

人工智能赋能酶工程

徐舒萍^{1,2}, 汪金秀^{1,2}, 孟帅奇^{1,2}, 邓秋风³, 张鹤渐^{1,2}, 王秀凤^{1,2}, 郭照^{1,2}, 崔海洋^{1,2}

(¹ 南京师范大学生命科学学院微生物改造技术全国重点实验室, 江苏 南京 210023; ² 南京师范大学大规模复杂系统数值模拟教育部重点实验室, 江苏 南京 210023; ³ 南京理工大学环境与生物工程学院, 江苏 南京 210094)

摘要: 酶工程是生物催化、合成生物学等研究的重要基础。天然酶具备较好的催化性能, 但用于实际生产的天然酶普遍存在稳定性不好、亲和力差、催化活性低的问题, 需要对天然酶进行改造。传统酶改造方法主要有理性设计和非理性设计, 能够得到一定的效果的突变体, 但试验周期长, 试错成本高。随着计算机技术的发展, 人工智能开始被用于酶设计。机器学习算法较为流行, 可通过随机森林、支持向量机等算法挖掘序列-功能数据中的规律从而指导理性设计, 缩小突变体目标范围等。深度学习则具有端到端建模能力, 可以从酶序列和结构中直接提取特征, 用于结构、功能和动态行为的预测。另外, 将人工智能方法和分子动力学模拟、量子化学计算方法结合起来可以对突变效应从物理上进行分析, 为筛选提供参考。酶设计研究从过去的经验尝试逐渐向数据驱动、模型驱动的理性设计演化, 在合成生物学、生物医药、绿色生物制造等领域也越来越受到重视。

关键词: 人工智能; 酶; 生物催化; 机器学习; 深度学习; 模拟

中图分类号: Q 814

文献标志码: A

文章编号: 0438-1157 (2026) 06-3224-15

Artificial intelligence-enabled enzyme engineering

XU Shuping^{1,2}, WANG Jinxiu^{1,2}, MENG Shuaiqi^{1,2}, DENG Qiufeng³, ZHANG Hejian^{1,2}, WANG Xiufeng^{1,2}, GUO Zhao^{1,2}, CUI Haiyang^{1,2}

(¹ College of Life Science, State Key Laboratory of Microbial Technology, Nanjing Normal University, Nanjing 210023, Jiangsu, China; ² Ministry of Education Key Laboratory of NSLSCS, Nanjing Normal University, Nanjing 210023, Jiangsu, China; ³ School of Environmental and Biological Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China)

Abstract: Enzyme engineering serves as a crucial foundation for research in biocatalysis, synthetic biology, and related fields. While natural enzymes possess excellent catalytic properties, those used in actual production often suffer from poor stability, low affinity, and low catalytic activity, necessitating their modification. Traditional enzyme modification methods primarily include rational design and non-rational design, which can yield mutants with certain benefits but with lengthy testing cycles and high trial-and-error costs. With the advancement of computer technology, artificial intelligence has begun to be applied to enzyme design. Machine learning algorithms are widely

收稿日期: 2025-10-11 修回日期: 2026-01-28

通信作者: 崔海洋 (1990—), 男, 博士, 教授, hcui@nnu.edu.cn

第一作者: 徐舒萍 (1999—), 女, 博士研究生, shupingxu@nnu.edu.cn; 汪金秀 (1991—), 女, 博士, 博士后, 90946@njnu.edu.cn; 孟帅奇 (1992—), 男, 博士, s.meng@biotec.rwth-aachen.de

基金项目: 煤炭重大专项青年科学家项目 (2025ZD1701600); 江苏省基础研究计划项目 (BK20253050); 微生物改造技术全国重点实验室开放课题 (M2025-17)

引用本文: 徐舒萍, 汪金秀, 孟帅奇, 邓秋风, 张鹤渐, 王秀凤, 郭照, 崔海洋. 人工智能赋能酶工程[J]. 化工学报, 2026, 77(6): 3224–3238

Citation: XU Shuping, WANG Jinxiu, MENG Shuaiqi, DENG Qiufeng, ZHANG Hejian, WANG Xiufeng, GUO Zhao, CUI Haiyang. Artificial intelligence-enabled enzyme engineering[J]. CIESC Journal, 2026, 77(6): 3224–3238

used. Methods such as random forests and support vector machines can uncover patterns in sequence-function data to guide rational design and narrow down the scope of potential mutants. On the other hand, deep learning offers end-to-end modeling capabilities, enabling the direct extraction of features from enzyme sequences and structures for predicting structure, function, and dynamic behavior. Furthermore, combining AI methods with molecular dynamics simulations and quantum chemistry calculations allows for a physical analysis of mutation effects, providing valuable insights for screening. Enzyme design research is evolving from past empirical trial-and-error approaches toward data-driven and model-driven rational design, and is gaining increasing attention in fields such as synthetic biology, biomedicine, and green biomanufacturing.

Keywords: artificial intelligence; enzyme; biocatalysis; machine learning; deep learning; simulation

引 言

酶作为高效、专一的生物催化剂,是驱动细胞代谢、分子组装和信号传导等关键过程的核心元件,也是合成生物学领域重要的基石^[1]。随着合成生物学与工业生物技术的快速融合,以酶为基础的生物催化系统已广泛应用于医药合成、精细化工、生物能源及环境工程等多个工业领域^[2]。不过,天然酶是为适应自然环境进化出来的,结构和功能并不是为工业、医药或环境治理场景“量身定制”的。具体表现为底物谱太窄、在高温有机溶剂或极端pH条件下稳定性不足、对复杂工业体系的耐受性有限,以及在某些工艺条件下催化效率不够理想^[3]。因此,通过人工手段对酶性能进行优化来克服天然酶的局限并满足工业应用需求是推动合成生物学与生物催化技术发展的关键方向。

围绕这一目标,酶工程形成了几条相对成熟的技术路线,包括定向进化、半理性设计、理性设计等。后两者通常被归为“知识驱动”策略:研究人员需要对酶的三维结构、催化机理以及关键残基有比较深入的理解,然后利用分子模拟、定点突变等工具,围绕活性位点或功能相关区域做针对性改造,以实现稳定性、活性或选择性的定向提升。但现实情况是,大多数酶的高分辨率结构和完整催化机理仍不清晰,真正能做到“心中有数地设计”的案例并不多,这也限制了理性和半理性设计在实际场景中的普适性。定向进化则提供了不一样的思路,即在实验室条件下模拟自然进化过程,构建突变体文库并借助高通量筛选技术获得优势变体^[4]。Arnold正是凭在定向进化领域的开创性工作获得2018年诺贝尔化学奖^[5]。然而,这种方法需要巨量文库和多轮筛选,周期长、成本高。更关键的是,在维数极高、近乎“天文级”的序列空间中,仅靠随机或半随

机突变加筛选,很难系统性地找到最有价值的突变组合——很多可能的优良解,实际上从未有机会被“看到”。

人工智能(artificial intelligence, AI)技术的兴起为酶工程提供了全新突破口^[6],其中机器学习(machine learning, ML)与深度学习(deep learning, DL)技术是目前较为常见的方法。AI的主要流程是以大量实验数据为基础学习酶序列、结构以及功能之间的关系,建立序列-结构-功能之间的预测模型,从而预测潜在的重要突变位点及突变对酶功能的影响^[7](图1)。当然除了突变体预测之外,AI还可辅助定向进化过程中的文库设计、突变体筛选、功能预测及结构建模等多个环节^[8]。另一方面,将AI与自由能分析、分子动力学(molecular dynamics, MD)模拟、量子化学计算等计算生物学工具结合,可以进一步地利用酶催化反应的分子机制,为催化位点设计与路径优化提供理论支撑^[9]。总之,AI为酶改造的创新提供了主要驱动力,本文系统综述了AI在酶工程领域的研究进展,并对机器学习、深度学习及计算生物学方法在酶工程相关应用方面展开了讨论,旨在展示人工智能在酶工程研究中的应用潜力,并探讨其在工业生物技术、医药开发及绿色制造等领域的应用前景。

1 机器学习赋能酶工程

酶工程作为生物技术领域的一个重要分支,旨在通过改变蛋白质的氨基酸序列,实现对酶功能的定向改造或优化。现如今,机器学习已成为推动酶工程的关键技术^[10]。常用的手段包括监督学习、无监督学习、半监督学习和强化学习等(图2)。监督学习常用于构建序列-功能预测模型,预测突变对酶功能或性能的影响。无监督学习(如聚类分析)能够在没有标签的情况下对大规模蛋白质序列或

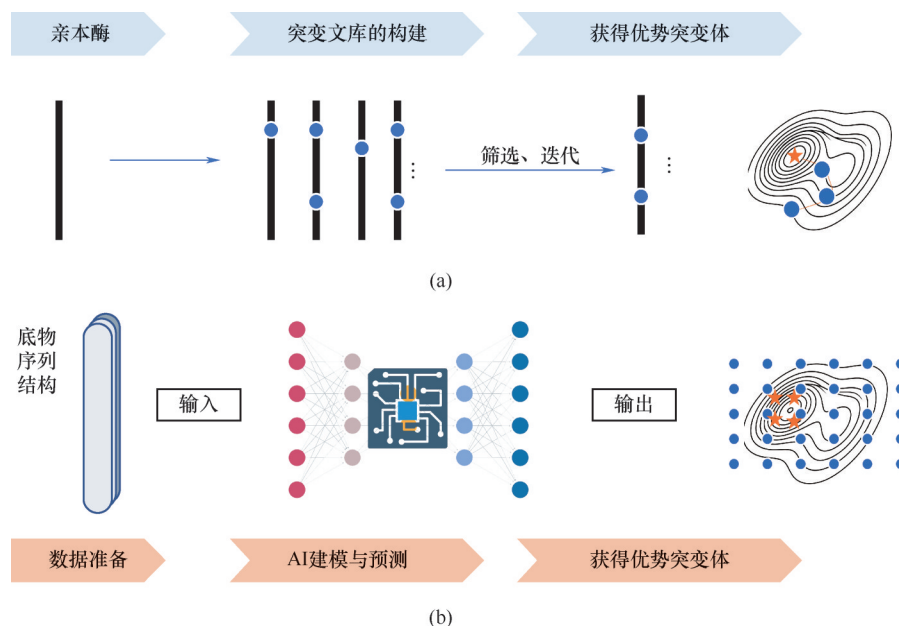


图1 传统酶工程(a)与人工智能驱动的酶工程(b)

Fig.1 Traditional enzyme engineering (a) and AI-driven enzyme engineering (b)

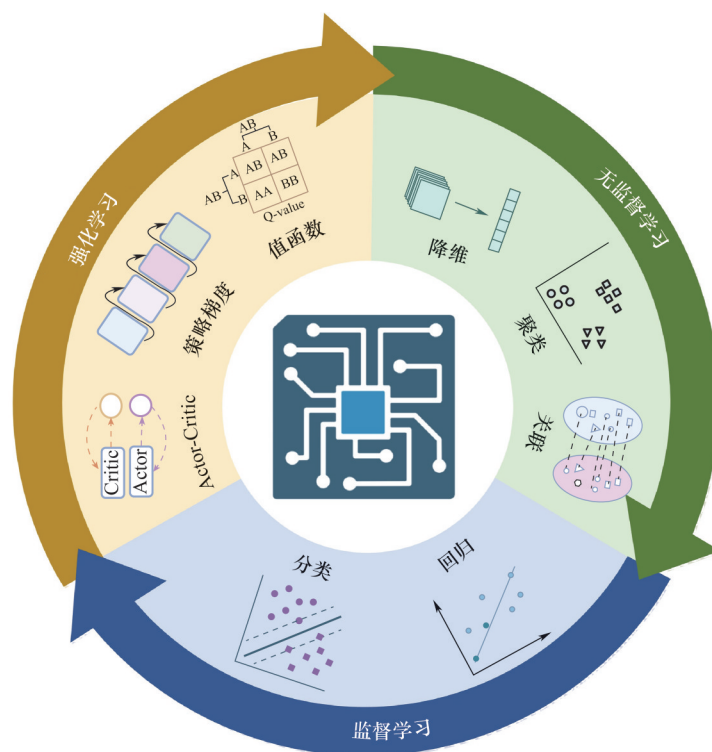


图2 机器学习策略的分类

Fig.2 Classification of machine learning strategies

结构进行分组,帮助识别潜在的进化关系和功能模块,为突变位点选择提供参考。半监督学习则可以在标注数据有限的情况下使用。强化学习的机制是依靠奖励信号驱动找到更优的氨基酸替换策略,适用于迭代式的序列优化和结构设计。在实际应

用中,研究人员常采用偏最小二乘(partial least squares, PLS)、随机森林(random forest, RF)、梯度提升树(gradient boosting decision tree, GBDT)、支持向量机(support vector machine/support vector regression, SVM/SVR)以及高斯过程回归(Gaussian process

regression, GPR)等算法^[9]。这些算法在不同数据类型和需求上有不同的体现。树模型(如RF、GBDT)对特征尺度不敏感,但能学习较复杂的非线性关系,因此,在只有中等数据量情况下,可用树模型预测突变体活性/稳定性并对候选序列进行排序。相反的,SVM/SVR在小样本但特征较多时往往具有较好的泛化性,适用于“高/低活性”分类或连续指标回归,但效果较依赖核函数与参数设置。另一方面,GPR适合与主动学习/贝叶斯优化结合,因为其除了给出预测值外还能提供预测不确定性。总体而言,数据少且希望以较少实验迭代优化时可优先考虑GPR+主动学习;追求稳健、快速筛选时RF/GBDT更合适。

在酶工程里,机器学习真正要学的,其实就是一件事:目标蛋白的“序列突变→功能表现”这条映射关系,也就是通常说的蛋白质适应度景观(protein fitness landscape)^[11]。该概念最早源自进化生物学^[12],可以粗略地想象成:把庞大的蛋白序列空间看成一片高维地形,每一个序列是地形上的一个点,而这个点的“海拔”就对应某项功能性能,比如催化活性、热稳定性或溶解性等。适应度(fitness)则是对这些性能的一个定量刻画。所以机器学习模型的流程就是:给它一批已经测过功能的突变体样本,让模型在这片“地形”上学会地势的大致起伏——也就是从有限数据中推断整个适应度景观的拓扑结构。掌握了这个“地形图”之后,就可以去预测那些还没做实验的突变体大概会处在什么高度,从而在设计阶段就提前筛掉不太可能有用的方案,大幅提升蛋白改造的效率和针对性^[13]。目前已有多项研究使用高斯过程回归、随机森林等算法对适应度景观进行建模。结果表明,这类模型不仅在单点突变预测中表现优异,还能一定程度上解析突

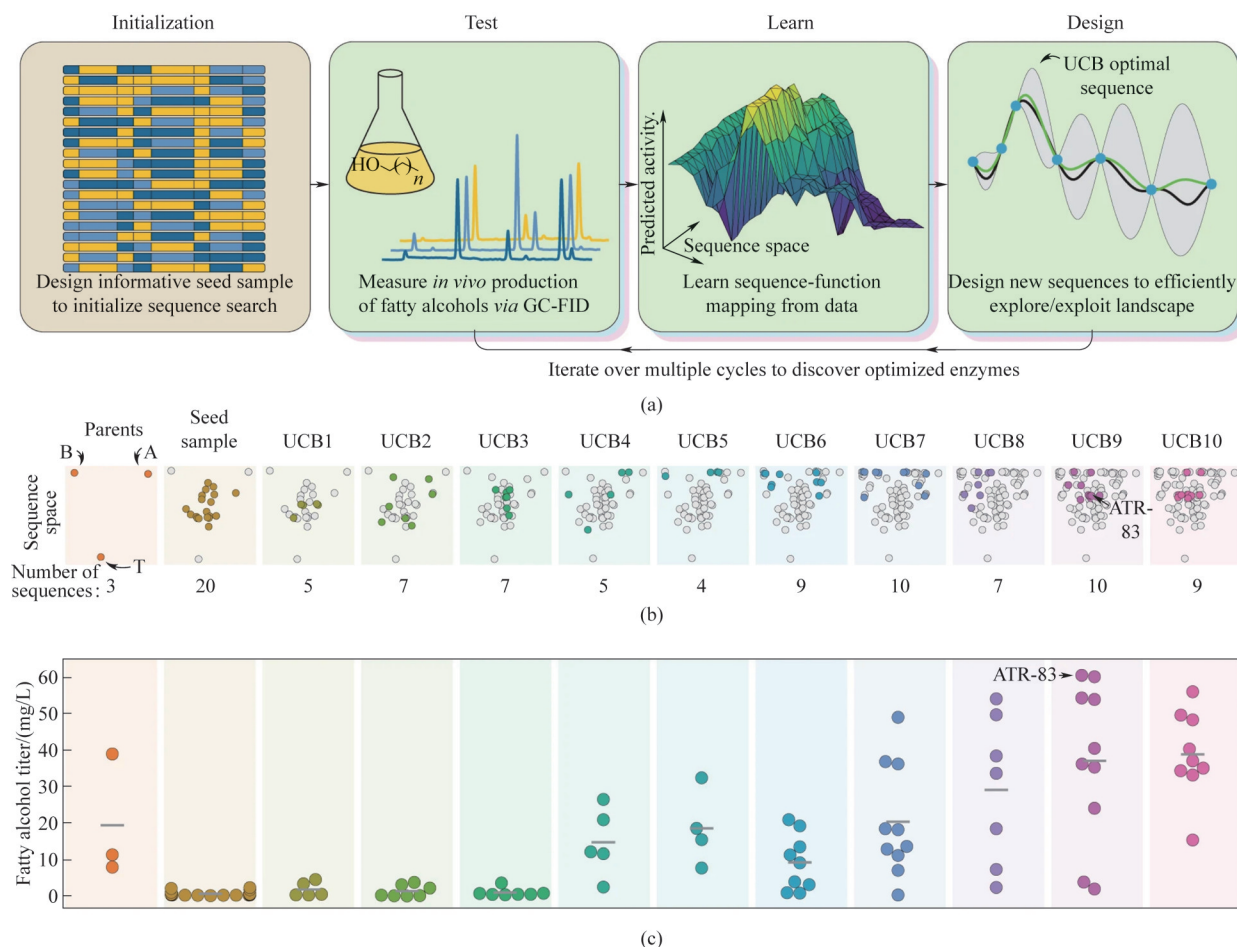
变间的上位效应,为复杂组合突变的理性设计提供依据^[14]。

表1总结了自1990年以来机器学习在蛋白质工程领域的应用进展。在早期阶段,机器学习最初被用于预测蛋白质的二级结构。进入21世纪后,机器学习逐渐拓展到酶工程研究。受小分子定量结构活性关系(QSAR)方法启发,研究者提出了一种用于酶优化的策略,即蛋白质序列-活性关系(ProSAR)算法^[14],2002年,Aita等^[17]从45条序列-活性数据出发,用“位点适应度+成对适应度”的显式参数化模型进行拟合,通过数值优化后分析14个单点突变的适应性贡献,并且在经过实验验证后发现其中两个位点的预测结果与实验数据吻合度较高。2007年Fox等^[22]用偏最小二乘法(PLS)算法指导了一个细菌卤代丙烷脱卤酶的定向进化,该酶参与合成降胆固醇药物阿伐他汀的生物催化工序,进化后的突变体包括35个突变位点,使氰化工序的体积产率提高了约4000倍,真正达到了商业性生物催化工序设计标准的要求。同时期Liao等^[23]使用该策略仅用95个湿实验突变体便把耐热蛋白酶K的活性提高了20倍。近年来,机器学习模型逐渐成为蛋白变体筛选的常用工具。例如Pires等^[18]利用图特征与梯度提升回归树(GBDT)建模来预测蛋白质突变的影响。Folkman等^[19]集成了5个专用支持向量机(SVM),实现仅用序列来预测突变导致的热力学稳定性变化(ΔG)。2021年,Greenhalgh等^[20]设计了一种结合高斯朴素贝叶斯分类器与高斯过程回归的机器学习策略(图3)。通过多轮“设计-测试-学习”循环,构建了嵌合酶库,然后使用上置信界(upper-confidence bound,UCB)标准,通过主动学习机制迭代筛选出高活性变体。最终获得的工程酶ATR-83在体内脂肪酸醇产量达到天然酶的近两倍,且其催

表1 机器学习在蛋白质工程领域的经典案例(1990—2024年)

Table 1 Classic cases of machine learning in the field of protein engineering (1990—2024)

年份	模型/算法	原理
1990	PROMIS (符号规则) ^[15]	归纳逻辑产生可读的“序列→二级结构”规则
1992	Golem (ILP) ^[16]	关系描述+一阶逻辑,原子距离与化学属性
2002	位点-成对适应度线性分解 ^[17]	位点适应度+成对适应度;最小二乘拟合
2008	ProSAR (线性加和) ^[14]	单点突变效应线性叠加,PLS回归
2014	mCSM (GBDT) ^[18]	图特征编码原子距离模式,梯度提升树
2016	EASE-MM ^[19]	二级结构+ASA自适应投票
2021	GNB-GP混合模型 ^[20]	高斯朴素贝叶斯分类“活性/非活性”+高斯过程回归效价(GP titer)
2024	主动学习 ^[21]	蛋白语言模型表征+高斯过程迭代

图3 机器学习加速的蛋白质序列优化^[20]

(a)序列空间搜索策略的概述(首先通过设计一组多样化的序列来初始化搜索,然后通过多次设计-测试-学习循环来高效地理解和优化体内脂肪醇的生产);(b)UCB优化十轮过程中的序列空间可视化(每一轮用不同的颜色表示;3种亲本酶位于这个嵌合序列空间的顶点,所有嵌合体都落在亲本酶的包络内;UCB 优化开始时广泛采样整个空间,但很快收敛到高活性区域);(c)序列优化过程中体内脂肪醇的产量(每个点表示单个序列在序列优化阶段的平均脂肪醇生产量,水平灰条表示该轮序列优化过程中的平均产量)

Fig.3 ML-accelerated protein sequence optimization^[20]

化效率显著提升。2024年,Matsushita等^[21]利用机器学习模型在大规模数据中捕捉复杂的序列-功能关系,可以在较短时间内有效预测出具有改良功能的酶变体。在其针对羟基苯甲酸羟化酶的改造实验中,筛选出了活性提升100倍变体PHBH-MWNL。

机器学习为酶工程提供了前所未有的机遇。即使在有限的数据条件下,各类机器学习模型仍能通过聚类、分类和排序等手段,从序列或结构中挖掘出不少潜在特征,再结合不同算法预测突变体的性能,用于辅助筛选和优化设计。可以肯定的是,在实际应用中仍有不少问题。首先,与模型表现最直接的因素就是数据。高质量且覆盖范围足够广的实验数据依然比较匮乏,这直接限制了模型的泛化能力。很多模型在“自己那一套数据集”上表现很好,一换到新体系就明显失灵。其次是蛋白

质本身的高度复杂性,并且蛋白的实验噪声也会增加训练难度。但机器学习模型仍是广大研究者关注的热点,其可解释性正逐渐成为一个核心议题。尤其是在揭示模型预测与蛋白质生物化学性质间关系方面。随着计算资源的持续提升和实验数据的积累,机器学习将继续在酶工程领域扮演关键角色。

2 深度学习赋能酶工程

深度学习本质上归属机器学习的一个子方向,但两者在基本思路与实现机制上仍有明显差别^[2]。如图4所示,传统的机器学习流程更偏向“手工模式”:研究人员通常需要根据已有的生物学和结构知识,从蛋白质序列或结构中人工设计并提取显式特征,然后再把这些特征打包给随机森林、支持向

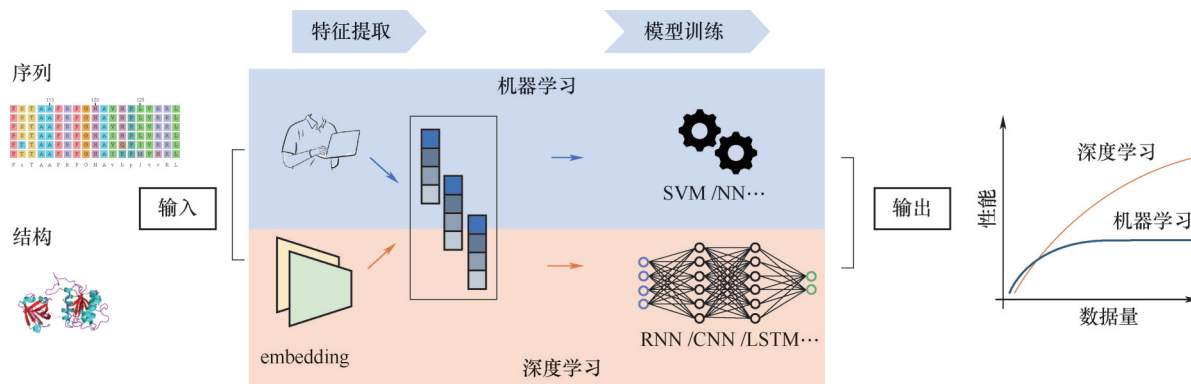


图4 机器学习与深度学习的区别:深度学习可直接实现(输入)端到(输出)端的机器学习

Fig.4 The difference between machine learning and deep learning is that deep learning can directly implement end-to-end machine learning (input to output)

量机或高斯过程回归等算法进行建模与预测。因此,机器学习模型的表现很大程度上与特征质量绑定,特征选得准不准、合不合理往往直接决定了效果上限。相比之下,深度学习则更多地受神经网络启发,形成了另外一种不一样的路:它可以直接从大规模的序列或结构数据中自动学习到有用特征,而不再强依赖研究者提前定义特征。研究者当然仍然需要做一些设计,比如网络结构、损失函数、超参数之类,但不再需要事无巨细地去手动构造每一个特征。一个代表性的例子就是图片识别。机器学习的做法是首先需要人工提取边缘、颜色分布、比例、纹理等特征,再用决策树、SVM等模型训练,最终得到输出结果。而使用深度学习时,卷积神经网络等模型可以在端到端的框架下自动学习图像中的多层次特征,研究者主要负责搭建和调参,模型则在训练中逐步完成特征提取和决策,最后直接

输出预测结果^[24]。神经网络的做法直接将特征提取过程从“显式”变为了“隐式”,这也是机器学习和深度学习在实际应用中最直观的差异。

自然界中的蛋白质本身自带大量复杂的理化信息,这些信息几乎散落在每一个氨基酸以及它们之间的相互作用中,也正是神经网络学习的主要“养料”。表2列举了几类典型的深度学习模型,分别依托不同类型的输入信息展开建模。一类是只需蛋白质一级序列的自然语言模型。如ESM (evolutionary scale models)^[25]和ProtCNN^[30]类似的蛋白质语言模型,仅提供氨基酸序列,就能自动学习出进化模式、保守区域和潜在功能位点。依托这些隐含表征,模型不仅可以用于蛋白质功能预测,在很多情况下甚至能对三维结构给出相当可靠的推断。另一类是基于蛋白质三级结构的深度学习模型。以Chem3DLLM^[38]类似的多模态大语言模型为

表2 蛋白质相关的典型深度学习模型

Table 2 Typical deep learning models related to proteins

使用信息	神经网络	深度学习模型
一级序列	Transformer	ESM 系列 ^[25]
		ProtTrans ^[26]
	双向长短期记忆网络(bi-directional long short-term memory, Bi-LSTM)	ECNet ^[27]
		DeepGOPlus ^[28]
	卷积神经网络(convolutional neural network, CNN)	NetSurfP ^[29]
		ProtCNN ^[30]
三级结构	Transformer	DeepSEA ^[31]
		AlphaFold2 ^[32]
	图神经网络(graph neural network, GNN)	DeepFR ^[33]
	卷积神经网络(CNN)	ProteinMPNN ^[34]
	等变图神经网络(equivariant graph neural network, ENG)	MutCompute ^[35]
	图神经网络(GNN)和注意力机制	LGN ^[36]
		EPP-GT ^[37]

例,它以蛋白质的三维结构当作条件输入,在此基础上生成与该靶标空间构象相匹配的候选小分子三维结构,用于辅助药物分子设计。类似地,Lu等^[35]使用PDB数据库中的大量结构,训练了一个3D自监督的卷积网络模型MutCompute。该模型会对每一个残基构造一个局部微环境,随后识别与局部微环境特征不匹配的“不舒适残基”,据此提示其作为潜在的突变位点。另外,如果在序列和结构之外整合实验数据(如活性测定、稳定性测试、结合实验等),模型的可靠性和可用范围往往还能再上一个台阶。实验数据一方面为深度学习模型提供了高质量的标注信号,另一方面也能在模型仅依据序列或结构做推断时,对某些系统性偏差起到“纠偏”的作用。通过这种多源信息的整合,模型在功能预测、序列优化甚至定量评估变体效应方面,通常能获得更稳健也更接近实际应用需求的表现。

随着生物技术的发展,越来越多的酶序列被挖掘与揭示,蛋白质数据库(如Uniprot, Pfam, BRENDA, CAZy等)以及实验数据的不断扩充,让AI模型得以基于大量批注数据上进行训练,用于提取学习高维“序列-结构-功能”的关系特征,该特征不仅可以辅助显著提高定向进化的效率,还可以实现酶的精准设计。进一步地,越来越多深度学习模型用于酶活性、酶功能预测等方面。

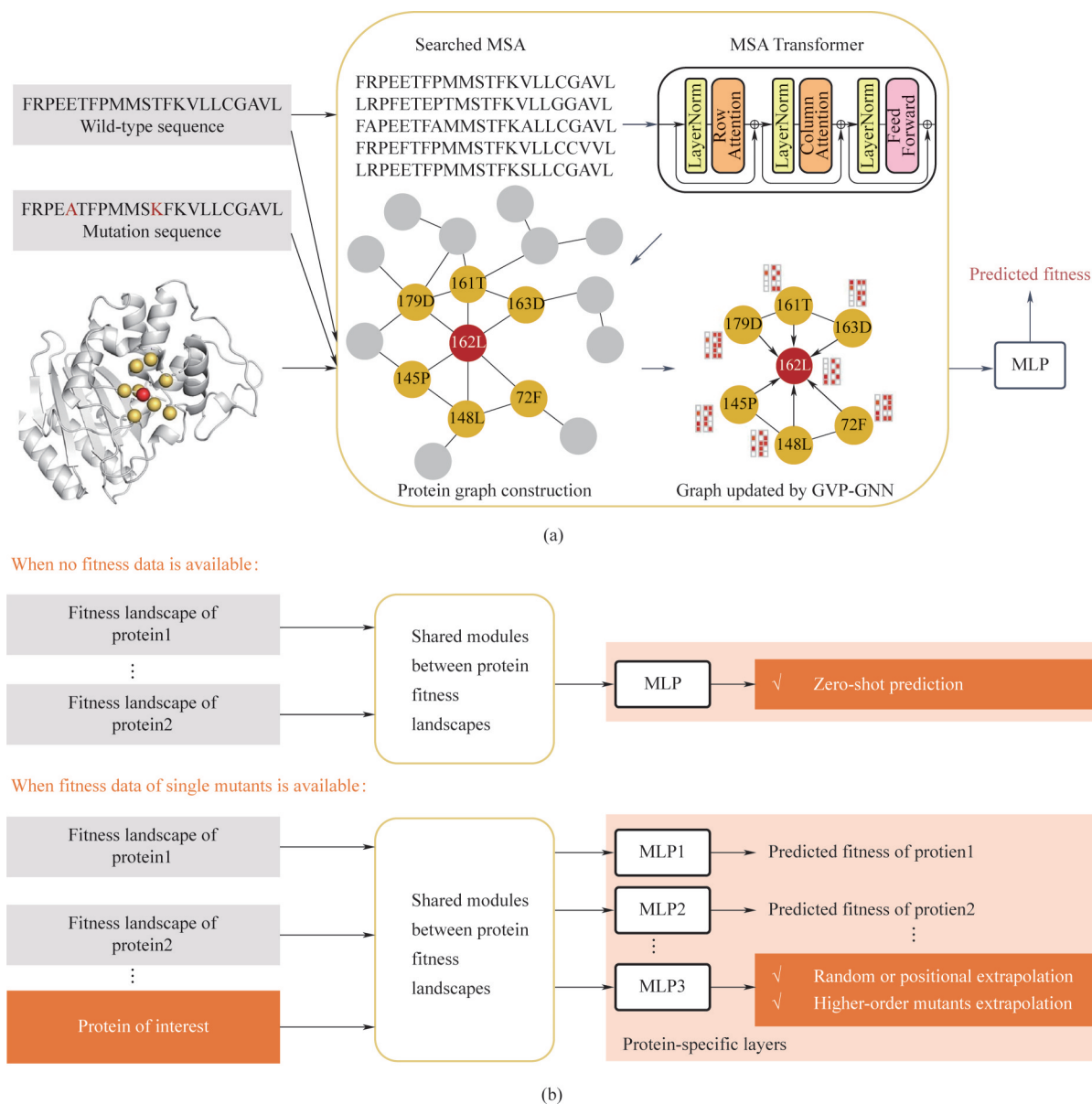
2.1 深度学习模型辅助定向进化与突变体筛选

深度学习在定向进化中的作用,可以理解为“先算后做”的工作模式:模型首先基于已有实验数据进行训练学习,捕捉突变与功能之间的非线性的复杂关系,进而指导研究人员聚焦于最可能产生有益效果的突变区域。这种范式避免了定向进化盲目式的突变,而是把实验资源聚焦于潜在的有益突变区域,从整体上提高定向进化的效率。Chen等^[11]的工作就是一个典型的例子,他们提出了一种新型的深度神经网络模型GVP-MSA(图5)。该模型有效地将序列与结构信号整合到了同一框架内,利用预先训练的蛋白质语言模型处理目标蛋白的多序列比对(MSA)信息,结合图神经网络等提取蛋白质三维结构信息。为了验证这种模型在实际场景中的实用性,该团队设计了多种测试场景,包括单点突变效应的预测、在全新蛋白质上进行零样本突变效应预测以及基于单点突变效应来推断多点突变的总体效应。这些场景基本对应了定向进化实验中不同阶段的实际需求。结果显示,GVP-MSA在

上述三类测试场景中都表现出较好的预测能力。这项作为如何利用机器学习,尤其是深度学习,来辅助定向进化提供了一个颇具启发性的范例。通过在序列突变空间中进行更有指导性的搜索,研究人员能够更快锁定具有改良或全新功能的候选序列,大幅缩短实验探索周期,也在一定程度上降低了实验成本。

2.2 深度学习驱动的酶理性设计

在理性设计方面,通过整合结构预测、机器学习和大规模分子模拟等多模态信息,目前已在酶的酶热稳定性设计、底物选择性设计等多方面取得了成果,更有部分成果达到了工业应用水平^[39]。热稳定性是工业化的重要指标,通过对多模态信息的分类与分析,可高效预测点突变或突变组合对活性及稳定性的影响,克服传统方法中复杂的上位效应。Meng等^[40]利用基于结构的深度学习模型来识别TfCut2中负责提高水解活性和增强稳定性的残基。机器学习指导设计确定了新的有益位置(L32E、S35E、H77Y、R110L、S113E、T237Q、R245Q和E253H),对这些位置进行评估和逐步重组,最终产生有益变体L32E/S113E/T237Q。与TfCut2 WT相比,TfCut2变体的PET解聚能力有所提高[无定形PET薄膜提高了2.9倍;结晶PET粉末(结晶度大于40%)提高了5.3倍]。在耐热性方面,变体L32E/S113E/T237Q的半失活温度提高了5.7℃。Cao等^[41]开发的DeepDDG,一种基于SRP神经网络的方法用于预测点突变对蛋白质稳定性的影响,则利用氨基酸保守性、序列上临近氨基酸的种类、氨基酸残基对在结构上的距离等特征,在超过5700个实验数据点上训练,并在独立测试集上实现皮尔森相关系数0.48~0.56,性能优于其他11种对比方法。Dauparas等^[34]通过使用预先训练的ProteinMPNN嵌入作为迁移学习的输入特征,利用了序列恢复和稳定性优化任务之间的知识重叠,提出了ThermoMPNN^[42],可以仅在已知蛋白质结构的情况下快速准确地预测稳定性变化。Innov'SAR则结合了结构活性关系(structure-activity relationship, SAR)和深度学习,通过构建酶-底物相互作用模型,预测突变对酶催化活性的影响。Zakary等^[43]以柠檬烯环氧化物水解酶为模型酶,以Innov'SAR为ML平台(包括数字信号过程),实现了蛋白质的高鲁棒性,可以抵抗伴随有害聚集的解析。

图5 深度神经网络模型GVP-MSA框架^[11]

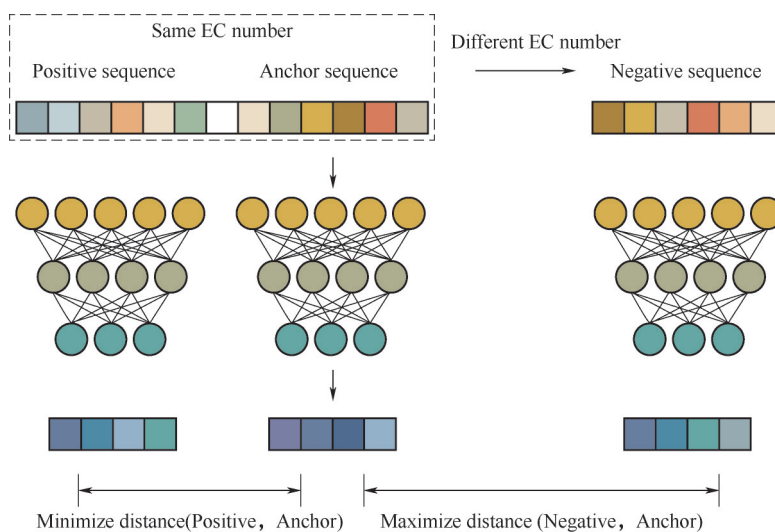
(a)蛋白质特异性模型GVP-MSA概述(该模型架构以野生型序列、突变序列和野生型蛋白质结构作为输入。首先,基于蛋白质序列数据库Uniref90中的野生型序列搜索MSA。使用蛋白质语言模型MSA Transformer从MSA中提取进化上下文,并将输出表示添加到生成蛋白质图的节点特征中。然后,使用两层GVP-GNN更新图。结合进化和结构上下文来学习特定蛋白质的适应性景观);(b)蛋白质特异性模型GVP-MSA概述(该模型架构以野生型序列、突变序列和野生型蛋白质结构作为输入。首先,基于蛋白质序列数据库Uniref 90中的野生型序列搜索MSA。使用蛋白质语言模型MSA Transformer从MSA中提取进化上下文,并将输出表示添加到生成蛋白质图的节点特征中。然后,使用两层GVP-GNN更新图。结合进化和结构上下文来学习特定蛋白质的适应性景观)

Fig.5 Framework of the GVP-MSA deep neural network model^[11]

2.3 酶功能分类与理化性质预测

由于神经网络在学习序列、结构与功能之间的关系方面的优越性,深度神经网络模型被用以预测酶的EC编号、底物特异性、蛋白质溶解度等。DeepEC^[44]以蛋白质序列作为输入,并预测EC数作为输出,使用3个卷积神经网络(CNN)作为EC编号预测的主要引擎,并对CNN无法分类的EC数进行同源性分析,但其也存在局限性,比如都是基于多

标签分类框架,会受到生物学中常见训练数据集有限和不平衡的影响。Yu等^[45]提出的CLEAN模型在这一方面做出了优化(图6),通过采用对比学习框架,有效优化了酶功能预测中因数据不平衡和标注有限导致的泛化能力不足问题。与传统多标签分类模型(如DeepEC)不同,CLEAN并不直接学习“序列-EC编号”的分类,而是通过对比学习,在嵌入空间中将相同功能的酶序列拉近,不同功能的推远,

图6 CLEAN 基于对比学习的框架^[45]

从上到下为 CLEAN 的训练过程。训练数据集中的每个参考序列(锚点)被采样一个具有相同 EC 编号的序列(正样本)和一个具有不同 EC 编号的序列(负样本)。通过向模型提供具有挑战性的负本来提高训练效率,优先选择与锚点具有较小欧几里得距离的负序列嵌入

Fig.6 The contrastive learning-based framework of CLEAN^[45]

表3 深度学习在不同分类的酶中的应用

Table 3 Application of deep learning in different classes of enzymes

EC 编号	工程化酶	模型名称	应用范围
EC1 氧化还原酶	苹果酸脱氢酶 (MDH)	ProteinGAN ^[46]	组合突变及序列从头设计
	血红素类蛋白质	CSA ^[47]	序列从头设计
	甲基化酶 HaeIII	ESM-1v ^[48]	单点突变
EC2 转移酶	聚对苯二甲酸乙二酯水解酶 PETase	MutCompute ^[35]	单点突变
	TEM-1β-内酰胺酶	Low-N ^[49] ; ECNet ^[27] ; MutCompute ^[50]	单点及组合突变
EC3 水解酶	抗菌肽(AMPs)	LSTM ^[51]	(短)序列从头设计
	赖氨酸修饰酶	DeepAcet ^[52]	翻译后修饰
	酪氨酸,色氨酸,半胱氨酸修饰酶	DeepNitro ^[53]	翻译后修饰
	TIM 桶	ESM-1v ^[48]	单点突变
EC4 裂合酶	—	—	—
EC5 异构酶	异构酶	DeEPn ^[54]	单点突变
EC6 连接酶	连接酶 E3	ProteinMPNN ^[55]	组合突变
EC7 转运酶	转运酶	DeEPn ^[54]	单点突变

从而学习到一个功能感知的序列表示空间。这种范式使其能够更有效地利用稀少类别的数据,因此,在面对训练集中出现次数极少的 EC 编号时,依然能保持高预测精度(在小样本 EC 测试集上 F1 分数达 0.817),显著提升了泛化能力。截至目前,已衍生出针对不同催化类型的酶的深度学习模型(表3)。

在酶活预测方面, Li 等^[56]提供了一种名为 DLKcat 的深度学习方法,成功预测了 300 多种酵母的基因组尺度 k_{cat} 数据。该模型仅从底物结构和蛋白质序列出发,用于任意代谢酶的高通量 k_{cat} 预测。DLKcat 不仅能够捕捉突变酶的 k_{cat} 变化,还能识别对 k_{cat} 值有强烈影响的氨基酸残基。类似的工作还有 CatPred^[57],不同的是该模型探索了多种网络架构和

特征表示方法,包括引入预训练的蛋白质语言模型特征和预训练的三维结构特征。在溶解度预测方面,Wang等^[58]提出了一种名为DeepSoluE的深度学习预测模型,主要用于预测大肠杆菌中的蛋白质溶解度。该模型创新性地将物理化学特征和氨基酸上下文信息结合起来,用以描述蛋白序列的功能特征。随后使用遗传算法选择最佳特征子集,再用长短期记忆网络(LSTM)提取特征信息并进行二分类任务。Chen等^[59]同样开发了一种基于序列的蛋白质溶解度预测方法GraphSol,主要通过预测蛋白质的接触图,结合注意力图卷积神经网络,以结构感知的方法预测蛋白质的溶解度。结果显示GraphSol明显优于其他基于序列的方法。这在一定程度上也说明,随着人工神经网络可解释架构和特征重要性分析方法的不断发展,能够以更清晰的方式理解酶的结构-功能关系,而不仅仅停留在“黑箱预测”的层面。

2.4 基于生成式模型的酶从头设计

RFdiffusion是当前应用最为广泛的蛋白质从头设计工具,它基于扩散模型开发,不仅能高精度地生成结构,还能在一定程度上“可编程”地控制目标形状和功能^[60]。在疫苗开发和药物递送领域,RFdiffusion设计出的结构展现出重要应用潜力。例如,研究者们已成功设计出具有精确对称性的多聚体蛋白纳米笼。这些纳米笼可自组装形成二十面体或八面体构型,其内部空腔可用于封装小分子或核酸,在经过表面工程化修饰之后可进行靶向识别或递送^[60]。在治疗性蛋白结合剂领域,RFdiffusion可以很好地解决传统方法存在的难以靶向高极性、低疏水性蛋白界面的难题。最近提出的CycleDesigner框架创新性地改造了RFdiffusion,使其适应环肽拓扑结构。之后结合HighFold进行构象预测,成功设计出针对特定靶标、兼具高稳定性和高亲和力的环肽结合剂^[61]。

RFdiffusion在酶设计方面同样展现了强大的设计功能:先把活性中心“钉死”,再让模型补全剩下的结构。即预先设定催化三联体或辅因子结合位点的空间坐标,再由模型围绕这些关键残基生成稳定的三维支架。随着RFdiffusion3(RFD3)的提出,这一思路进一步拓展至全原子水平的多组分复合物设计,能够同时建模蛋白质与配体、金属离子或核酸的相互作用^[62]。在蛋白-多肽相互作用设计领域,RFdiffusion也取得了重要进展。在最近的针对

具有螺旋倾向的生物活性肽类激素(如胰高血糖素样肽-1, GLP-1)的设计工作中,研究者将参数化生成方法与深度学习相结合,借助RFdiffusion等模型设计出对这些肽激素具有高亲和力的结合蛋白^[63]。这类结合蛋白既可以被用作长效、高灵敏度的检测探针,也可以通过调控肽激素在体内的半衰期,间接参与代谢调节与疾病干预。从应用视角看,这实际上是在为多肽药物配套设计“可编程调控模块”。

综上所述,酶工程正迈向一个全数据驱动的理性设计新阶段。这个转变并不是单一技术突破造成的,而是多条技术路线汇聚的结果。结构方面,AlphaFold等高精度结构预测工具为理性设计提供了可靠的结构基础^[32]。功能预测方面,CLEAN等模型通过对比学习增强了序列-功能映射的泛化能力,提升了功能预测的准确性^[45]。在这样的技术背景下,能够整合多模态数据、强化学习算法以及自动化实验流程的AI智能体(Agent)系统开始逐步成型。例如,能根据结构约束自动设计序列的ProteinMPNN^[34]、能实现蛋白质全新设计的RFdiffusion^[64],以及无须人工干预的通用型人工智能驱动的自主酶工程平台iBioFoundry^[65]等系统。这些系统正在把“从序列生成、结构预测到功能筛选与验证”的整个流程,尽可能打通并自动化,实现真正意义上的高通量、闭环优化。

3 计算生物学辅助酶工程

酶的功能与其氨基酸序列、二级结构、三维结构存在复杂的关联。传统计算生物学在解析这些关系中发挥了重要的作用,不仅能辅助研究人员发掘潜在的功能位点,还能通过分子动力学模拟(molecular dynamics simulation, MD)、量子力学/分子力学(quantum mechanics/molecular mechanics, QM/MM)等方法揭示突变体的机制性变化,从而为蛋白质的理性设计提供理论依据。随着AI的快速发展,这些传统模拟方法生成的数据逐渐成为深度学习模型的重要训练素材。

近年来,基于AI的对接评分函数和配体构象的模型越来越多。研究人员通过迁移学习、图神经网络这些技术来优化分子配对预测,虚拟筛选的效率与命中率也因此提高不少。Lai等^[66]提出了一个基于图Transformer的新型交互感知模型Interformer,引入了交互感知混合密度网络来精确描述蛋白质

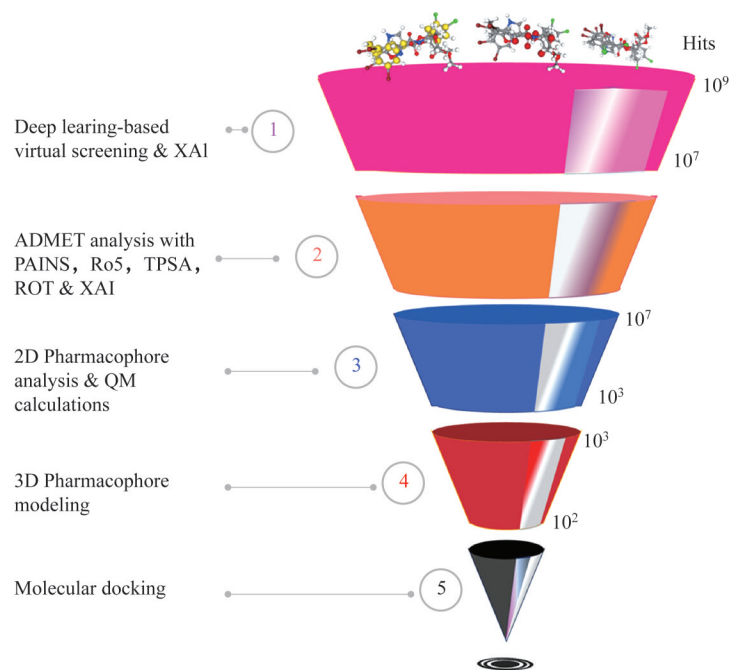
和配体之间的非共价相互作用,在 Posebusters 和 PDBbind 基准测试中表现最好。Luttens 等^[67]用 CatBoost 分类器配合波形预测框架来高效筛选超大规模化学库,在 35 亿化合物的测试中把计算成本降低到 1/1000 左右,并找到了一些有多个靶点活性的化合物。FABind+则在口袋预测和对接模块中加了置换损失,再结合置信度模型与高效采样策略,不用额外训练就能同时做回归与采样两种任务^[68]。Luo 等^[69]提出的 ApoDock 将机器学习驱动的条件侧链包装和传统采样方法结合起来,用混合密度网络进行评分,用 AlphaFold2 与 ESMFold 建模结构进行对接时,成功率比现有方法提升 28.5%,在蛋白质-配体结合研究中十分有前景。上述研究从不同角度突破了传统分子对接的局限,说明 AI 在分子对接方面确实很有潜力,也给未来的药物发现提供了更智能、更高效的工具。

分子动力学模拟在揭示蛋白质和其他生物分子的功能机制、结构基础以及小分子、肽和蛋白质的设计和优化方面很有价值。这些模拟记录下原子运动的细节,把时间分辨率拉得很高,从而明白蛋白质和生物分子具体怎么运动,这往往对理解蛋白质功能和配体的结合很关键^[70]。另外,蛋白质和溶剂分子之间的相互作用十分复杂,模拟蛋白质在不同溶剂中的状态有助于在溶剂甚至一些非常规介质中做出更好的生物催化效果。这几年 AI 算法在处理时间序列数据上有了突破,研究人员开始试着将 MD 轨迹数据与循环神经网络(RNN)或 Transformer 架构结合起来,预测蛋白质的动力学行为,以及其在不同溶剂中的构象适应能力,这样可以加快在非天然环境下优化酶的效率。蛋白质在溶液中通常呈现多种构象,分子动力学模拟本来是用来生成具有真实概率分布的构象集合的,但即使是个简单的问题,所需的计算资源也非常庞大。Lewis 等^[71]开发了一种名为 BioEmu 的生成式深度学习模拟器,把 AlphaFold 编码、扩散模型与大规模 MD/实验数据结合起来,通过三阶段训练高效生成蛋白质平衡构象集合。BioEmu 在单张 GPU 上每小时能生成数千个独立结构,不管是功能构象捕捉、平衡分布重建还是预测折叠稳定性,均优于传统 MD,速度提升 4~5 个数量级,能量误差低于 1 kcal/mol (1 cal = 4.18 J)。此外,其在复杂蛋白构象模拟及药物发现(如结合口袋、变构机制识别)方面都展现出不错的应用潜力,也为基因组规模的动态蛋白

质研究提供了可扩展路径。

解析酶反应机制始终是酶工程领域的核心。然而酶反应过程高度复杂,底物化学键断裂与生成、关键氨基酸残基的动态运动,均发生在飞秒至皮秒的时间尺度^[72]。目前,MD 与 QM/MM 的组合方法被公认为是研究酶催化机理可信度最高的计算模拟方法之一。基于此,越来越多的工作开始尝试将机器学习方法引入传统的 QM/MM 模拟流程,以提升计算效率并扩展其应用范围。Platero-Rochart 等^[73]构建了一种引导式 QM/MM 分子动力学模拟与机器学习耦合的技术方案,并将其应用于 *Ideonella sakaiensis* 来源的 MHET 酶催化 MHET 水解反应的研究。该研究以 steered QM/MM 模拟轨迹为训练数据集,构建了基于核岭回归(KRR)的机器学习模型,并结合 ElasticNet 回归与 Coulomb matrix(CM)结构表征进行特征优化。结果表明,该模型能够较为准确地预测反应过程中拉拽功与能垒变化,平均绝对误差低于 3 kcal/mol,相关评分高于 0.90。类似的思路也出现在计算药物设计领域。Madushanka 等^[74]通过将 XTB 半经验量子力学方法深度整合到 3D 药效团分析过滤器中。如图 7 所示,该团队提出的 SmartCADD 工作流构建了一套多层筛选体系,将深度学习驱动的初步筛选方法与基于经典力场及量子信息的结构过滤策略相结合,形成了一个无缝的端到端过滤管道。其中,QM 优化不仅为形状、距离及对齐得分提供可靠结构基础,还与 AI 初筛、ADMET 过滤及分子对接协同作用。实践表明,这种多信息融合的筛选策略能够在早期阶段有效减少假阳性候选分子,从而提高候选化合物(例如 HIV 抑制剂)的结构匹配度与活性预测可靠性,也在一定程度上降低了后续实验验证的成本与时间投入。

随着计算资源的持续提升以及 QM/MM 模拟数据的不断积累,相关数据集正逐渐呈现出更加多样化的特征,这也为数据驱动方法的进一步发展提供了基础。这些数据配合人工智能在上下游生物信息和化学信息挖掘中的能力,以及高性能计算硬件的持续发展,QM/MM 方法将在突变体筛选、活性优化及蛋白质设计等领域发挥重要作用。目前已发展出多种基于计算生物学的通用型设计策略,例如 Knowvolution^[75]、2GenReP 及 InSiReP^[76]、CompassR^[77]等。其中的突变位点选择、重组位点搭配过程也越来越多地借助 AI 决策系统,如强化学习或图搜索算法,从大量可能组合中快速收敛至最优设计方案,

图7 SmartCADD 筛选设计流程^[74]Fig.7 SmartCADD filter design flow^[74]

极大提高研发效率。几年来,很多综述已经对该部分进行了完整综述^[9, 78-79]。其中一个代表性的例子是 Weigmann 等^[80]提出的端到端的 AI 引导型 3DMEngineering 方法,通过三阶段流程(起始蛋白面板设计、AI 驱动的热点残基筛选、迭代突变与优化)系统性提升了转氨酶的催化活性。该方法有效地结合了结构背景、共同进化与文献信号识别关键突变位点,通过 AI 预测新型突变实现了超越传统序列-活性关系的优化。

4 讨论与展望

近年来,AI 与生物催化的深度融合,确实在悄悄重塑蛋白质工程的研究方式和“做事路径”。过去更多依赖经验和逐步试错的流程,如今越来越多地被数据驱动和模型引导所取代。本文尝试系统梳理机器学习、深度学习以及计算生物学在酶设计与优化中的前沿应用。AI 不仅显著提升了突变体设计的效率与精准度,更在理解酶催化机制、预测复杂性状和推动理性设计方面展现出强大潜力。值得一提的是,AI 的学习对象逐步从手工提取特征向多模态信息变化,正说明 AI 学习高维、非线性数据的优势在被逐渐放大。

不过,如果从实际落地的角度来讲,其实际应用仍面临若干关键挑战。最核心也是目前绕不开

的问题,还是数据本身:高质量、标准化、注释完整的实验数据并不多,这直接卡住了模型的泛化和可靠性。要破这个局,大致有几条路子可以并行推进。一方面,需要逐步形成跨平台的数据标准,让不同实验室、不同项目产生的数据至少能“说同一种话”,便于沉淀成共享的知识库。另一方面,借助实验室自动化、微流控和高通量筛选,把设计到测量的流程尽可能自动化、连续化,持续产出可追溯的数据流,而不是零散的点状结果。同时,在模型端也要更重视不确定性和数据质量,尽量让模型不仅给出一个预测值,还能告诉实验人员“我有多确定”,从而更聪明地分配实验资源。至于数据分散、各自为战的问题,未来大概率要靠新的协作机制来慢慢化解,在保护知识产权的前提下,让更多数据以某种安全的方式参与到模型训练中。总体来看,AI 在酶工程和生物催化中的定位,正在从工具型角色转向真正意义上的“设计合作者”,只要数据和基础设施跟得上,它的影响力还会继续加码。

“黑箱”是大多数 AI 模型另一大特性,其预测结果与生物物理机制之间的不明确性大大制约了其在机制驱动设计中的应用。当前多数方法仍集中于单一性质预测(如稳定性或活性),多目标协同优化及动态环境下的酶功能预测能力仍有待加强。但随着数据、算法与实验技术的协同进步,如果上述问题得到解决,AI 将有望彻底改变理解与设计生

物系统的思维方式,为合成生物学、绿色制造及生物医药等领域带来深远影响。

参考文献

- [1] Bornscheuer U T, Huisman G W, Kazlauskas R J, et al. Engineering the third wave of biocatalysis[J]. *Nature*, 2012, **485** (7397): 185–194.
- [2] Yi D, Bayer T, Badenhorn C P S, et al. Recent trends in biocatalysis[J]. *Chemical Society Reviews*, 2021, **50**(14): 8003–8049.
- [3] Sheldon R A, Brady D. The limits to biocatalysis: pushing the envelope[J]. *Chemical Communications*, 2018, **54**(48): 6088–6104.
- [4] Packer M S, Liu D R. Methods for the directed evolution of proteins[J]. *Nature Reviews Genetics*, 2015, **16**(7): 379–394.
- [5] Zhou Y, Zhu L Y, Zou P. A brief introduction to Nobel prize in chemistry 2018: harnessing the power of evolution[J]. *University Chemistry*, 2019, **34**(1): 1–6.
- [6] Vaissier Welborn V, Head-Gordon T. Computational design of synthetic enzymes[J]. *Chemical Reviews*, 2019, **119**(11): 6613–6630.
- [7] 徐沛, 汪卫华, 宁洪伟, 等. 人工智能辅助的酶分子改造应用进展[J]. *生物工程学报*, 2024, **40**(6): 1728–1741.
Xu P, Wang W H, Ning H W, et al. Progress in the application of artificial intelligence-assisted molecular modification of enzymes [J]. *Chinese Journal of Biotechnology*, 2024, **40**(6): 1728–1741.
- [8] Yang J, Li F Z, Arnold F H. Opportunities and challenges for machine learning-assisted enzyme engineering[J]. *ACS Central Science*, 2024, **10**(2): 226–241.
- [9] Kouba P, Kohout P, Haddadi F, et al. Machine learning-guided protein engineering[J]. *ACS Catalysis*, 2023, **13**(21): 13863–13895.
- [10] Wang X Y, Li A N, Li X J, et al. Empowering protein engineering through recombination of beneficial substitutions[J]. *Chemistry – A European Journal*, 2024, **30**(16): e202303889.
- [11] 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.
- [12] Romero P A, Arnold F H. Exploring protein fitness landscapes by directed evolution[J]. *Nature Reviews Molecular Cell Biology*, 2009, **10**(12): 866–876.
- [13] 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.
- [14] Fox R J, Huisman G W. Enzyme optimization: moving from blind evolution to statistical exploration of sequence – function space[J]. *Trends in Biotechnology*, 2008, **26**(3): 132–138.
- [15] King R D, Sternberg M J E. Machine learning approach for the prediction of protein secondary structure[J]. *Journal of Molecular Biology*, 1990, **216**(2): 441–457.
- [16] Muggleton S, King R D, Stenberg M J E. Protein secondary structure prediction using logic-based machine learning[J]. *Protein Engineering, Design and Selection*, 1992, **5**(7): 647–657.
- [17] Aita T, Hamamatsu N, Nomiya Y, et al. Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution[J]. *Biopolymers*, 2002, **64**(2): 95–105.
- [18] Pires D E V, Ascher D B, Blundell T L. mCSM: predicting the effects of mutations in proteins using graph-based signatures[J]. *Bioinformatics*, 2014, **30**(3): 335–342.
- [19] Folkman L, Stantic B, Sattar A, et al. EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models[J]. *Journal of Molecular Biology*, 2016, **428**(6): 1394–1405.
- [20] Greenhalgh J C, Fahlberg S A, Pfleger B F, et al. Machine learning-guided acyl-ACP reductase engineering for improved *in vivo* fatty alcohol production[J]. *Nature Communications*, 2021, **12**: 5825.
- [21] Matsushita T, Kishimoto S, Hara K, et al. Functional enhancement of flavin-containing monooxygenase through machine learning methodology[J]. *ACS Catalysis*, 2024, **14**(9): 6945–6951.
- [22] 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.
- [23] Liao J, Warmuth M K, Govindarajan S, et al. Engineering proteinase K using machine learning and synthetic genes[J]. *BMC Biotechnology*, 2007, **7**(1): 16.
- [24] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *Nature*, 2015, **521**(7553): 436–444.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence[J]. *Bioinformatics*, 2020, **36**(2): 422–429.
- [29] Høie M H, Kiehl E N, Petersen B, et al. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning[J]. *Nucleic Acids Research*, 2022, **50**(W1): W510–W515.
- [30] Bileschi M L, Belanger D, Bryant D H, et al. Using deep learning to annotate the protein universe[J]. *Nature Biotechnology*, 2022, **40** (6): 932–937.
- [31] Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning – based sequence model[J]. *Nature Methods*, 2015, **12**(10): 931–934.
- [32] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021, **596**(7873): 583–589.
- [33] Gligorijević V, Renfrew P D, Kosciółek T, et al. Structure-based protein function prediction using graph convolutional networks[J]. *Nature Communications*, 2021, **12**: 3168.
- [34] Dauparas J, Anishchenko I, Bennett N, et al. Robust deep learning-based protein sequence design using ProteinMPNN[J]. *Science*, 2022, **378**(6615): 49–56.
- [35] Lu H Y, Diaz D J, Czarnecki N J, et al. Machine learning-aided engineering of hydrolases for PET depolymerization[J]. *Nature*,

- 2022, **604**(7907): 662–667.
- [36] Zhou B X, Zheng L R, Wu B H, et al. Protein engineering with lightweight graph denoising neural networks[J]. Journal of Chemical Information and Modeling, 2024, **64**(9): 3650–3661.
- [37] Tian Y, Zhang D C, Xing H D, et al. Computational glycosyltransferases masked deoxynivalenol toxicity and halted FHB spread in wheat grains[J]. Journal of Advanced Research, 2026, **83**: 129–139.
- [38] Jiang L, Sun S Z, Qi B Q, et al. Chem3DLLM: 3D multimodal large language models for chemistry[EB/OL]. arXiv: 2508.10696v1, 2025. <https://arxiv.org/html/2508.10696v1>.
- [39] 崔颖璐, 吴边. 符合工程化需求的生物元件设计[J]. 中国科学院院刊, 2018, **33**(11): 1150–1157.
- Cui Y L, Wu B. Biological components design for engineering requirements[J]. Bulletin of the Chinese Academy of Sciences, 2018, **33**(11): 1150–1157.
- [40] Meng S Q, Li Z Y, Zhang P, et al. Deep learning guided enzyme engineering of *Thermobifida fusca* cutinase for increased PET depolymerization[J]. Chinese Journal of Catalysis, 2023, **50**: 229–238.
- [41] Cao H L, Wang J X, He L P, et al. DeepDDG: predicting the stability change of protein point mutations using neural networks[J]. Journal of Chemical Information and Modeling, 2019, **59**(4): 1508–1514.
- [42] Dieckhaus H, Brocidiaco M, Randolph N Z, et al. Transfer learning to leverage larger datasets for improved prediction of protein stability changes[J]. Proceedings of the National Academy of Sciences of the United States of America, 2024, **121**(6): e2314853121.
- [43] Zakary S, Oyewusi H, Huyop F. Dehalogenases for pollutant degradation in brief: a mini review[J]. Journal of Tropical Life Science, 2021, **11**(1): 17–24.
- [44] Ryu J Y, Kim H U, Lee S Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, **116**(28): 13996–14001.
- [45] Yu T H, Cui H Y, Li J C, et al. Enzyme function prediction using contrastive learning[J]. Science, 2023, **379**(6639): 1358–1363.
- [46] Repecka D, Jauniskis V, Karpus L, et al. Expanding functional protein sequence spaces using generative adversarial networks[J]. Nature Machine Intelligence, 2021, **3**(4): 324–333.
- [47] Çarklı Yavuz B, Yurtay N, Ozkan O. Prediction of protein secondary structure with clonal selection algorithm and multilayer perceptron[J]. IEEE Access, 2018, **6**: 45256–45261.
- [48] Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design[J]. Nature Communications, 2022, **13**: 4348.
- [49] 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.
- [50] Shroff R, Cole A W, Diaz D J, et al. Discovery of novel gain-of-function mutations guided by structure-based deep learning[J]. ACS Synthetic Biology, 2020, **9**(11): 2927–2935.
- [51] Müller A T, Hiss J A, Schneider G. Recurrent neural network model for constructive peptide design[J]. Journal of Chemical Information and Modeling, 2018, **58**(2): 472–479.
- [52] Wu M Q, Yang Y X, Wang H, et al. A deep learning method to more accurately recall known lysine acetylation sites[J]. BMC Bioinformatics, 2019, **20**(1): 49.
- [53] Xie Y B, Luo X T, Li Y P, et al. DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning[J]. Genomics, Proteomics & Bioinformatics, 2018, **16**(4): 294–306.
- [54] Semwal R, Aier I, Tyagi P, et al. DeEPn: a deep neural network based tool for enzyme functional annotation[J]. Journal of Biomolecular Structure and Dynamics, 2021, **39**(8): 2733–2743.
- [55] Kao H W, Lu W L, Ho M R, et al. Robust design of effective allosteric activators for Rsp5 E3 ligase using the machine learning tool ProteinMPNN[J]. ACS Synthetic Biology, 2023, **12**(8): 2310–2319.
- [56] Li F R, Yuan L, Lu H Z, et al. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction[J]. Nature Catalysis, 2022, **5**(8): 662–672.
- [57] Boorla V S, Maranas C D. CatPred: a comprehensive framework for deep learning *in vitro* enzyme kinetic parameters[J]. Nature Communications, 2025, **16**: 2072.
- [58] Wang C, Zou Q. Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with DeepSoluE[J]. BMC Biology, 2023, **21**(1): 12.
- [59] Chen J W, Zheng S J, Zhao H Y, et al. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map[J]. Journal of Cheminformatics, 2021, **13**(1): 7.
- [60] Anishchenko I, Pellock S J, Chidyausiku T M, et al. *De novo* protein design by deep network hallucination[J]. Nature, 2021, **600**(7889): 547–552.
- [61] Zhang C H, Xu Z Y, Lin K, et al. CycleDesigner: leveraging CycRFDiffusion and HighFold to design cyclic peptide binders for specific targets[J]. Journal of Chemical Information and Modeling, 2025, **65**(12): 6155–6165.
- [62] Butcher J, Krishna R, Mitra R, et al. *De novo* design of all-atom biomolecular interactions with RFDiffusion3[PP/OL]. bioRxiv (2025-11-19) [2026-01-04]. <http://doi.org/10.1101/2025.09.18.676967>.
- [63] Vázquez Torres S, Leung P J Y, Venkatesh P, et al. *De novo* design of high-affinity binders of bioactive helical peptides[J]. Nature, 2024, **626**(7998): 435–442.
- [64] Watson J L, Juergens D, Bennett N R, et al. *De novo* design of protein structure and function with RFDiffusion[J]. Nature, 2023, **620**(7976): 1089–1100.
- [65] Singh N, Lane S, Yu T H, et al. A generalized platform for artificial intelligence-powered autonomous enzyme engineering[J]. Nature Communications, 2025, **16**: 5648.
- [66] Lai H, Wang L Y, Qian R Y, et al. Interformer: an interaction-aware model for protein-ligand docking and affinity prediction[J]. Nature Communications, 2024, **15**: 10223.
- [67] Luttens A, Cabeza de Vaca I, Sparring L, et al. Rapid traversal of vast chemical space using machine learning-guided docking screens[J]. Nature Computational Science, 2025, **5**(4): 301–312.
- [68] Gao K Y, Pei Q Z, Zhang G B, et al. FABind+: enhancing molecular docking through improved pocket prediction and pose generation[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2025: 330–341.
- [69] Luo D, Qu X Y, Lu D X, et al. Ligand-conditioned side chain packing for flexible molecular docking[J]. Journal of Chemical

- Theory and Computation, 2025, **21**(3): 1494–1505.
- [70] Hollingsworth S A, Dror R O. Molecular dynamics simulation for all[J]. *Neuron*, 2018, **99**(6): 1129–1143.
- [71] Lewis S, Hempel T, Jiménez-Luna J, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning[J]. *Science*, 2025, **389**(6761): eadv9817.
- [72] Osuna S, Jiménez-Osés G, Noey E L, et al. Molecular dynamics explorations of active site structure in designed and evolved enzymes[J]. *Accounts of Chemical Research*, 2015, **48**(4): 1080–1089.
- [73] Platero-Rochart D, Krivobokova T, Gastegger M, et al. Prediction of enzyme catalysis by computing reaction energy barriers *via* steered QM/MM molecular dynamics simulations and machine learning[J]. *Journal of Chemical Information and Modeling*, 2023, **63**(15): 4623–4632.
- [74] Madushanka A, Laird E, Clark C, et al. SmartCADD: AI-QM empowered drug discovery platform with explainability[J]. *Journal of Chemical Information and Modeling*, 2024, **64**(17): 6799–6813.
- [75] Cheng F, Zhu L L, Schwaneberg U. Directed evolution 2.0: improving and deciphering enzyme properties[J]. *Chemical Communications*, 2015, **51**(48): 9760–9772.
- [76] Cui H Y, Davari M D, Schwaneberg U. Recombination of compatible substitutions by 2GenReP and InSiReP[M]//*Enzyme Engineering*. New York, NY: Springer US, 2021: 71–81.
- [77] Cui H Y, Cao H, Cai H Y, et al. Computer-assisted recombination (CompassR) teaches us how to recombine beneficial substitutions from directed evolution campaigns[J]. *Chemistry — A European Journal*, 2020, **26**(3): 643–649.
- [78] Siedhoff N E, Schwaneberg U, Davari M D. Machine learning-assisted enzyme engineering[M]//*Methods in Enzymology*. Amsterdam: Elsevier, 2020: 281–315.
- [79] Listov D, Goverde C A, Correia B E, et al. Opportunities and challenges in design and optimization of protein function[J]. *Nature Reviews Molecular Cell Biology*, 2024, **25**(8): 639–653.
- [80] Weigmann K F G, Heijl S, Vrolijk B, et al. Navigating the sequence–function landscape: AI-driven discovery of unseen and synergistic mutations in an amine transaminase[J]. *ACS Catalysis*, 2025, **15**(17): 15121–15131.