特约评述

DOI: 10.12211/2096-8280.2025-044

AI+定向进化赋能蛋白改造及优化

宋成治[1]，林一瀚[1, 2, 3]
（[1] 北京大学定量生物学中心，北京大学-清华大学生命科学联合中心，北京大学前沿学科交叉研究院，北京 100871；[2] 北京大学教育部细胞增殖与分化重点实验室，北京大学生命科学学院，北京 100871；[3] 北京大学成都前沿交叉生物技术研究院，四川 成都 610213）

摘要：定向进化是合成生物学领域的核心底层技术之一。通过在实验室中模拟自然界发生的进化过程，定向进化利用功能筛选从大量的突变序列文库中不断获得性能提升的蛋白序列，帮助实现野生型蛋白难以实现的功能。近年来不断发展的机器学习、蛋白语言模型等人工智能（artificial intelligence， AI）方法进一步拓展了该技术的使用场景和工作效率，帮助其在酶、抗体、生物传感器等的改造中取得优异表现。本文总结了传统定向进化在突变文库构建和功能筛选过程中使用的典型策略，并对近年来开发的高效连续定向进化平台进行介绍，进一步对定向进化技术存在的序列空间有限、容易陷入局部最优等一系列问题进行探讨。快速迭代的机器学习模型与定向进化相结合，一方面能够缓解序列空间的探索局限性，另一方面能够从起始序列设计、中间文库优化、功能信息提取等多个维度对定向进化的实验流程进行完善，帮助实现更加高效的蛋白改造尝试。为明确定向进化结合机器学习的应用潜力，本文重点展示了机器学习辅助定向进化的代表案例。最后，简要探讨了该领域的潜在挑战和未来发展方向。

关键词：定向进化；机器学习；蛋白改造；蛋白语言模型；合成生物学

中图分类号：Q816　　**文献标志码：**A

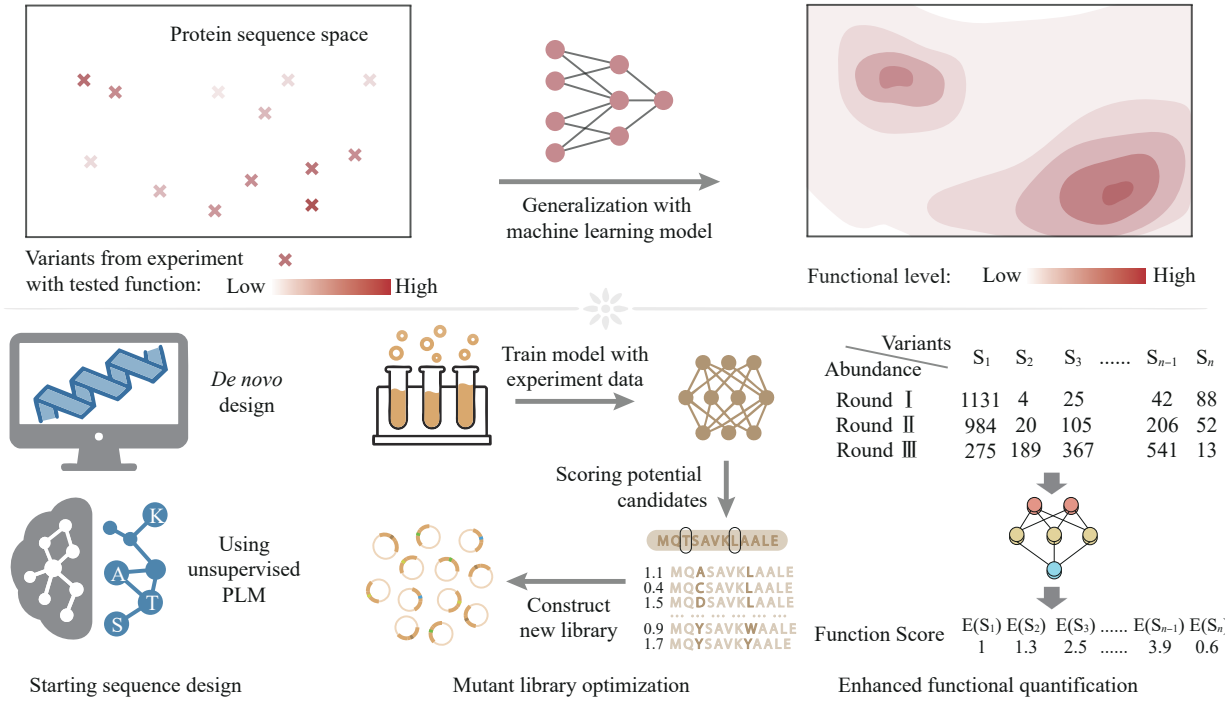
AI-enabled directed evolution for protein engineering and optimization

SONG Chengzhi[1]，LIN Yihan[1, 2, 3]
（[1]*Center for Quantitative Biology*，*Peking-Tsinghua Center for Life Sciences*，*Academy for Advanced Interdisciplinary Studies*，*Peking University*，*Beijing 100871*，*China*；[2]*The MOE Key Laboratory of Cell Proliferation and Differentiation*，*School of Life Sciences*，*Peking University*，*Beijing 100871*，*China*；[3]*Chengdu Academy for Advanced Interdisciplinary Biotechnologies*，*Peking University*，*Chengdu 610213*，*Sichuan*，*China*）

Abstract: Directed evolution is one of the core enabling technologies in synthetic biology. By recapitulating evolutionary processes that occur in nature within the laboratories, directed evolution employs functional screening to continually isolate variants with improved performance from large mutant libraries for functions that are difficult to

收稿日期：2025-05-12 修回日期：2025-06-06
基金项目：国家重点研发计划（2020YFA0906900）
引用本文：宋成治，林一瀚. AI+定向进化赋能蛋白改造及优化[J]. 合成生物学，2025，6(3)：617-635
Citation：SONG Chengzhi，LIN Yihan. AI-enabled directed evolution for protein engineering and optimization[J]. Synthetic Biology Journal，2025，6(3)：617-635

achieve with wild-type proteins. In recent years, rapidly advancing artificial intelligence (AI) approaches—such as machine learning and protein language models—have further expanded both the range of applications and the operational efficiency of directed evolution, yielding unprecedented achievements in the engineering of enzymes, antibodies, biosensors, and more. In this review, we first outline classic strategies and emerging techniques for mutagenesis and functional selection in traditional directed evolution, followed by an in-depth examination of various continuous directed evolution systems. We highlight common limitations of directed evolution, emphasizing issues such as constrained search space and susceptibility to local optima. Combining rapidly iterated AI methods with directed evolution offers promising solutions to these challenges. Protein language models, in particular, leverage learned patterns from experimental variants alongside fundamental protein properties, providing superior predictive accuracy for unexplored mutants and facilitating the extrapolation of sequence-function relationships to broader sequence space. AI-based methods enhance directed evolution workflows from multiple perspectives. *De novo* protein design and unsupervised protein language models aid in generating functional starting sequences with targeted sequence diversity. Machine learning models trained on experimental data enable the construction of optimized mutant libraries tailored for subsequent selection rounds. Additionally, models derived from statistical physics and dynamical systems help extract detailed functional information from data acquired across multiple selection rounds. Collectively, these machine learning approaches significantly enhance the overall efficiency of directed evolution. To illustrate the transformative potential of machine learning-assisted directed evolution, we discuss exemplary cases of protein function improvement and modification. Lastly, we briefly address ongoing challenges and future directions in this rapidly evolving and promising research area.



Machine learning systematically optimizes multiple stages of directed evolution

Keywords: directed evolution; machine learning; protein engineering; protein language model; synthetic biology

1　定向进化与蛋白功能改造

定向进化作为实验方法的提出可追溯至 20 世纪 60 年代[1]。尽管随后便出现了其在蛋白领域的应用[2]，但直到易错 PCR 技术的出现[3]，定向进化技术才开始大量用于蛋白功能的改造与优化。1991 年，Arnold 等[4]首先使用定向进化在实验室条件下提升了枯草杆菌蛋白酶在有机溶剂二甲基甲酰胺中的酶活性，在多轮筛选后，含有三处突变的序列变体的活性相比于野生型提升了 38 倍。如今，定向进化已成为蛋白工程的常用工具，广泛应用于酶、抗体、转录因子等多类蛋白的改造及优化任务中[5-12]。

1.1　定向进化方法的优势与一般步骤

相比于其他方法[13-15]，定向进化并不需要提前预知所优化蛋白的结构信息或功能机制，在可以进行有效筛选的前提下，该方法能够应用于众多结构未知或功能机制不清的蛋白质。通过对序列文库进行多轮迭代进化，定向进化能对起始蛋白序列附近的序列空间进行高效探索，帮助研究人员快速定位显著提升功能的关键突变。同时，定向进化对所需提升的蛋白性质并无特定限制，对结合强度、催化活性、底物选择性、热稳定性等各类性质均能进行优化提升，具有较强灵活性与通用性。传统定向进化流程主要包括突变文库构建和功能筛选。在从序列库中富集得到功能提升的序列后，可根据得到的优化序列设计新一轮的突变文库进一步筛选，经过不断迭代，得到性能大幅提升的目标蛋白（图 1）。

突变文库的产生方法主要分为体外构建和体内构建两类[16-17]。由 Goeddel 等开发的易错 PCR 技术无疑是当前应用最广泛的体外实验方法之一[3]（图 1）。低保真 DNA 聚合酶能够在 PCR 过程中引入大量点突变，通过调整反应体系的镁离子和锰离子浓度，延长 PCR 的扩增轮数，该方法可实现单碱基 10^{-3} 的突变引入频率；使用诱变核酸类似物可进一步将突变率提升至 10^{-2} 到 10^{-1} 量级[18]。尽管传统的基于 *Taq* 聚合酶的易错 PCR 反应存在 AT 到 GC 间替换以及 AT 到 TA 间颠换的偏好性，但这种突变谱的偏倚可以通过将 *Taq* 酶与其他类型的聚合酶组合使用得到缓解[19]。与易错 PCR 相比，定点饱和突变（site-saturation mutagenesis， SSM）能够在特定位置引入所有的 19 种氨基酸突变类型[20-21]，在已知结构信息时，针对特定序列区间构

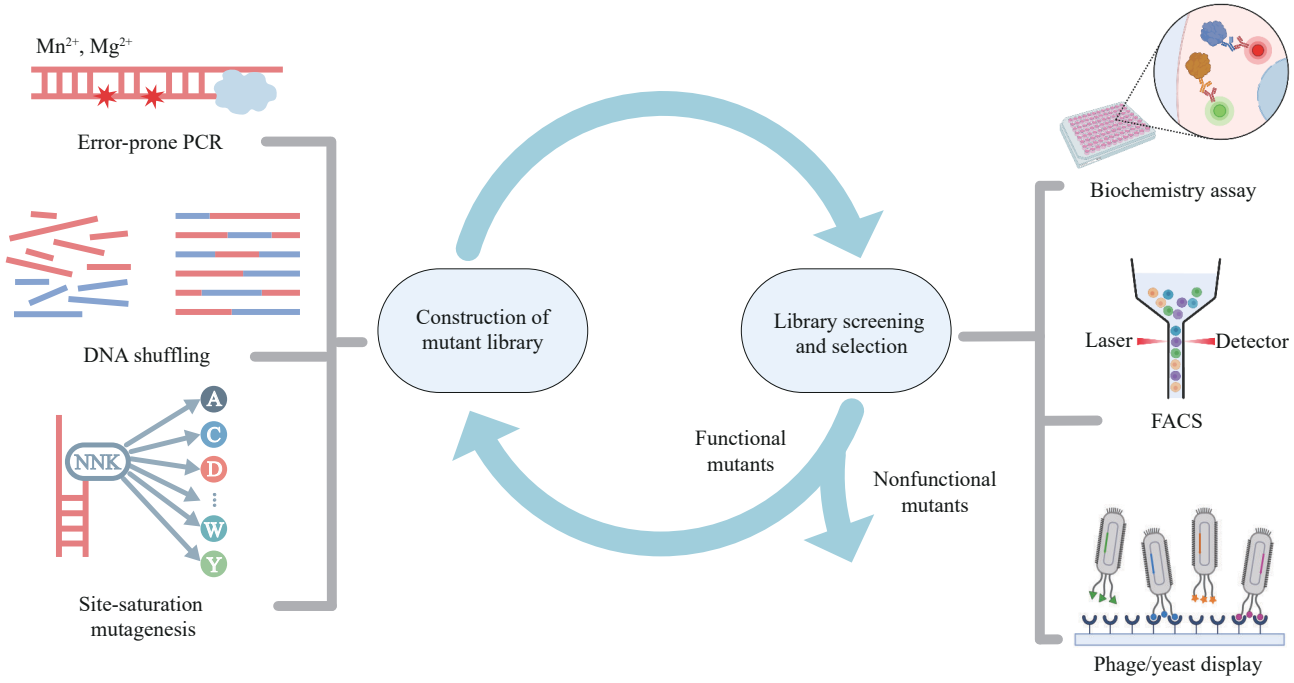


图 1　定向进化的一般步骤与相关技术方法（使用 BioRender 绘制）

Fig. 1　Schematic workflow for directed evolution and related methodologies (created in BioRener)

建多位点饱和突变文库可加速筛选进程，更快获得功能提升的序列变体。DNA重排是另一类广泛使用的体外突变方法，通过模拟自然界同源重组过程，该方法可对不同区段有益突变进行快速组合[22-23]。部分DNA重排方法还能实现不依赖同源序列的重组过程[24-25]，大大扩充了突变引入和组合的实现手段。

相比于体外突变，体内突变能够直接在宿主细胞内完成诱变过程，无需进行额外的分子克隆和转化操作，在实验流程上更加便捷。早期的体内突变文库构建主要通过使用诱变剂处理，或利用具有DNA错配修复机制缺陷的菌株实现[26-27]。但这些方法均会在全基因组层面造成影响，并可能大幅降低宿主细胞的遗传稳定性。该局限性一方面会对突变效率带来限制，另一方面还会因非目标突变而绕过实验设定的筛选压力，影响定向进化的最终效果。该问题的一种潜在解决方案是通过正交的增变质粒或增变菌株来实现目标基因特异性的体内诱导突变。T7噬菌体中的DNA依赖型RNA聚合酶（T7 RNAP）能够特异性结合T7启动子并对其下游基因进行转录，通过将T7 RNAP和胞嘧啶脱氨酶融合，Moore等[28]有针对性地在T7启动子下游的目的基因中引入了CG到TA的突变。与上述基于T7 RNAP的技术相似，OrthoRep通过在宿主细胞内引入酿酒酵母的核外复制系统来实现正交的诱导突变[29-30]。此外，也有方法基于CRISPR体系实现靶向诱变[31-32]。比如，EvolvR方法通过Cas9 nickase定位目的基因，并利用所融合的易错DNA聚合酶在随后的修复和复制过程中引入突变[31]。借助不同类型的DNA聚合酶，EvolvR可实现从10^{-7}到10^{-3}不等的突变效率。相比之下，T7-DIVA技术则结合了上述正交复制体系和CRISPR体系的优势[33]，一方面通过融合有胞嘧啶脱氨酶的T7 RNAP来实现突变产生，另一方面又利用dCas9在目标基因下游形成障碍物来终止转录过程，从而更加精确地指定了突变产生的序列区间。Yi等[34]开发的TADR系统利用噬菌体蛋白CisA对目标质粒特定起始序列进行切割，并使用由Rep解旋酶引导的易错T5 DNA聚合酶在目的基因引入突变。通过对辅助质粒中的解旋酶和T5 DNA聚合酶进行序列优化，该系统可实现目的基因序列突变频率最高20万倍的提升，同时还可在一轮突变筛选中引入多位点复杂突变。除上述列举的方法外，还有一些方法基于逆转录元件[35]或基因重组[36]等技术实现了体内的突变文库构建。相关技术方法的开发和完善是目前定向进化领域的研究热点之一。

在完成突变文库的构建后，便需要对文库中的序列进行筛选和检测（图1）。在文库规模较小时，核磁共振、色谱、质谱等传统的化学表征便可直接对突变序列的功能表现进行检测。Arnold等在对细胞色素P450酶的定向进化实验中[37-38]，通过气相色谱检测的方式对从文库中挑选出的92条序列变体进行了环丙烷生成效率的检测，最终获得了具有优异催化活性和立体选择性的P450突变体。需要注意的是，上述案例是在已知文库中大量序列的功能特点后，从中有针对性地挑选出了少数序列进行了下游更精细的生化表征。然而，在大多数定向进化案例中，在没有结构和功能先验知识的情况下，往往仍要从较大的突变文库中进行筛选才更有可能得到功能提升的进化序列。不断发展的测序技术和荧光激活细胞分选（FACS）技术为突变文库的高通量检测和表征提供了便利的条件。对于蛋白功能直接与细胞存活相关的案例[39-40]，可通过一段时间培养后收集细胞直接进行测序的方式获得功能提升的突变序列。而对于功能本身不直接与细胞活性挂钩的目标蛋白，则可通过酵母展示[10, 41]、构建荧光或抗性报告系统[42-43]等方式将其功能转化为荧光强度或细胞数量等能够进行直接测量的指标，随后结合流式细胞分选等技术对其功能进行表征，利用高通量测序对富集到的突变体进行序列鉴定。

1.2 连续定向进化

传统定向进化方法在多轮的进化过程中需要不断进行新文库的生成和筛选，重复的操作流程不仅造成更高的时间和人力成本投入，同时还极大限制了进化实验所能进行的迭代轮数[44-45]。近年来不断发展的连续定向进化方法旨在将原先需要大量人工干预才能完成的多轮进化过程转移到实验体系内自发进行。在体系构建完成后，突变的

产生和筛选自动发生，长时间、大规模的进化筛选也因此更具可能。这里我们对部分病毒介导的连续定向进化体系进行重点介绍，同时我们也将对其他类型的连续定向进化平台做简单概述。

噬菌体介导的连续定向进化系统PACE是较早实现连续突变和筛选的实验平台之一[46]［图2(a)］。它将M13丝状噬菌体的感染能力与需要进化的目标蛋白功能偶联，通过不断稀释替换培养体系内的宿主细胞实现对拥有高侵染能力的噬菌体的筛选，进而得到功能提升的蛋白序列。如图2(a)所示，噬菌体复制和感染所需的基因均被编码于筛选质粒SP中，而其中感染宿主细胞必需的衣壳蛋白pⅢ被需要进化的目标蛋白替换，编码pⅢ蛋白的基因则被置于辅助质粒AP内并和诱变质粒MP共同转化至宿主细胞。诱变质粒能够不断促进SP质粒中的目的基因序列发生突变，当突变的目标蛋白能够行使其预期功能时，AP质粒中对应的启动子才可被激活并开启下游gⅢ基因的表达，帮助产生具有感染能力的新一代噬菌体。研究者利用8天超过200轮的进化过程，成功得到了能够识别T3启动子以及能够借助CTP起始转录的T7 RNAP变体[46]。目前，PACE已被运用于多类蛋白的功能改造和优化任务中，进化得到了拥有全新特异性的蛋白酶[47-48]、性能更优的Cas9系统和碱基编辑器[49-50]、序列简短而高效的降解因子[51]等。

另一项研究工作利用Sindbis病毒实现了哺乳动物细胞内的连续定向进化[52]。在这一名为VEGAS的实验平台中［图2(b)］，研究人员将Sindbis病毒的结构蛋白基因从病毒基因组中去除，并通过调控质粒（pSSG）将这部分基因转移至宿主细胞内。目的基因则被插入病毒基因组中（pTSin），借助该RNA病毒复制过程中的高出错率不断引入突变。当与病毒基因组共同表达的目的基因编码蛋白能够正常发挥功能时，宿主细胞内病毒结构蛋白基因的表达将被激活，从而产生更多携带该突变目的基因序列的病毒颗粒，实现对目标蛋白的功能筛选。例如，在对G蛋白偶联受体（G protein-coupled receptor， GPCR）的筛选试

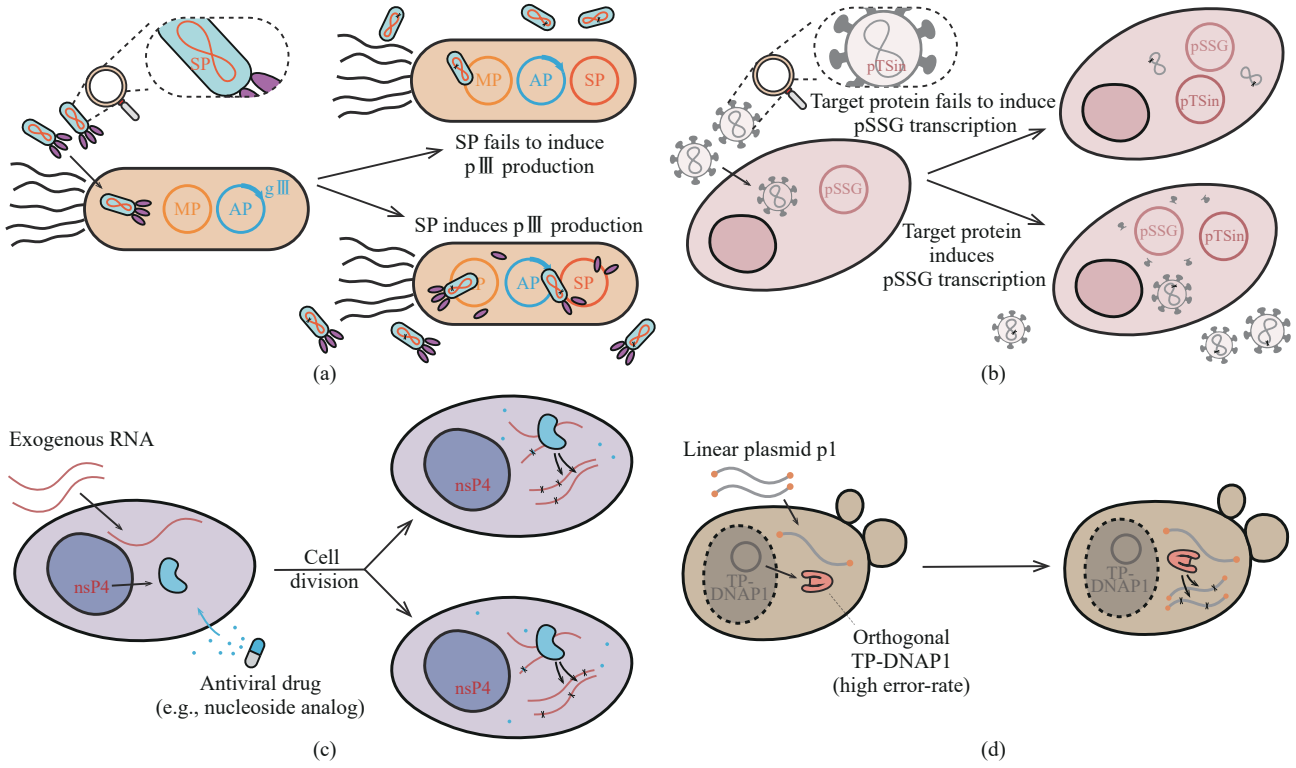


图2 部分连续定向进化平台原理示意图

［(a)～(d)分别为PACE[46]、VEGAS[52]、REPLACE[42]、OrthoRep系统[29]，图中ᚷ用于表示在目标片段中出现的突变。］

Fig. 2 Illustration of continuous directed evolution platforms

[Panels (a)–(d) correspond to PACE[46], VEGAS[52], REPLACE[42], and OrthoRep[29], respectively. The symbol ᚷ denotes mutation events.]

验中，研究人员将病毒结构蛋白基因置于含有血清响应元件（serum response element， SRE）的启动子下游。当目标GPCR蛋白（MRGPRX2）能成功激活时，该GPCR将激活MAPK通路，并通过该通路中ERK蛋白的磷酸化最终促使血清响应因子（serum response factor， SRF）与SRE的结合，进而开启病毒结构蛋白的表达。通过在三天的筛选过程中将吗啡配体的浓度由1 μmol/L调整至100 nmol/L，并最后降低至10 nmol/L，VEGAS系统成功进化得到了无需配体作用便可持续激活的GPCR变体；从10^7量级的cDNA文库开始，经过不到一周的进化筛选，获得了多条能有效提高GPCR与Gα蛋白结合稳定性的纳米抗体。然而，虽然上述方法实现了哺乳动物细胞内的连续定向进化，但该方法具有较高的细胞毒性，难以长时间稳定维持；同时目的基因在筛选过程中容易丢失，造成实验筛选失败[53]；更重要的是，实验过程中能够产生具有侵染能力的活病毒，其潜在的生物危害使得相关实验流程难以在普通实验室开展[54-55]。

另一连续定向进化系统REPLACE成功解决了上述问题[42]［图2（c）］。与VEGAS中利用完整Sindbis病毒进行进化筛选不同，REPLACE体系将病毒结构蛋白全部去除，利用含有目的基因的病毒RNA在细胞增殖过程中的不断复制对有益突变进行富集。具体而言，该系统通过在病毒基因组中引入点突变，并将含有Sindbis病毒RNA依赖型RNA聚合酶（RNA-dependent RNA polymerase， RdRp）的nsP4片段转移至宿主细胞稳定表达，大幅提高了病毒RNA与宿主细胞的兼容性，实现了病毒RNA的稳定复制。该系统能够通过一周的细胞培养获得10^8量级的病毒RNA拷贝，为长时间、大规模的序列筛选创造了有利条件。相比于依赖大量nsP4自身的低保真转录过程引入突变，REPLACE系统将宿主细胞内稳定表达的nsP4片段控制在较低水平，通过加入不同浓度针对nsP4的诱变剂来人为调控目的基因中突变的引入速度。相比于nsP4自身的突变谱，诱变剂莫诺拉韦（Molnupiravir）影响下的突变谱在不同替换类型上更加平衡。同时，由于该系统无需在宿主细胞中引入病毒结构蛋白，RNA拷贝数量的维持和扩增可在不进行完整病毒组装的情况下在宿主细胞内发生，大大降低了潜在的生物危害。REPLACE系统在多个进化案例中均取得了突出效果：从初始序列出发，经过两周的细胞培养与两轮细胞分选，成功将EGFP与StayGold两种绿色荧光蛋白的发射光谱显著蓝移；经过7天的连续进化筛选，获得了针对不同MEK1抑制剂具有高度耐受性的多种MEK1突变体；以原始PadR转录因子为起点，经三周进化，实现了其转录调控动态范围的大幅提升。

在病毒介导的连续定向进化平台之外，正交复制系统介导的实验体系也同样备受关注。其中，OrthoRep平台通过酿酒酵母的核外复制系统来实现不影响宿主基因组的目的基因突变诱导[29]［图2（d）］。该系统使用TP-DNAP1质粒编码正交型易错DNA聚合酶，此聚合酶能够在含有目的基因的线性化质粒p1内高效引入突变。研究者利用该系统对疟原虫二氢叶酸还原酶（*Pf*DHFR）的耐药机制进行探究，在90组独立定向进化实验中，不仅得到了多条常见的耐药性产生路径，还发现了少数摆脱上位效应约束的新突变类型[29]。将OrthoRep与酵母展示技术相结合，Wellner等[56]搭建了AHEAD系统（autonomous hypermutation yeast surface display）对SARS-CoV-2抗体进行进化筛选。研究人员成功获得与病毒RBD区域结合能力大幅提高的多条纳米抗体序列，部分抗体实现了925倍的中和活性提升。Rix等通过定向进化进一步将OrthoRep系统的单碱基替换率提升至1.7×10^{-4}量级[57]，在3个月的进化筛选后，该平台能够产生上千条序列差异超过60个氨基酸的蛋白突变体，其序列间差异性已超出小鼠与人类同源基因序列差异的中位值。此外，还有一系列连续定向进化平台通过CRISPR介导等方式实现突变的连续产生和筛选[31，39，58-59]。这些连续定向进化平台不仅使蛋白功能优化更加高效，其在长时间进化过程中积累的大量突变信息也将帮助我们理解蛋白序列与功能的映射规律，启发我们更好地对蛋白功能进行改造与设计。

1.3 定向进化方法的局限性

尽管经过良好设计的定向进化实验能够产生

功能提升的蛋白突变体，但实验中筛选检测的突变序列数量相比于整个蛋白序列空间仍只是冰山一角。对于一个含有300个氨基酸的典型蛋白质，其双突变的组合数约为1600万，而其三突变组合数更是达到300亿的量级。虽然当前的检测手段已在通量上得到明显提升，但高阶突变的全覆盖检测仍远远超出了当前技术的适用范畴。与此同时，上位效应等因素的存在进一步限制了定向进化实验所能探索的序列范围：对于存在较强上位效应的位点组合，定向进化的演化路径大概率会停止于序列空间的局部最优位置，而很难到达全局最优点[60-61]。上位效应还会导致不同类型的有利突变在序列空间中被大幅度分隔开来[62]，而在定向进化的实验过程中，突变的引入往往只能以单个碱基替换的方式依次进行，在有限的进化轮数内，单次实验所能探索到的有效极值点数量也因此十分有限。另外，对于大部分高通量检测方法，蛋白功能的强弱常常需要转化为序列丰度、荧光强度等间接测量指标，尽管良好的实验设计能够保证二者间较强的相关性，但额外的转换过程不可避免地引入了更多的测量噪声，使得对于突变体功能的评估需要更加谨慎。

2 机器学习助力探索更大序列空间

定向进化技术的一大局限在于实验历经的突变类型十分有限。在有限的文库规模和检测通量下，研究人员不得不在覆盖序列的深度或广度上做出取舍，这便不可避免地使诸多潜在的强功能突变体无法被进化过程所富集。

近年来快速发展的机器学习等人工智能技术有力推动了不同科学领域的进步，通过对已有数据进行分析学习，该类方法能够提取数据中存在的特征与规律，并将这些“知识”运用于模型未曾接触过的新样本中，对其性质进行推理和判断。人工智能模型的工作原理正好适应了定向进化技术扩展序列搜索空间的需求。理论上说，模型在利用实验数据进行训练后，有望对任意类型的突变组合进行功能预测，进而摆脱定向进化实验的技术局限，填补定向进化实验中未曾探索到的序列区域存在的空白。Fox等[63-64]较早使用统计学习方法对定向进化数据进行分析，借助其表征蛋白序列-功能对应关系的线性模型ProSAR，研究人员在使用多轮实验数据对模型参数进行回归拟合后，成功得到了氰化反应产率提升近4000倍的卤醇脱卤酶[64]。在其优化过程中，ProSAR利用60000条实验突变提取信息，从30个突变位点约10^9量级的潜在突变组合中筛选得到了最终的优化突变体，极大拓展了实验方法的探索空间。在另一项研究中[65]，Romero等使用242条嵌合型细胞色素P450酶突变体的热稳定性信息对一高斯过程模型进行训练，该模型不仅能够对新序列的热稳定性进行较为准确的估计，在结合额外的突变体功能数据后，还能对新序列的活性信息进行简单评估。

传统机器学习方法虽然能在相对简单的预测任务中拥有不错表现，但其泛化能力往往只限于对训练集中出现的已有突变类型进行简单组合，无法获取更为普适的位点突变倾向性或更为复杂的突变关联信息[66]。由于这类模型需要依赖于先验实验结果，因而难以用来指导完整的进化过程。近年来不断发展的蛋白语言模型（protein language models， PLM）为这一难点的突破带来了新的契机。计算机领域的大语言模型借助大量文本数据进行训练，在对庞大数据集分析处理后，这类模型能够一定程度掌握自然语言中词汇和句子的组合模式与搭配规律，从而在此基础上完成文本翻译、语义理解、会话生成等复杂任务。与自然语言类似，不同蛋白质均由其序列中的一个个氨基酸构成。如果将各类蛋白比作一个个句子，那么蛋白序列中的氨基酸就是组成句子的单词。采用与自然语言处理类似的方式，通过大量的蛋白序列对模型进行训练，便可以获得针对蛋白质的语言模型[67]。当前不同公共数据库中已有海量的蛋白序列信息（UniProt， NCBI， GenBank等），通过对这些数据进行分析学习，蛋白语言模型或许能够像自然语言模型一样，提取得到蛋白序列的构成规律，并将这些知识用于更精确的功能预测和序列设计任务中[68]。

目前为止，已有一系列不同类型的蛋白语言模型被提出，对不同PLM的简要汇总参见表1。

UniRep作为较早出现的蛋白语言模型之一[71, 92]，通过在UniRef50数据集[93]的2400万蛋

表1 蛋白质语言模型汇总

Table 1 Summary for protein language models

蛋白质语言模型 Protein language model	年份 Year	参数量 Parameters	训练数据规模 Training data size	架构 Architecture
ProtVec[69]	2015	—	Swiss-Prot 0.55M sequences	Word2Vec
SeqVec[70]	2019	93M	UniRef50 33M sequences	BiLSTM
UniRep[71]	2019	18M	UniRef50 24M sequences	mLSTM
TAPE (Transformer)[72]	2019	38M	Pfam 31M protein domains seqeunces	Encoder-only
ProGen[73-74]	2020	1.2B	280M sequences from 5 sources	Decoder-only
ESM-1b[75]	2021	650M	UniRef50 27.1M sequence	Encoder-only
ESM-1v[76]	2021	650M	UniRef90 98M sequence	Encoder-only
MSA Transformer[77]	2021	100M	26M multiple sequence alignments	Axial Transformer
ProteinBERT[78]	2022	16M	UniRef90 106M sequences	Encoder-only
ProtGPT2[79]	2022	738M	UniRef50 45M sequences	Decoder-only
ProtT5 (ProtTrans)[80]	2022	3B(XL)/11B(XXL)	BFD + UniRef50 2.3B sequences	Encoder-Decoder
ESM-2[81]	2023	8M～15B	UniRef50 60M sequences	Encoder-only
ProGen2[82]	2023	151M～6.4B	UniRef90+BFD30 1B sequences	Decoder-only
Ankh[83]	2023	450M(base);1.15B(large)	UniRef50 45M sequences	Encoder-Decoder
PoET[84]	2023	604M	UniRef50 29M sets of homologous sequence	Decoder-only
ESM3[85]	2024	1.4B～98B	3.15B protein sequences, 236M protein structures, and 539M proteins’function annotations	Encoder-only
CARP[86]	2024	600k～640M	UniRef50 42M sequences	CNN
ProLLaMA[87]	2024	7B	UniRef50 48M sequences	Decoder-only
xTrimoPGLM[88]	2024	1B～100B	UniRef90 + ColabFoldDB 939M protein sequences	Transformer
Prot42[89]	2025	500M(base);1.1B(large)	UniRef50 57M sequences	Decoder-only
T5ProtChem[90]	2025	102M	UniRef50 52M protein sequences and PubChem 97M SMILES sequences	Encoder-Decoder
LC-PLM[91]	2025	1.4B	UniRef50 63M sequences	BiMamba

白序列上进行训练，能够输出有效反映序列二级结构、残基物化性质、序列物种来源等信息的蛋白编码。该模型采用LSTM架构[94]，在无需额外数据输入的情况下便可对天然或人造蛋白序列的稳定性进行直接预测；在使用少量荧光蛋白的亮度信息进行微调后，UniRep还可对与训练集差异较大的突变序列的荧光强度进行较为准确的预估。当前最为流行的ESM系列模型在泛化能力上显著提升[75-77, 81, 85, 95]，其中ESM-1v使用9800万蛋白序列对其transformer架构中的6.5亿参数进行训练，在突变体的无监督功能预测上超出先前模型的表现[76]。在使用目标蛋白的少量同源序列对模型进行微调后，其表现还可进一步优化。ESM-2通过不同参数量模型的比较表明了参数量上升能够给模型性能带来持续的增强[81]。使用150亿参数的ESM-2模型得到的序列编码进行蛋白结构预测，ESMFold可得到原子级精度的高质量蛋白结构，其准确性比肩相同时期最优的结构预测模型AlphaFold2[96]，而所需的计算时间却比AlphaFold2减少了近一个数量级。与此同时，ESM-2在突变体功能预测的表现上也展现出相比先前模型的明显进步[97]。最新推出的ESM3模型则在序列信息之外考虑蛋白结构和功能注释等其他模态信息能够对模型编码带来的帮助[85]，通过使用30亿蛋白序列、2亿蛋白结构、5亿蛋白功能注释信息进行模型训练，ESM3获得了泛化能力的大幅提升。该模型在GFP序列重新设计任务中，成功产生了与天然GFP序列相似度仅有58%的全新突变序列，

而这样的序列差异性在自然界中需要5亿年的进化积累才能产生。此外，其他一些PLM也展现出了较为突出的模型性能[79, 82-83, 86]。其中，Ankh模型通过对预训练数据集构成、训练过程、模型结构等进行优化，在参数量较小、运算资源消耗较少的前提下，同样实现了优异的模型泛化能力，在不同下游预测任务中均达到当时的最优水平[83]。

人工智能技术尤其是蛋白语言模型的发展为定向进化方法探索更大序列空间创造了有利条件。借助模型的泛化能力，定向进化实验能够进一步提升筛选效率，定位序列类型更加丰富的功能蛋白。

3 基于AI的实验流程优化

基于前述分析，我们看到蛋白语言模型在扩展定向进化序列搜索空间方面的巨大潜力。然而，AI技术对定向进化的赋能作用远不止于此——它正在从实验流程的每一个环节入手，系统性地提升定向进化的效率和成功率。下面我们将从起始序列优化、中间文库设计和功能信息提取三个方面，详细阐述AI方法如何对定向进化实验流程进行全方位的改进。

3.1 起始序列的优化选取

尽管定向进化实验在初始筛选阶段的选择压一般相对较小，但为使初始序列能够在体系中扩繁并积累更多突变，实验的进化起点必须具有一定的功能活性。对于蛋白功能的优化，野生型序列常常已满足要求，可作为进化起点直接使用；但对于蛋白功能的改造，诸如换用不同反应底物、识别新型酶切位点、适应不同反应环境条件等需求，野生型序列本身通常不具备相应功能。传统的理性设计或突变筛选依赖于较多的先验知识和额外的实验检测[38, 48, 98]，人工智能技术则为我们提供了起始序列选取的其他途径。

近年来，功能蛋白的从头设计备受人们关注，比如，Glögl等[99]设计出能够高亲和力结合肿瘤坏死因子受体TNFR1的拮抗剂和激动剂，Torres等[100]构建了可以有效中和蛇毒3FTx的新型蛋白，Yeh等[101]设计了能够选择性催化人造荧光素DTZ和h-CTZ发光的荧光素酶等。从头设计方法目前正应用于越来越多的蛋白类型，这无疑为起始序列选取提供了更多选择。在Kipnis等[102]的研究中，他们通过计算方法从头设计出一种以β桶作为结构骨架的逆醇醛缩合反应催化酶。从设计序列出发进行定向进化后，他们得到的人造酶RAβb-16.2最终展现出反应活性和手性选择性的大幅提升。除此之外，借助蛋白语言模型对功能的无监督评估，模型预筛得到的高潜力突变体也能为起始序列选取提供参考。在Ding等提出的MODIFY算法中[103]，借助多个模型的综合打分以及基于结构的序列过滤，具有较强功能的突变序列能够被有效识别并被加入备选序列文库。另外，为探索更大序列空间，还可通过计算方法在起始序列中引入不同程度的序列差异。在Fram等的工作中[104]，他们通过同源序列比对的方法生成了相应的最大熵模型，使用该模型进行不同步数的采样，即可获得与原始序列具有指定序列相似性水平的突变体。与此类似，前述Modify算法也可通过调节预期功能强度和序列多样性间的权重配比，生成更具差异性的起始序列文库[103]。

3.2 中间文库设计的优化

对于非连续定向进化实验，在后续多轮的筛选过程中仍需要对每轮用于检测的突变文库进行重新设计与合成。在检测通量有限的条件下，结合已经获得的序列功能信息对后续筛选文库进行合理设计可以帮助提升筛选效率，获得更多功能更强的突变序列。

在一项计算模拟中[105]，研究人员使用GB1蛋白4个位点的深度突变扫描数据[106]对三种不同的文库生成方案进行了比较。前两种方案均通过对所测实验数据中的单位点最优突变进行不同方式的组合来获得下一轮筛选的序列文库；第三种方案则通过使用相同数量的数据对一简单的神经网络进行训练，依据模型对所有可能序列进行打分，挑选出其中打分最高的序列组成后续筛选文库。经过600次的重复模拟，前两种传统的文库构建方案分别在4.0%和4.9%的试验中找到了全局最优序

列，而使用简单机器学习模型的方案找到最优序列的概率提升至8.2%。与此同时，为找到相近功能强度的序列变体，计算方法所需的测试数据量相比传统方案可减少约30%。

在另一项计算模拟工作中[107]，研究人员通过使用四个不同蛋白语言模型对突变序列进行无监督打分，采用基于打分排序和序列聚类采样的文库构建方法，在总共四轮、每轮仅对12个突变体进行检测的进化筛选过程后，成功找到了诸多实际功能强度位居前5%的序列突变体。Yang等的工作采用主动学习的策略[108]，在序列文库构建时对模型置信度较低的区域进行额外采样，这一方法实现了序列优化效果近8倍的提升。在Zhou等的工作中[109]，他们提出基于检索增强和排序学习的小样本模型训练框架FSFP，在使用少量实验数据的条件下，便可获得模型预测精度的显著提高。他们将该方法运用于Phi29 DNA聚合酶的热稳定性研究，实现了高打分序列25%的阳性率提升。此外，在Wittmann等的工作中[110]，研究人员进一步对不同模型、不同采样方法在前述GB1数据内得到的筛选结果进行了更全面的比较。他们发现在序列文库内减少零功能或弱功能突变序列能够显著提升实验的整体表现，优化后的进化筛选流程能够在2000次模拟中以99.70%的概率找到全局最优点。

以上案例展现出基于AI方法的中间文库设计对提升定向进化实验效率带来的巨大帮助。

3.3 提升功能信息提取准确性

在许多定向进化实验中，突变体的功能强弱往往被默认为是筛选后该序列对应的测量指标高低（对于大多数高通量检测手段，最终的测量指标常常是突变体的序列丰度）[29，46，50，111-112]。然而，序列丰度本身会受到筛选前文库中不同序列数量差异的影响，对于连续定向进化，特定突变体最终的拷贝数量还与其在进化过程中出现的时间早晚有关。虽然终末时间点的序列丰度能够定性反映突变体的功能情况，但这种描述方法在定量层面并不准确。也有许多研究将筛选前后特定突变序列的相对富集程度作为其功能强弱的指征[60，106，113-115]，尽管通过这一方法获得的结果通常与对突变体功能直接进行生化表征的实际测量值存在较高相关性[60，114]，但其准确性仍有很大提升空间。

Fernandez等[116]使用Potts模型对多轮筛选过程中的测序丰度数据进行拟合，通过突变体对应的能量高低对其功能强弱进行描述。该方法能够结合多轮筛选的序列丰度信息对突变体功能进行更加准确的提取，在对数据中的噪声鲁棒性更强的同时，还能对蛋白内存在的上位效应进行简单预测。Sesta等[117]进一步考虑突变产生的详细动力学过程，并从多轮采样结果整体出发对模型进行拟合，在测试数据集上实现了对突变体功能更准确的评估。此外，在Shen等的工作中[118]，通过构建机器学习模型Evoracle，他们成功从常规短读长测序结果中提取得到了精度较高的完整突变体的功能打分。

3.4 小结

AI方法的应用能够对定向进化实验流程的不同阶段进行优化。一方面，它能在有限的实验通量下实现更加高效的进化筛选，在获得功能提升更为显著的突变体的同时，兼顾筛选文库中的序列差异性，获得不同类型的改造序列；另一方面，它能在检测精度有限的条件下从实验数据中提取得到更加准确的功能信息，帮助我们更好地把握蛋白序列与功能间的对应关系，为进一步的序列改造和优化提供便利。随着相关研究的不断深入，不依赖于湿实验的虚拟定向进化正逐渐成为可能[119-120]。但由于当前模型在完全不依赖实验数据条件下对复杂突变组合的功能推理能力仍比较有限[57]，因而其在准确性上仍有很大的提升空间。此外，虚拟定向进化还面临诸多技术挑战：首先，高质量功能表征数据的标注成本较高，限制了训练数据的规模和多样性；其次，现有深度学习模型的“黑盒”特性使得预测结果缺乏生物学可解释性，难以为实验设计提供机理指导；最后，不同蛋白家族间的功能预测模型通用性不足，需要针对特定目标进行重新训练。

4 AI辅助定向进化的实际应用

机器学习辅助定性进化（machine learning-

assisted directed evolution， MLDE）能够有效结合定向进化和人工智能算法各自的优势，对目标蛋白进行更加高效的功能提升或改造［图3（a）］。这一方法已被运用于各类蛋白的优化任务中，取得了诸多重要成果。

在酶的优化方面，Wu等[105]利用两轮的定向进化对Rma-NOD酶催化一新型卡宾插入反应过程中产物对映异构体的选择性进行了提升。在使用Rma-NOD中7个突变位点的124个突变体对模型进行训练后，经过两轮的进化筛选，他们分别得到了对*S*型对映体选择性达93%以及对*R*型对映体选择性达79%的突变序列。在Jiang等的研究中[121]，通过使用PLM与随机森林预测层构成的机器学习模型EVOLVEpro，他们在六轮进化后得到了比野生型T7 RNAP翻译活性高出约57倍的T7 RNAP变体（epT7）［图3（b）］；在九轮进化后得到了活性上升至筛选前2.6倍的Bxb1整合酶变体。Yang等[108]采用主动学习的策略使模型在文库生成过程中主动选取小批次之前未曾涉及的突变类型，在对五个活性位点的两轮进化实验后，他们将一非天然环丙烷化反应的产率从原先的12%提升至93%。Landwehr等[122]借助无细胞筛选平台收集了酰胺合成酶McbA的1216个单突变体的功能信息，并对其增强岭回归模型进行训练。在四轮的进化筛选后，他们成功得到了催化活性提升42倍的McbA突变体。Jiang等开发了温度感知蛋白质语言模型PRIME[123]，并将其运用于多种酶的AI辅助定向进化，充分展现了蛋白质语言模型在相关优化任务中的普适性。该模型通过较少的实验轮数和较低的实验通量，即可实现多个突变位点的有效组合，进化效率显著优于传统方法。

针对抗体的优化方面，Hie等[124]在不考虑任何关于抗体或抗原结构信息的情况下，通过使用ESM1b和ESM-1v中多个模型的组合[75-76]，对甲型流感病毒、埃博拉病毒、SARS-CoV-2病毒对应的7种不同抗体进行了序列优化。经过两轮的进化筛选，他们通过对20个或更少的突变体进行测试获得的数据，得到了亲和力提升7～160倍不等的新抗体序列。其中部分突变体在热稳定性和中和活性上也表现出明显的提升。在Shanker等的研究中[125]，他们借助ESM-IF1模型在突变筛选过程中综合考虑抗体的结构信息[95]。在尽可能维持原先蛋白结构的情况下，研究人员在序列中优先引入更加符合天然蛋白序列特点的突变类型。在对SARS-CoV-2的两种不同抗体LY-CoV1404和SA58进行两轮进化筛选后，优化抗体能够对原先逃逸的病毒突变株系BQ.1.1和XBB.1.5分别产生25倍的中和活性提升［图3（c）］和37倍的亲和力提升［图3（d）］。尽管只对30个左右的突变抗体序列进行了筛选，但其中绝大部分的序列都展现出一定的活性提升。在考虑抗原抗体复合物整体的结构信息后，模型对突变体功能的预测准确性还可进一步上升。

对于其他类型蛋白的优化，Bedbrook等[126]使用163个视紫红质（ChR）突变体的已知功能信息首先对一高斯过程分类模型进行训练，用于对突变体功能有无进行判定。随后，他们选用这些数据中有详细功能表征的突变体分别对三个高斯过程回归模型进行训练，用于对突变体的光电流强度、波长敏感性和失活速率进行预测。利用这些模型，他们对120 000条突变序列进行活性和响应性质的两轮筛选，对得到的28个高打分序列进行试验测试，并最终得到了多个对不同波长光刺激具有高敏感性，且在光电流强度上有明显提升的ChR突变体。在Ma等的工作中[59]，通过使用定向进化不同阶段富集得到的82个突变体序列对模型EvoAI进行训练，研究人员成功从复杂的高阶突变序列文库中筛选得到抑制效果提升10～38倍的AmeR转录因子突变体［图3（e）］；与此同时，若仅使用低阶突变的深度突变扫描数据（DMS）对模型进行训练，得到的筛选序列在抑制效果上则明显减弱。此外，前述的EVOLVEpro模型同样被用于先导编辑器等多个其他蛋白的功能优化任务中[121]，还有其他一些工作使用MLDE对荧光蛋白、5-羟色胺传感器等蛋白进行改造[127-128]，也取得了不错的效果。

AI辅助定向进化可针对酶、抗体、转录因子、生物传感器等诸多蛋白类型灵活部署。通过将机器学习与定向进化方法各自的优势相融合，这一技术路线能够大幅提升传统实验的筛选效率，赋能各类目标蛋白的改造与优化。在进行AI模型选择时，同类型模型中参数量较大的蛋白语言模型

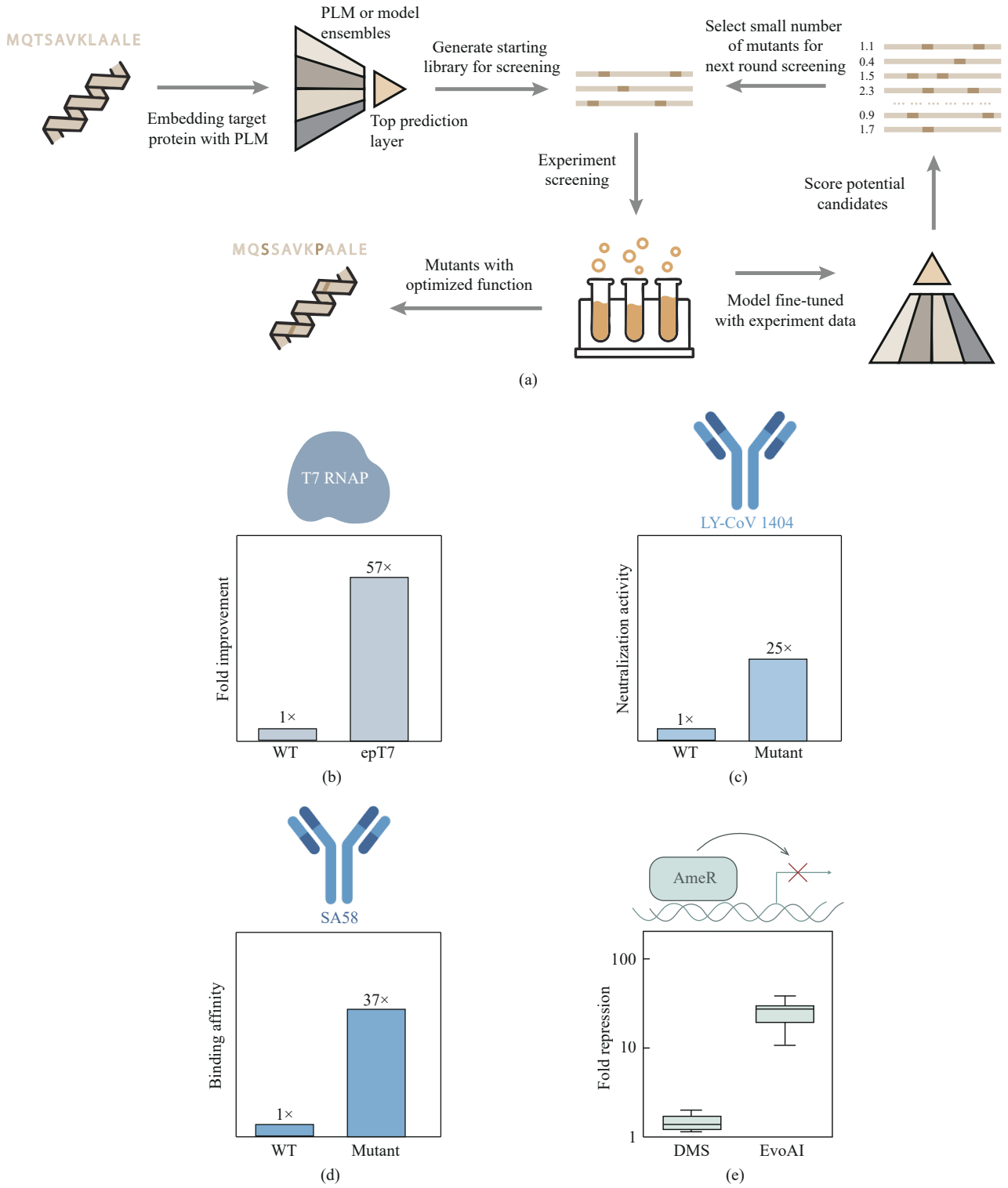


图3 机器学习辅助定向进化（MLDE）的一般流程和案例展示

（a）MLDE的一般流程；（b）～（e）MLDE在酶[121]、抗体[125]和转录因子[59]优化改造中的应用。（e）中分别展示了根据DMS和EvoAI方法得到的打分前十的AmeR突变体的抑制倍数提升情况

Fig. 3 Workflow for machine learning-assisted directed evolution (MLDE) and case studies

(a)MLDE workflow; (b)–(e) Applications of MLDE to the optimization of enzymes[121], antibodies[125], and transcription factors[59]. Panel (e) corresponds to fold-increase in inhibitory activity for the top-10 AmeR mutants, as ranked by models trained on DMS data or by EvoAI.

在性能上通常会有更良好的表现。需要注意的是，随着参数量的上升，模型训练和预测所需消耗的计算资源和时间也会大幅增加[81, 129]。与此同时，在模型规模达到一定量级后，参数数量的进一步提升对模型预测精度的改善往往有限。以ESM-2为例[81]，在模型参数从8M增加到650M的过程中，其在突变体功能评估、结构预测等下游任务中的表现均获得了显著提升（如在CASP14中提升0.14，在CAMEO中提升0.22）；但当参数数量进一步增加至15B时，模型获得的性能提升相比之下变得较为微弱（在CASP14中提升0.04，在CAMEO中提升0.02）。类似现象在ProtT5、ProGen2等模型的部分评估中也有出现。因而在选用蛋白语言模型时，在优先使用效率更高的新架构模型的同时，可根据计算资源状况对模型参数量进行灵活选取。另外，模型的选择还要与研究者本身关注的问题和所能采集的数据规模相结合。如果希望通过少量实验测量较快获得功能提升的优化序列，EVOLVEpro[122]等小样本模型便提供了不错的参考；如果希望通过相对较多的数据对目标蛋白的序列-功能关系进行更加精细的预测，寻找功能提升倍数更强的突变类型，则基于ECNet[130]、SESNet[131]等模型进行训练和突变体筛选或将更加有效。

5 总结与展望

定向进化目前已成为蛋白工程的常用工具，通过对自然进化过程的实验室模拟，它能在无需过多先验知识的情况下对蛋白功能进行高效的提升。虽然已历经多年的发展，新技术的涌现仍不断完善着这一经典的实验技术。其中，高通量检测和连续定向进化系统的进步正持续为这一传统实验技术注入新的活力。针对定向进化技术存在的固有缺陷，当前飞速发展的人工智能方法能够有效应对这些不足。以机器学习为代表的计算工具一方面能够拓展定向进化实验中有限的序列采样空间；另一方面又能从起始序列设计、中间文库生成、功能信息提取等多个方面对定向进化的实验流程进行改进。定向进化与人工智能方法的结合已在各个类型蛋白的改造与优化任务中发挥重要作用，产生了许多兼具基础研究和应用生产价值的重要成果。

日益完善的连续定向进化技术为蛋白功能改造提供了更多选择，通过灵活控制筛选条件，研究者能够在持续的实验进化中更方便地获得先前蛋白并不具备的新功能。之前的研究已通过该实验方案在蛋白酶特异性底物替换[48]、抗体新靶标结合能力提升[56]、碱基编辑器的编辑谱拓宽[50]等诸多情境中实现了蛋白功能的改造。在此基础上，AI方法与实验方法的深度融合，可以通过高潜力的初始序列选取、自动化的筛选压力控制、更精准的突变数据分析等环节，将0→1的功能改造和1→100的功能提升在同一框架下完成，更快更高效地实现目标功能的提升与强化。

定向进化结合人工智能方法的蛋白优化流程目前仍处在快速的发展迭代阶段，在实验和模型层面均需要更多的探索与尝试。随着检测技术的进步，高通量的功能表征数据正快速积累，但相应的分析方法仍相对匮乏。目前大部分MLDE方法均从“小样本”的前提出发，尽可能通过少量的数据检测达到更好的蛋白优化效果。尽管这些方法能够从几百甚至几十个的少量数据点中找到序列优化较为高效的路径，但却无法对高通量数据中蕴含的复杂信息进行有效提取。对于蛋白中存在的部分高阶突变组合和复杂上位效应，仅通过当前的无监督模型或少量的数据点训练仍无法对其性质进行准确的评估[60, 66, 132]。同时，不断普及的连续定向进化技术在数据的分析处理上仍缺少有效工具。已有研究表明，当前蛋白语言模型的泛化能力尚不足以对实验中的复杂突变体功能进行无监督预测[57]。也有少部分工作尝试对连续定向进化中的信息进行压缩提取[59]，但其在处理方法上仍采用类似先前小样本中的训练思路，并未能对实验产生的众多功能序列进行有效利用。大量相关数据中蕴藏的宝贵信息仍待进一步挖掘。在实验技术方面，当前定向进化的功能筛选仍时常依赖于针对特定案例的报告系统设计。对于一些功能特殊的蛋白，仍无法有效地对其功能进行高通量检测，探索构建普适性的报告系统对于实验方法在更多场景中的应用无疑具有重要意义。

尽管我们反复强调人工智能方法对定向进化

短板的弥补，但与此同时，这种帮助的关系其实也是相互的。正如蛋白语言模型诞生之初同样依赖于大量的蛋白序列数据一样，通过定向进化不断积累的大量功能表征结果为我们理解序列-功能的准确映射关系提供了数据上的支持。随着近年来自动化实验技术在定向进化领域的引入[133]以及高通量连续定向进化平台的不断完善和普及，我们相信，公共数据库中不同蛋白突变体的功能表征数据会在未来以更快速度增长。这些数据的积累或将成为打破当前功能预测模型局限性的关键，帮助模型实现无监督预测能力的持续提升。与此同时，在数据积累之外，我们同样需要在算法和方法学层面寻求突破：①发展基于物理化学原理的半监督学习方法，减少对大规模标注数据的依赖[134-135]；②构建具有更强可解释性的机器学习模型，提供突变位点重要性评分和预测依据[136]；③建立跨蛋白家族的迁移学习框架，提高模型在不同蛋白系统中的泛化能力[137]。通过数据积累与方法学创新的协同发展，真正意义上高精度的“虚拟定向进化”将成为可能。AI与定向进化方法的有机融合，将开启蛋白功能改造与优化的新时代。

参 考 文 献

[1] LERNER S A, WU T T, LIN E C. Evolution of a catabolic pathway in bacteria[J]. Science, 1964, 146(3649): 1313-1315.

[2] HALL B G. Experimental evolution of a new enzymatic function. Kinetic analysis of the ancestral (*ebg*^{o}) and evolved (*ebg*$^{+}$) enzymes[J]. Journal of Molecular Biology, 1976, 107 (1): 71-84.

[3] LENUG D W, CHEN E, GOEDDEL D V. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction[J]. Technique JMCMB, 1989, 1: 11-15.

[4] CHEN K Q, ARNOLD F H. Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media[J]. Nature Biotechnology, 1991, 9(11): 1073-1077.

[5] KLENK C, SCRIVENS M, NIEDERER A, et al. A Vaccinia-based system for directed evolution of GPCRs in mammalian cells[J]. Nature Communications, 2023, 14: 1770.

[6] HUFFMAN M A, FRYSZKOWSKA A, ALVIZO O, et al. Design of an *in vitro* biocatalytic cascade for the manufacture of islatravir[J]. Science, 2019, 366(6470): 1255-1259.

[7] TOURNIER V, TOPHAM C M, GILLES A, et al. An engineered PET depolymerase to break down and recycle plastic bottles[J]. Nature, 2020, 580(7802): 216-219.

[8] SARAI N S, FULTON T J, O’MEARA R L, et al. Directed evolution of enzymatic silicon-carbon bond cleavage in siloxanes[J]. Science, 2024, 383(6681): 438-443.

[9] RAPPAZZO C G, TSE L V, KAKU C I, et al. Broad and potent activity against SARS-like viruses by an engineered human monoclonal antibody[J]. Science, 2021, 371(6531): 823-829.

[10] BANACH B B, PLETNEV S, OLIA A S, et al. Antibody-directed evolution reveals a mechanism for enhanced neutralization at the HIV-1 fusion peptide site[J]. Nature Communications, 2023, 14: 7593.

[11] TABEBORDBAR M, LAGERBORG K A, STANTON A, et al. Directed evolution of a family of AAV capsid variants enabling potent muscle-directed gene delivery across species [J]. Cell, 2021, 184(19): 4919-4938.e22.

[12] LIN R, ZHOU Y T, YAN T, et al. Directed evolution of adeno-associated virus for efficient gene delivery to microglia[J]. Nature Methods, 2022, 19(8): 976-985.

[13] CLARKE J, FERSHT A R. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation[J]. Biochemistry, 1993, 32(16): 4322-4329.

[14] REA V, KOLKMAN A J, VOTTERO E, et al. Active site substitution A82W improves the regioselectivity of steroid hydroxylation by cytochrome P450 BM3 mutants as rationalized by spin relaxation nuclear magnetic resonance studies[J]. Biochemistry, 2012, 51(3): 750-760.

[15] MAITI A, BUFFALO C Z, SAURABH S, et al. Structural and photophysical characterization of the small ultra-red fluorescent protein[J]. Nature Communications, 2023, 14: 4155.

[16] PACKER M S, LIU D R. Methods for the directed evolution of proteins[J]. Nature Reviews Genetics, 2015, 16(7): 379-394.

[17] WANG Y J, XUE P, CAO M F, et al. Directed evolution: methodologies and applications[J]. Chemical Reviews, 2021, 121(20): 12384-12444.

[18] ZACCOLO M, WILLIAMS D M, BROWN D M, et al. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues[J]. Journal of Molecular Biology, 1996, 255(4): 589-603.

[19] VANHERCKE T, AMPE C, TIRRY L, et al. Reducing mutational bias in random protein libraries[J]. Analytical Biochemistry, 2005, 339(1): 9-14.

[20] DENNIG A, SHIVANGE A V, MARIENHAGEN J, et al. OmniChange: the sequence independent method for simultaneous site-saturation of five codons[J]. PLoS One,

2011, 6(10): e26222.

[21] PÜLLMANN P, ULPINNIS C, MARILLONNET S, et al. Golden Mutagenesis: an efficient multi-site-saturation mutagenesis approach by Golden Gate cloning with automated primer design[J]. Scientific Reports, 2019, 9: 10932.

[22] ZHAO H M, GIVER L, SHAO Z X, et al. Molecular evolution by staggered extension process (StEP) *in vitro* recombination [J]. Nature Biotechnology, 1998, 16(3): 258-261.

[23] COCO W M, LEVINSON W E, CRIST M J, et al. DNA shuffling method for generating highly recombined genes and evolved enzymes[J]. Nature Biotechnology, 2001, 19(4): 354-359.

[24] SIEBER V, MARTINEZ C A, ARNOLD F H. Libraries of hybrid proteins from distantly related sequences[J]. Nature Biotechnology, 2001, 19(5): 456-460.

[25] BITTKER J A, LE B V, LIU J M, et al. Directed evolution of protein enzymes using nonhomologous random recombination [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(18): 7011-7016.

[26] GREENER A, CALLAHAN M, JERPSETH B. An efficient random mutagenesis technique using an *E. coli* mutator strain [J]. Molecular Biotechnology, 1997, 7(2): 189-195.

[27] BADRAN A H, LIU D R. Development of potent *in vivo* mutagenesis plasmids with broad mutational spectra[J]. Nature Communications, 2015, 6: 8425.

[28] MOORE C L, PAPA L J 3RD, SHOULDERS M D. A processive protein *Chimera* introduces mutations across defined DNA regions *in vivo*[J]. Journal of the American Chemical Society, 2018, 140(37): 11560-11564.

[29] RAVIKUMAR A, ARZUMANYAN G A, OBADI M K A, et al. Scalable, continuous evolution of genes at mutation rates above genomic error thresholds[J]. Cell, 2018, 175(7): 1946-1957.e13.

[30] RAVIKUMAR A, ARRIETA A, LIU C C. An orthogonal DNA replication system in yeast[J]. Nature Chemical Biology, 2014, 10(3): 175-177.

[31] HALPERIN S O, TOU C J, WONG E B, et al. CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window[J]. Nature, 2018, 560(7717): 248-252.

[32] TOU C J, SCHAFFER D V, DUEBER J E. Targeted diversification in the *S. cerevisiae* genome with CRISPR-guided DNA polymerase I[J]. ACS Synthetic Biology, 2020, 9 (7): 1911-1916.

[33] ÁLVAREZ B, MENCÍA M, DE LORENZO V, et al. *In vivo* diversification of target genomic sites using processive base deaminase fusions blocked by dCas9[J]. Nature Communications, 2020, 11: 6436.

[34] YI X, KHEY J, KAZLAUSKAS R J, et al. Plasmid hypermutation using a targeted artificial DNA replisome[J]. Science Advances, 2021, 7(29): eabg8712.

[35] CROOK N, ABATEMARCO J, SUN J, et al. *In vivo* continuous evolution of genes and pathways in yeast[J]. Nature Communications, 2016, 7: 13051.

[36] CARR P A, WANG H H, STERLING B, et al. Enhanced multiplex genome engineering through co-operative oligonucleotide co-selection[J]. Nucleic Acids Research, 2012, 40(17): e132.

[37] LEWIS J C, ARNOLD F H. Catalysts on demand: selective oxidations by laboratory-evolved cytochrome P450 BM3[J]. Chimia, 2009, 63(6): 309.

[38] COELHO P S, BRUSTAD E M, KANNAN A, et al. Olefin cyclopropanation *via* carbene transfer catalyzed by engineered cytochrome P450 enzymes[J]. Science, 2013, 339(6117): 307-310.

[39] CHEN H Q, LIU S, PADULA S, et al. Efficient, continuous mutagenesis in human cells using a pseudo-random DNA editor [J]. Nature Biotechnology, 2020, 38(2): 165-168.

[40] STIFFLER M A, HEKSTRA D R, RANGANATHAN R. Evolvability as a function of purifying selection in TEM-1 β-lactamase[J]. Cell, 2015, 160(5): 882-892.

[41] STARR T N, GREANEY A J, HILTON S K, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding[J]. Cell, 2020, 182(5): 1295-1310.e20.

[42] MA L, LIN Y H. Orthogonal RNA replication enables directed evolution and Darwinian adaptation in mammalian cells[J]. Nature Chemical Biology, 2025, 21(3): 451-463.

[43] COLLINS C H, LEADBETTER J R, ARNOLD F H. Dual selection enhances the signaling specificity of a variant of the quorum-sensing transcriptional activator LuxR[J]. Nature Biotechnology, 2006, 24(6): 708-712.

[44] MORRISON M S, PODRACKY C J, LIU D R. The developing toolkit of continuous directed evolution[J]. Nature Chemical Biology, 2020, 16(6): 610-619.

[45] MOLINA R S, RIX G, MENGISTE A A, et al. *In vivo* hypermutation and continuous evolution[J]. Nature Reviews Methods Primers, 2022, 2: 36.

[46] ESVELT K M, CARLSON J C, LIU D R. A system for the continuous directed evolution of biomolecules[J]. Nature, 2011, 472(7344): 499-503.

[47] PACKER M S, REES H A, LIU D R. Phage-assisted continuous evolution of proteases with altered substrate specificity[J]. Nature Communications, 2017, 8: 956.

[48] BLUM T R, LIU H, PACKER M S, et al. Phage-assisted evolution of botulinum neurotoxin proteases with reprogrammed specificity[J]. Science, 2021, 371(6531): 803-810.

[49] MILLER S M, WANG T N, RANDOLPH P B, et al. Continuous evolution of *Sp*Cas9 variants compatible with non-G PAMs[J]. Nature Biotechnology, 2020, 38(4): 471-481.

[50] RICHTER M F, ZHAO K T, ETON E, et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity[J]. Nature Biotechnology, 2020, 38(7): 883-891.

[51] MERCER J A M, DECARLO S J, ROY BURMAN S S, et al. Continuous evolution of compact protein degradation tags regulated by selective molecular glues[J]. Science, 2024, 383 (6688): eadk4422.

[52] ENGLISH J G, OLSEN R H J, LANSU K, et al. VEGAS as a platform for facile directed evolution in mammalian cells[J]. Cell, 2019, 178(3): 748-761.e17.

[53] DENES C E, COLE A J, TRAN M T N, et al. The VEGAS platform is unsuitable for mammalian directed evolution[J]. ACS Synthetic Biology, 2022, 11(10): 3544-3549.

[54] KIMMAN T G, SMIT E, KLEIN M R. Evidence-based biosafety: a review of the principles and effectiveness of microbiological containment measures[J]. Clinical Microbiology Reviews, 2008, 21(3): 403-425.

[55] ARTIKA I M, MA'ROEF C N. Laboratory biosafety for handling emerging viruses[J]. Asian Pacific Journal of Tropical Biomedicine, 2017, 7(5): 483-491.

[56] WELLNER A, MCMAHON C, GILMAN M S A, et al. Rapid generation of potent antibodies by autonomous hypermutation in yeast[J]. Nature Chemical Biology, 2021, 17(10): 1057-1064.

[57] RIX G, WILLIAMS R L, HU V J, et al. Continuous evolution of user-defined genes at 1 million times the genomic mutation rate[J]. Science, 2024, 386(6722): eadm9073.

[58] TIAN R Z, REHM F B H, CZERNECKI D, et al. Establishing a synthetic orthogonal replication system enables accelerated evolution in *E. coli*[J]. Science, 2024, 383(6681): 421-426.

[59] MA Z Y, LI W J, SHEN Y H, et al. EvoAI enables extreme compression and reconstruction of the protein sequence space [J]. Nature Methods, 2025, 22(1): 102-112.

[60] JOHNSTON K E, ALMHJELL P J, WATKINS-DULANEY E J, et al. A combinatorially complete epistatic fitness landscape in an enzyme active site[J]. Proceedings of the National Academy of Sciences of the United States of America, 2024, 121(32): e2400439121.

[61] WEINREICH D M, DELANEY N F, DEPRISTO M A, et al. Darwinian evolution can follow only very few mutational paths to fitter proteins[J]. Science, 2006, 312(5770): 111-114.

[62] PODGORNAIA A I, LAUB M T. Pervasive degeneracy and epistasis in a protein-protein interface[J]. Science, 2015, 347 (6222): 673-677.

[63] FOX R, ROY A, GOVINDARAJAN S, et al. Optimizing the search algorithm for protein engineering by directed evolution [J]. Protein Engineering Design and Selection, 2003, 16(8): 589-597.

[64] FOX R J, DAVIS S C, MUNDORFF E C, et al. Improving catalytic function by ProSAR-driven enzyme evolution[J]. Nature Biotechnology, 2007, 25(3): 338-344.

[65] ROMERO P A, KRAUSE A, ARNOLD F H. Navigating the protein fitness landscape with Gaussian processes[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(3): E193-E201.

[66] OTWINOWSKI J, PLOTKIN J B. Inferring fitness landscapes by regression produces biased estimates of epistasis[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(22): E2301-E2309.

[67] OFER D, BRANDES N, LINIAL M. The language of proteins: NLP, machine learning & protein sequences[J]. Computational and Structural Biotechnology Journal, 2021, 19: 1750-1758.

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

[69] ASGARI E, MOFRAD M R K. Continuous distributed representation of biological sequences for deep proteomics and genomics[J]. PLoS One, 2015, 10(11): e0141287.

[70] HEINZINGER M, ELNAGGAR A, WANG Y, et al. Modeling aspects of the language of life through transfer-learning protein sequences[J]. BMC Bioinformatics, 2019, 20(1): 723.

[71] ALLEY E C, KHIMULYA G, BISWAS S, et al. Unified rational protein engineering with sequence-based deep representation learning[J]. Nature Methods, 2019, 16(12): 1315-1322.

[72] RAO R, BHATTACHARYA N, THOMAS N, et al. Evaluating protein transfer learning with TAPE[C]//Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019, 32: 9689-9701[2025-06-03]. https://proceedings. neurips. cc/paper_files/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html.

[73] MADANI A, KRAUSE B, GREENE E R, et al. Large language models generate functional protein sequences across diverse families[J]. Nature Biotechnology, 2023, 41(8): 1099-1106.

[74] MADANI A, MCCANN B, NAIK N, et al. ProGen: language modeling for protein generation[EB/OL]. arXiv, 2020: 2004.03497. (2020-03-08) [2025-06-03]. https://arxiv. org/abs/2004.03497v1.

[75] RIVES A, MEIER J, SERCU T, et al. 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.

[76] MEIER J, RAO R, VERKUIL R, et al. Language models enable zero-shot prediction of the effects of mutations on protein function[C/OL]//Advances in Neural Information

Processing Systems 34 (NeurIPS 2021), 2021, 34: 29287-29303[2025-06-03]. https://proceedings.neurips.cc/paper_files/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

[77] RAO R M, LIU J, VERKUIL R, et al. MSA transformer[C]// Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, 139: 8844-8856[2025-06-03]. https://proceedings.mlr.press/v139/rao21a.html.

[78] BRANDES N, OFER D, PELEG Y, et al. ProteinBERT: a universal deep-learning model of protein sequence and function [J]. Bioinformatics, 2022, 38(8): 2102-2110.

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

[80] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: toward understanding the language of life through self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 7112-7127.

[81] LIN Z M, AKIN H, RAO R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.

[82] NIJKAMP E, RUFFOLO J A, WEINSTEIN E N, et al. ProGen2: exploring the boundaries of protein language models [J]. Cell Systems, 2023, 14(11): 968-978.e3.

[83] ELNAGGAR A, ESSAM H, SALAH-ELDIN W, et al. Ankh: optimized protein language model unlocks general-purpose modelling[EB/OL]. arXiv, 2023: 2301.06568. (2023-01-16) [2025-06-03]. https://arxiv.org/abs/2301.06568v1.

[84] TRUONG T F JR, BEPLER T. PoET: a generative model of protein families as sequences-of-sequences[C/OL]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023: 2306.06156[2025-06-03]. https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4366126eba252699b280e8f93c0ab2f-Abstract-Conference.html.

[85] HAYES T, RAO R, AKIN H, et al. Simulating 500 million years of evolution with a language model[J]. Science, 2025, 387(6736): 850-858.

[86] YANG K K, FUSI N, LU A X. Convolutions are competitive with transformers for protein sequence pretraining[J]. Cell Systems, 2024, 15(3): 286-294.e2.

[87] LV L, LIN Z Y, LI H, et al. ProLLaMA: a protein large language model for multi-task protein language processing[J]. IEEE Transactions on Artificial Intelligence, 2025, PP(99): 1-12.

[88] CHEN B, CHENG X Y, LI P, et al. xTrimoPGLM: unified 100B-scale pre-trained transformer for deciphering the language of protein[EB/OL]. arXiv, 2024: 2401.06199. (2024-01-11)[2025-06-03]. https://arxiv.org/abs/2401.06199v2..

[89] SAYEED M A, TEKIN E, NADEEM M, et al. Prot42: a novel family of protein language models for target-aware protein binder generation[EB/OL]. arXiv, 2025: 2504.04453. (2025-04-06)[2025-06-03]. https://arxiv.org/abs/2504.04453v2.

[90] KELLY T, XIA S, LU J Y, et al. Unified deep learning of molecular and protein language representations with T5ProtChem[J]. Journal of Chemical Information and Modeling, 2025, 65(8): 3990-3998.

[91] WANG Y H, WANG Z C, SADEH G, et al. LC-PLM: long-context protein language modeling using bidirectional mamba with shared projection layers[EB/OL]. arXiv, 2025:: 2411.08909. (2024-10-29) [2025-06-03]. https://doi. org/10.48550/arXiv.2411.08909.

[92] BISWAS S, KHIMULYA G, ALLEY E C, et al. Low-N protein engineering with data-efficient deep learning[J]. Nature Methods, 2021, 18(4): 389-396.

[93] SUZEK B E, WANG Y Q, HUANG H Z, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches[J]. Bioinformatics, 2015, 31(6): 926-932.

[94] RADFORD A, JOZEFOWICZ R, SUTSKEVER I. Learning to generate reviews and discovering sentiment[EB/OL]. arXiv, 2017: 1704.01444. (2017-04-05) [2025-06-03]. https://doi. org/10.48550/arXiv.1704.01444.

[95] HSU C, VERKUIL R, LIU J, et al. Learning inverse folding from millions of predicted structures[C/OL]//Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022, 162: 8946-8970[2025-06-03]. https://proceedings.mlr.press/v162/hsu22a.html.

[96] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[97] NOTIN P, KOLLASCH A, RITTER D, et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design [C/OL]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023: 64331-64379[2025-06-03]. https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

[98] BADRAN A H, GUZOV V M, HUAI Q, et al. Continuous evolution of *Bacillus thuringiensis* toxins overcomes insect resistance[J]. Nature, 2016, 533(7601): 58-63.

[99] GLÖGL M, KRISHNAKUMAR A, RAGOTTE R J, et al. Target-conditioned diffusion generates potent TNFR superfamily antagonists and agonists[J]. Science, 2024, 386 (6726): 1154-1161.

[100] VÁZQUEZ TORRES S, BENARD VALLE M, MACKESSY S P, et al. *De novo* designed proteins neutralize lethal snake venom toxins[J]. Nature, 2025, 639(8053): 225-231.

[101] YEH A H W, NORN C, KIPNIS Y, et al. *De novo* design of luciferases using deep learning[J]. Nature, 2023, 614(7949):

774-780.

[102] KIPNIS Y, CHAIB A O, VOROBIEVA A A, et al. Design and optimization of enzymatic activity in a *de novo* β-barrel scaffold[J]. Protein Science, 2022, 31(11): e4405.

[103] DING K, CHIN M, ZHAO Y L, et al. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering[J]. Nature Communications, 2024, 15: 6392.

[104] FRAM B, SU Y, TRUEBRIDGE I, et al. Simultaneous enhancement of multiple functional properties using evolution-informed protein design[J]. Nature Communications, 2024, 15: 5141.

[105] WU Z, JENNIFER KAN S B, LEWIS R D, et al. 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.

[106] WU N C, DAI L, OLSON C A, et al. Adaptation in protein fitness landscapes is facilitated by indirect paths[J]. eLife, 2016, 5: e16965.

[107] CHU H Y, FONG J H C, THEAN D G L, et al. Accurate top protein variant discovery *via* low-N pick-and-validate machine learning[J]. Cell Systems, 2024, 15(2): 193-203.e6.

[108] YANG J, LAL R G, BOWDEN J C, et al. Active learning-assisted directed evolution[J]. Nature Communications, 2025, 16: 714.

[109] ZHOU Z Y, ZHANG L, YU Y X, et al. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning[J]. Nature Communications, 2024, 15: 5566.

[110] WITTMANN B J, YUE Y S, ARNOLD F H. Informed training set design enables efficient machine learning-assisted directed protein evolution[J]. Cell Systems, 2021, 12(11): 1026-1045.e7.

[111] THURONYI B W, KOBLAN L W, LEVY J M, et al. Continuous evolution of base editors with expanded target compatibility and improved activity[J]. Nature Biotechnology, 2019, 37(9): 1070-1079.

[112] HU J H, MILLER S M, GEURTS M H, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity[J]. Nature, 2018, 556(7699): 57-63.

[113] JUDGE A, SANKARAN B, HU L Y, et al. Network of epistatic interactions in an enzyme active site revealed by large-scale deep mutational scanning[J]. Proceedings of the National Academy of Sciences of the United States of America, 2024, 121(12): e2313513121.

[114] OLSON C A, WU N C, SUN R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain[J]. Current Biology, 2014, 24(22): 2643-2651.

[115] LIU G, ZENG H Y, MUELLER J, et al. Antibody complementarity determining region design using high-capacity machine learning[J]. Bioinformatics, 2020, 36(7): 2126-2133.

[116] FERNANDEZ-DE-COSSIO-DIAZ J, UGUZZONI G, PAGNANI A. Unsupervised inference of protein fitness landscape from deep mutational scan[J]. Molecular Biology and Evolution, 2021, 38(1): 318-328.

[117] SESTA L, UGUZZONI G, FERNANDEZ-DE-COSSIO-DIAZ J, et al. AMaLa: analysis of directed evolution experiments *via* annealed mutational approximated landscape[J]. International Journal of Molecular Sciences, 2021, 22(20): 10908.

[118] SHEN M W, ZHAO K T, LIU D R. Reconstruction of evolving gene variants and fitness from short sequencing reads[J]. Nature Chemical Biology, 2021, 17(11): 1188-1198.

[119] ALVAREZ S, NARTEY C M, MERCADO N, et al. *In vivo* functional phenotypes from a computational epistatic model of evolution[J]. Proceedings of the National Academy of Sciences of the United States of America, 2024, 121(6): e2308895121.

[120] DI BARI L, BISARDI M, COTOGNO S, et al. Emergent time scales of epistasis in protein evolution[J]. Proceedings of the National Academy of Sciences of the United States of America, 2024, 121(40): e2406807121.

[121] JIANG K Y, YAN Z Q, DI BERNARDO M, et al. Rapid *in silico* directed evolution by a protein language model with EVOLVEpro[J]. Science, 2025, 387(6732): eadr6006.

[122] LANDWEHR G M, BOGART J W, MAGALHAES C, et al. Accelerated enzyme engineering by machine-learning guided cell-free expression[J]. Nature Communications, 2025, 16: 865.

[123] JIANG F, LI M C, DONG J J, et al. A general temperature-guided language model to design proteins of enhanced stability and activity[J]. Science Advances, 2024, 10(48): eadr2641.

[124] HIE B L, SHANKER V R, XU D, et al. Efficient evolution of human antibodies from general protein language models[J]. Nature Biotechnology, 2024, 42(2): 275-283.

[125] SHANKER V R, BRUUN T U J, HIE B L, et al. Unsupervised evolution of protein and antibody complexes with a structure-informed language model[J]. Science, 2024, 385(6704): 46-53.

[126] BEDBROOK C N, YANG K K, ROBINSON J E, et al. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics[J]. Nature Methods, 2019, 16(11): 1176-1184.

[127] UNGER E K, KELLER J P, ALTERMATT M, et al. Directed evolution of a selective and sensitive serotonin sensor *via* machine learning[J]. Cell, 2020, 183(7): 1986-2002.e26.

[128] SAITO Y, OIKAWA M, NAKAZAWA H, et al. Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins[J]. ACS Synthetic Biology, 2018, 7(9): 2014-2022.

[129] CHENG X Y, CHEN B, LI P, et al. Training compute-optimal protein language models[C/OL]//Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024, 37:

69386-69418[2025-06-03]. https://proceedings. neurips. cc/paper_files/paper/2024/hash/8066ae1446b2bbccb5159587cc3b3bcc-Abstract-Conference.

[130] LUO Y N, JIANG G D, YU T H, et al. ECNet is an evolutionary context-integrated deep learning framework for protein engineering[J]. Nature Communications, 2021, 12: 5743.

[131] LI M C, KANG L Q, XIONG Y, et al. SESNet: sequence-structure feature-integrated deep learning method for data-efficient protein engineering[J]. Journal of Cheminformatics, 2023, 15(1): 12.

[132] DIECKHAUS H, KUHLMAN B. Protein stability models fail to capture epistatic interactions of double point mutations[EB/OL]. bioRxiv, 2024: 2024.08.20.608844. (2024-08-20) [2025-06-03]. https://biorxiv.org/lookup/doi/10.1101/2024.08.20.608844.

[133] YU T H, BOOB A G, SINGH N, et al. *In vitro* continuous protein evolution empowered by machine learning and automation[J]. Cell Systems, 2023, 14(8): 633-644.

[134] GELMAN S, JOHNSON B, FRESCHLIN C, et al. Biophysics-based protein language models for protein engineering[EB/OL]. bioRxiv, 2024: 2024.03.15.585128. (2024-03-15) [2025-06-03]. https://biorxiv.org/lookup/doi/10.1101/2024.03.15.585128.

[135] OLIVARES-GIL A, BARBERO-APARICIO J A, RODRÍGUEZ J J, et al. Semi-supervised prediction of protein fitness for data-driven protein engineering[J]. Journal of Cheminformatics, 2025, 17(1): 88.

[136] VIG J, MADANI A, VARSHNEY L R, et al. BERTology meets biology: interpreting attention in protein language models[EB/OL]. arXiv, 2020: 2006.15222. (2020-06-26)[2025-06-03]. https://arxiv.org/abs/2006.15222v3.

[137] CHEN L, ZHANG Z H, LI Z H, et al. Learning protein fitness landscapes with deep mutational scanning data from multiple sources[J]. Cell Systems, 2023, 14(8): 706-721.e5.

通讯作者：林一瀚（1983—），男，教授，博士生导师。研究方向为定量系统生物学、合成生物学。

E-mail：yihan.lin@pku.edu.cn

第一作者：宋成治（1998—），男，博士研究生。研究方向包括系统生物学、合成生物学、生物物理。

E-mail：czsong@stu.pku.edu.cn