• AI 驱动底层技术 •

王猛 中国科学院天津工业生物技术研究所研究员、博士生导师。主要研究方向为合成生物学、高通量自动化技术等。领导团队建立的高通量编辑与筛选平台实验室初步实现软、硬件装备和生物技术的整合，建立高通量自动化合成生物改造技术体系，完成大肠杆菌、谷氨酸棒杆菌、酿酒酵母、枯草芽孢杆菌等多种模式合成生物的高通量自动化改造，最大通量达到 9 000 位点/月以上。在 *Nature Communications*、*Science Advances*、*Trends in Biotechnology*、*ACS Catalysis*、*Metabolic Engineering* 等国际期刊上发表 55 篇文章，申请专利 14 项。

基于人工智能的 CRISPR-Cas 系统的设计、挖掘与改造

毛雨丰[1,2,3]，储光芸[1,2]，梁庆玲[1,2]，刘叶[1,2]，杨毅[1,2]，廖小平[1,2,3]，王猛[1,2,3*]

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
3 低碳合成工程生物学全国重点实验室，天津 300308

毛雨丰, 储光芸, 梁庆玲, 刘叶, 杨毅, 廖小平, 王猛. 基于人工智能的 CRISPR-Cas 系统的设计、挖掘与改造[J]. 生物工程学报, 2025, 41(3): 949-967.

MAO Yufeng, CHU Guangyun, LIANG Qingling, LIU Ye, YANG Yi, LIAO Xiaoping, WANG Meng. Artificial intelligence-assisted design, mining, and modification of CRISPR-Cas systems[J]. Chinese Journal of Biotechnology, 2025, 41(3): 949-967.

摘 要：随着合成生物学的兴起，CRISPR-Cas 系统作为基因编辑的核心工具在医药、农业和工业生物技术等领域展现了巨大潜力。本文综述了人工智能(artificial intelligence, AI)技术在 CRISPR-Cas 系统设计、挖掘与改造中的应用进展。AI 技术，特别是机器学习，通过分析高通量测序数据，优化 sgRNA 设计、提升编辑效率、预测脱靶效应。本文讨论了 AI 在单链引导 RNA (single guide RNA, sgRNA)设计与评估中的应用，并对基于机器学习的 CRISPR 阵列、Cas 蛋白的注释与挖掘，以及 AI 在 CRISPR 相关的基因编辑关键蛋白改造中的潜力也进行了重点探讨。这些研究不仅提高了基因编辑的效率和精确性，还为基因组工程开辟了新的可能性，也为实现智能化和精准化的基因组编辑奠定了基础。

关键词：合成生物学；基因编辑；CRISPR-Cas；单链引导 RNA 设计；酶的挖掘与改造

资助项目：中国科学院战略性先导科技专项(XDC0110201)；国家自然科学基金(32101186, 32301273, 12326611)
This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDC0110201) and the National Natural Science Foundation of China (32101186, 32301273, 12326611).
*Corresponding author. E-mail: wangmeng@tib.cas.cn
Received: 2024-10-31; Accepted: 2025-02-18; Published online: 2025-02-19

Artificial intelligence-assisted design, mining, and modification of CRISPR-Cas systems

MAO Yufeng[1,2,3], CHU Guangyun[1,2], LIANG Qingling[1,2], LIU Ye[1,2], YANG Yi[1,2], LIAO Xiaoping[1,2,3], WANG Meng[1,2,3*]

1 Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
2 National Center of Technology Innovation for Synthetic Biology, Tianjin 300308, China
3 State Key Laboratory of Engineering Biology for Low-carbon Manufacturing, Tianjin 300308, China

Abstract: With the rapid advancement of synthetic biology, CRISPR-Cas systems have emerged as a powerful tool for gene editing, demonstrating significant potential in various fields, including medicine, agriculture, and industrial biotechnology. This review comprehensively summarizes the significant progress in applying artificial intelligence (AI) technologies to the design, mining, and modification of CRISPR-Cas systems. AI technologies, especially machine learning, have revolutionized sgRNA design by analyzing high-throughput sequencing data, thereby improving the editing efficiency and predicting off-target effects with high accuracy. Furthermore, this paper explores the role of AI in sgRNA design and evaluation, highlighting its contributions to the annotation and mining of CRISPR arrays and Cas proteins, as well as its potential for modifying key proteins involved in gene editing. These advancements have not only improved the efficiency and precision of gene editing but also expanded the horizons of genome engineering, paving the way for intelligent and precise genome editing.

Keywords: synthetic biology; gene editing; CRISPR-Cas; sgRNA design; enzyme mining and modification

基因组编辑和基因表达调控是合成生物学研究中改造或再造生命的重要手段之一。近年来,随着基因组编辑技术,特别是CRISPR-Cas[1]系统的发展与应用，研究人员如今能够在基因组的特定位点进行快速、精确、高效的遗传修改，实现基因的敲除、插入或替换，大幅提升了合成生物学研究的效率。CRISPR-Cas 系统以其高效、精准和易操作的特点，广泛应用于基础分子生物学、医学研究、工业菌种改造、农业品种改良以及环境保护等领域。CRISPR 技术源自细菌的噬菌体免疫机制，主要可以分为 3 个阶段：(1) 适应阶段[2]。外源 DNA 入侵时，Cas1 和 Cas2 复合物识别并切割特定 DNA 片段，将其整合到 CRISPR 阵列中，形成新的间隔序列。此过程由原型间隔区相邻基序(protospacer adjacent motif, PAM)序列介导，PAM 在不同物种中有所差异，帮助区分自身与外源 DNA。(2) CRISPR RNA (crRNA)表达阶段[3]。CRISPR 阵列被转录为前体 CRISPR RNA (pre-crRNA)，并与 tracrRNA 互补配对，触发核酸酶切割生成成熟的 crRNA，指导特定的核酸酶识别外源 DNA。(3) 干扰阶段[4]。成熟的 crRNA 与 tracrRNA 共同引导 Cas 蛋白(如 Cas9)寻找并识别特定的靶序列。一旦匹配成功，并且附近存在可用的 PAM 序列，Cas 蛋白即可切割目标 DNA，破坏外源遗传物质。这种由 RNA 引导的靶向 DNA 切割具有高度的灵活性和可编程性。结合宿主的同源重组或非同源末端连接修

复机制，该方法能够实现对目标 DNA 序列的精准编辑。随后研究人员对系统进行了改造，利用短 RNA 连接子(linker)将 crRNA 和 tracrRNA 连接为单一的向导 RNA (single guide RNA, sgRNA)来引导 Cas9 特异性切割靶标 DNA，使系统更加精简和高效。

目前，CRISPR-Cas 系统主要分为 2 大类[5]：I 类系统，包括 I 型、III 型和 IV 型，使用多亚基蛋白的效应复合体来切割核酸；II 类系统，一般由具有多个结构域的单一蛋白发挥作用，包括 II 型、V 型和 VI 型。相比于 I 类系统，II 类系统通常依赖单一蛋白来执行功能，因其简洁性而在基因编辑技术中更为常用。目前，II 型的 Cas9 蛋白[例如来源于酿脓链球菌(*Streptococcus pyogenes*)的 spCas9]是基因组编辑领域中最为广泛使用的 Cas 系统[6]。此外，其他物种和类型的 Cas 蛋白也被陆续发现并应用于基因组编辑，如金黄色葡萄球菌 Cas9 (SaCas9[7])、V 型的 Cas12 蛋白[8]以及 VI 型的 Cas13 家族蛋白[9]。这些蛋白作为替代选项，提供了多样化的活性、PAM 识别能力、RNA 靶向特性以及传递便利性[1]，适应不同应用场景的需求。此外，通过改造 Cas 蛋白或与其他功能蛋白(如脱氨酶、逆转录酶、DNA 甲基化酶等)融合，多种 Cas 衍生技术被开发出来，包括 CRISPRi[10]、CRISPRa[11]、碱基编辑[12-13]、引导编辑[14]和表观基因组编辑[15]等。这些技术不产生双链断裂即可实现基因抑制、激活、碱基替换、短片段插入或删除，或通过 DNA 甲基化/组蛋白修饰对目标基因进行调控，以满足更广泛的细胞工程改造需求。CRISPR-Cas 系统的有效性依赖于其组分(sgRNA、Cas 酶以及耦合的其他酶)的特性。为了获得高效的 CRISPR-Cas 系统，通常需要对 sgRNA、Cas 酶以及脱氨酶等其他功能元件进行精心设计和优化。

近年来，人工智能(artificial intelligence，AI)，特别是机器学习技术的进步，为 CRISPR-Cas 系统中 sgRNA 的设计与评估、关键酶元件的发掘与定向进化带来了全新的机遇(图 1)。结合高通量深度测序的编辑数据，AI 技术不仅能够优化 sgRNA 设计，提升 CRISPR-Cas 系统的编辑效率，还能有效预测和评估脱靶效应[16]；此外，基于 AI 尤其是深度学习模型(如 AlphaFold[17]等)的蛋白质三维结构预测工具为基于结构相似性分析的关键酶发掘与定向进化提供了强大助力。

目前，已有多篇综述全面回顾了 sgRNA 设计与脱靶评估中的机器学习模型算法原理与应用效果[16,18-19]，以及 AI 技术在酶设计领域的应用及其相关原理[20-22]。因此，本文将重点从生物学视角综述机器学习在基因编辑功能元件设计中的应用，简要介绍 AI 技术在 sgRNA 设计中的应用，重点探讨 AI 技术在关键酶元件的注释、发掘与改造中的作用，并展望未来的发展方向，以期为基于 AI 的基因组编辑的智能化和精准化奠定基础。

1 sgRNA 的设计与评估

在特定的 Cas 核酸酶系统(如 Cas9、Cpf1 等)中，sgRNA 设计的重点在于 pre-crRNA 序列的构建。理论上，只要 sgRNA 的 5′端前 20 bp 序列与目标 DNA 序列完全互补，sgRNA-Cas9 复合物就应能够结合该位点并完成切割。然而，研究结果显示，不同的 sgRNA 序列在编辑效率上存在显著差异[19,23-24]。sgRNA 的编辑效率(on-target)不仅受到 PAM 种类与位置，以及 sgRNA 与靶序列之间碱基配对的影响，还受到 DNA 序列特征[25]、靶序列特定位置的碱基偏好性[26-28]以及下游的 DNA 序列[26,29]、染色质特征[26,30]和一些表观遗传因素[31]等的影响。而在特异性(off-target)评估方面，除了目标序列在参考基因组中的特异性[32-33]外，还受到染色质状态[34-35]、

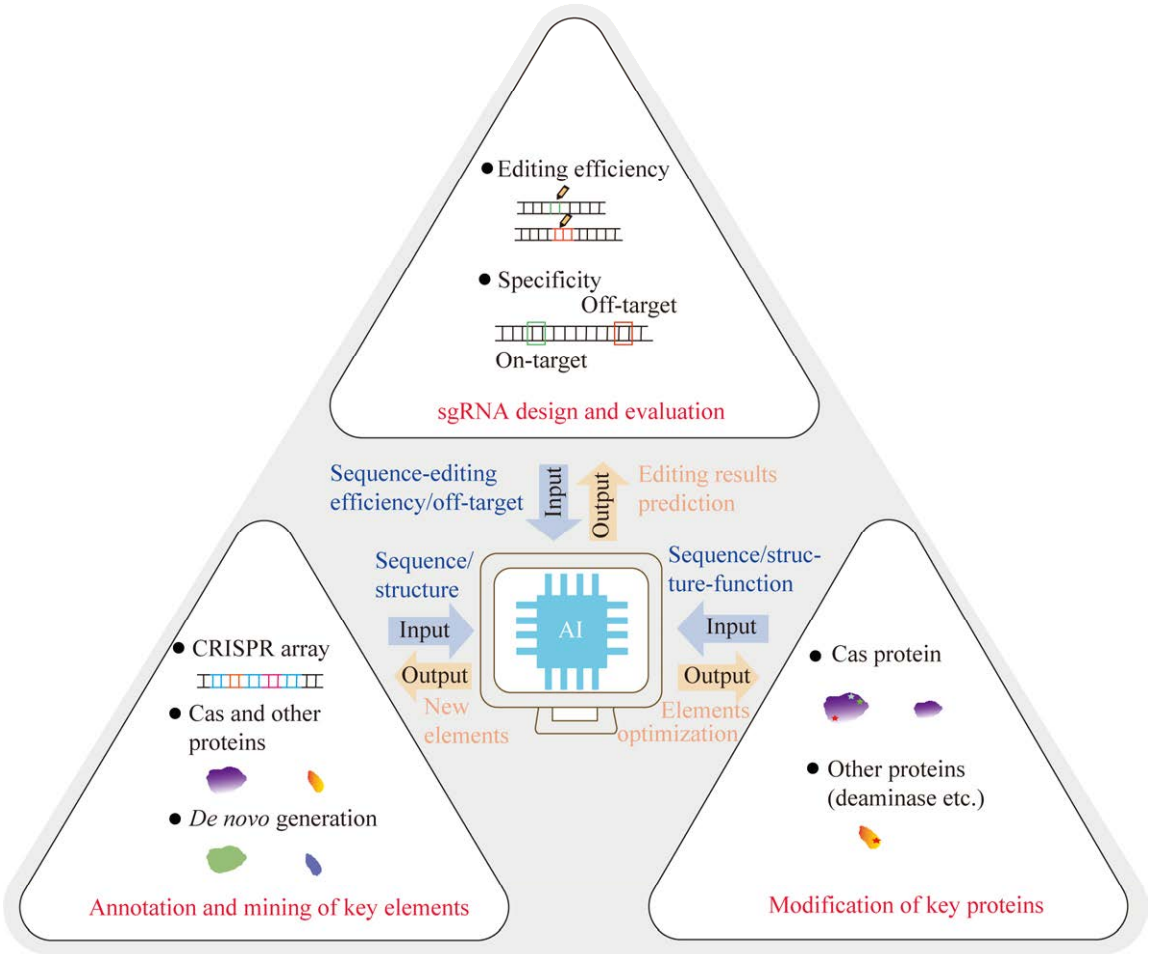

图 1　人工智能指导的 CRISPR-Cas 系统的设计、挖掘与改造

Figure 1　Design, discovery and engineering of CRISPR-Cas systems guided by artificial intelligence.

Cas 酶特征[36-37]等因素的影响。鉴于 sgRNA 设计中存在诸多影响因素，设计具有高编辑效率和特异性的 sgRNA 仍然是一项挑战。

近年来，计算生物学和生物信息学的研究人员开发了一系列高效的工具和方法，以加速这一进程。通常采用 2 种基本方法[19,38]：一类是基于假设先验的方法，这些工具综合考虑多个因素来评估 sgRNA；另一类是基于机器学习的方法(表 1)，通过利用已报道或实验测得的数据训练模型，以预测 sgRNA 的可用性。基于假设先验的方法高度依赖理论模型，可能忽视某些重要的生物学特征，导致无法反映全面的特征图景。相比之下，基于大量实验数据的机器学习方法能够捕捉更复杂的特征和模式，预示着 sgRNA 设计的未来发展趋势。

1.1　基于机器学习的 sgRNA 的编辑效率评估

用于 CRISPR 同源重组技术的 sgRNA 机器学习训练的数据集通常可分为功能筛选数据集和 sgRNA-靶点对合成库数据集[52]。功能筛选数据集主要依赖于细胞表型变化(如药物耐受性、细胞生长等)来推断编辑效果，代表性数据集包括 Rule set 2[30]、DeepCRISPR[50]等。这些数据集易于获取，并且可直接用于基因型-表型关联

表 1　代表性的基于机器学习的 guide RNA 设计工具

Table 1　Representative machine learning-based guide RNA design tools

Tool name	Accession link	Year	Machine learning function	Cas type	Data Source	Reference
Azimuth 2.0	Web tool: None Source code: https://github.com/MicrosoftResearch/Azimuth/releases/tag/v2.0	2016	On-target prediction	Cas9	Human	[30]
DeepSpCas9	Web tool: http://deepcrispr.info/DeepSpCas9/ Source code: https://github.com/MyungjaeSong/Paired-Library	2019	On-target prediction	Cas9	Human	[39]
DeepHF	Web tool: http://www.deephf.com/ Source code: https://github.com/izhangcd/DeepHF	2019	On-target prediction	Cas9	Human	[40]
C-RNNCrispr	Web tool: None Source code: https://github.com/Peppags/C_RNNCrispr	2020	On-target prediction	Cas9	Human	[41]
Be-Hive	Web tool: https://www.crisprbehive.design/ Source code: https://github.com/maxwshen/be_predict_efficiency	2020	On-target prediction	Cas9-ABE/CBE	Human/mouse	[42]
DeepPE	Web tool: https://www.crisprbehive.design/ Source code: https://github.com/maxwshen/be_predict_efficiency	2021	On-target prediction	Cas9-revertase	Human	[43]
CAELM	Web tool: None Source code: https://github.com/YQLiCAS/BE4max	2022	On-target prediction	Cas9-CBE	Human/mouse	[44]
Elevation-aggregation	Web tool [a]: https://crispr.ml/ Source code [a]: http://research.microsoft.com/en-us/projects/crispr	2018	Off-target prediction	Cas9	Human	[45]
CNN_std	Web tool: None Source code: https://github.com/MichaelLinn/off_target_prediction	2018	Off-target prediction	Cas9	Human	[46]
CRISPR-Net	Web tool: None Source code: https://codeocean.com/capsule/9553651/tree/v1	2020	Off-target prediction	Cas9	Human	[47]
Quantitative CRISPRi Design	Web tool: http://www.thu-big.com/sgRNA design/Quantitative CRISPRi Design/ Source code: https://github.com/fenghuibao/CRISPR mismatch analysis	2021	Off-target prediction	dCas9	Bacteria	[48]
CRISTA	Web tool [a]: https://crista.tau.ac.il/ Source code: None	2017	On/off-target prediction	Cas9	Human	[49]
DeepCRISPR	Web tool [a]: http://www.deepcrispr.net/ Source code: https://github.com/bm2-lab/DeepCRISPR	2018	On/off-target prediction	Cas9	Human	[50]
CRISPRon/off	Web tool: https://rth.dk/resources/crispr/ Source code: https://github.com/RTH-tools/crispron	2022	On/off-target prediction	Cas9	Human	[51]

[a]: The web tools or source code of CRISTA, Elevation-aggregation, and DeepCRISPR were inaccessible during testing from October to December 2024.

分析，但缺乏详细的编辑结果，并容易受到细胞存活率的影响[53]。相较之下，通过构建sgRNA-靶点对合成库数据集能够获得单核苷酸分辨率的定量编辑结果，为基因编辑[如插入缺失(indel)率和编辑模式]提供更精确的信息，因此更适合机器学习训练。主流数据集包括CINDEL[54]、DeepSpCas9[39]等。然而，构建这种数据集的实验成本较高，导致数据来源有限，且大多数数据来自人类和小鼠等高等动物细胞系，细菌来源的数据较少，不同物种或实验条件下的结果可能存在较大差异[55]。这种数据稀疏性和异质性使得许多工具只能使用有限的数据集进行训练，预测结果可能与实际情况偏差较大。例如，Konstantakos 等[16]对 DeepCRISPR、DeepHF 等 8 种基于机器学习的 sgRNA 设计工具/模型和 1 种基于假设先验的工具 E-CRISP 进行了系统的测评；他们选取了公开的、独立的、来自不同物种和不同收集方式的 6 个数据集，结果显示各模型的预测能力在不同测试集上差异显著，没有单一模型在所有数据集上表现优于其他模型，但基于机器学习的工具表现通常优于假设先验的工具。

深度学习特别是集成学习、迁移学习等技术的发展，为应对数据稀疏性和异质性挑战带来了新的机遇。集成学习是一种将多个模型的预测结果组合起来以提高整体预测准确性的技术。集成学习通过堆叠、加权平均或投票等方法，将多个神经网络的预测结果进行集成，以减少单个模型的偏差。集成学习在 CRISPR-Cas 系统的研究中被广泛应用，例如，清华大学汪小我团队[56]使用了 16 个实验 sgRNA 数据集对 15 种公共 sgRNA 设计算法进行了全面的比较分析，确定了表现最佳的算法，并进一步实施了各种计算策略构建了集成模型。验证分析表明，新的集成模型在预测不同实验条件下 sgRNA 效率方面比任何单一算法单独使用都有更好的性能。这种框架结合了多种深度学习模型的优势，有效提升了基因编辑关键元件的预测能力和适应性。迁移学习是一种对预训练模型进行调整以解决新任务的机器学习技术。这种方法以基于初始数据集训练得到的模型作为起点，然后根据新任务的特定数据集对模型参数进行微调，是另一条应对基因编辑数据的稀疏性和异质性挑战的策略。Ham 等[53]使用基于 *ccdB* 致死基因的双质粒正选择系统，构建了基于 TevSpCas9 的 279 条和基于 SpCas9 的 303 条 sgRNA 的活性的高质量数据集；直接基于这些数据集构建的模型的 Spearman 相关系数分别为 0.308 和 0.417；随后，他们利用报道的大肠杆菌编辑效率数据集(约 4 万条 sgRNA 的 Guo-eSpCas9 数据集和约 4.5 万条 sgRNA 的 Guo-spCas9 数据集)建立了基础模型；对于 TevSpCas9 数据集，使用 Guo-eSpCas9 数据集作为基础模型进行迁移学习后，Spearman 相关系数提升至 0.630；对于 SpCas9 数据集，联合使用 Guo SpCas9 数据集和 Guo-eSpCas9 数据集作为基础模型进行迁移学习后，Spearman 相关系数提升至 0.627。这些结果显著高于之前报道的模型[39-41,47,55]在相同数据集上的最佳预测结果(Spearman 相关系数 0.52)。此外，Ham 等[53]还构建了针对啮齿枸橼酸杆菌(*Citrobacter rodentium*)(约3.1万条)和肠炎沙门氏菌(*Salmonella enterica*)(296 条)的 sgRNA 数据集，2 种迁移模型在这 2 个数据集上的预测的 Spearman 相关系数均达到了 0.612–0.678 的较高水平，证明了迁移模型的预测精准度与泛化能力。

针对其他 CRISPR 衍生技术，例如碱基编辑技术和引导编辑技术，也有相应的机器学习模型(例如 Be-Hive[42]、DeepPE[43]等)被开发出来用于评估编辑效率。这些模型同样面临数据

稀疏性和异质性的挑战。此外，sgRNA 编辑效率的数据集大多数仅针对基因组上少数靶位点或者事先插入的一段异源序列，这使得目标序列的多样性受到限制，不利于实际编辑效率的预测。为此，研究者们开始关注构建大规模的体内的基因组原位编辑数据集。例如，中国科学院天津工业生物技术研究所的王猛团队联合所内张学礼团队和毕昌昊团队[44]基于哺乳动物细胞高通量自动化基因编辑平台测定了 1 134 个哺乳动物细胞 293T 原位编辑数据，创新性地在机器学习模型构建中整合了目标序列的染色体特征，首次提出染色质可及性对编辑结果的影响权重为靶点序列的 1/6。该模型可以比现有模型更准确地预测细胞原位靶点的碱基编辑结果，也为进一步优化 sgRNA 的编辑效率评估提供了一个自动化辅助构建原位编辑数据集的可行参考。

1.2 基于机器学习的 sgRNA 的特异性评估

用于 CRISPR-Cas9 系统 sgRNA 特异性评估的数据集来源大致可以分为基于全基因组测序的数据集和双链断裂标记数据集。全基因组测序是评估基因编辑脱靶效应的常用策略，通过扫描整个基因组，可以检测未知的脱靶突变、大片段插入、缺失以及基因组结构变异[57-58]。然而，由于测序深度和覆盖度的限制，这些数据集在识别低频突变时可能存在偏差。此外，全基因组测序的成本较高，数据获取也较为有限。相比之下，双链断裂标记的数据集通过标记双链断裂(double strand breaks, DSB)并富集相关区域以检测脱靶效应。例如 Digenome-seq[59]、CIRCLE-seq[60]、BLESS[61]和 GUIDE-seq[62]等技术利用生物标记测定 DSB，能够提供精确的 Cas 蛋白切割位点信息，并且不依赖于预测模型。这类数据集在收集识别低频脱靶位点(如频率为 0.01%–0.1%)方面表现出色[59-62]。然而，体外数据集(如 Digenome-seq 和 CIRCLE-seq)因不受细胞内环境的影响，无法完全模拟体内的脱靶效应。体内数据集如 BLESS 只能捕捉瞬时断裂，而 GUIDE-seq 则依赖 dsODN 转化效率。这些局限性导致不同实验条件的数据集的适用性有所差异。

总体而言，基于全基因组测序和双链断裂标记的数据集各有优势：前者适用于大规模未知脱靶效应的全面评估，后者则更适合高灵敏度的精确检测。这两类数据集在基于机器学习的脱靶效应预测工具的开发中均发挥了重要作用。基于这些数据集已经开发了多种机器学习(CRISTA[49]、Elevation[45]等)和深度学习(CNN_std[46]、CRISPR-Net[47]、DeepCRISPR[50]、CRSIPRon/off[51]等)工具。然而，如同编辑效率预测所面临的挑战，脱靶评估数据集也存在数据稀疏性和异质性的问题。因此，提升模型在不同实验条件下的泛化能力仍然是亟待解决的难题。

对于其他 CRISPR 衍生技术，特别是 CRISPRi 和 CRISPRa，由于不同错配容忍度导致脱靶结合比脱靶切割更容易发生[63-64]，预测脱靶切割的模型无法直接用于预测脱靶结合。因此，需要专门测定用于预测脱靶结合的数据集[48,65-66]。例如，清华大学张翀团队[48]在细菌中开发了针对 CRISPRi 系统的筛选系统；通过将 sgRNA 与目标 DNA 序列的结合亲和力与细菌细胞生长相耦合，采用 *sacB* (蔗糖致死基因)作为反向筛选标记，在蔗糖选择性培养条件下，sgRNA 与目标 DNA 的结合亲和力越高，越能抑制 *sacB* 转录，维持细胞生长。基于该系统，研究人员绘制了 dCas9 在带有特定错配的 sgRNA 引导下与 DNA 靶点的体内结合亲和力全景图，并基于该数据集训练了深度学习模型，开发了面向 CRISPRi 的 sgRNA 设计和脱靶评估的

在线工具 Quantitative CRISPRi Design[48]。

2 基因编辑关键元件的注释、挖掘与生成

CRISPR-Cas 系统的注释与挖掘主要涉及 2 个方面：CRISPR 阵列的识别以及 Cas 蛋白、碱基脱氨酶等功能蛋白的识别与挖掘。CRISPR 阵列由重复序列和间隔序列交替排列构成，对其准确识别是预测和解析 CRISPR-Cas 系统的首要步骤。识别 CRISPR 阵列的关键在于检测重复序列的特征，包括长度、序列保守性及间隔序列的多样性。而 Cas 等功能蛋白的鉴定与挖掘则需要综合考虑其序列、结构和进化特征，以便准确识别和区分不同的功能蛋白。

近年来，计算生物学和生物信息学的研究人员开发了一系列高效的工具和方法[67-70]，以加速 CRISPR 阵列的识别和 Cas 等功能蛋白的注释。通常采用 2 种基本方法：一类是基于序列特征检测的方法，这些工具通过分析序列特征、长度和保守性来识别 CRISPR 阵列或 Cas 等功能蛋白；另一类是基于机器学习的方法(表 2)，通过利用大量已知的 CRISPR 阵列和 Cas 等功能蛋白序列/结构数据训练模型，以提高识别的准确性和效率。

表 2 用于 CRISPR-Cas 系统注释的机器学习工具

Table 2 Machine learning tools for CRISPR-Cas system annotation

Tool name	Accession link	Year	Annotatable units	Reference
CRISPRidentify	Web tool: None Source code: https://github.com/BackofenLab/CRISPRidentify	2021	CRISPR array	[71]
CRISPRclassify	Web tool: https://shiny.posit.co/ Source code: None	2021	CRISPR array	[72]
CRISPRcasIdentifier	Web tool: None Source code: https://github.com/BackofenLab/CRISPRcasIdentifier	2020	Cas gene	[73]
CASPredict	Web tool: http://i.uestc.edu.cn/caspredict/cgi-bin/CASPredict.pl Source code: https://github.com/shanshan1996/caspredict	2021	Cas gene	[74]
CRISPRcasStack	Web tool: https://bioinfor.nefu.edu.cn/CRISPRCasStack/ Source code: https://github.com/yrjia1015/CRISPRCasStack	2022	Cas gene	[75]
CRISPR-Cas-Docker	Web tool [a]: https://www.crisprcasdocker.org Source code: https://github.com/hshimlab/CRISPR-Cas-Docker	2023	Cas gene	[76]
CRISPRCasTyper	Web tool: https://cctyper.crispr.dk/ Source code: https://github.com/Russel88/CRISPRCasTyper/tree/	2020	CRISPR array (based on XGBoost) and Cas gene (Non-ML)	[77]
CRISPRloci	Web tool: https://rna.informatik.uni-freiburg.de/CRISPRloci/ Source code: https://github.com/BackofenLab/CRISPRloci	2021	CRISPR array (based on CRISPRidentify) and Cas gene (based on CRISPRcasIdentifier)	[78]
CRISPRtracrRNA	Web tool: None Source code: https://github.com/BackofenLab/CRISPRtracrRNA	2022	CRISPR array (based on CRISPRidentify) and Cas gene (based on CRISPRcasIdentifier)	[79]
CRISPRimmunity	Web tool: http://www.microbiome-bigdata.com/CRISPRimmunity/index/ Source code: https://github.com/HIT-ImmunologyLab/CRISPRimmunity	2023	CRISPR array (based on PILER-CR, CRT, CRISPRCasFinder, CRISPRidentify) and Cas gene (HMMscan)	[80]

[a]: The web tool for CRISPR-Cas-Docker was inaccessible during testing from October to December 2024.

基于序列特征检测的方法依赖于明确的序列生物学特征，能够快速识别 CRISPR 阵列的重复序列和间隔序列。然而，当主要影响因素未被充分考量时，这类方法的表现可能不尽如人意。相比之下，机器学习特别是深度学习方法能够从大量实验数据中捕捉更复杂的特征和模式，不仅提升了 CRISPR 阵列的识别准确性，还在 Cas 蛋白的鉴定中展现了更强的适应性和预测能力。这预示着未来 CRISPR-Cas 系统预测将更加依赖于数据驱动的方法。

2.1 基于机器学习的 CRISPR 阵列的注释

CRISPR 阵列转录出包含重复-间隔单元的 pre-crRNA，随后转录物被加工为成熟的 CRISPR RNA (crRNA)或 sgRNA。CRISPR 阵列的鉴定主要依赖于针对重复-间隔单元的识别。传统的方法主要是利用基于局部比对算法的搜索工具[81] (basic local alignment search tool, BLAST)、隐马尔可夫模型[82] (hidden Markov model, HMM)等工具，参考已知的 CRISPR 阵列数据，通过一些预定义参数基于序列相似性进行搜索。相关的工具包括 CRISPRFinder[83]、CRISPRCasFinder[84]、CRISPRDetect[85]等，这类工具通常较为简便，但其主要缺陷在于无法从给定的基因组数据中识别新的阵列。此外，由于物种间间隔序列高度变异，简单的搜索策略可能导致大量的假阴性结果。一些研究尝试通过搜索引入的序列模式来识别阵列的重复区域。这种策略通常使用 *k*-mer 和基于图的方法，克服了基于参考阵列的预测方法的不足。这类方法需要首先找到最大重复对，然后合并这些重复形成共识直接重复，再根据重复和间隔的长度及其他特征(如重复之间的相似性)，预定义的评分函数进行排名。相关的工具包括 CRISPR Recognition Tool[86]和 PILER-CR[87]等。近年来，测序技术的进步加速了宏基因组数据的积累，生成了大量的公共宏基因组数据。研究人员针对宏基因组中 CRISPR 阵列的分析，开发了专门的工具，例如 CRASS[88]和 MetaCRISPR[89]。

然而，由于 CRISPR 阵列中可变间隔区域的异质性，人工推测的评分函数难以进行精准预测。因此，开发具有预测能力的数据驱动模型对于解决此类复杂问题至关重要。弗赖堡大学的 Rolf Backofen 团队[71]开发了基于机器学习的 CRISPR 阵列鉴定的工具 CRISPRidentify。该工具包括 3 个主要步骤：(1) 检测重复序列单元并构建候选阵列。(2) 提取多个与阵列相关的特征。(3) 基于提取的特征进行机器学习分类。CRISPRidentify 为每个候选阵列生成一个特征向量，一旦特征向量与重复区域关联，经过训练的分类器将决定阵列候选是否可能是 CRISPR 阵列，并计算该可能性的评分。在同一测试集上，CRISPRidentify 在召回率(真阳性率，TPR)、特异性(真阴性率，TNR)、准确率、平衡准确率和 Matthews 相关系数这个方面均优于 CRT、CRISPRCasFinder 和 CRISPRDetect，尤其在特异性方面具备显著优势，充分展现了基于机器学习进行 CRISPR 阵列鉴定的应用潜力。鉴于 CRISPRidentify 的优越性能，后续开发的 CRISPR-Cas 系统注释与鉴定功能的 CRISPRtracrRNA[79]、CRISPRloci[78]和 CRISPRimmunity[80]均使用 CRISPRidentify 进行 CRISPR 阵列的识别。除了 CRISPRidentify 及其衍生工具，Nethery 等[72]提出了一种基于梯度提升决策树模型 XGBoost 的工具 CRISPRclassify，可用于对 CRISPR 位点进行检测和分类。该方法通过分析 CRISPR 重复序列，能够在不依赖 Cas 基因的情况下识别 CRISPR 位点，发现基于 Cas 基因的分类方法无法处理的位点。此外，还有一些工具虽然使用传统方法进行 CRISPR

阵列的搜索，但在后续的阵列分类过程中引入了机器学习技术。例如，CRISPRCasTyper[77]采用了梯度提升决策树(XGBoost)模型来预测CRISPR 阵列的亚型，这有效提升了对孤儿CRISPR 阵列和远缘阵列的分类能力。

2.2 基于机器学习的Cas等功能蛋白的注释与挖掘

针对 Cas 家族相关蛋白的注释，除了经典的 BLAST 方法，利用隐马尔可夫模型(HMM)捕捉 Cas 家族蛋白序列中的保守区域，可以提高识别的准确性。常用的工具包括 MacSyFinder[90]和 HMMCAS[91]。然而，这种方法依赖于 Cas 蛋白数据库，因此在数据库不完整或存在偏差时可能会影响识别结果。此外，对于序列高度多样化的 Cas 蛋白(如 Cas7 和 Cas8)，基于 HMM 的方法可能识别效果不佳[73-74]。

弗赖堡大学的 Rolf Backofen 团队[73]开发了一个 Cas 亚型分类工具 CRISPRcasIdentifier，使用 HMM 提取的 CRISPR 盒数据特征(编码 Cas 蛋白的基因)进行训练，并评估了 3 种机器学习模型支持向量机(support vector machines, SVM)、分类与回归树(classification and regression tree, CART)和极端随机树(extra trees classifier, ERT)的训练效果。其在公共 CRISPR 基准数据集[5]上的测试获得了 0.91 的 F 分数和 0.89 的平衡准确性，远高于其他 5 种工具[73]。Yang 等[74]开发了一个基于 SVM 模型的 Cas 蛋白预测工具 CASPredict，使用从蛋白质序列中提取的 400 种二肽组成特征，在 Cas 蛋白与非 Cas 蛋白分类上的准确率达到 84.84%。Zhang 等[75]基于集成学习框架开发了 Cas 蛋白识别工具 CRISPRcasStack，其在独立数据集上的准确率、Matthews 相关系数、敏感性等相较于 CASPredict 具有一定优势。尽管这些模型取得了令人鼓舞的结果，但它们通常需要大量的特征工程和领域特定知识。

深度学习的出现彻底改变了蛋白质数据建模和分析领域，特别是 AlphaFold[17,92-93]系列深度学习模型的出现使得蛋白质三维结构的精准预测成为可能。Park 等[76]开发了一种在线工具 CRISPR-Cas-Docker，提供 2 种预测特定 crRNA 序列最优 Cas 蛋白的方法：基于结构的方法(*in silico* docking)和基于序列的方法(机器学习分类)。基于结构的方法利用 AlphaFold2 预测候选 Cas 蛋白的结构，并在此基础上模拟其与特定 crRNA 的对接情况。基于序列的方法：采用 K-最近邻(k-nearest neighbor, KNN)算法进行模型训练，通过分析数据库中的 CRISPR 阵列与 Cas 系统类型之间的关系，实现根据 CRISPR 阵列特征预测其对应的 Cas 系统类型。性能评估结果显示，该工具整体预测准确率达到了 92.3%。对于主要类别(数据点较多的情况)，F1 分数(一种综合衡量精确率和召回率的指标)超过 0.89[76]。2023 年，CRISPR 基因编辑技术奠基人之一张锋团队[94]开发了一种新的搜索算法——基于快速局部敏感哈希的聚类算法(FLSHclust)。使用该算法对 3 个主要的公共数据库进行挖掘，从中识别出了 188 种新型 CRISPR 系统；随后他们结合 AlphaFold2 的结构预测结果，对其中 4 个系统进行了详细表征与解析，极大地丰富了 CRISPR-Cas 系统的多样性[94]。此外，由于 I 类 CRISPR-Cas 系统(如 I 型、III 型)依赖多亚基效应复合体(如 Cascade 复合体)完成 DNA 靶向与切割，其注释与设计面临更高的复杂度。张锋团队[94]结合 FLSHeLust 算法与 AlphaFold2 结构预测，从宏基因组数据中挖掘出新型 I-F 型系统 Cas8-HNH 系统(Cas8-HNH、Cas7、Cas5 和 Cas3 复合体)以及 I-E 型的 Cas5-HNH 系统(Cas5-HNH、Cas7、Cas5、Cas3 和 Cas1)。此类结构驱动的分析方法为多亚基系统的注释与挖掘提供了新思路。

2024 年，诺贝尔奖得主、CRISPR 基因编辑技术奠基人之一 Jennifer Doudna 教授团队[95]整合了 AlphaFold2 与传统结构比对程序，开发出了一种自动化结构检索方法，发现了 Cas13 的祖先——Cas13an，并进一步解析了 Cas13an 的结构及其作用机制，将 Cas13 的起源追溯到与防御相关的核糖核酸酶(ribonuclease)。

基于深度学习的蛋白结构预测也推动了其他 CRISPR 衍生技术，如碱基编辑技术的发展。中国科学院遗传与发育生物学研究所的高彩霞团队[96]使用 AlphaFold2 对代表性的脱氨功能序列进行了三维结构预测，基于预测的三维结构进行了蛋白质多重比对与聚类，成功将潜在的脱氨酶划分为 20 类；除已报道的 APOBEC/AID 胞嘧啶脱氨酶外，他们发现了 5 类具有全新序列和结构的蛋白具有胞嘧啶脱氨酶活性，并以此为基础成功开发出了一系列新型碱基编辑工具。基于类似的策略，中国农业科学院深圳农业基因组研究所的左二伟团队[97]扩大了进行三维结构预测的候选脱氨酶序列范围，从 1 483 个胞嘧啶脱氨酶的聚类中选择了 272 个具有代表性的脱氨酶进行碱基编辑活性检测，成功开发出多种高效、无序列偏好性的新型胞嘧啶碱基编辑工具。

2.3 基于蛋白语言模型的 Cas 等功能蛋白的从头生成

随着深度学习技术，特别是蛋白质语言模型(protein language model, pLM)的发展，新型蛋白质元件的创制正逐渐从对微生物的基因组信息的探索和挖掘转向从头生成[98]。pLM 是一种先进的深度学习模型，它借鉴了自然语言处理技术来解析蛋白质序列。pLM 将氨基酸序列视作一种特殊的“语言”，并通过分析序列中的模式来学习其内在的“语法”和“语义”。其核心思想在于蛋白质序列中的氨基酸残基间复杂的相互作用与自然语言中单词和句子的结构关系相似。通过深入学习这些关系，pLM 能够预测蛋白质的功能、结构和动态特性。蛋白质结构预测结果可以被纳入 pLM 的训练数据中，这种结合结构信息的训练策略能够极大地增强 pLM 对蛋白质序列与结构之间复杂关系的捕捉能力，提高其在预测蛋白质特性时的准确性。通过这种方式，pLM 不仅能够理解蛋白质序列的线性信息，还能够洞察其三维结构的深层含义，为蛋白质科学提供了一个强大的分析和预测工具。

在近期发表的一篇 *Science* 论文中，斯坦福大学和加州大学伯克利分校的研究团队以 80 000 个细菌和古菌基因组(共计 3 000 亿个 DNA 碱基对)为基础训练开发了蛋白语言模型 Evo[99]。针对 CRISPR-Cas 系统，研究团队利用 82 430 个 CRISPR-Cas 序列信息微调了 Evo 模型，并在模型中增加了特定的提示标识如 Cas9、Cas12 和 Cas13 以便后续可以根据这些标识设计特定蛋白家族的 CRISPR-Cas 系统。虽然 Evo 设计的 CRISPR-Cas9 系统尚未通过实验验证，但其预测的结构与天然蛋白质的结构非常相似，这为进一步探索提供了一个有希望的起点。近期，Profluent 公司在 BioRxiv 公布的研究[100]中，通过挖掘 26.2 Tb 的微生物基因组和宏基因组数据构建了包含超百万 CRISPR-Cas 操作单元的 CRISPR-Cas Atlas；并基于此数据库训练了大语言模型 ProGen2，生成了 200 多个新的 CRISPR-Cas9 蛋白序列。其中，最具前景的 OpenCRISPR-1 的 Cas9 类蛋白序列与 SpCas9 相比有 400 个突变，与任何已知天然 Cas 蛋白相比有近 200 个突变，在基因组靶点编辑和脱靶检测中，表现出相当高的编辑效率和更高的特异性[100]。此外，他们还生成了一系列新型脱氨酶，其中 PF-DEAM-1 和 PF-DEAM-2 在实验

中显示出与 ABE8.20 相当的碱基编辑效率；通过引入突变，OpenCRISPR-1 可改造成类似 Cas9 nickase 的切口酶，并与 PF-DEAM-1 或 PF-DEAM-2 结合，实现碱基编辑功能[100]。

3 基因编辑关键酶的改造

对 CRISPR-Cas 系统核心酶元件的改造在提高编辑效率、减少脱靶效应以及扩展基因编辑应用范围方面至关重要。近年来，基因编辑关键蛋白的改造主要依赖于定向进化或(半)理性设计的定点改造方法[1,101]。定向进化的基本思路是通过构建大规模随机突变文库并使用高通量实验筛选有益变体，无需依赖对酶催化机制、结构和特定突变影响的深入理论认知。然而，高通量实验的设计和实施对于许多蛋白质来说仍是挑战。因此，(半)理性设计方法凭借对蛋白质结构、功能和折叠机制的深入理解，通过设计小而精的突变文库或定点突变来减少实验筛选任务量，成为了另一种主流蛋白改造策略。然而，以上策略通常需要多轮迭代改造，周期较长，并且依赖于丰富的先验知识以及对蛋白质及其与其他分子的相互作用的深入理解。

为了进一步提高效率和精度，机器学习被引入到基因编辑蛋白改造的工作流程中[22,102]。通过对已有数据的训练来预测突变的效果可以减少实验筛选工作量。一般而言将从具有广泛突变空间代表性的多样化突变变体开始，通过多轮测试-学习-设计循环过程，利用机器学习分析序列-功能景观，提升高性能突变体的识别效率从而缩短实验周期。尽管这种策略已在多种蛋白的改造中显示出优良的性能，但在 Cas 相关改造中尚未被广泛采用。香港大学黄兆麟团队[103-104]率先尝试将这种策略应用于 Cas 蛋白的改造。2022 年，他们结合前期开发的基于条形码的无缝组合 DNA 组装技术 CombiSEAL[105] 和 Arnold 团队开发的机器学习辅助定向进化方法 MLDE[106-107]，针对金黄色葡萄球菌 Cas9 核酸酶(SaCas9)，生成了 10 个用于 Cas9 工程的多域组合突变文库的计算和实验数据集并进行了交叉验证，证明了结合机器学习的工程方法可以将实验筛选工作量减少 95%[103]；与单纯依靠定向进化技术相比，高性能突变体的富集率提高了约 7.5 倍。2024 年，为了进一步提升高性能突变体的富集能力，该团队将零样本预测(zero-shot prediction)和多轮低数量样本选择(low-N sampling)结合，通过对少量预测的最优变体进行实验来引导主动学习(active learning)的迭代，开发了 TopVIP 预测算法[104]；利用改进的机器学习策略，通过 4 轮的挑选和验证(每轮 12 个变体)，在组合文库中筛选出最优 1%突变体的准确率高达 92.6%。

AlphaFold2[17]显著加快了 CRISPR 系统关键酶的改造进程。南方科技大学朱健康院士团队[108]利用 AlphaFold2 预测了 Cas12i3 的三维结构，并将其与同家族的 Cas12i1/i2 蛋白结构进行比对，确定了 Cas12i3 与核酸互作的关键氨基酸位点；基于这些预测位点，团队构建了一个包含 150 个点突变的文库，并通过荧光报告系统筛选出 26 个活性提高 1.5 倍以上的 Cas12i3 变体，并通过后续的突变组合构建出高活性与特异性的变体 Cas-SF01。美国佛罗里达大学的 Nguyen 等[109]同时利用传统的蛋白同源建模工具 SWISS-MODEL[110]和基于深度学习的 AlphaFold2 预测了来源于短芽孢杆菌(*Brevibacillus* sp.) SYP-B805 的 Cas12b (BrCas12b)的结构。为了提高蛋白的热稳定性，其利用深度学习稳定性预测工具 DeepDDG[111]和半理性设计工具 HotSpotWizard 3.0[112]，识别并设计了 35 个潜在的突变位点；最终发现 16 个突变体的熔解温度(T_m)有所提升，其中 5 个突变体的 T_m 提高了超

过 2 ℃。对比 AlphaFold 和 SWISS-MODEL 预测准确性可以发现，基于 SWISS-MODEL 的预测准确率为 52% (13/25)，而 AlphaFold 的预测准确率相对更高，可以达到 60% (15/25)[109]。

基于这一策略，军事医学研究院的王升启、舒文杰团队联合南京医科大学的张军团队以及之江实验室的朱世强团队[113]，开发了一个参数规模约 6.95 亿的多模态蛋白质深度表征学习模型；该模型以 AlphaFold 蛋白质结构数据库中的约 1.6 亿个蛋白质的序列和结构信息为训练数据，系统地学习了蛋白质序列的特征分布和空间折叠规律。在多模态深度表征学习模型的基础上，研究团队进一步开发了一种用于零样本 (zero-shot) 预测蛋白质突变效应的方法 ProMEP。在对 tRNA 特异性腺苷脱氨酶 TadA 进行改造的过程中，在利用 ProMEP 预测 top 10 的有利突变和不利突变时，显示出有利突变的预测准确率在 50%–70%之间，而不利突变的预测准确率达到了 100%[113]。此外，基于 ProMEP 设计的 TadA 点突变体构建的 ABE 碱基编辑器在 A5/A6 位置上的 A-to-G 编辑效率和旁编辑效应与经典的 ABE9 相当，且在脱靶率方面优于 ABE8e[113]。这些成果展示了 ProMEP 在蛋白质工程中的潜力，特别是在提高基因编辑工具性能方面的应用前景。

除了活性、特异性和稳定性优化，Cas 蛋白的小型化也是当前基因编辑领域的研究热点[114]。小型化不仅可解决腺相关病毒(adeno-associated virus, AAV)载体容量的限制，还可减少宿主细胞代谢负担，降低免疫原性，提升递送效率。更小的 Cas 蛋白为整合多功能元件提供了空间，增强了基因编辑的灵活性，同时可能提高靶向特异性，减少脱靶效应。吉林大学李占军团队[115]通过结合 AlphaFold2 预测结构信息，提出了一种基于核酸-蛋白相互作用、动态构象重组和同源保守性的蛋白质小型化策略，成功设计了多种紧凑型 Cas13 变体，平均精简了约 30%的序列，且其 RNA 结合和切割活性与野生型酶相当。吉林大学袁泓明团队[116]提出了一种基于网络服务的快速通用发现(web-based fast generic discovery, WFG) 策略，结合 AlphaFold2 预测结构信息，发现并小型化多条新型单链 DNA 脱氨酶(Sdds)，平均精简了约 40%的序列。通过这一策略，团队成功设计并构建了多套更紧凑的 Sdd-CBE。

4 展望

近年来，基因编辑技术正迅速发展，AI 在 sgRNA 的设计与评估、新型 CRISPR-Cas 的注释与挖掘、编辑关键蛋白的设计与改造等方面的应用已成为推动这一领域前进的关键力量，但仍然面临着诸多挑战。

在 sgRNA 的设计与评估方面，尽管 AI 技术已显著提升 sgRNA 的设计效率和准确性，但数据的稀疏性和异质性仍是限制模型预测和泛化能力的主要挑战。未来的研究应集中于构建更为多样化和大规模的训练数据集，并结合迁移学习、集成学习等技术开发更加先进的 AI 模型，以增强模型的泛化能力。此外，sgRNA 的设计不仅仅需要考虑编辑效率，还需有效预测脱靶效应，这对模型的精准性提出了更高的要求。在 CRISPR 系统的应用中，sgRNA 设计的目标可能需要扩展到多重 guide RNA 阵列的优化。研究表明，合理调整 guide RNA 阵列中的重复序列长度和间隔序列的排列可以提高 guide RNA 的稳定性，并减少脱靶效应[117]。例如，优化重复序列的长度(如控制在 20–36 nt 之间)能够维持高效编辑的同时，减少因过长的重复序列导致的转录和表达不稳定问题。此外，sgRNA 的表达水平以及 Cas 蛋白的表达量均会

对实际的编辑效率造成较大影响。但随着相关数据的不断积累并结合 AI 技术，未来有望实现多目标优化策略，进一步提升 guide RNA 设计的整体效率和安全性。针对特定细胞系或物种的定制化 sgRNA 设计也将成为未来的研究热点。

CRISPR-Cas 系统的多样性和复杂性为发现新型关键元件(例如 Cas 蛋白和修饰酶)提供了巨大的潜力。然而，底层专利的封锁仍然是我国在基因编辑领域面临的主要障碍。随着 AI 技术的快速发展，特别是 AlphaFold 等深度学习工具的兴起，蛋白质结构预测和功能注释变得更加高效。利用 AI 驱动的新型蛋白聚类方法，研究人员能够快速发现具有独特功能的酶元件，从而规避现有专利的限制，实现基因编辑领域的“弯道超车”，为我国生物技术领域的发展奠定坚实基础。需要特别指出的是，随着蛋白语言模型的发展，基于语言模型从头设计和生成全新基因编辑关键蛋白元件的新时代即将开启。

在基因编辑关键蛋白的改造方面，Cas 酶等基因编辑关键蛋白的定向进化和功能改造是提升编辑效率、减少脱靶效应的重要途径。开发新的 AI 框架来指导 Cas 酶等基因编辑关键蛋白定向进化实验，减少实验迭代次数，提高改造效率，将是未来研究的重要方向。同时，随着蛋白质语言模型的发展，关键蛋白的改造将更加精准、智能。

另外，借助自动化设施，基因编辑的全流程将实现高通量并行操作，从 sgRNA 设计、实验构建到结果分析的各个环节均可自动化完成。这样不仅显著缩短了实验周期，还能持续输出标准化、大规模、高质量的数据，有望解决编辑数据稀疏性与异质性的挑战。在关键蛋白的挖掘与优化方面，自动化的蛋白质表达、纯化和功能测试将大大加速重要酶元件的改造过程，有望为机器学习模型的训练提供源源不断的高质量数据支持。未来，通过自动化设施与 AI 技术的深度融合，不仅能够提升基因编辑的效率，还有望大幅缩短技术迭代的时间，使基因编辑尽早进入智能化、精准化发展的新时代。

作者贡献声明

毛雨丰：初稿写作、稿件润色修改；储光芸、梁庆玲、刘叶、杨毅、廖小平：稿件润色修改；王猛：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] PACESA M, PELEA O, JINEK M. Past, present, and future of CRISPR genome editing technologies[J]. Cell, 2024, 187(5): 1076-1100.
[2] RATH D, AMLINGER L, RATH A, LUNDGREN M. The CRISPR-Cas immune system: Biology, mechanisms and applications[J]. Biochimie, 2015, 117: 119-128.
[3] DELTCHEVA E, CHYLINSKI K, SHARMA CM, GONZALES K, CHAO YJ, PIRZADA ZA, ECKERT MR, VOGEL J, CHARPENTIER E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III[J]. Nature, 2011, 471(7340): 602-607.
[4] STERNBERG SH, REDDING S, JINEK M, GREENE EC, DOUDNA JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9[J]. Nature, 2014, 507(7490): 62-67.
[5] MAKAROVA KS, WOLF YI, IRANZO J, SHMAKOV SA, ALKHNBASHI OS, BROUNS SJJ, CHARPENTIER E, CHENG D, HAFT DH, HORVATH P, MOINEAU S, MOJICA FJM, SCOTT D, SHAH SA, SIKSNYS V, TERNS MP, VENCLOVAS Č, WHITE MF, YAKUNIN AF, YAN W, et al. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants[J]. Nature Reviews Microbiology, 2020, 18(2): 67-83.
[6] JINEK M, CHYLINSKI K, FONFARA I, HAUER M, DOUDNA JA, CHARPENTIER E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity[J]. Science, 2012, 337(6096): 816-821.
[7] ANN RAN F, CONG L, YAN WX, SCOTT DA, GOOTENBERG JS, KRIZ AJ, ZETSCHE B, SHALEM

O, WU XB, MAKAROVA KS, KOONIN EV, SHARP PA, ZHANG F. *In vivo* genome editing using *Staphylococcus aureus* Cas9[J]. Nature, 2015, 520(7546): 186-191.
[8] ZETSCHE B, GOOTENBERG JS, ABUDAYYEH OO, SLAYMAKER IM, MAKAROVA KS, ESSLETZBICHLER P, VOLZ SE, JOUNG J, van der OOST J, REGEV A, KOONIN EV, ZHANG F. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system[J]. Cell, 2015, 163(3): 759-771.
[9] SHMAKOV S, ABUDAYYEH OO, MAKAROVA KS, WOLF YI, GOOTENBERG JS, SEMENOVA E, MINAKHIN L, JOUNG J, KONERMANN S, SEVERINOV K, ZHANG F, KOONIN EV. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems[J]. Molecular Cell, 2015, 60(3): 385-397.
[10] GILBERT LA, LARSON MH, MORSUT L, LIU ZR, BRAR GA, TORRES SE, STERN-GINOSSAR N, BRANDMAN O, WHITEHEAD EH, DOUDNA JA, LIM WA, WEISSMAN JS, QI LS. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes[J]. Cell, 2013, 154(2): 442-451.
[11] MAEDER ML, LINDER SJ, CASCIO VM, FU YF, HO QH, KEITH JOUNG J. CRISPR RNA-guided activation of endogenous human genes[J]. Nature Methods, 2013, 10(10): 977-979.
[12] KOMOR AC, KIM YB, PACKER MS, ZURIS JA, LIU DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage[J]. Nature, 2016, 533(7603): 420-424.
[13] NISHIDA K, ARAZOE T, YACHIE N, BANNO S, KAKIMOTO M, TABATA M, MOCHIZUKI M, MIYABE A, ARAKI M, HARA KY, SHIMATANI Z, KONDO A. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems[J]. Science, 2016, 353(6305): aaf8729.
[14] ANZALONE AV, RANDOLPH PB, DAVIS JR, SOUSA AA, KOBLAN LW, LEVY JM, CHEN PJ, WILSON C, NEWBY GA, RAGURAM A, LIU DR. Search-and-replace genome editing without double-strand breaks or donor DNA[J]. Nature, 2019, 576(7785): 149-157.
[15] NUÑEZ JK, CHEN J, POMMIER GC, ZACHERY COGAN J, REPLOGLE JM, ADRIAENS C, RAMADOSS GN, SHI QM, HUNG KL, SAMELSON AJ, POGSON AN, KIM JYS, CHUNG A, LEONETTI MD, CHANG HY, KAMPMANN M, BERNSTEIN BE, HOVESTADT V, GILBERT LA, WEISSMAN JS. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing[J]. Cell, 2021, 184(9): 2503-2519.e17.
[16] KONSTANTAKOS V, NENTIDIS A, KRITHARA A, PALIOURAS G. CRISPR-Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning[J]. Nucleic Acids Research, 2022, 50(7): 3616-3637.
[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(7873): 583-589.
[18] SHERKATGHANAD Z, ABDAR M, CHARLIER J, MAKARENKOV V. Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review[J]. Briefings in Bioinformatics, 2023, 24(3): bbad131.
[19] 王远立, 啜国晖, 闫继芳, 石雷, 刘琦. 计算机辅助 CRISPR 向导 RNA 设计[J]. 生物工程学报, 2017, 33(10): 1744-1756.
WANG YL, CHUAI GH, YAN JF, SHI L, LIU Q. *In silico* CRISPR-based sgRNA design[J]. Chinese Journal of Biotechnology, 2017, 33(10): 1744-1756 (in Chinese).
[20] HUANG PS, BOYKEN SE, BAKER D. The coming of age of *de novo* protein design[J]. Nature, 2016, 537(7620): 320-327.
[21] NOTIN P, ROLLINS N, GAL Y, SANDER C, MARKS D. Machine learning for functional protein design[J]. Nature Biotechnology, 2024, 42(2): 216-228.
[22] 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).
[23] GRAHAM DB, ROOT DE. Resources for the design of CRISPR gene editing experiments[J]. Genome Biology, 2015, 16: 260.
[24] CHUAI GH, WANG QL, LIU Q. *In silico* meets *in vivo*: towards computational CRISPR-based sgRNA design[J]. Trends in Biotechnology, 2017, 35(1): 12-21.
[25] DOUDNA JA, CHARPENTIER E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9[J]. Science, 2014, 346(6213): 1258096.
[26] XU H, XIAO TF, CHEN CH, LI W, MEYER CA, WU Q, WU D, CONG L, ZHANG F, LIU JS, BROWN M, SHIRLEY LIU X. Sequence determinants of improved CRISPR sgRNA design[J]. Genome Research, 2015, 25(8): 1147-1157.
[27] WANG T, WEI JJ, SABATINI DM, LANDER ES. Genetic screens in human cells using the CRISPR-Cas9 system[J]. Science, 2014, 343(6166): 80-84.
[28] WU XB, SCOTT DA, KRIZ AJ, CHIU AC, HSU PD, DADON DB, CHENG AW, TREVINO AE, KONERMANN S, CHEN SD, JAENISCH R, ZHANG F, SHARP PA. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells[J]. Nature Biotechnology, 2014, 32(7): 670-676.
[29] DOENCH JG, HARTENIAN E, GRAHAM DB, TOTHOVA Z, HEGDE M, SMITH I, SULLENDER M, EBERT BL, XAVIER RJ, ROOT DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation[J]. Nature Biotechnology, 2014, 32(12): 1262-1267.
[30] DOENCH JG, FUSI N, SULLENDER M, HEGDE M, VAIMBERG EW, DONOVAN KF, SMITH I, TOTHOVA Z, WILEN C, ORCHARD R, VIRGIN HW, LISTGARTEN J, ROOT DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9[J]. Nature Biotechnology, 2016, 34(2): 184-191.

[31] CHARI R, MALI P, MOOSBURNER M, CHURCH GM. Unraveling CRISPR-Cas9 genome engineering parameters *via* a library-on-library approach[J]. Nature Methods, 2015, 12(9): 823-826.
[32] HSU PD, SCOTT DA, WEINSTEIN JA, ANN RAN F, KONERMANN S, AGARWALA V, LI YQ, FINE EJ, WU XB, SHALEM O, CRADICK TJ, MARRAFFINI LA, BAO G, ZHANG F. DNA targeting specificity of RNA-guided Cas9 nucleases[J]. Nature Biotechnology, 2013, 31(9): 827-832.
[33] KIM D, KIM S, KIM S, PARK J, KIM JS. Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq[J]. Genome Research, 2016, 26(3): 406-415.
[34] SINGH R, KUSCU C, QUINLAN A, QI YJ, ADLI M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction[J]. Nucleic Acids Research, 2015, 43(18): e118.
[35] KIM D, KIM JS. DIG-seq: a genome-wide CRISPR off-target profiling method using chromatin DNA[J]. Genome Research, 2018, 28(12): 1894-1900.
[36] MORGENS DW, WAINBERG M, BOYLE EA, URSU O, ARAYA CL, KIMBERLY TSUI C, HANEY MS, HESS GT, HAN K, JENG EE, LI A, SNYDER MP, GREENLEAF WJ, KUNDAJE A, BASSIK MC. Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens[J]. Nature Communications, 2017, 8: 15178.
[37] AGUIRRE AJ, MEYERS RM, WEIR BA, VAZQUEZ F, ZHANG CZ, BEN-DAVID U, COOK A, HA G, HARRINGTON WF, DOSHI MB, KOST-ALIMOVA M, GILL S, XU H, ALI LD, JIANG GZ, PANTEL S, LEE Y, GOODALE A, CHERNIACK AD, OH C, et al, TSHERNIAK A, HAHN WC. Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting[J]. Cancer Discovery, 2016, 6(8): 914-929.
[38] 杨毅, 毛雨丰, 杨春贺, 王猛. 面向微生物遗传操作的编辑序列设计工具的研究进展[J]. 合成生物学, 2023, 4(1): 30-46.
YANG Y, MAO YF, YANG CH, WANG M. Recent progress in computational tools for designing editing sequences used in microbial genetic manipulations[J]. Synthetic Biology Journal, 2023, 4(1): 30-46 (in Chinese).
[39] KIM HK, KIM Y, LEE S, MIN S, BAE JY, CHOI JW, PARK J, JUNG D, YOON S, KIM HH. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance[J]. Science Advances, 2019, 5(11): eaax9249.
[40] WANG DQ, ZHANG CD, WANG B, LI B, WANG Q, LIU D, WANG HY, ZHOU Y, SHI LM, LAN F, WANG YM. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning[J]. Nature Communications, 2019, 10(1): 4284.
[41] ZHANG GS, DAI ZM, DAI XH. C-RNNCrispr: prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks[J]. Computational and Structural Biotechnology Journal, 2020, 18: 344-354.
[42] ARBAB M, SHEN MW, MOK B, WILSON C, MATUSZEK Ż, CASSA CA, LIU DR. Determinants of base editing outcomes from target library analysis and machine learning[J]. Cell, 2020, 182(2): 463-480.e30.
[43] KIM HK, YU G, PARK J, MIN S, LEE S, YOON S, KIM HH. Predicting the efficiency of prime editing guide RNAs in human cells[J]. Nature Biotechnology, 2021, 39(2): 198-206.
[44] LI SW, AN JJ, LI YQ, ZHU XG, ZHAO DD, WANG LX, SUN YH, YANG YZ, BI CH, ZHANG XL, WANG M. Automated high-throughput genome editing platform with an AI learning *in situ* prediction model[J]. Nature Communications, 2022, 13(1): 7386.
[45] LISTGARTEN J, WEINSTEIN M, KLEINSTIVER BP, SOUSA AA, KEITH JOUNG J, CRAWFORD J, GAO K, HOANG L, ELIBOL M, DOENCH JG, FUSI N. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs[J]. Nature Biomedical Engineering, 2018, 2(1): 38-47.
[46] LIN JC, WONG KC. Off-target predictions in CRISPR-Cas9 gene editing using deep learning[J]. Bioinformatics, 2018, 34(17): i656-i663.
[47] LIN JC, ZHANG ZL, ZHANG SX, CHEN JY, WONG KC. CRISPR-net: a recurrent convolutional network quantifies CRISPR off-target activities with mismatches and indels[J]. Advanced Science, 2020, 7(13): 1903562.
[48] FENG HB, GUO JH, WANG TM, ZHANG C, XING XH. Guide-target mismatch effects on dCas9-sgRNA binding activity in living bacterial cells[J]. Nucleic Acids Research, 2021, 49(3): 1263-1277.
[49] ABADI S, YAN WX, AMAR D, MAYROSE I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action[J]. PLoS Computational Biology, 2017, 13(10): e1005807.
[50] CHUAI GH, MA HH, YAN JF, CHEN M, HONG NF, XUE DY, ZHOU C, ZHU CY, CHEN K, DUAN B, GU F, QU S, HUANG DS, WEI J, LIU Q. DeepCRISPR: optimized CRISPR guide RNA design by deep learning[J]. Genome Biology, 2018, 19(1): 80.
[51] ANTHON C, CORSI GI, GORODKIN J. CRISPRon/off: CRISPR/Cas9 on- and off-target gRNA design[J]. Bioinformatics, 2022, 38(24): 5437-5439.
[52] FONG JHC, WONG ASL. Advancing CRISPR/Cas gene editing with machine learning[J]. Current Opinion in Biomedical Engineering, 2023, 28: 100477.
[53] HAM DT, BROWNE TS, BANGLOREWALA PN, WILSON TL, MICHAEL RK, GLOOR GB, EDGELL DR. A generalizable Cas9/sgRNA prediction model using machine transfer learning with small high-quality datasets[J]. Nature Communications, 2023, 14(1): 5514.
[54] KIM HK, SONG M, LEE JN, VIPIN MENON A, JUNG S, KANG YM, CHOI JW, WOO E, KOH HC, NAM JW, KIM H. *In vivo* high-throughput profiling of CRISPR-Cpf1 activity[J]. Nature Methods, 2017, 14(2): 153-159.
[55] GUO JH, WANG TM, GUAN CG, LIU B, LUO C, XIE Z, ZHANG C, XING XH. Improved sgRNA design in bacteria *via* genome-wide activity profiling[J]. Nucleic Acids Research, 2018, 46(14): 7052-7069.
[56] CHEN YH, WANG XW. Evaluation of efficiency prediction algorithms and development of ensemble model for CRISPR/Cas9 gRNA selection[J]. Bioinformatics, 2022, 38(23): 5175-5181.

[57] CHO SW, KIM S, KIM Y, KWEON J, KIM HS, BAE SS, KIM JS. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases[J]. Genome Research, 2014, 24(1): 132-141.
[58] FU YF, FODEN JA, KHAYTER C, MAEDER ML, REYON D, KEITH JOUNG J, SANDER JD. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells[J]. Nature Biotechnology, 2013, 31(9): 822-826.
[59] KIM D, BAE SS, PARK J, KIM E, KIM S, YU HR, HWANG J, KIM JI, KIM JS. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells[J]. Nature Methods, 2015, 12(3): 237-243.
[60] TSAI SQ, NGUYEN NT, MALAGON-LOPEZ J, TOPKAR VV, ARYEE MJ, KEITH JOUNG J. CIRCLE-seq: a highly sensitive *in vitro* screen for genome-wide CRISPR-Cas9 nuclease off-targets[J]. Nature Methods, 2017, 14(6): 607-614.
[61] CROSETTO N, MITRA A, SILVA MJ, BIENKO M, DOJER N, WANG Q, KARACA E, CHIARLE R, SKRZYPCZAK M, GINALSKI K, PASERO P, ROWICKA M, DIKIC I. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing[J]. Nature Methods, 2013, 10(4): 361-365.
[62] TSAI SQ, ZHENG ZL, NGUYEN NT, LIEBERS M, TOPKAR VV, THAPAR V, WYVEKENS N, KHAYTER C, JOHN IAFRATE A, LE LP, ARYEE MJ, KEITH JOUNG J. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases[J]. Nature Biotechnology, 2015, 33(2): 187-197.
[63] STERNBERG SH, LaFRANCE B, KAPLAN M, DOUDNA JA. Conformational control of DNA target cleavage by CRISPR-Cas9[J]. Nature, 2015, 527(7576): 110-113.
[64] BOYLE EA, ANDREASSON JOL, CHIRCUS LM, STERNBERG SH, WU MJ, GUEGLER CK, DOUDNA JA, GREENLEAF WJ. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(21): 5461-5466.
[65] HU WX, RONG Y, GUO Y, JIANG F, TIAN W, CHEN H, DONG SS, YANG TL. ExsgRNA: reduce off-target efficiency by on-target mismatched sgRNA[J]. Briefings in Bioinformatics, 2022, 23(5): bbac183.
[66] CHENG XL, LI ZX, SHAN RC, LI ZH, WANG SN, ZHAO WC, ZHANG H, CHAO LM, PENG J, FEI T, LI W. Modeling CRISPR-Cas13d on-target and off-target effects using machine learning approaches[J]. Nature Communications, 2023, 14(1): 752.
[67] ZHANG YW, ZHAO GF, AHMED FYH, YI TF, HU SY, CAI T, LIAO Q. *In silico* method in CRISPR/cas system: an expedite and powerful booster[J]. Frontiers in Oncology, 2020, 10: 584404.
[68] ALKHNBASHI OS, MEIER T, MITROFANOV A, BACKOFEN R, VOß B. CRISPR-Cas bioinformatics[J]. Methods, 2020, 172: 3-11.
[69] SHARMA S, MURMU S, DAS R, TILGAM J, SAAKRE M, PAUL K. A review on bioinformatics advances in CRISPR-Cas technology[J]. Journal of Plant Biochemistry and Biotechnology, 2023, 32(4): 791-807.
[70] LIU ZL, LIU JY, YANG ZH, ZHU LY, ZHU ZM, HUANG H, JIANG L. Endogenous CRISPR-Cas mediated *in situ* genome editing: State-of-the-art and the road ahead for engineering prokaryotes[J]. Biotechnology Advances, 2023, 68: 108241.
[71] MITROFANOV A, ALKHNBASHI OS, SHMAKOV SA, MAKAROVA KS, KOONIN EV, BACKOFEN R. CRISPRidentify: identification of CRISPR arrays using machine learning approach[J]. Nucleic Acids Research, 2021, 49(4): e20.
[72] NETHERY MA, KORVINK M, MAKAROVA KS, WOLF YI, KOONIN EV, BARRANGOU R. CRISPRclassify: repeat-based classification of CRISPR loci[J]. The CRISPR Journal, 2021, 4(4): 558-574.
[73] PADILHA VA, ALKHNBASHI OS, SHAH SA, de CARVALHO ACPLF, BACKOFEN R. CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems[J]. GigaScience, 2020, 9(6): giaa062.
[74] YANG SS, HUANG J, HE BF. CASPredict: a web service for identifying Cas proteins[J]. PeerJ, 2021, 9: e11887.
[75] ZHANG TJ, JIA YR, LI HF, XU DL, ZHOU J, WANG GH. CRISPRCasStack: a stacking strategy-based ensemble learning framework for accurate identification of Cas proteins[J]. Briefings in Bioinformatics, 2022, 23(5): bbac335.
[76] PARK HM, WON J, PARK Y, ANZAKU ET, VANKERSCHAVER J, van MESSEM A, de NEVE W, SHIM H. CRISPR-Cas-Docker: web-based *in silico* docking and machine learning-based classification of crRNAs with Cas proteins[J]. BMC Bioinformatics, 2023, 24(1): 167.
[77] RUSSEL J, PINILLA-REDONDO R, MAYO-MUÑOZ D, SHAH SA, SØRENSEN SJ. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci[J]. The CRISPR Journal, 2020, 3(6): 462-469.
[78] ALKHNBASHI OS, MITROFANOV A, BONIDIA R, RADEN M, TRAN VD, EGGENHOFER F, SHAH SA, ÖZTÜRK E, PADILHA VA, SANCHES DS, de CARVALHO ACPLF, BACKOFEN R. CRISPRloci: comprehensive and accurate annotation of CRISPR-Cas systems[J]. Nucleic Acids Research, 2021, 49(W1): W125-W130.
[79] MITROFANOV A, ZIEMANN M, ALKHNBASHI OS, HESS WR, BACKOFEN R. CRISPRtracrRNA: robust approach for CRISPR tracrRNA detection[J]. Bioinformatics, 2022, 38(Supplement_2): ii42-ii48.
[80] ZHOU FX, YU XR, GAN R, REN K, CHEN CG, REN CY, CUI M, LIU YC, GAO YY, WANG SY, YIN MY, HUANG TJ, HUANG ZW, ZHANG F. CRISPRimmunity: an interactive web server for CRISPR-associated Important Molecular events and Modulators Used in geNome edIting Tool identifYing[J]. Nucleic Acids Research, 2023, 51(W1): W93-W107.
[81] ALTSCHUL SF, GISH W, MILLER W, MYERS EW, LIPMAN DJ. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[82] EDDY SR. Profile hidden Markov models[J]. Bioinformatics, 1998, 14(9): 755-763.
[83] GRISSA I, VERGNAUD G, POURCEL C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats[J]. Nucleic Acids Research, 2007, 35(Web server issue): W52-W57.
[84] COUVIN D, BERNHEIM A, TOFFANO-NIOCHE C, TOUCHON M, MICHALIK J, NÉRON B, ROCHA EPC, VERGNAUD G, GAUTHERET D, POURCEL C. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins[J]. Nucleic Acids Research, 2018, 46(W1): W246-W251.
[85] BISWAS A, STAALS RHJ, MORALES SE, FINERAN PC, BROWN CM. CRISPRDetect: a flexible algorithm to define CRISPR arrays[J]. BMC Genomics, 2016, 17: 356.
[86] BLAND C, RAMSEY TL, SABREE F, LOWE M, BROWN K, KYRPIDES NC, HUGENHOLTZ P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats[J]. BMC Bioinformatics, 2007, 8: 209.
[87] EDGAR RC. PILER-CR: fast and accurate identification of CRISPR repeats[J]. BMC Bioinformatics, 2007, 8: 18.
[88] SKENNERTON CT, IMELFORT M, TYSON GW. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data[J]. Nucleic Acids Research, 2013, 41(10): e105.
[89] LEI JK, SUN YN. Assemble CRISPRs from metagenomic sequencing data[J]. Bioinformatics, 2016, 32(17): i520-i528.
[90] ABBY SS, NÉRON B, MÉNAGER H, TOUCHON M, ROCHA EPC. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems[J]. PLoS One, 2014, 9(10): e110726.
[91] CHAI GS, YU M, JIANG LX, DUAN YC, HUANG J. HMMCAS: a web tool for the identification and domain annotations of Cas proteins[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 16(4): 1313-1315.
[92] 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(7792): 706-710.
[93] 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(8016): 493-500.
[94] ALTAE-TRAN H, KANNAN S, SUBERSKI AJ, MEARS KS, ESRA DEMIRCIOGLU F, MOELLER L, KOCALAR S, OSHIRO R, MAKAROVA KS, MACRAE RK, KOONIN EV, ZHANG F. Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering[J]. Science, 2023, 382(6673): eadi1910.
[95] YOON PH, ZHANG ZY, LOI KJ, ADLER BA, LAHIRI A, VOHRA K, SHI HL, 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.
[96] 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.
[97] XU K, FENG H, ZHANG HH, HE CF, KANG HF, YUAN TL, SHI L, ZHOU CK, HUA GY, CAO YQ, ZUO ZR, ZUO EW. Structure-guided discovery of highly efficient cytidine deaminases with sequence-context independence[J]. Nature Biomedical Engineering, 2025, 9: 93-108.
[98] 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.
[99] NGUYEN E, POLI M, DURRANT MG, KANG B, KATREKAR D, LI DB, BARTIE LJ, THOMAS AW, KING SH, BRIXI G, SULLIVAN J, NG MY, LEWIS A, LOU A, ERMON S, BACCUS SA, HERNANDEZ-BOUSSARD T, RÉ C, HSU PD, HIE BL. Sequence modeling and design from molecular to genome scale with Evo[J]. Science, 2024, 386(6723): eado9336.
[100] RUFFOLO JA, NAYFACH S, GALLAGHER J, BHATNAGAR A, BEAZER J, HUSSAIN R, RUSS J, YIP J, HILL E, PACESA M, MEESKE AJ, CAMERON P, MADANI A. Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. bioRxiv, 2024. https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1.
[101] KOVALEV MA, DAVLETSHIN AI, KARPOV DS. Engineering Cas9: next generation of genomic editors[J]. Applied Microbiology and Biotechnology, 2024, 108(1): 209.
[102] 曲戈, 朱彤, 蒋迎迎, 吴边, 孙周通. 蛋白质工程: 从定向进化到计算设计[J]. 生物工程学报, 2019, 35(10): 1843-1856.
QU G, ZHU T, JIANG YY, WU B, SUN ZT. Protein engineering: from directed evolution to computational design[J]. Chinese Journal of Biotechnology, 2019, 35(10): 1843-1856 (in Chinese).
[103] THEAN DGL, CHU HY, FONG JHC, CHAN BKC, ZHOU P, KWOK CCS, CHAN YM, MAK SYL, CHOI GCG, HO JWK, ZHENG ZL, WONG ASL. Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities[J]. Nature Communications, 2022, 13(1): 2219.

[104] CHU HY, FONG JHC, THEAN DGL, ZHOU P, FUNG FKC, HUANG YH, WONG ASL. Accurate top protein variant discovery *via* low-N pick-and-validate machine learning[J]. Cell Systems, 2024, 15(2): 193-203.e6.
[105] CHOI GCG, ZHOU P, YUEN CTL, CHAN BKC, XU F, BAO SY, CHU HY, THEAN D, TAN K, WONG KH, ZHENG ZL, WONG ASL. Combinatorial mutagenesis en masse optimizes the genome editing activities of SpCas9[J]. Nature Methods, 2019, 16(8): 722-730.
[106] WITTMANN BJ, YUE YS, ARNOLD FH. Informed training set design enables efficient machine learning-assisted directed protein evolution[J]. Cell Systems, 2021, 12(11): 1026-1045.e7.
[107] WU Z, JENNIFER KAN SB, LEWIS RD, WITTMANN BJ, ARNOLD FH. Machine learning-assisted directed protein evolution with combinatorial libraries[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(18): 8852-8858.
[108] DUAN ZQ, LIANG YF, SUN JL, ZHENG HJ, LIN T, LUO PY, WANG MG, LIU RH, CHEN Y, GUO SH, JIA NN, XIE HT, ZHOU ML, XIA MH, ZHAO KJ, WANG SH, LIU N, JIA YL, SI W, CHEN QT, et al. An engineered Cas12i nuclease that is an efficient genome editing tool in animals and plants[J]. The Innovation, 2024, 5(2): 100564.
[109] NGUYEN LT, RANANAWARE SR, YANG LG, MACALUSO NC, OCANA-ORTIZ JE, MEISTER KS, PIZZANO BLM, SANDOVAL LSW, HAUTAMAKI RC, FANG ZR, JOSEPH SM, SHOEMAKER GM, CARMAN DR, CHANG LW, RAKESTRAW NR, ZACHARY JF, GUERRA S, PEREZ A, JAIN PK. Engineering highly thermostable Cas12b *via de novo* structural analyses for one-pot detection of nucleic acids[J]. Cell Reports Medicine, 2023, 4(5): 101037.
[110] WATERHOUSE A, BERTONI M, BIENERT S, STUDER G, TAURIELLO G, GUMIENNY R, HEER FT, de BEER TAP, REMPFER C, BORDOLI L, LEPORE R, SCHWEDE T. SWISS-MODEL: homology modelling of protein structures and complexes[J]. Nucleic Acids Research, 2018, 46(W1): W296-W303.
[111] CAO HL, WANG JX, HE LP, QI YF, ZHANG JZ. DeepDDG: predicting the stability change of protein point mutations using neural networks[J]. Journal of Chemical Information and Modeling, 2019, 59(4): 1508-1514.
[112] SUMBALOVA L, STOURAC J, MARTINEK T, BEDNAR D, DAMBORSKY J. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information[J]. Nucleic Acids Research, 2018, 46(W1): W356-W362.
[113] CHENG P, MAO C, TANG J, YANG S, CHENG Y, WANG WK, GU QX, HAN W, CHEN H, LI SH, CHEN YF, ZHOU JL, LI WJ, PAN AM, ZHAO SW, HUANG XX, ZHU SQ, ZHANG J, SHU WJ, WANG SQ. Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering[J]. Cell Research, 2024, 34: 630-647.
[114] WU HM, SUN YX, WANG YM, LUO LQ, SONG YZ. Advances in miniature CRISPR-Cas proteins and their applications in gene editing[J]. Archives of Microbiology, 2024, 206(5): 231.
[115] ZHAO FY, ZHANG T, SUN XD, ZHANG XY, CHEN LT, WANG HJ, LI JZ, FAN P, LAI LX, SUI TT, LI ZJ. A strategy for Cas13 miniaturization based on the structure and AlphaFold[J]. Nature Communications, 2023, 14(1): 5545.
[116] DENG JC, LI XY, YU H, YANG L, WANG ZR, YI WF, LIU Y, XIAO WY, XIANG HY, XIE ZC, LV DM, OUYANG HS, PANG DX, YUAN HM. Accelerated discovery and miniaturization of novel single-stranded cytidine deaminases[J]. Nucleic Acids Research, 2024, 52(18): 11188-11202.
[117] GAWLITT S, LIAO CY, ACHMEDOV T, BEISEL CL. Shortened CRISPR-Cas9 arrays enable multiplexed gene targeting in bacteria from a smaller DNA footprint[J]. RNA Biology, 2023, 20(1): 666-680.