索新的蛋白质设计方法来提高效率和准确性，成为当前研究的热点[3]。

近年来，人工智能(artificial intelligence, AI)技术的快速发展为蛋白质挖掘、评估、设计带

来了新的机遇(图 1)。特别是深度学习 AI 模型，通过学习大量的蛋白质序列和结构数据，能够预测蛋白质的功能和特性。蛋白质语言模型(protein language model, PLM)是这一领域的一个创新应用，它借鉴了自然语言处理中的技术，将氨基酸序列视作“词汇”，通过学习蛋白质的“语言”来预测其结构和功能。例如，ESM-1v[4]和 SESNet[5]等模型通过结合无监督学习和监督学习，显著提高了蛋白质突变预测的准确性。这些模型不仅能够预测突变对蛋白质功能的影响，还能够指导实验设计，减少实验成本。AI 技术的应用不仅提高了蛋白质设计的效率，还为理解蛋白质的复杂性提供了新的视角。此外，基于 AI 的蛋白质从头设计技术正在成为该领域的一个新趋势。这种技术利用深度学习模型，根据目标需求从头开始设计全新的蛋白质，这些全新的蛋白质可以扩展已有蛋白质的序列、结构和功能空间。通过在从头设计过程加入辅助调节控制信息，可以实现引导定向生成符合所需结构和功能的蛋白质，例如与特定蛋白结合的结合剂、特异性小分子结合蛋白及催化特定反应的酶等。随着 AI 技术的不断进步，未来蛋白质设计将更加智能化、精准化，为生命科学和工业应用带来革命性的变化。

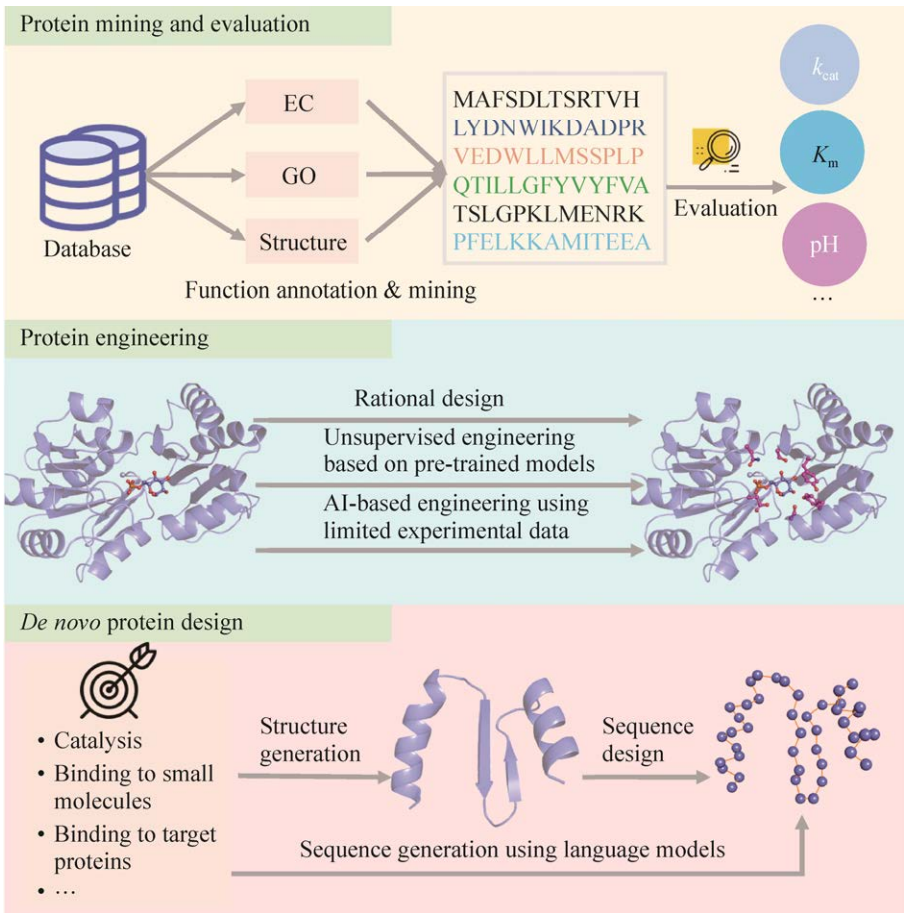

图 1　蛋白元件挖掘、改造和从头设计

Figure 1　Mining and evaluation, engineering, and *de novo* design of proteins. EC: Enzyme commission, used to describe the catalytic activity and reaction type of enzymes. GO: Gene ontology, used to describe the functions of genes and proteins.

目前已有一些关于 AI 蛋白设计的综述，例如 2022 年西班牙赫罗纳大学的 Ferruz 等[6]发表了关于语言模型可控蛋白设计的综述，讨论了生成性语言模型对蛋白质设计的可预见性影响。2023 年康里奇等[7]介绍了酶工程的主要发展历程，并梳理了 AI 助力酶工程的研究进展。2024 年张锦雄等[8]对大语言模型在蛋白质设计中的应用进行了综述，主要从模型架构方面进行分类介绍。2024 年刘南等[9]对人工智能时代下的蛋白质从头设计进行了综述，着重介绍了最新算法模型和存在的问题。2024 年 Notin 等[10]综述了机器学习功能蛋白设计的最新进展、应用挑战以及未来趋势。相比于已发表的其他综述，本文在 AI 酶设计方面根据实际使用场景分为了无需实验数据的无监督改造和有实验数据的改造；在从头酶设计方面从结构设计、序列设计、功能描述直接到序列的端到端设计进行细致地分类和介绍；此外，还对面向生物制造的蛋白元件挖掘、评估的最新研究进展和新兴技术方法进行了全面介绍。本文详细综述了蛋白元件挖掘、评估、改造以及从头设计的最新研究进展，以期为相关领域的科研工作提供指导。

1 蛋白元件的注释、挖掘与评估

蛋白是构建生物催化系统的重要功能单元，是利用人工生物体系作为催化剂以合成工业化学品、药物和功能材料的必要条件。在生物工程应用中，为了满足工业生物制剂的多样化需求，需从自然界中大规模挖掘新的蛋白元件，这涉及蛋白元件的功能注释、挖掘与评估。几十年来，计算生物学和生物信息学领域的专家学者开发了一系列高质量且高效的方法(表 1)，以加速这一进程。

表 1 蛋白功能注释与挖掘评估

Table 1 Protein function annotation, discovery and evaluation

Name	Link	Year	Classification	Availability
SVM-Prot 2016	http://bidd.group/cgi-bin/svmprot/svmprot.cgi	2016	EC	No
DEEPre	http://www.cbrc.kaust.edu.sa/DEEPre/index.html	2018	EC	Yes
ECPred	https://ecpred.kansil.org/	2018	EC	Yes
DEEPEC	https://bitbucket.org/kaistsystemsbiology/deepec/src/master/	2019	EC	Yes
CLEAN	https://clean.platform.moleculemaker.org/configuration	2023	EC	Yes
ECRECer	https://ecrecer.biodesign.ac.cn/	2023	EC	Yes
CAFA2	–	2016	GO	No
GOstruct 2.0	https://sourceforge.net/projects/strut/	2017	GO	Yes
DeepGO	https://github.com/bio-ontology-research-group/deepgo	2018	GO	Yes
NetGO	https://issubmission.sjtu.edu.cn/netgo/	2019	GO	No
DeepFRI	https://beta.deepfri.flatironinstitute.org/	2021	GO	Yes
DeepGO-SE	https://github.com/bio-ontology-research-group/deepgo2	2024	GO	Yes
ESP	https://github.com/AlexanderKroll/ESP	2023	Enzyme-substrate binding	Yes
Km_prediction	https://github.com/AlexanderKroll/KM_prediction	2021	K_m	Yes
DLKcat	https://github.com/SysBioChalmers/DLKcat	2022	k_{cat}	Yes
TurNuP	https://turnup.cs.hhu.de/	2023	k_{cat}	Yes
UniKP	https://github.com/Luo-SynBioLab/UniKP	2023	K_m/k_{cat}	Yes
DeepET	https://zenodo.org/records/6351465	2022	Optimal temperature	Yes
EpHod	https://github.com/jafetgado/EpHod	2023	pH	Yes
REME	https://reme.biodesign.ac.cn/	2024	Mining and evaluation	Yes

1.1 蛋白元件功能注释

目前对蛋白元件进行功能注释，主要以酶序列/结构数据为基础，将酶号(enzyme commission, EC)/基因本体(gene ontology, GO)编号作为预测目标。研究人员进行了连续不断的研究，取得了一系列的研究进展，如 2007 年上海交通大学与美国戈登生命科学研究所联合发布的 EzyPred[11] 采用自顶向下的策略将酶与非酶的预测准确性提高到了 91.3%，将一级 EC 号预测准确率提高至 93.7%。2009 年，美国生物技术高性能计算软件应用研究所的研究团队开发了 CatFam[12]，该方法首次引入错误预测率(false-positive rate, FPR)作为调控变量，当 FPR 设置为 1%时，CatFam 的预测精度达到 98.6%，召回率为 95.0%。相比其他方法，CatFam 在精度和召回率上均表现更为优异。然而，该方法能够预测的 EC 编号数量较为有限。2013 年，厦门大学的研究团队开发了多功能酶预测方法 MFEC[13]将蛋白功能注释扩展到多功能酶上。2016 年，重庆大学团队开发了 SVM-Prot 2016[14]将支持向量机(support vector machines, SVM)技术引入蛋白功能预测，并将预测的覆盖范围扩展至 54 个功能家族，EC 号增加至 192 个。2018 年，土耳其中东技术大学和欧洲生物信息研究所的联合团队开发了 ECPred[15]，该方法利用基于 EC 树结构的分层预测策略，能够预测 858 个 EC 编号，涵盖 6 个主类、55 个子类、163 个次子类和 634 个底物类。同年，沙特阿拉伯阿卜杜拉国王科技大学、美国伊利诺伊理工学院和中国杭州电子科技大学的联合团队推出了 DEEPre[16]，抛弃了传统的手工特征提取技术，采用端到端的策略，从酶序列的原始编码中提取卷积和序列特征，从而提升了预测性能，并在大规模数据集上实现了高效的 EC 号预测，覆盖范围扩展至 3 518 个 EC 编号。这些研究成功将 EC 号预测级别推进到完整的 4 级，所覆盖的 EC 号也越来越多，然而这些方法的预测精度普遍不高。

为了解决这些问题，2019 年 Lee 团队将深度学习方法引入到 EC 号预测工作中，使用 3 组卷积神经网络(convolutional neural network, CNN)将预测精度提高到 92%[17]。虽然使用该方法可以获得较高的注释精度，但是该方法的注释召回率只有 45%，在实际使用中仍显不足[17]。为了进一步提高注释性能，2023 年伊利诺伊大学厄巴纳-香槟分校的赵惠民团队使用蛋白语言模型与对比学习方法将召回率提高到 81%[18]。然而该方法无法区分酶与非酶，会引入大量假阳性结果导致挖掘性能不佳。为了得到一个具有均衡预测性能的 EC 注释工具，2023 年中国科学院天津工业生物技术研究所生物设计中心发布了一个基于蛋白语言预训练模型与层次深度学习方法的 EC 号预测工具 ECRECer，该方法的预测准确率、预测精度、召回率均超过 80%，是目前预测性能较为均衡的一个方法[19]。

类似于 EC 编号，GO 是另外一种广泛使用的功能注释目标，可以从细胞组件、分子功能和生物过程等更精细、更全面的角度对元件功能进行刻画，专家学者也对此进行了大量研究。Fang 等[20]构建了基于蛋白质结构域的本体数据库 DcGO，该数据库实现了蛋白质与 GO 号的关联检索。然而，单纯的数据库无法对新发掘的蛋白质功能进行预测。为了实现对新蛋白质的功能预测，Kahanda 等[21]设计了新的算法 Gostruct 2.0，该方法将结构化支持向量机(SVM)与条件约束引入到预测方法中，在 GO 子本体(分子功能、生物过程和细胞组成)中均取得了显著提高，特别是在生物过程子本体中，AUC 从 58%提升至 71%。尽管 GOstruct 2.0 在处理已注释蛋白质的新功能预测任务上表现出色，但仍需解决模型复杂度和参数优化等问题。

为了提供统一的蛋白质功能注释评估方法，Function-SIG 研究组发布了一系列评估数据、方法与平台，包括 CAFA1、CAFA2 和 CAFA3[22]，大大促进了蛋白元件功能注释社区的发展。鉴于深度学习的快速发展，研究人员将神经网络引入到 GO 预测中。例如，DeepGO[23]训练了 3 个模型，每个模型对应一个 GO 子本体，使用嵌入层和卷积神经网络(CNN)来学习蛋白质序列的特征，取得了较好的预测性能，其中在酵母蛋白功能预测中的平均预测精度达到了 46%。NetGO (NetGO 2.0)[24]将循环神经网络和基于学习的排序方法引入大规模蛋白质功能预测中，在功能注释的关键评估(critical assessment of functional annotation, CAFA)中表现出色，尤其是在生物过程本体(biological process ontology, BPO)和细胞组分本体(cellular component ontology, CCO)方面。

DeepFRI 使用图卷积神经网络结合蛋白质序列与结构信息，通过从蛋白质语言模型中提取序列特征，并结合蛋白质结构的接触图来实现功能预测[25]。DeepFRI 能够处理大规模的蛋白质序列数据，并具有显著的去噪能力，即使在实验结构被蛋白质模型替代时也只有轻微的性能下降。DeepFRI 在预测 GO 术语和 EC 号码方面表现出色，尤其是在预测具有低序列同源性的蛋白质功能时。DeepGO-SE[26]利用预训练的大型语言模型来从蛋白质序列预测 GO 术语，该方法能够使用序列特征、本体论知识和蛋白质间相互作用信息来提高预测性能。由于 GO 到蛋白质的翻译过程较复杂且 GO 号数量更多(现有 42 093 个)，基于 GO 号的蛋白元件注释性能较 EC 编号普遍偏低，在 UniProtKB/Swiss-Prot 数据集上的注释精确率-召回率曲线下面积(area under the precision-recall curve, AUPRC)约为 75%，且不同物种中的注释准确率与精度差别较大。

1.2 蛋白元件挖掘

蛋白元件挖掘是生物信息学和结构生物学中的一个重要领域，它涉及从蛋白质数据库中识别和提取具有特定功能的蛋白质片段或结构域。这一过程对于理解生物分子的功能机制以及开发新型生物技术具有重要意义。传统的蛋白元件挖掘方法主要依赖于序列相似性、功能域相似性分析，而近年来，随着结构生物学和计算方法的发展，基于结构相似性的挖掘方法也逐渐兴起。

在序列相似性方面，研究者通过比对目标序列与已知功能的蛋白质序列，利用 BLAST 等工具寻找具有高度序列同源性的蛋白质，从而推测其可能的功能。这种方法简单快捷，但受限于已知序列的覆盖范围，对于低同源性或新出现的蛋白元件，其预测能力有限。功能域相似性分析则更进一步，它不仅考虑序列的相似性，还关注蛋白质中特定功能域的保守性。通过识别这些功能域，研究者可以预测蛋白质的生物学功能，即使在序列相似性不高的情况下。例如，通过识别锌指结构域、螺旋-环-螺旋结构域等，可以推测蛋白质可能参与的调控或结合过程[27]。

最新的结构相似性挖掘方法则利用了蛋白质三维结构信息，通过比较蛋白质的空间构象来识别功能相似的蛋白元件。AlphaFold[28]等深度学习模型能够预测蛋白质的三维结构，为结构相似性分析提供了强大的工具。高彩霞教授的研究团队利用 AI 辅助的大规模蛋白结构预测，建立了基于三级结构的高通量蛋白聚类方法，成功挖掘出一系列新型脱氨酶[29]。这些新型脱氨酶被用于开发具有我国自主知识产权的新型碱基编辑工具，有望打破国外在该领域的专利垄断，提升我国在生物技术

产业中的竞争力。

2024 年，诺贝尔化学奖得主、CRISPR 基因编辑技术奠基人之一的 Jennifer Doudna 教授团队，将基于 AlphaFold 2 的人工智能的蛋白质结构预测与传统结构比对程序结合，开发出了一种自动化结构检索，发现了 Cas13 的祖先——Cas13an[30]；与其他 Cas13 相比，Cas13an 的尺寸较小，仅 449 个氨基酸，是其他 Cas13 的 1/3 (Cas13a 是 1 159 个氨基酸)；并进一步解析了 Cas13an 的结构及其作用机制，与较大的 Cas13 不同，Cas13an 用于 crRNA 加工和 RNA 引导切割的是一个单一的活性位点，揭示了祖先核糖核酸酶(ribonuclease)结构域的 2 种活性模式。Cas13an 这种小型化的 Cas13 仍然功能齐全，为研究人员扩展了 RNA 编辑工具箱。

Van Kempen 等[31]提出了一种名为 3Di 的新型字母表，3Di 中的每个字母描述了氨基酸和其最近氨基酸之间的三级结构的相互作用，基于此开发的 Foldseek，通过将结构比对转化为 3Di 序列比对，将计算时间缩短了 4 到 5 个数量级，成为基于蛋白结构挖掘的有力工具。Cho 和 Bonneau 团队开发了深度学习模型 TM-Vec[32]，通过 2 个序列的嵌入表示的相似性估计其对应的结构比对 TM-score，可以在大型序列数据库中搜索结构-结构相似性，进行蛋白元件的快速挖掘。

此外，蛋白质语言模型的发展也为蛋白质挖掘提供了新的方法。PLMSearch[33]使用预先训练好的蛋白质语言模型的深度表征，并使用大量真实结构相似性来训练相似性预测模型，可以像 MMseqs2 一样在几秒钟内搜索数百万个查询目标蛋白质对，同时将灵敏度提高 3 倍以上，可与当前最先进的结构搜索方法相媲美。DHR[34]能够在不依赖序列比对的情况下，以更高的速度和灵敏度检测序列相似性低但仍具有结构或功能相似性的远程同源体，比 PSI-BLAST 和 DIAMOND 等传统方法快 22 倍，比 HMMER 快 28 700 倍。

1.3 蛋白元件评估

在酶挖掘过程中，对所挖掘的功能元件进行评估是不可或缺的一步。评估方法包括酶动力学参数，如 K_m (米氏常数)，表示酶与底物的亲和力，K_m 值越小，表示酶对底物的亲和力越高；k_{cat} (催化常数)，表示每个酶分子每秒转化底物分子的数目，反映酶的催化效率。此外，还包括最适生长温度和最适 pH 等参数，这些参数能够帮助研究人员更好地理解酶在不同条件下的表现和稳定性。

近年来，随着深度学习技术的发展，研究者提出了许多针对不同参数的预测工具。例如，Kroll 等[35]开发的 ESP 方法可以评估所挖掘的酶是否能催化对应底物，这一方法通过模拟酶和底物的相互作用，可以帮助生物学家从大量的候选蛋白中快速虚拟筛选可能的候选酶。KM_Prediction 方法于 2021 年被提出，该方法通过融合蛋白质的嵌入表示与底物的嵌入表示，预测其对底物的结合亲和力，这一技术对于筛选高效酶具有重要意义[36]。随后，2022 年，Li 等[37]开发了 DLKcat，这是一种基于卷积神经网络(CNN)的蛋白质表示和基于图卷积网络(graph convolutional network, GCN)的化合物表示的预测模型，能够预测酶的催化效率。2023 年，Kroll 等[38]开发的 TurNuP，不仅仅只关注底物，还实现了反应层面 k_{cat} 的评估。同年 Yu 等[39]开发了 UniKP，该方法在同一个框架内实现 k_{cat}、K_m 和催化效率的预测；与 DLKcat 相比，UniKP 的平均决定系数(R^2)提高了 20%。

此外，Li 等[40]开发的 DeepET 可评估最适生长温度，这一工具通过分析酶在不同温度下的活性变化，帮助研究人员确定最佳操作条

件。Gado 等[41]开发的 EpHod 实现了对最适 pH 的评估，这对于优化工业酶的应用环境具有重要意义。2024 年，中国科学院天津工业生物技术研究所生物设计中心发布的 REME 酶挖掘与评估平台，将上述酶评估工具集成到了一个统一的平台，简化了研究流程，大大提高了研究的准确性和效率，实现了元件的全面评估[42]。

2 蛋白元件的改造

蛋白作为生物催化系统的核心组件，在工业化学品、药物和功能材料的合成中发挥着至关重要的作用。但天然的蛋白往往不能满足工业应用对于性能的要求，需要对蛋白进行改造，以提高其催化效率、稳定性和选择性。围绕这一主题，研究人员采用包括理性设计、AI 酶改造等策略，开发了一系列高质量且高效的方法(表 2)，并通过计算设计以加速这一进程。

2.1 理性设计

蛋白质理性设计是现代生物技术和生物工程领域的重要研究方向之一。理性设计建立在对酶催化机制和结构与功能关系深刻理解的基础上，对潜在突变位点进行预测与模拟，进而考察氨基酸突变对目标蛋白稳定性、选择性以及与底物的结合和催化的影响。基于对蛋白质结构和功能的理解，借助理性设计进行有目的的改造，科学家们可以设计出具有特定功能的蛋白质，以满足各种应用需求。

理性设计通常聚焦于酶催化口袋的活性位点，能否正确识别活性位点是缩小突变体筛选范围的关键。基于酶催化过程的动态特征以及酶结构自身的运动特性，研究人员通过蛋白底物结合口袋分析、构象动力学分析[43]、B-因子分析等手段高效定位活性位点。通过计算设计技术，研究人员可以把实验试错范围缩小 3–4 个数量级[44]。这种方法不需要构建大规模的突变体库，节省了大量的人力、物力及财力，能在较短的时间内设计并得到性质改善的突变体。

中国科学院天津工业生物技术研究所研发团队针对磷酸酶的底物特异性进行理性设计，通过底物结合前后 B-因子分析定位影响底物特

表 2 AI 蛋白改造策略

Table 2 Methods for AI-based protein engineering

Name	Deep learning approach	GitHub
DeepSequence	Unsupervised learning	https://github.com/debbiemarkslab/DeepSequence
EVE	Unsupervised learning	https://github.com/OATML/EVE
ESM-1v	Unsupervised learning	https://github.com/facebookresearch/esm
ESM-IF1	Unsupervised learning	https://github.com/facebookresearch/esm
MSA-Transformer	Unsupervised learning	https://github.com/facebookresearch/esm
Tranception	Unsupervised learning	https://github.com/OATML-Markslab/Tranception
MutComput	Unsupervised learning	–
MutComputX	Unsupervised learning	https://github.com/danny305/MutComputeX
AlphaMissense	Unsupervised learning	https://github.com/google-deepmind/alphamissense
SaProt	Unsupervised learning	https://github.com/westlake-repl/SaProt
ProSST	Unsupervised learning	https://github.com/ai4protein/ProSST
ESM-1b	Supervised learning	https://github.com/facebookresearch/esm
ECNet	Supervised learning	https://github.com/luoyunan/ECNet
Low-N	Few-shot learning	https://github.com/churchlab/low-N-protein-engineering
SESNet	Few-shot learning	–
FSFP	Few-shot learning	https://github.com/ai4protein/Pro-FSFP

异性识别的关键氨基酸位点、通过底物结合口袋分析定位底物特异性结合位点、通过特异性序列保守性分析定位影响特异性的突变，借助Rosetta 酶设计工具，对选取的 15 个关键位点共设计了 46 个单点突变体，进一步的组合突变成功将底物广谱的磷酸酶改造成了高特异性酶(产率 94.0%–98.8%)，并同步将活性提升了5.4–11.8 倍[45]。利用得到的有益突变，构建的热力学驱动合成系统，实现了廉价的蔗糖和淀粉向果糖和甘露糖的高效转化，为工业生产奠定了基础。

通过对腈水解酶(SsNIT)底物结合口袋的分析，定位到 2 个影响立体选择性的关键氨基酸残基，W170 与 V198 处于互为镜像的位置，通过对 W170 和 V198 位点进行镜像置换，实现了对 3-异丁基戊二腈水解反应立体偏好性的翻转[46]。通过对醇脱氢酶 TbSADH 环区(loop)进行分析，将 84 位脯氨酸残基作为改造热点，通过调控酶分子动力学特性，提升酶分子的构象可塑性，经定点突变和迭代突变实验获得了催化活性及立体选择性均大幅提升的组合突变[47]。

2.2 基于 AI 预训练模型的无监督蛋白改造

无监督学习不需要特定的数据标签，从未标记的数据本身提取信息，学习数据的内在规律和模式。自然界中存在的上亿条蛋白氨基酸序列及大量序列中蕴藏的进化信息、根据序列使用深度学习模型 AlphaFold 2 以极高准确度预测的大量蛋白质结构及结构中蕴藏的空间相互作用等信息，为蛋白质无监督学习模型提供了数据基础。

基于自然语言处理(natural language processing, NLP)技术发展的语言模型，成为蛋白质无监督学习的一种重要手段。蛋白质语言模型(PLM)通过学习氨基酸序列的语义和语法规则，能够判断特定位置的氨基酸突变相对于野生型是否更符合自然界中蛋白质的通用规则，从而指导蛋白质的突变。由 Facebook AI 团队提出的通用蛋白质语言模型 ESM-1v[4]，通过对来自 Uniref90 的 9 800 万条蛋白质序列进行随机遮掩训练，要求模型预测被遮盖的位置上的氨基酸；该模型不需要任何实验数据或额外的监督式学习，仅从大量序列中学习捕获序列突变对功能的影响，在 41 个深度突变扫描数据集上零样本预测评估中，17 个超过基线方法。

为了在序列学习过程中加入序列进化约束信息，Riesselma 等[48]提出了 DeepSequence，该模型利用变分自编码器(variational auto-encoders, VAE)架构，能够学习每个蛋白质家族内的高阶非线性进化约束，并预测突变效应。进一步改进的 EVE 模型[49]对跨物种的大规模进化数据进行学习，在人类疾病相关基因变异致病性方面有出色的预测能力。Rao 等[50]提出的 MSA-Transformer 模型使用进化相关序列家族多重序列比对(multiple sequence alignment, MSA)取代单一蛋白质序列作为蛋白质语言模型的输入，在 Transformer 中新增行和列注意力机制学习 MSA 中序列的进化规律和氨基酸的共变信息，通过跨蛋白质家族的掩码语言模型(masked language model, MLM)变体进行训练，实现进化约束的蛋白质突变预测。Tranception[51]通过在大规模未对齐蛋白序列上训练，利用自回归预测技术以及推理阶段的同源序列检索功能，实现了蛋白突变适应度和多重变异影响的有效评估。

基于序列的设计方法难以全面捕捉结构和功能之间的关系，在序列上比较远的残基在三维结构中可以比较近，蛋白质结构能够提供比序列更多的特定氨基酸周围环境信息。ESM-IF1 模型[52]使用蛋白结构骨架坐标进行了增强，采用几何向量感知机(geometric vector

perceptrons, GVP)对通过 AlphaFold 2 预测的 1 200 万条序列的结构进行编码学习。斯坦福大学的研究表明 ESM-IF1 模型具有在无监督条件下指导各种蛋白质进化的能力，通过筛选并测试了约 30 种 SARS-CoV-2 病毒变体，得到了中和效力更高的治疗性抗体，证明了整合结构信息对识别有利突变的优势[53]。MutCompute 通过从 19 300 个蛋白质结构中提取特定残基周围肽原子构造微环境，使用 3D 卷积神经网络(3-dimensions convolutional neural network, 3DCNN)对 170 万个微环境进行学习，使模型能够对不符合进化环境的不稳定位点做出预测[54]。Lu 等[55]使用 MutCompute 对聚对苯二甲酸乙二酯(polyethylene terephthalate, PET)水解酶进行了突变设计，得到了具有优异催化活性和热稳定性的突变体 FAST-PETase，能在 7 d 内完全降解大多数 PET 制品。而 MutComputeX 是 MutCompute 的扩展版本[56]，被描述为“用于蛋白质-X 界面工程的自监督 3D 残差神经网络”。MutComputeX 的主要优势在于它能够考虑更广泛的分子相互作用，包括蛋白质与核酸、糖基、配体、辅因子或其他蛋白质的相互作用。这使得 MutComputeX 能够应用于更复杂的蛋白质工程任务，如界面工程和底物特异性工程。

综合考虑序列和结构信息的预训练模型在蛋白质突变效应预测中也取得了显著进展。谷歌 DeepMind 利用蛋白质语言建模和基于微调的 AlphaFold 构建了无监督致病性预测模型 AlphaMissense[57]，可以实现对蛋白质组范围内的错义变体效应准确预测。西湖大学开发的 SaProt 模型[58]利用 Foldseek 将蛋白结构处理成结构感知词表，通过将结构词表和氨基酸词表结合，从而让模型同时考虑蛋白质的序列和结构信息，在零样本突变预测能力上超过了之前的先进模型。Li 等[59]提出的 ProSST 模型通过结构量化模块将蛋白质结构转化为残基标记序列和结构标记序列，并通过解缠结注意力机制学习两者之间的关系，在零样本突变预测中达到了最先进水平。

2.3 结合实验数据的 AI 蛋白改造

无监督模型不需要经过额外训练即可直接在特定蛋白质上执行突变体预测任务，但在完全没有湿实验的情况下以零样本(zero-shot)预测蛋白质突变-性质的变化，往往精度较低。

监督学习可以通过在无监督学习的基础上进一步针对特定蛋白质的突变数据集进行训练，从已知突变结果中学习序列与功能之间的细微关联，从而对该特定蛋白未知突变进行更准确的预测。例如，基于 Transformer 的 ESM-1b 模型[60]，首先在大规模蛋白质序列数据集 UniRef50 上进行了广泛的预训练，捕捉蛋白质序列中的全局信息和模式。为了提高模型在特定任务上的表现，进一步对特定蛋白质的突变数据进行微调，这种方法使得模型能够深入理解蛋白质的精细特性，显著提升了预测的准确性。另外，ECNet 模型利用了无监督学习模型 TAPE 来编码蛋白质序列特征，同时从多序列比对(MSA)中学习残基间的相互依存关系，是一个利用进化环境预测特定蛋白质突变效果的有监督模型。使用 ECNet 对 TEM-1β-内酰胺酶低阶突变体的微调学习，预测得到了具有更强氨苄青霉素耐药性的高阶突变体[61]。

尽管监督学习模型在预测精度上具有优势，但它们通常需要大量的实验数据来训练，这在实践中可能意味着巨大的工作量和成本。为了解决这一问题，Biswas 等[62]基于在大型未标记蛋白序列上进行全局无监督预训练开发了 UniRep 模型，通过借助 UniRep 对功能蛋白序列基本特征的理解，减少了监督学习对实验数据的需求，仅使用少量功能变体(Low-N)即可实现对序列空间的大规模探索。SESNet 通过整合

局部和全局进化信息以及蛋白质的三维结构信息，并首先利用大量无监督模型的预测结果进行预训练，这一策略大幅度提高了模型对蛋白质突变的初始理解。通过这种数据增强策略，SESNet 在使用少量实验数据进行微调后，能够更准确地预测蛋白质突变体的功能和适应性，特别是在高阶突变体的预测中表现出色[5]。

2024 年，洪亮课题组提出了一种新的方法 FSFP[63]，综合利用元学习、排序学习和参数高效的微调，在只利用任意几十个湿实验数据下便可以微调蛋白质预训练模型并大幅提高对蛋白质突变-性质预测的效果。FSFP 方法先利用蛋白质预训练模型评估目标蛋白质与 ProteinGym 中的蛋白质的相似度，并从 ProteinGym 中取出与目标蛋白质最相近的 2 个蛋白质数据集作为元学习的 2 个辅助任务，同时利用 GEMME 对目标蛋白质的打分数据作为第 3 个辅助任务；最后利用排序学习损失函数和 Lora 训练方法，在极少量(几十个)的真实湿实验数据上训练蛋白质预训练模型；结果表明，即便是在原始的蛋白质预训练模型对突变-性质预测的 Spearman 相关性低于 0.1 的情况下，FSFP 方法只利用任意 20 个湿实验数据训练模型，也能将上述的预测相关性大幅提高到 0.5 以上[63]。

3 蛋白元件从头设计

蛋白质从头设计不依赖于现有的天然蛋白质，从零开始设计具有特定结构和功能的蛋白质，这一过程通常涉及从零开始设计蛋白质的结构和序列，以满足特定的功能需求。目前已有多种深度学习方法被开发出来用以生成蛋白质主链骨架结构，以及设计可稳定折叠到给定主链骨架结构的氨基酸序列，以大语言模型为主的方法实现了从“语言”到“语言”的端到端序列生成，不依赖结构信息或序列设计过程，直接生成新的序列。蛋白质从头设计致力于创造自然界不存在的、可成功折叠的、具有特定功能的蛋白质，相关计算工具正在不断涌现(表 3)，其设计效率、精度和成功率的不断提高，正在引领蛋白质设计和应用的革命。

3.1 结构生成

用于蛋白质结构预测的深度神经网络如 trRosetta[64]、AlphaFold 2、RoseTTAFold[65]等已经能够从给定氨基酸序列以极高的准确度预测蛋白质三级结构，这些模型隐含的对蛋白质结构的深刻理解使它们可以被用来进行蛋白质结构的从头设计，生成创新性的结构。基于这种思想，David Baker 课题组提出了基于 hallucination 的蛋白从头设计方法[66]，使用结构预测模型 trRosetta 评估序列的 2D 结构特征，将所有蛋白质数据银行(Protein Data Bank, PDB)结构 2D 特征平均信号作为随机噪音背景分布，在初始随机氨基酸序列空间通过蒙特卡洛随机点突变更新序列，并最大化预测结构与背景分布之间的差距来生成新的合理的蛋白质结构。虽然幻想生成的结构序列与天然蛋白序列差异很大，但因为 trRosetta 是基于大量 PDB 结构训练而成，而已有的 PDB 结构覆盖了蛋白质大部分可能的折叠空间，因此 trRosetta 生成的蛋白并没有跳出已知的蛋白折叠空间。另外幻想生成的序列还比较短，生成的结构比较简单。

在图像和文本生成领域成功应用的扩散模型也被用于蛋白质的从头设计，已有多个模型发布，如 FoldingDiff[67]、ProteinSGM[68]、RFdiffusion[69]、Chroma[70]等。扩散模型通过对蛋白质添加连续噪声训练神经网络，再通过去噪音逐渐还原出蛋白结构。为了增加从噪声中生成折叠成目标结构的成功率，David Baker 课题组通过微调结构预测模型 RoseTTAFold 作为扩散模型中的去噪网络开发了 RFdiffusion 模型[69]，

表 3 蛋白质从头设计方法

Table 3 Methods for *de novo* protein design

Name	Deep learning approach	Task	GitHub
Hallucination	Structural stability loss	Structure design	https://github.com/RosettaCommons/RFDesign
RFdiffusion	Structural diffusion model	Structure design	https://github.com/RosettaCommons/RFdiffusion
Chroma	Coordinate diffusion model	Structure design	https://github.com/generatebio/chroma
ProteinGenerator	Sequence space diffusion model	Structure design	https://github.com/RosettaCommons/protein_generator
RFdiffusionAA	Structural diffusion model	Structure design	https://github.com/baker-laboratory/rf_diffusion_all_atom
ABACUS-R	Local environment encoding and decoding	Sequence design	https://github.com/JasonWei2014/ABACUS-R
ProteinMPNN	Graph neural network, message passing	Sequence design	https://github.com/dauparas/ProteinMPNN
Frame2seq	Structure-conditioned masked language model	Sequence design	https://github.com/dakpinaroglu/Frame2seq
ProtGPT2	Large language model	Sequence generation	–
XTrimoPGLM	Large language model	Sequence generation	–
ProGen2	Large language model	Sequence generation	https://github.com/enijkamp/progen2
ProGen	Large language model	Sequence generation	https://github.com/salesforce/progen
ProLLaMA	Large language model	Sequence generation	https://github.com/PKU-YuanGroup/ProLLaMA
HelixProtX	Multimodal model	Sequence generation	https://github.com/PaddlePaddle/PaddleHelix/tree/dev/apps/helixprotx
ESM3	Multimodal model	Sequence generation	https://github.com/evolutionaryscale/esm

通过最小化每步预测和真实结构之间的均方误差损失来学习逆转噪声的每个过程，最终收敛于可设计的蛋白质主链结构；该模型除了用于无条件蛋白质单体生成，还可针对特定设计任务，利用折叠信息、功能基序等辅助调节信息进行拓扑约束的对称寡聚体设计、酶功能位点支架设计等，准确度和复杂性高于之前的hallucination，可生成长度达 600 个残基的结构。

Generate Biomedicines 公司基于扩散模型和图神经网络开发的 Chroma 蛋白生成模型[70]，在生成过程实现了更灵活多样的条件采样约束，通过对称性、亚结构、形状约束，甚至自然语言提示，引导定向生成符合所需结构特性和功能属性的蛋白质，从主要关注生成合理骨架到重点关注蛋白功能迈进一步；对 310 种蛋白的实验表征表明，Chroma 生成的新蛋白可以折叠和高度表达且具有良好的生物物理特性。将 RFdiffusion 基于结构的扩散改为基于序列的扩散发展的 ProteinGenerator 模型[71]，在定向生成方面可以使用更丰富多样的属性引导，如静电量、疏水性、氨基酸比例、位置特异性评分矩阵(position-specific scoring matrix, PSSM)等，还可基于多个构象扩散生成变构蛋白，同时生成蛋白结构和对应氨基酸序列。

David Baker 课题组 2023 年设计的荧光素酶是基于深度学习从头设计有催化功能的蛋白质的里程碑式工作[72]，该工作在幻想生成的基础上采用“family-wide hallucination”深度学习方法生成大量包含多种口袋形状的蛋白质结构，从头设计出了高活性和特异性催化小分子底物二苯基四嗪(diphenyltetrazolium, DTZ)的荧光素酶 LuxSit。随着 RoseTTAFold All-Atom[73]和

Alphafold 3[74]的出现，蛋白质领域进入全原子时代。在 RoseTTAFold 网络基础上，通过引入小分子等非蛋白成分开发的 RoseTTAFold All-Atom 进一步推动了 RFdiffusion 的升级，促成了 RFdiffusion All-Atom (RFdiffusionAA)的诞生。RFdiffusionAA 可以直接基于给定的非蛋白分子生成结合该分子的蛋白质，针对 3 种不同小分子设计的蛋白结合物展现了高效结合能力和大幅提升的稳定性[73]。结构生成不断走向定制化、结构和序列同步设计，全原子的发展进一步推进了从头设计结合蛋白或催化的酶蛋白的工业级应用。

3.2 序列设计

蛋白质序列设计又被称为固定骨架设计，当确定了蛋白质三维骨架结构之后，序列设计致力于找到能折叠成该目标结构的氨基酸序列。不同于蛋白质结构预测，序列设计中氨基酸的预测优化与周围蛋白环境密切相关，具有环境依赖性。中国科学技术大学刘海燕课题组提出的 ABACUS-R 算法[75]将中心氨基酸残基的化学和空间结构环境映射后解码为中心残基氨基酸类型等多种特征，在全部或部分序列从头设计任务中能够得到各个位点环境依赖的最适氨基酸类型；从头设计的 57 条序列中，49 条通过实验验证能够可溶表达且折叠成超高热稳定性单体。

ProteinMPNN 借助图神经网络获取局部结构信息[76]，氨基酸节点和边之间的消息传递机制使节点能够迭代地与它们的邻居交换信息，获取每个残基周围的环境结构信息，进而预测出与之对应的氨基酸序列，序列恢复率达到 52.4%，高于 Rosetta 的 32.9%。实验证实，无论是 hallucination 还是 RFdiffusion 生成的蛋白结构，其对应的序列成功率都很低，所以这些方法往往会在生成骨架结构后借助 ProteinMPNN 方法重新进行序列设计优化。ProteinMPNN 作为解决逆向折叠的有效手段，在单体蛋白、环状同源多聚体、纳米颗粒、蛋白-蛋白结合物等多种序列设计任务中取得了高成功率。在 ProteinMPNN 基础上进一步发展的 LigandMPNN 模型能够有效模拟非蛋白质分子，如小分子、核苷酸和金属等，在序列恢复、侧链设计和亲和力优化上均获得了领先的效果[77]。

最近发布的 Frame2seq 是基于结构条件掩蔽语言模型[78]，序列设计准确率与 ProteinMPNN 接近，但设计速度快 6.2 倍；实验测试的 26 条从头设计序列中，22 个可溶、17 个形成稳定折叠，与天然序列同一性最低为 0。总的来说，这些基于 AI 的序列设计方法在实验中有较高的成功率，成为从头设计符合预期结构和功能的蛋白质的一种有效手段。

3.3 端到端序列生成

蛋白氨基酸序列可以看作是一种蛋白质语言，且序列中包含了蛋白结构与功能信息。通过对大量蛋白质序列的学习，蛋白质语言模型能够学习到蛋白质序列中氨基酸排列和进化规律，像大语言模型 ChatGPT 生成新语句一样，生成具有结构和功能的新序列。基于大语言模型开发的从头序列生成方法可以不使用明确的结构信息，也不用进行基于结构的序列设计优化过程，仅基于学习到的蛋白质序列模式，一步端到端从功能描述直接生成不同于自然界的、稳定的功能性新序列，为蛋白质从头设计提供了新的策略。

Ferruz 等[79]受到自然语言模型 GPT 自回归训练方式的启发，提出了蛋白自回归语言模型 ProtGPT2，在涵盖整个蛋白空间的约 5 000 万序列上训练后，能够以高通量方式从头生成全新蛋白质序列。ProtGPT2 生成的序列在采样蛋白质未知区域空间的同时，保留了天然蛋

白质的氨基酸偏好性、二级结构含量及球形度等关键特征。百图生科与清华大学联合研发的 xTrimoPGLM 蛋白质语言模型[80]，能够生成与自然蛋白质结构类似的新蛋白质序列，此外，因模型具有同时解决蛋白质理解和生成 2 大类任务的统一预训练框架，模型还能进行蛋白结构的预测。

为了进一步在模型训练中纳入更深层次的生物学上下文信息，Madani 等基于包含条件标签的自回归模型(conditional transformer language, CTRL)提出了 ProtGen 条件蛋白语言模型，可以生成跨不同家族的功能性蛋白序列[81]。训练过程使用大量蛋白质家族、生物过程和分子功能等属性控制标签，用于控制特定功能的蛋白质序列生成；进一步使用特定家族(5 个溶菌酶家族)序列对模型进行微调，实验证实生成的序列与天然溶菌酶具有相似催化性能，且与已知序列相似性可低至 31%[81]。后续发展的 ProGen2 将参数从 12 亿扩展到 64 亿，在基因组、元基因组及免疫组序列数据上进行训练，在捕捉序列进化信息的分布、新序列生成和蛋白质适应度预测方面达到最先进的性能，且不需要进一步微调[82]。

最近开发的 ProLLaMA 是第一个能够同时处理多个蛋白质语言处理(protein language processing, PLP)任务的蛋白质大语言模型[83]，在无条件蛋白质序列生成任务中，ProLLaMA 在 pLDDT 和 TM-score 等常用指标上均达到了最先进水平；在可控蛋白质生成中，根据蛋白功能的文本描述，ProLLaMA 可从头生成具有所需功能的与天然蛋白质一样强大的新型蛋白质。

另外，混合文本、蛋白质序列、蛋白质结构的多模态生成模型正在快速发展，多模态模型既能生成蛋白结构也能生成序列。比如百度推出的大型多模态模型系统 HelixProtX[84]实现了蛋白质生成任务的统一，包括序列、结构和描述的生成，不仅能够从氨基酸序列生成功能描述，还能从文本描述生成蛋白质序列和结构。Evolutionary Scale AI 发布的多模态生成语言大模型 ESM3[85]，在 2.78 亿个天然蛋白质的数据上进行了训练，不仅能够理解和生成蛋白质序列，还能综合考虑蛋白质的结构与功能。多模态融合模型能够提供更加全面和精准的理解和预测，是未来的一个重要研究方向。

4 展望

近年来，蛋白元件的挖掘、改造和从头设计领域取得了显著进展。通过深度学习和机器学习算法，研究人员能够根据开发需求，快速定位候选蛋白并预测蛋白质的稳定性、亲和力以及其他关键性质，这些技术在药物设计和工业生物催化中展现出巨大潜力。AI 技术的应用不仅提高了蛋白质设计的效率，还为理解蛋白质的复杂性提供了新的视角。随着算法的不断优化和计算能力的提高，可以预见到将出现更加精准和高效的蛋白质设计方法。

未来，AI 和语言模型在蛋白质科学中的作用将变得更加重要。AI 将能够处理更大规模的生物数据，构建更为复杂的模型，以预测蛋白质之间的相互作用和动态变化。此外，语言模型将更加深入地理解蛋白质的“语言”，从而在没有大量实验数据支持的情况下，也能设计出具有预期功能的蛋白质。此外，随着 AI 技术在蛋白质设计领域的不断进步，未来有望实现全原子按需功能蛋白的生成。这意味着，通过精确控制蛋白质的每一个原子，可以设计出具有特定功能的蛋白质，满足特定应用的需求。这种全原子级别的控制将使得蛋白质设计更加精确和灵活，能够针对不同的生物化学过程和环境条件，定制出最优的蛋白质解决方案。

另外，与自动化设施的结合将成为蛋白质设计的重要趋势。通过 AI 设计和实验验证的迭代过程，可以快速筛选出有益的候选蛋白，并进一步优化蛋白质的性能。这种策略不仅能够加速蛋白质的优化过程，还能够减少资源的浪费。同时，与自动化设施的结合将使得蛋白质设计和测试过程更加高效。自动化的蛋白质表达、纯化和功能测试平台将与 AI 模型紧密集成，实现从设计到实验验证的无缝对接，大大缩短研究周期，提高研发效率，推动蛋白质工程向智能化、精准化发展。

作者贡献声明

刘翠：方案设计、初稿写作；史振坤：初稿写作、提供材料；马红武：监督指导、经费支持；廖小平：方案设计、监督指导、稿件润色修改。

作者利益冲突公开声明

作者声明没有任何可能会影响本文所报告工作的已知经济利益或个人关系。

REFERENCES

[1] 赵国屏. 合成生物学: 从“造物致用”到产业转化[J]. 生物工程学报, 2022, 38(11): 4001-4011.
ZHAO GP. Synthetic biology: from “build-for-use” to commercialization[J]. Chinese Journal of Biotechnology, 2022, 38(11): 4001-4011 (in Chinese).

[2] LIU HY, CHEN Q. Computational protein design with data-driven approaches: recent developments and perspectives[J]. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2023, 13(3): e1646.

[3] HUANG PS, BOYKEN SE, BAKER D. The coming of age of *de novo* protein design[J]. Nature, 2016, 537(7620): 320-327.

[4] MEIER J, RAO R, VERKUIL R, LIU J, RIVES A. Language models enable zero-shot prediction of the effects of mutations on protein function[J]. Advances in neural information processing systems 2021, 34: 29287-29303.

[5] LI MC, KANG LQ, XIONG Y, WANG YG, FAN GS, TAN P, HONG L. SESNet: sequence-structure feature-integrated deep learning method for data-efficient protein engineering[J]. Journal of Cheminformatics, 2023, 15(1): 12.

[6] FERRUZ N, HÖCKER B. Controllable protein design with language models[J]. Nature Machine Intelligence, 2022, 4(6): 521-532.

[7] 康里奇, 谈攀, 洪亮. 人工智能时代下的酶工程[J]. 合成生物学, 2023, 4(3): 524-534.
KANG LQ, TAN P, HONG L. Enzyme engineering in the age of artificial intelligence[J]. Synthetic Biology Journal. 2023, 4(3): 524-534 (in Chinese).

[8] 张锦雄, 孟雪莉, 陈燕, 韦松键, 吕丽兰, 胡小春. 大语言模型在蛋白质设计中的应用综述[J]. 基因组学与应用生物学 2024, 43(8): 1303-1320.
ZHANG JX, MENG XL, CHEN Y, WEI SJ, LÜ LL, HU XC. A review for the application of large language model in protein design[J]. Genomics and Applied Biology, 2024, 43(8): 1303-1320 (in Chinese).

[9] 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 N, JIN XC, YANG CZ, WANG ZY, MIN XP, GE SX. *De novo* protein design in the age of artificial intelligence[J]. Chinese Journal of Biotechnology, 2024, 40(11): 3912-3929 (in Chinese).

[10] NOTIN P, ROLLINS N, GAL Y, SANDER C, MARKS D. Machine learning for functional protein design[J]. Nature Biotechnology, 2024, 42(2): 216-228.

[11] SHEN HB, CHOU KC. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses[J]. Biochemical and Biophysical Research Communications, 2007, 364(1): 53-59.

[12] YU CG, ZAVALJEVSKI N, DESAI V, REIFMAN J. Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases[J]. Proteins, 2009, 74(2): 449-460.

[13] ZOU Q, CHEN WC, HUANG Y, LIU XR, JIANG Y. Identifying multi-functional enzyme by hierarchical multi-label classifier[J]. Journal of Computational and Theoretical Nanoscience, 2013, 10(4): 1038-1043.

[14] LI YH, XU JY, TAO L, LI XF, LI S, ZENG X, CHEN SY, ZHANG P, QIN C, ZHANG C, CHEN Z, ZHU F, CHEN YZ. SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity[J]. PLoS One, 2016, 11(8): e0155290.

[15] DALKIRAN A, RIFAIOGLU AS, MARTIN MJ, CETIN-ATALAY R, ATALAY V, DOĞAN T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature[J]. BMC Bioinformatics, 2018, 19(1): 334.

[16] LI Y, WANG S, UMAROV R, XIE BQ, FAN M, LI LH, GAO X. DEEPre: sequence-based enzyme EC number prediction by deep learning[J]. Bioinformatics, 2018, 34(5): 760-769.

[17] RYU JY, KIM HU, LEE SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(28): 13996-14001.

[18] YU TH, CUI HY, LI JC, LUO YN, JIANG GD, ZHAO HM. Enzyme function prediction using contrastive learning[J]. Science, 2023, 379(6639): 1358-1363.

[19] SHI ZK, DENG R, YUAN QQ, MAO ZT, WANG RY, LI HR, LIAO XP, MA HW. Enzyme commission number prediction and benchmarking with hierarchical

dual-core multitask learning framework[J]. Research, 2023, 6: 0153.
[20] FANG H, GOUGH J. DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more[J]. Nucleic Acids Research, 2013, 41(Database issue): D536-D544.
[21] KAHANDA I, BEN-HUR A. GOstruct 2.0: automated protein function prediction for annotated proteins[C]// Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Boston Massachusetts, USA. ACM, 2017: 60-66.
[22] ZHOU NH, JIANG YX, BERGQUIST TR, LEE AJ, KACSOH BZ, CROCKER AW, LEWIS KA, GEORGHIOU G, NGUYEN HN, HAMID MN, DAVIS L, DOGAN T, ATALAY V, RIFAIOGLU AS, DALKIRAN A, CETIN ATALAY R, ZHANG C, HURTO RL, FREDDOLINO PL, ZHANG Y, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens[J]. Genome Biology 2019, 20: 244.
[23] KULMANOV M, KHAN MA, HOEHNDORF R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier[J]. Bioinformatics, 2018, 34(4): 660-668.
[24] YAO SW, YOU RH, WANG SJ, XIONG Y, HUANG XD, ZHU SF. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information[J]. Nucleic Acids Research, 2021, 49: W469-W475.
[25] GLIGORIJEVIĆ V, RENFREW PD, KOSCIOLEK T, LEMAN JK, BERENBERG D, VATANEN T, CHANDLER C, TAYLOR BC, FISK IM, VLAMAKIS H, XAVIER RJ, KNIGHT R, CHO K, BONNEAU R. Structure-based protein function prediction using graph convolutional networks[J]. Nature Communications, 2021, 12: 3168.
[26] KULMANOV M, GUZMÁN-VEGA FJ, DUEK ROGGLI P, LANE L, AROLD ST, HOEHNDORF R. Protein function prediction as approximate semantic entailment[J]. Nature Machine Intelligence, 2024, 6(2): 220-228.
[27] 刘卓, 张飞, 赵心清, 白凤武. 锌指蛋白及人工锌指蛋白对微生物代谢影响的研究进展[J]. 生物工程学报, 2014, 30: 331-340.
LIU Z, ZHANG F, ZHAO XQ, BAI FW. Effects of zinc-finger proteins and artificial zinc-finger proteins on microbial metabolisms: a review[J]. Chinese Journal of Biotechnology, 2014, 30(3): 331-340 (in Chinese).
[28] JUMPER J, EVANS R, PRITZEL A, GREEN T, FIGURNOV M, RONNEBERGER O, TUNYASUVUNAKOOL K, BATES R, ŽÍDEK A, POTAPENKO A. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596: 583-589.
[29] HUANG JY, LIN QP, FEI HY, HE ZX, XU H, LI YJ, QU KL, HAN P, GAO Q, LI BS, LIU GW, ZHANG LX, HU JC, ZHANG R, ZUO EW, LUO YL, RAN YD, QIU JL, ZHAO KT, GAO CX. Discovery of deaminase functions by structure-based protein clustering[J]. Cell, 2023, 186(15): 3182-3195.e14.
[30] YOON PH, ZHANG Z, LOI KJ, ADLER BA, LAHIRI A, VOHRA K, SHI H, RABELO DB, TRINIDAD M, BOGER RS, AL-SHIMARY MJ, DOUDNA JA. Structure-guided discovery of ancestral CRISPR-Cas13 ribonucleases[J]. Science, 2024, 385(6708): 538-543.
[31] VAN KEMPEN M, KIM SS, TUMESCHEIT C, MIRDITA M, LEE J, GILCHRIST CLM, SÖDING J, STEINEGGER M. Fast and accurate protein structure search with Foldseek[J]. Nature Biotechnology, 2024, 42(2): 243-246.
[32] HAMAMSY T, MORTON JT, BERENBERG D, CARRIERO N, GLIGORIJEVIC V, BLACKWELL R, STRAUSS CEM, LEMAN JK, CHO K, BONNEAU R. TM-Vec: template modeling vectors for fast homology detection and alignment[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2022.07.25.501437.
[33] LIU W, WANG ZY, YOU RH, XIE CH, WEI H, XIONG Y, YANG JY, ZHU SF. Author Correction: PLMSearch: protein language model powers accurate and fast sequence search for remote homology[J]. Nature Communications, 2024, 15: 7766.
[34] HONG L, HU ZH, SUN SQ, TANG XR, WANG JM, TAN QX, ZHENG LZ, WANG S, XU S, KING I, GERSTEIN M, LI Y. Fast, sensitive detection of protein homologs using deep dense retrieval[J]. Nature Biotechnology, 2024. DOI: 10.1038/s41587-024-02353-6.
[35] KROLL A, RANJAN S, ENGQVIST MKM, LERCHER MJ. A general model to predict small molecule substrates of enzymes based on machine and deep learning[J]. Nature Communications, 2023, 14(1): 2787.
[36] KROLL A, ENGQVIST MKM, HECKMANN D, LERCHER MJ. Deep learning allows genome-scale prediction of Michaelis constants from structural features[J]. PLoS Biology, 2021, 19(10): e3001402.
[37] LI FR, YUAN L, LU HZ, LI G, CHEN Y, ENGQVIST MKM, KERKHOVEN EJ, NIELSEN J. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction[J]. Nature Catalysis, 2022, 5(8): 662-672.
[38] KROLL A, ROUSSET Y, HU XP, LIEBRAND NA, LERCHER MJ. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning[J]. Nature Communications, 2023, 14(1): 4139.
[39] YU H, DENG HX, HE JH, KEASLING JD, LUO XZ. UniKP: a unified framework for the prediction of enzyme kinetic parameters[J]. Nature Communications, 2023, 14(1): 8211.
[40] LI G, BURIC F, ZRIMEC J, VIKNANDER S, NIELSEN J, ZELEZNIAK A, ENGQVIST MKM. Learning deep representations of enzyme thermal adaptation[J]. Protein Science, 2022, 31(12): e4480.
[41] GADO JE, KNOTTS M, SHAW AY, MARKS D, GAUTHIER NP, SANDER C, BECKHAM GT. Deep learning prediction of enzyme optimum pH[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2023.06.22.544776.
[42] SHI ZK, WANG DH, LI Y, DENG R, LIN JW, LIU C, LI HR, WANG RY, ZHAO MQ, MAO ZT, YUAN QQ, LIAO XP, MA HW. REME: an integrated platform for reaction enzyme mining and evaluation[J]. Nucleic

Acids Research, 2024, 52W299-W305.
[43] LIU BB, QU G, LI JK, FAN WC, MA JN, XU Y, NIE Y, SUN ZT. Conformational dynamics-guided loop engineering of an alcohol dehydrogenase: capture, turnover and enantioselective transformation of difficult-to-reduce ketones[J]. Advanced Synthesis & Catalysis, 2019, 361(13): 3182-3190.
[44] 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.
[45] TIAN C, YANG JG, LIU C, CHRN P, MEN Y, MA HW, SUN YX, MA YH. Engineering substrate specificity of HAD phosphatases and multienzyme systems development for the thermodynamic-driven manufacturing sugars[J]. Nature Communications, 2022, 13(1): 3582.
[46] YU SS, LI JL, YAO PY, FENG JH, CUI YF, LI JJ, LIU XT, WU QQ, LIN JP, ZHU DM. Inverting the enantiopreference of nitrilase-catalyzed desymmetric hydrolysis of prochiral dinitriles by reshaping the binding pocket with a mirror-image strategy[J]. Angewandte Chemie International Edition, 2021, 60(7): 3679-3684.
[47] QU G, BI YX, LIU BB, LI JK, HAN X, LIU WD, JIANG YY, QIN ZM, SUN ZT. Unlocking the stereoselectivity and substrate acceptance of enzymes: proline-induced loop engineering test[J]. Angewandte Chemie International Edition, 2022, 61(1): e202110793.
[48] RIESSELMAN AJ, INGRAHAM JB, MARKS DS. Deep generative models of genetic variation capture the effects of mutations[J]. Nature Methods, 2018, 15: 816-822.
[49] FRAZER J, NOTIN P, DIAS M, GOMEZ A, MIN JK, BROCK K, GAL Y, MARKS DS. Disease variant prediction with deep generative models of evolutionary data[J]. Nature, 2021, 599(7883): 91-95.
[50] RAO RM, LIU J, VERKUIL R, MEIER J, CANNY J, ABBEEL P, SERCU T, RIVES A. MSA transformer[C]//International Conference on Machine Learning (PMLR), 2021: 8844-8856.
[51] NOTIN P, DIAS M, FRAZER J, MARCHENA-HURTADO J, GOMEZ AN, MARKS D, GAL Y. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval[C]//International Conference on Machine Learning (PMLR), 2022: 16990-17017.
[52] HSU C, VERKUIL R, LIU J, LIN Z, HIE B, SERCU T, LERER A, RIVES A. Learning inverse folding from millions of predicted structures[C]. International Conference on Machine Learning (PMLR), 2022: 8946-8970.
[53] SHANKER VR, BRUUN TUJ, HIE BL, KIM PS. Unsupervised evolution of protein and antibody complexes with a structure-informed language model[J]. Science, 2024, 385(6704): 46-53.
[54] KULIKOVA AV, DIAZ DJ, LOY JM, ELLINGTON AD, WILKE CO. Learning the local landscape of protein structures with convolutional neural networks[J]. Journal of Biological Physics, 2021, 47(4): 435-454.
[55] LU HY, DIAZ DJ, CZARNECKI NJ, ZHU CZ, KIM W, SHROFF R, ACOSTA DJ, ALEXANDER BR, COLE HO, ZHANG Y, LYND NA, ELLINGTON AD, ALPER HS. Machine learning-aided engineering of hydrolases for PET depolymerization[J]. Nature, 2022, 604(7907): 662-667.
[56] D'OELSNITZ S, DIAZ DJ, KIM W, ACOSTA DJ, DANGERFIELD TL, SCHECHTER MW, MINUS MB, HOWARD JR, DO H, LOY JM, ALPER HS, ZHANG YJ, ELLINGTON AD. Biosensor and machine learning-aided engineering of an amaryllidaceae enzyme[J]. Nature Communications, 2024, 15(1): 2084.
[57] CHENG J, NOVATI G, PAN J, BYCROFT C, ŽEMGULYTĖ A, APPLEBAUM T, PRITZEL A, WONG LH, ZIELINSKI M, SARGEANT T, SCHNEIDER RG, SENIOR AW, JUMPER J, HASSABIS D, KOHLI P, AVSEC Ž. Accurate proteome-wide missense variant effect prediction with AlphaMissense[J]. Science, 2023, 381(6664): eadg7492.
[58] SU J, HAN C, ZHOU Y, SHAN J, ZHOU X, YUAN F. SaProt: protein language modeling with structure-aware vocabulary[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2023.10.01.560349.
[59] LI M, TAN Y, MA X, ZHONG BZT, YU HQ, ZHOU ZY, OUYANG WL, ZHOU BX, HONG L, TAN P. ProSST: protein language modeling with quantized structure and disentangled attention[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2024.04.15.589672.
[60] 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.
[61] 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(1): 5743.
[62] BISWAS S, KHIMULYA G, ALLEY EC, ESVELT KM, CHURCH GM. Low-N protein engineering with data-efficient deep learning[J]. Nature Methods, 2021, 18(4): 389-396.
[63] ZHOU ZY, ZHANG L, YU YX, WU BH, LI MC, HONG L, TAN P. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning[J]. Nature Communications, 2024, 15(1): 5566.
[64] DU ZY, SU H, WANG WK, YE LS, WEI H, PENG ZL, ANISHCHENKO I, BAKER D, YANG JY. The trRosetta server for fast and accurate protein structure prediction[J]. Nature Protocols, 2021, 16(12): 5634-5651.
[65] 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.
[66] 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(7889): 547-552.
[67] WU KE, YANG KK, van den BERG R, ALAMDARI S, ZOU JY, LU AX, AMINI AP. Protein structure generation *via* folding diffusion[J]. Nature Communications, 2024, 15(1): 1059.
[68] LEE JS, KIM J, KIM PM. ProteinSGM: score-based generative modeling for *de novo* protein design[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2022.07.13.499967.
[69] 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, Vázquez-Torres S, Lauko A, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2022.12.09.519842
[70] INGRAHAM JB, BARANOV M, COSTELLO Z, BARBER KW, WANG WJ, ISMAIL A, FRAPPIER V, LORD DM, NG-THOW-HING C, van VLACK ER, TIE S, XUE V, COWLES SC, LEUNG A, RODRIGUES JV, MORALES-PEREZ CL, AYOUB AM, GREEN R, PUENTES K, OPLINGER F, et al. Illuminating protein space with a programmable generative model[J]. Nature, 2023, 623(7989): 1070-1078.
[71] LISANZA SL, GERSHON JM, TIPPS S, ARNOLDT L, HENDEL S, SIMS JN, LI X, BAKER D. Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2023.05.08.539766.
[72] 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(7949): 774-780.
[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, DiMAIO F, BAKER D. Generalized biomolecular modeling and design with RoseTTAFold All-Atom[J]. Science, 2024, 384(6693): eadl2528.
[74] ABRAMSON J, ADLER J, DUNGER J, EVANS R, GREEN T, PRITZEL A, RONNEBERGER O, WILLMORE L, BALLARD AJ, BAMBRICK J, BODENSTEIN SW, EVANS DA, HUNG CC, O'NEILL M, REIMAN D, TUNYASUVUNAKOOL K, WU Z, ŽEMGULYTĖ A, ARVANITI E, BEATTIE C, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024, 630: 493-500.
[75] 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(7): 451-462.
[76] 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.
[77] DAUPARAS J, LEE GR, PECORARO R, AN L, ANISHCHENKO I, GLASSCOCK C, BAKER D. Atomic context-conditioned protein sequence design using LigandMPNN[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2023.12.22.573103.
[78] AKPINAROGLU D, SEKI K, GUO A, ZHU E, KELLY MJS, KORTEMME T. Structure-conditioned masked language models for protein sequence design generalize beyond the native sequence space[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2023.12.15.571823.
[79] FERRUZ N, SCHMIDT S, HÖCKER B. ProtGPT2 is a deep unsupervised language model for protein design[J]. Nature Communications, 2022, 13(1): 4348.
[80] CHEN B, CHENG X, LI P, GENG Y-A, GONG J, LI S, BEI Z, TAN X, WANG B, ZENG X. xTrimoPGLM: unified 100B-scale pre-trained transformer for deciphering the language of protein[J/OL]. arXiv. [2024-08-01]. https://doi.org/10.48550/arXiv.2401.06199.
[81] MADANI A, KRAUSE B, GREENE ER, SUBRAMANIAN S, MOHR BP, HOLTON JM, OLMOS JL Jr, 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(8): 1099-1106.
[82] 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.
[83] LV L, LIN ZY, LI H, LIU YY, CUI JX, CHEN CY-C, YUAN L, TIAN YH. ProLLaMa: a protein large language model for multi-task protein language processing[J/OL]. arXiv. [2024-08-01]. https://doi.org/10.48550/arXiv.2402.16445.
[84] CHEN Z, CHEN T, XIE C, XUE Y, ZHANG X, ZHOU J, FANG X. Unifying sequences, structures, and descriptions for any-to-any protein generation with the large multimodal model HelixProtX[J/OL]. arXiv. [2024-08-01]. https://doi.org/10.48550/arXiv.2407.09274.
[85] HAYES T, RAO R, AKIN H, SOFRONIEW NJ, OKTAY D, LIN Z, VERKUIL R, TRAN VQ, DEATON J, WIGGERT M. Simulating 500 million years of evolution with a language model[J/OL]. bioRxiv. [2024-08-01]. https://doi.org/10.1101/2024.07.01.600583.