

综述

人工智能在酶工程中的应用与进展

孙萌^{1#}, 王曼璐^{1#}, 李雨泽², 袁波², 陈晨曦², 王梁华³, 陈涵^{3*}, 孙铭娟^{3*}¹海军军医大学基础医学院学员一大队, 上海 200433; ²海军军医大学基础医学院学员四大队, 上海 200433;³海军军医大学基础医学院生物化学与分子生物学教研室, 上海 200433)

摘要: 近年来, 随着人工智能技术的快速发展, 其在酶工程领域的应用展现出巨大潜力, 特别是在酶分子定向进化方面。人工智能技术辅助酶分子设计为酶定向进化提供了预测指导, 为酶工程提供了新的设计策略和工具。酶分子定向进化的基本流程包括构建基因突变数据库、模拟突变条件和筛选理想酶蛋白突变体, 并提取其基因以实现酶的连续进化。本文介绍了酶分子定向进化的基本技术背景和原理, 重点阐述了人工智能在酶分子设计及定向进化领域的应用, 包括相关算法、数据库建立、线上平台技术路线, 并对该领域的未来发展进行了展望。

关键词: 酶工程; 定向进化; 人工智能; 机器学习; 蛋白质设计

Application and research progress of artificial intelligence in enzyme engineering

SUN Meng^{1#}, WANG Manlu^{1#}, LI Yuze², YUAN Bo², CHEN Chenxi²,WANG Lianghua³, CHEN Han^{3*}, SUN Mingjuan^{3*}¹First Students Brigade, School of Basic Medical Science, Naval Medical University, Shanghai 200433, China;²Fourth Students Brigade, School of Basic Medical Science, Naval Medical University, Shanghai 200433, China;³Department of Biochemistry and Molecular Biology, School of Basic Medical Science, Naval Medical University, Shanghai 200433, China)

Abstract: In recent years, with the rapid development of artificial intelligence (AI) technology, its application in the field of enzyme engineering has demonstrated tremendous potential, particularly in the directed evolution of enzyme molecules. AI-assisted enzyme molecule design provides predictive guidance for directed enzyme evolution, offering new design strategies and tools for enzyme engineering. The basic process of directed evolution of enzyme molecules includes constructing a gene mutation database, simulating mutation conditions, screening for ideal enzyme protein mutants, and extracting their genes to achieve continuous enzyme evolution. This work introduces the basic technical background and principles of directed evolution of enzyme molecules, focuses on the application of AI in the field of enzyme molecule design and directed evolution, including related algorithms, database establishment, and online platform technology routes, and the future development prospects.

Key Words: enzyme engineering; directed evolution; artificial intelligence; machine learning; protein design

收稿日期: 2024-08-16

基金项目: 国家自然科学基金面上项目(82173732)

#共同第一作者: 孙萌, E-mail: sm1844399597@163.com; 王曼璐, E-mail: KIMUAN@outlook.com

*通信作者: 孙铭娟, E-mail: sunmj@smmu.edu.cn; 陈涵, E-mail: nmuchenhan@foxmail.com

长期自然进化使得天然酶在原生环境下具有高效和稳定等特点, 但从生物体中提取的天然酶在体外通常很不稳定, 其生产成本高昂、应用局限性较多, 很少适合直接用于催化过程。为了获得具有理想特性的酶, 研究者开发了酶的定向进化技术。酶的定向进化技术模拟了酶的自然进化过程, 通过随机突变和高通量筛选, 以获得具有期望特性的酶突变体。近年来, 出现了高度依赖高通量策略的定向进化, 其以高效率和高选择性处理非天然底物, 并在商业可行的工艺条件下开发了多种新型生物催化。然而, 定向进化成本高、耗时长, 限制了生物催化在许多工业领域的潜在应用过程。如AA9家族的裂解性多糖单氧酶(lytic polysaccharide monooxygenases, LPMOs)尽管能够提取, 却难耐高温, 效率较低, 所以对高效耐热、可工业化酶的需求非常迫切。又由于利用筛选标记的遗传操作系统慢、周期长, 实验结果并不理想。

随着人工智能(artificial intelligence, AI)技术的兴起, 基于大量生物信息数据, AI辅助酶分子设计运用机器学习算法预测酶的属性 and 功能, 指导酶的理性设计和定向进化, 使得酶分子设计和优化更加高效和精确。研究者一般会通过序列比对和功能注释信息分析酶的催化位点和结合位点, 动力学模拟结合Lambda predict预测单个氨基酸突变的影响以及确定突变方向, 并通过高通量筛选出符合进化方向的突变体, 以期获得更强结合能力和更优催化活性的功能酶类。酶工程利用的人工智能辅助策略是由数据驱动的新策略。通过学习已有数据中的信息, 建立起输入属性(如基因序列)到输出属性(如发挥功能)的映射关系, 对物理或生物层面的基础信息需求量小, 从而大大减少了工作量^[1]。

1 酶分子定向进化概念的提出及酶工程的发展

20世纪60年代, 美国分子生物学家Sol Spiegelman使用RNA噬菌体Q首次在分子水平实现了对单一分子的定向改造。在1981年, Hall^[2]成功通过酶的定向进化提高了大肠杆菌k12的第二半乳糖苷酶的底物专一性, 得到了想要的突变酶。同样, 在1986年, Liao等^[3]通过对特定高温下的突变

库进行筛选, 得到了能在特定温度下稳定存在的突变酶。在1993年有学者首先提出“酶分子定向进化”这个概念, 也许最早提出使用易错PCR(error-prone PCR)的方法来改造天然酶。1994年, Stemmer等^[4]首创DNA改组技术(DNA shuffling), 这也是目前构建突变基因库的常用方法之一。除此之外, 基因突变库的建立方法还有SeSaM、StEP、ITCHY、RACHITT、SCRATCHY、PACE、STEMs等。

近几年来, 随着计算机技术和人工智能的发展, 通过算法构建突变基因库成为一种新的高效的方式, 如通过AI限制突变库大小, 以此来缩小搜索范围^[5]。结合酶工程的原理, AI技术通过算法能够预测和设计酶的突变体, 从而优化酶的稳定性、活性和底物特异性等。考虑到酶的定向进化成本高、耗时长等局限性, AI辅助酶设计可以为定向进化提供理论基础和预测指导, 同时定向进化为AI建模及预测提供了实验数据和验证机会。

2 常规酶分子定向进化的操作流程

常规酶分子定向进化是一个循序往复的过程, 如图1所示, 通过不断地构建酶基因位点上的随机突变和筛选所需的相关功能基因突变来获得人们所需要的酶分子。定向进化是一个迭代重复的过程, 首先选定初始目标蛋白, 向待改造酶的基因中引入突变, 使基因组多样化, 然后制定并运行高效蛋白表达和筛选策略, 对筛选得到的优良突变体基因进行多次循环的突变和筛选, 直到获得满意的性能, 如提高的酶活性、提高的抗体结合力、酶底物特异性的改变, 甚至创造全新的反应类型等。

2.1 酶基因突变库的建立

为获取更高效的酶, 首先需要构建出具有足够大样本量的酶基因突变库。在自然情况下, 由于基因的突变主要源于基因组复制过程中的缺陷, 以及紫外光等活性物质对DNA的损伤作用, 野生型基因通常以很低的突变率进行突变。为了获得大量不同点位的突变样本, 目前科学研究中普遍采用了以下几种方法来获得高质量的突变文库。

2.1.1 易错PCR

易错PCR是一种应用广泛的体外随机突变技

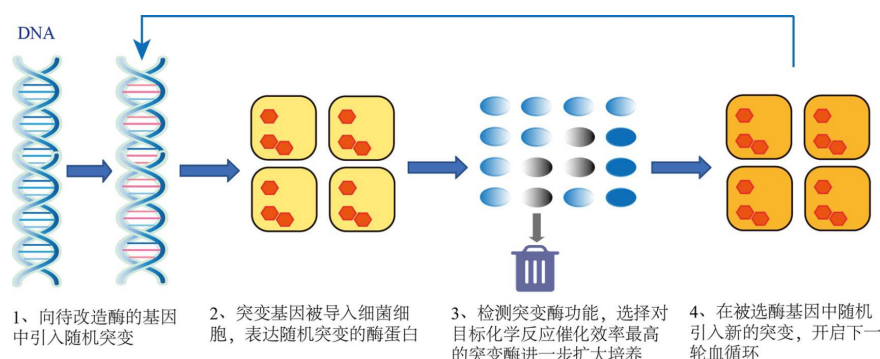


图1 常规酶分子定向进化操作流程示意图

术，通常是指利用或改善PCR反应条件，如金属离子 Mg^{2+} 含量，以及反应条件中的dNTP含量等，来调节PCR化学反应中的突变频率^[6]，从而大大降低了化学反应的保真性，使错误碱基对序列能以一个相对较高的频率添加到扩增基因中，通过多次循环就能得到多种不同点位突变的DNA群组。

2.1.2 交错延伸PCR技术

交错延伸技术PCR(overlap extension PCR)又称重叠延伸PCR技术，其通过重叠引物扩增实现基因突变，在分子克隆、蛋白质工程、基因编辑以及合成生物学中有广泛应用。在设计好包含突变的引物后，第一轮用一对引物来实现对目标基因片段的扩增，第二轮再使用一个引物与第一轮PCR产物的一个片段结合，另一个引物与另一个片段结合从而实现交错结合。利用PCR扩增将重叠片段连接成一个完整的突变基因，最后将其整合入载体，验证突变的有效性。交错延伸PCR技术可以高效、灵活、精准地实现多种基因和蛋白质的设计与优化，但由于其复杂性及相对较低的克隆效率使得其在应用中存在一定程度的局限。

2.1.3 体外随机引发重组技术

体外随机引发重组技术(*in vitro* random mutagenesis)是一种用于生成多样化基因变异并直接在体外验证突变效果的实验方法。其主要步骤包括：设计DNA模板、通过化学试剂或酶引入随机突变、PCR扩增突变后的DNA、将突变体克隆直接转入表达载体，筛选出具有特定表型及功能的突变体。体外随机引发重组技术操作简单、突变随机，可通过高通量筛选快速评估。但由于突变的不确定性以及产生的大量样本，导致该技术工作效率较低、筛选较为复杂，目前应用已不再

广泛。

2.1.4 DNA改组

通过脱氧核糖核酸酶 I (deoxyribonuclease I, DNase I)将目的基因切割为不同大小的片段，在PCR复性过程中，通过不同片段间具有的部分同源性发生错配，引入突变，通过重组使基因融合，从而创造出不同的突变子。通过反复的改组，还可以积累优势突变，富集正突变，去除负突变，提高突变文库的丰度，创造新基因和获得期望功能的蛋白质^[7]。

2.1.5 CRISPR/Cas介导的定向进化

CRISPR/Cas系统是一种原核生物的免疫系统，为了对抗外来基因的侵袭，为细胞创造了获得性免疫，该体系能够鉴定出外源DNA，并把它切断，从而沉默外源基因的表达。正因为这样精确的靶向功能，该系统被开发成一种高效的基因编辑工具，可以实现对靶向基因或区域进行编辑。根据小向导RNA(small guide RNA, sgRNA)引导的Cas蛋白或Cas融合蛋白的特性，可以使该区域基因产生双链断裂、单链缺刻和单碱基替换^[8]等变化。再由修复机制、错配机制或化学催化机制引入突变，实现突变文库的构建。一方面，可以通过CRISPR产生双链断裂进行突变库的建立，利用Cas9-sgRNA复合物对DNA双链的切割作用产生双链断裂(double-strand break, DSB)，然后一般通过非同源末端连接(non-homologous end joining, NHEJ)途径和同源性修复(homology directed repair, HDR)途径完成DNA修复。向大肠杆菌底盘系统、酵母底盘系统、植物与底盘系统的DNA基因组中引入多样性的突变^[9]。另一方面，CRISPR系统可与具有诱变功能的蛋白质联用，借助CRISPR系统对基因

组产生DSB后修复途径,但往往会对宿主本身造成损伤,并且受到需要提供HDR供体模板的限制;而利用Cas9的引导功能,进一步开发碱基编辑工具、基于碱基编辑的定向进化技术、Cas-DNA聚合酶定向进化技术可以解决这些问题^[10]。

2.2 高通量筛选

在获取足够量的样本之后,要采取的工作是对获得的不同点突变基因进行表达并筛选,得到表达更高效的酶。常规的突变文库在处理较大突变文库时费时费力,为此人们结合计算机技术,开发了新一代的高通量筛选平台,这大大增加了获得所需突变特性基因的机会,并减少了时间和成本。受体结合分析法、酶活性测定法、细胞分子测定法、细胞活性测定法、代谢物质测定法、基因产物测定法等方法,均已广泛用于药物高通量筛选中——以平板或微孔板形式作为实验工具载体,以智能的自动化操作系统执行实验过程,用计算机对数以千计的样品数据进行分析处理^[11],借助如光学测定技术、放射性检测技术、荧光检测技术等得出准确性较高的实验结果。

2.2.1 光学筛选技术

光学筛选技术是指通过光学方法(荧光、发光、吸收等)实现对细胞或分子的检测和分析,由于其具有高灵敏度和快速数据获取能力,适用于筛选表达特定基因或突变体的生物样品,目前在高通量筛选中应用广泛。

荧光筛选是最常见的光学筛选方法之一。通过在目标分子上标记荧光探针,研究人员可以在特定波长下高效识别出表达特定表型的样本。同时,结合微孔板技术和荧光暗背景筛选技术,自动化和基于荧光的高通量筛选可以提高筛选速度并缩小规模,与流式细胞荧光分选技术(flow cytometry sorting, FACS)相当。

发光筛选技术利用生物发光或化学发光反应,基于紫外/可见光谱生成可检测的信号,分为直接和间接检测方法,常用于评估酶活性、细胞活力或其他生物过程的变化。直接方法可以通过测量吸光度进行筛选,无明显吸光特征可通过加入金属离子螯合、与酶促或化学反应耦联来检测。这种方法在噪声背景较低的情况下,提供了高灵敏度的检测手段。

现代成像技术(如共聚焦显微镜和自动化成像系统)结合光学筛选,能够提供细胞或组织的高分辨率图像,进一步增强筛选的准确性。图像软件分析,可以定量评估细胞特征,识别出具有所需表型的样本^[12]。

2.2.2 流式细胞荧光分选技术

流式细胞荧光分选技术能检测并区分出活性明显改变或底物特异性显著改变的酶变体^[6],具有极高的敏感度和分辨率(10^8 突变体/天)。将酶活性表现型的编码基因偶联,使酶活性转化为可探测的荧光信息,在通过流式细胞仪探头时,仪器通过荧光强度为每个细胞加上不同的电荷,经过高压静电场的作用进行分选,从而达到将表达不同酶活性的细胞进行筛选分离的效果。

2.2.3 液滴微流控技术

在流式细胞荧光分选技术基础上,人们将液滴微流控技术与之结合(图2)。液滴微流控技术是在微观尺度通道中,利用流线剪切力和表面张力之间的相互作用,把连续流体切割分离成离散的纳升级或以下体积的液滴的一类微纳米工艺。利用该液滴微流控技术可以产生大量反应条件均一的单分散液滴,极大减小了试剂消耗,凭借10 kHz的产生频率,能够满足每天 10^8 量级的高通量筛选。通过向微液滴中加入特异荧光底物,根据反应前后荧光强度的变化达到对酶活性的定性分析,从而实现筛选工作^[13]。

3 人工智能在酶分子定向进化中的应用

为了满足生产需要,最早出现在酶分子定向进化中以改变酶分子性质的设计方法是酶分子的理性设计,一般指了解酶活性部位的结构和它们在酶功能中的作用,改变酶分子中个别氨基酸残基和结构域来改变酶的特异性^[14]。与理性设计不同,非理性设计不需要完全了解相关酶的结构和功能关系,依托于高通量筛选,在特定条件下鉴别突变酶的所需性质,使酶的快速设计成为可能。半理性设计介于非理性设计和理性设计之间,克服了两者的缺点,在对蛋白质的理化性质、三维结构、构效关系、催化机理等信息有一定理解的基础上,在计算机辅助下,对活性中心或者活性口袋的热点氨基酸进行定点突变、饱和突变、组合

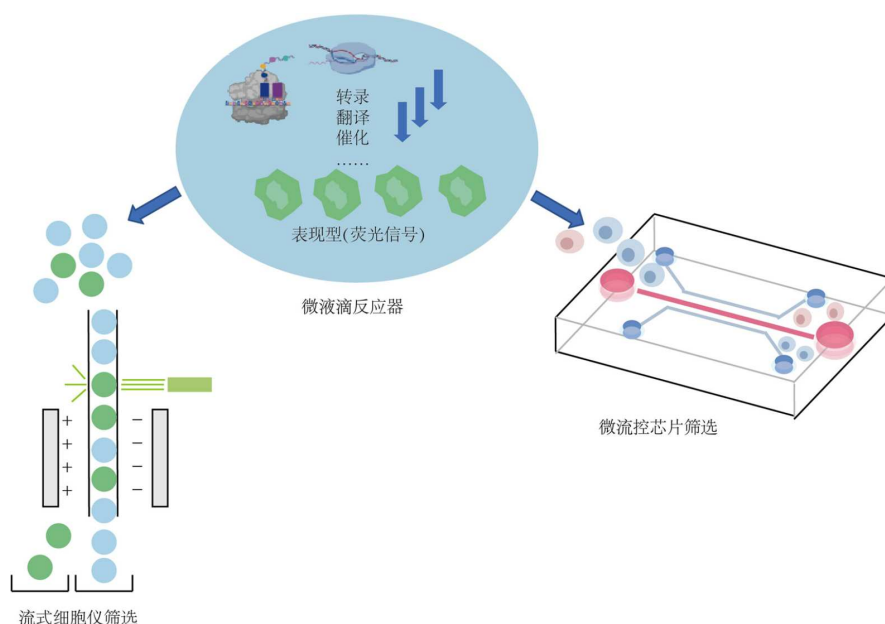


图2 液滴微流控技术示意图

突变技术^[15]。随着计算机技术的发展,人工智能在酶工程的分支用途越来越广泛。作为一种功能强大、高效和多用途的工具,机器学习及人工智能可以直接从数据中建立模型进行预测,该应用属于酶分子定向进化的技术方法。考虑到人工智能的辅助与半理性设计的适用条件相符且应用范围最为广泛,所以将其归类于半理性设计。

从被用于预测蛋白质二级结构到被用于预测蛋白质的高级结构及特征属性,机器学习算法已然成为从实验数据中学习并指导蛋白质设计的有利工具。“大数据”作为这些算法的训练集,有助于预测新的和改进的突变体,有助于在特定位点上诱变,从而加速蛋白质的定向进化^[16]。

针对酶分子定向进化的瓶颈——筛选工作量,人工智能助力酶定向进化的技术手段应运而生。其主要借助适当的计算机手段指导突变体及相关数据库的设计、构建预测模型、进行模型训练和评估。表1列举了近5年来机器学习指导蛋白质设计的应用实例,涉及蛋白质的热稳定性、催化活性、对映体选择性、光敏性及可溶性等多个方面^[17]。

3.1 人工智能辅助的酶分子设计策略概述

在人工智能辅助的蛋白分子设计策略中,通常是根据已有的数据,用多种机器算法构建“输入特性-输出特性”的映射联系。根据训练的实验数据是否具备标记信息(如目标蛋白的一个或多个属

性),机器学习一般可划分为监督学习和无监督学习^[27],在酶分子定向进化中的机器学习一般属于监督学习。其主要步骤为提取蛋白质特征属性后,将交付机器学习算法进行学习并生成可以描述数据模型的目标函数,并对蛋白质序列进行虚拟进化,通过训练和测试评估效能,最终给出预测结果。图3展示了机器学习指导酶分子设计的流程。其中有三个关键点:构建序列函数模型、分子描述符和机器算法。分子描述符和算法对预测模型的准确性影响较大,同时也是人工智能策略开拓的重点方向。

3.1.1 构建序列函数模型

机器学习一般分为经典机器学习和深度学习^[28]。前者较依赖于研究者提取的特征,可能会被设定好的特征值限制而忽略数据中重要的隐藏信息^[29]。而后者是通过深度神经网络,分层处理数据,能减少噪声干扰、挖掘隐藏信息,所以适用于从大量数据中发现复杂结构^[27],然后构建深度学习训练模型,将数据集分为训练集、测试集、验证集,分别用于构建序列函数模型、评价模型的性能、调整参数并选择最佳的模型结构。机器学习中的算法种类繁多且更新迭代速度快,如经典机器学习中常见的PLS回归^[30]、支持向量机^[31]、决策树/随机森林^[32]和贝叶斯网络^[33]等,以及深度学习中的变分自编码器、卷积神经网络和循环神

表 1 近5年机器学习指导蛋白质设计的应用实例

年份	蛋白名称	改造性能	算法及模型	参考文献
2021	C族和G族链球菌中抗体结合蛋白G	结合亲和力	贝叶斯优化	[18]
	酰基-ACP还原酶	催化活性	高斯过程(GP)	[19]
2022	卤代烷烃脱卤素酶和氟化酶	溶解度	深度神经网络ProteinMPNN	[20]
	PET水解酶	热稳定性、催化活性	MutCompute算法	[21]
2023	抑制剂厄洛替尼与表皮生长因子受体(EGFR)激酶结合的结构域PDB 1M17	蛋白质-配体相互作用	卷积神经网络、Alphaspace	[22]
	RhlA鼠李糖脂生物表面活性剂的关键生物合成酶	产物选择性	贝叶斯优化引导进化论(BO-EVO)算法	[18]
2024	生物活性螺旋肽高亲和力结合剂	结合亲和力	RFdiffusion、ProteinMPNN	[23]
	DprE1抑制剂	结合作用	深度神经网络、深度学习模型	[24]
	果胶裂解酶	热稳定性	AlphaFold2、RoseTTAFold	[25]
	SH-2蛋白酪氨酸磷酸酶2抑制剂	结合亲和力	XGBoost、SHAP分析	[26]

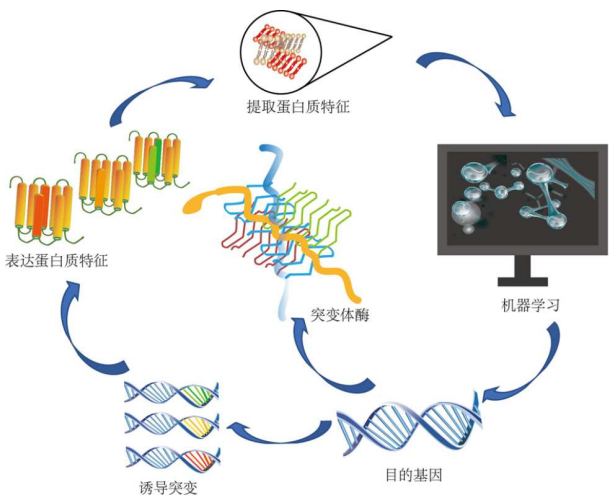


图3 人工智能辅助的酶分子设计策略示意图

神经网络等，这些算法均可辅助蛋白分子设计^[29]。

深度学习模型也被称为神经网络，堆叠多个由非线性激活函数连接的线性层，这使它们能够从结构化输入中提取高级特征。神经网络非常适合处理带有大量标记数据集的任务。它们已被应用于蛋白质-核酸结合、蛋白质-主要组织相容性复合物结合的结合位点预测，蛋白质-配体结合后的溶解度、热稳定性、亚细胞定位、立体结构等的预测。图4展示了选择建模蛋白质序列-函数关系算法的一般启发式。如果需要估计模型的不确定性，GPs是最简单的现成解决方案。除此之外，线性模型提供了一个简单的、可解释的基线。如果线性模型不够准确，随机森林、增强树或支持向量机对于<10 000个示例的数据集是有效的，而神经网络

一般在较大的数据集上提供最佳性能^[13]。

3.1.2 分子描述符

分子描述符是将分子的物理化学性质信息转换成计算机可处理的数字形式的工具。在机器学习指导酶分子定向进化中，关键是建立适用于具有各自特征的蛋白质序列的描述符体系。分子描述符一般有4种类型，包括基于氨基酸序列的描述符、基于结构的描述符、使用嵌入式的序列描述符以及突变指示描述符，表2列举了常用的描述符及其具体应用^[16]。最常见的独热编码描述符的工作原理为：多个酶突变体序列若在某一相同位点上包含多种不同的氨基酸，则可以用一个多维向量表示该目标位点的所有氨基酸，也可以根据氨基酸的物理性质(如电荷、体积等)进行编码，表示特定氨基酸序列^[27]。

Arnold团队使用高斯回归比较发现，使用嵌入式描述符训练的模型预测能力与独热编码、氨基酸性质或字符串核错配描述符的预测能力相近，甚至超过这些模型。因此，即使嵌入的维度数量级较低，它们仍能够实现准确的预测。蛋白质特征工程工具包和氨基酸指数描述符的预测能力表现较差。

3.1.3 算法

所有的基于人工智能的预测都需要借助算法。机器学习算法在解决蛋白质预测等复杂问题方面具有良好的应用前景。目前，虽然有多种机器学习算法，但无法保证一种算法在所有的学习任务中都是最优的。面对不同的问题，需要尝试不同

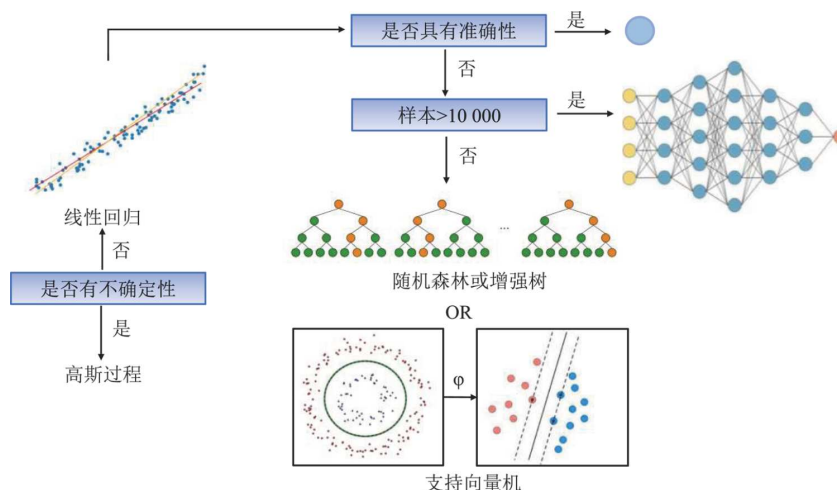


图4 深度学习模型流程示意图

表2 机器学习指导酶设计中常用的分子描述符

类别	描述	应用	参考文献
基于序列的描述符	谱核函数	氨基酸组成	[34]
	一维深度卷积神经网络	预测二级结构和溶剂可及性	[35]
	几何描述符	扭转角密度和氨基酸距离密度直方图	[36]
	联合三元组描述符	K间隔氨基酸对和联合三联体	[37]
	空间和化学特征	三维网格蛋白配体结构	[38]
	Protor包中的氨基酸组成描述符	蛋白质溶解度和活性	[39]
	识别	预测模型	[40]
	z-标度	生物活性肽的设计	[41]
	主成分评分疏水、立体和电子性质的矢量	肽序列结构变异性的参数化	[42]
	S-scales	二级结构预测	[43]
	蛋白质特征工具包(ProFET)	特征工程捕获	[44]
	拓扑标度	研究多肽定量构效关系	[45]
	结构拓扑标度	研究多肽及类似物定量序列活性模型	[46]
	氨基酸广义信息因子分析标度	研究阳离子抗菌肽定量的构效关系	[47]
	分子表面加权整体不变分子	预测两个肽组的活性	[48]
	蛋白质指纹图谱	分析结构活性	[49]
	BLOSUM矩阵	揭示氨基酸空间的基本结构和维数	[50]
	字符串核错配	预测蛋白质属性	[50]
	独热表示	预测蛋白质属性	[50]
	氨基酸物理化学性质AAIndex	预测蛋白质属性	[50]
基于结构的描述符	sPairs	预测蛋白质的三维结构	[51]
嵌入式	蛋白质向量	从结构蛋白预测无序蛋白	[52]
突变指示	突变识别	建立预测模型	[17]

的模型和算法来找到最优解。基于相似的功能特性，表3列举了机器学习常用的算法^[27]。

选择算法的一般准则是以线性模型作为基线，受复杂程度和所需训练集大小等因素影响^[17]，深

度学习时间可能会延长，从而影响计算速度。因此，在具体的预测任务中，需要全面考虑准确性、计算速率和预测难度等因素，再选择合适的算法。

表 3 机器学习常用算法列表

算法/方法	模型	特征	参考文献
贝叶斯算法	贝叶斯网络	为变量之间的多种关系建模(如线性、非线性、组合、随机)	[53]
贝叶斯优化引导进化算法	贝叶斯网络	指导多轮机器人实验, 探索组合诱变库的蛋白质适应度景观	[18]
基于实例方法	K近邻	假设每个实例对应于n维空间中的一个点	[54]
	径向基函数网络	对非线性分离的超曲面进行建模	[55]
	自组织映射神经网络	保留输入空间的拓扑特征	[56]
	支持向量机	适用于原始空间中非线性分离, 而无需在变换后的高维空间中定位每个点	[57]
决策树算法	决策树	用于对输入数据进行分类或预测值	[58]
集成算法	引导聚集算法	在同一数据集的不同子样本上训练多个模型, 降低预测的方差	[59]
	提升方法	将弱分类器融合之后形成一个强分类器, 减小预测模型偏差	[60]
	模型融合	训练一系列模型, 以加权方式线性组合这些模型, 得到最终预测模型	[61]
降维算法	主成分分析	无监督学习下通过转换矩阵, 将原始数据向量投影到新的特征空间中来提取特征	[62]
	线性判别分析	有监督学习下通过转换矩阵, 将原始数据向量投影到新的特征空间中来提取特征	[56]
	t-分布领域嵌入算法	将原始高维输入空间中识别的邻域信息保留在低维输出空间中	[57]
人工神经网络	深度神经网络	具有多个隐藏层的人工神经网络	[28]
	前馈神经网络	从输入端直接指向隐藏层, 然后再向前指向输出层, 而无需循环或反向连接	[57]
	循环神经网络	识别序列数据并对其建模	[63]
	玻尔兹曼机	每个神经元通过能量函数与其他所有神经元相连	[59]
	变分自编码器	提高他们从数据中提取有用特征的能力	[60]
	对抗自编码器	提高对数据特征的提取能力	[61]
	卷积神经网络	用于输入数据可以表示为图像或类似图像的情况	[62]
蒙特卡洛算法	随机采样模型	通过大量独立试验进行近似计算, 生成大量随机独立样本, 并基于大数定律和样本量的增加提高估计值的稳定性和准确性	[64]
	数值积分	用于数值积分的蒙特卡洛方法, 是用对应域的随机点数量来近似积分值, 适合高维空间的积分	[65]
	马尔可夫链(MCMC)	利用马尔可夫链来依次生成一连串相关的样本, 用于逼近复杂分布的样本	[66]

3.2 相关的数据库和线上平台

在人工智能时代, 数据库和在线平台是研究者不可或缺的资源, 它们为酶分子定向进化提供了丰富有力的数据支持和高效快捷的分析工具。数据库包含了大量关于蛋白质的功能、结构和突变效应等数据。在线平台则提供了从特征提取到预测模型构建的全套工具, 使得研究者能够利用人工智能算法对蛋白质序列进行分析和预测。综合利用这些数据库和在线平台, 不仅可以提高研究效率, 还能通过机器学习算法增强对蛋白质功能和特性的预测能力, 帮助研究者优化酶设计策略。

3.2.1 数据库

目前, 已有许多数据库收集并整理了来源于文献或实验的数据, 涉及蛋白质的序列、结构、功

能、突变稳定性和溶解度等特征属性。即使其中有大量的蛋白质序列缺少生物学信息功能注释, 也可以通过人工智能算法如深度学习从这些序列中挖掘隐藏信息, 高效提取蛋白质特征属性。这些数值指标可作为下一个步骤中从“已知特性”到“预期属性”的顶层预测模型的输入数据。有人工智能支撑的数据库可以存储大量的已知和未知的蛋白质信息, 优化了酶分子设计的方法途径。[表4](#)是常见数据库汇总表。

3.2.2 线上平台

除数据库外, 目前已有许多线上平台或工具包来帮助研究者们获得蛋白质序列中的特征属性以及使用人工智能算法的工具。[表5](#)是基于蛋白质序列的特征生成工具汇总表。

表4 常见数据库汇总表

名称	类型	数目/大小	特点
UniProtKB	蛋白质序列、功能信息、研究论文索引的蛋白质数据库	UniProtKB/Swiss-Prot包括560 000多条手动注释的蛋白质序列, UniProtKB/TrEMBL则包含了2亿多条自动注释的蛋白质序列	已有学者利用该数据库提供的大量蛋白质序列信息, 利用自然语言处理技术成功构建预测模型
Protein Data Bank	生物大分子三维结构的数据库	145 000多个来源于X射线单晶衍射、核磁共振、电子衍射等实验手段确定的蛋白质、DNA、RNA等生物大分子结构	该数据库为蛋白质结构预测模型构建提供了大量的初始数据
Pro ThermDB	蛋白质信息、结构信息、实验条件、文献信息和实验热力学数据库	>32 000条	突变体数据中突变类型包括野生单点突变和多点突变
FileProtDB	蛋白质稳定性数据的数据库	242个蛋白质的6 715个变体数据	手动管理, 仅包含单点突变体蛋白质数据
SoluProtMut DB	突变体蛋白质溶解度的数据库	917条突变数据	手动管理, 数据已经针对机器学习应用进行了整理
ProtaBank	蛋白质工程数据的数据库	700多种蛋白质的1 800 000多个突变体	手动输入, 不仅储存各种类型的突变信息, 还提供完整的序列信息

表5 基于蛋白质序列的特征生成工具汇总表

名称	主要功能	类型	参考文献
BioGPS	提供了许多用于生物信息学的工具, 包括蛋白质序列分析	网页平台数据库	[67]
美国国家生物技术信息中心(NCBI)	提供了大量的生物信息学资源和工具, 包括蛋白质序列数据库和BLAST等工具	网页平台数据库	[68]
欧洲生物信息学研究所(EBI)	提供了大量的生物信息学资源和工具, 包括蛋白质序列数据库和InterPro等工具	网页平台数据库	[69]
Swiss-Prot	提供高质量的蛋白质序列和人工注释信息	网页平台数据库	[63]
PseAAC	从蛋白质序列生成氨基酸的疏水性、亲水性、侧链质量、 α -COOH的pK值、 α -NH ₃ 的pK值以及25 °C时的P值6种特征网页平台	网页平台数据库	[70]
PROFEAT	从蛋白质序列生成多种氨基酸分子的结构和物理化学特征, 包含11种不同类型分子描述符	网页平台数据库	[71]
Pse-in-One	从蛋白质序列生成多种氨基酸分子的结构和物理化学特征, 包含8种分子描述符	网页平台数据库	[72]
POSSUM	从蛋白质序列生成21种基于PSSM矩阵的特征	网页平台数据库	[73]
Alphafold2.0	机器学习有关蛋白质结构的物理和生物学知识, 利用多序列比对, 融入深度学习算法的设计中	网页平台数据库	[74]
ExPASy	提供一系列工具(ProtParam、ProtScale、PeptideMass), 用于分析蛋白质序列的物理化学性质	网页平台数据库	[74]
SOLart	支持从特征提取、预测模型构建到性能评估的完整流程, 但是用户不能获得特征信息	网页平台数据库	[75]
BioGRED	以蛋白质为核心的互作数据库	网页平台数据库	[76]
propy	从蛋白质序列生成多种氨基酸分子的结构和物理化学特征, 包含13种不同类型分子描述符	Python工具包	[77]
protr/ProtrWeb	从蛋白质序列生成多种氨基酸分子的结构和物理化学特征, 包含22种分子描述符	R工具包/网页平台	[78]
Repi	从蛋白质序列生成多种氨基酸分子的结构和物理化学特征, 包含10种分子描述符	R/Bioconductor工具包	[79]
PSIPred	使用多种机器学习方法来预测蛋白质的二级结构。	Python库	[80]
PyMod	提供多种蛋白质评估模型质量的指标和可视化工具。	Python库	[81]

4 应用与未来的发展前景

目前, 基于人工智能辅助的酶分子进化已经逐步走向应用, 人工智能的计算能力与预测水平大大节省了人力及物力成本。与定向进化方法相比, 神经网络学习高维信息的能力逐渐提高, 学习对象从蛋白质基本的手工特征转换为以蛋白质序列

和结构为主的原生信息。同时, 高精度的预测模型可以更快地找到最优突变体, 从而减少实验周期。已有很多研究展示出生物信息学与实验结合带来的巨大优势, 且用途广泛。在合适的采样方法的引导下, 深度神经网络模型可以提供与酶的活性、热稳定性甚至是选择性等功能改造相关的

建议, 具有较高可信度。但作为一种数据驱动的计算方法仍有其局限性, 其预测结果受到训练集和采样策略的影响和限制。目前, 在预测精度和学习蛋白质上位性等方面人工智能技术还存在瓶颈, 期望在未来能构建出具备高泛化能力和快速采样能力的高性能神经网络模型^[82]。AI在酶分子进化模型的构建、突变酶体结构和功能的预测、药物及新酶的发现等领域应用广泛。在近几年的研究与应用中有较多案例利用机器学习技术来处理大量的生物信息学数据, 并通过预测模型来指导实验设计。以下列举几例应用案例。

4.1 基于逐步非线性回归的血管紧张素转化酶抑制肽QSAR建模——高活性

该技术通过不同子描述符表达肽序列, 利用SSNR进行特征选择, 最后使用ICM(模型)加权整合, 获得较为准确的性质预测, 考虑了不同分子描述符所具有的本质差异, 非线性回归模型SVR在整体上优于参比模型PLSR与MLR。基于该结果分析了高活性肽及其氨基酸的偏好性, 为人工合成潜在高活性ACE抑制肽提供了可能的序列组合^[83]。

4.2 金属酶靶向位点进化——结构预测

金属酶是含有一种或几种金属离子作为辅基的结合酶的统称, 广泛参与生理和病理过程, 是非常重要的药物作用靶标群。金属结合位点的电子和几何结构具有复杂性、多变性和动态性等特点, 计算机设计能省去大量人力、物力成本^[84]。Stepniewska-Dziubinska等^[85]考虑了活性位点处的金属离子, 提出了基于三维卷积神经网络的Pafnucy模型。该模型用于预测蛋白-配体的结合亲和力, 具有良好的预测效果。Jiménez等^[86]基于原子占有率提出KDEEP模型, 进一步提高了预测的准确性。随着人工智能的进一步发展, 基于图神经网络模型的靶标-配体结合亲和力预测方法还考虑了金属离子, 如Jiang等^[87]采用分子内图卷积模块、分子间图卷积模块、图池化模块等, 以复合物为输入提出Interaction GraphNet模型, 通过全连接输出预测结合亲和力。

4.3 基于机器学习构建嗜热菌丝体酶模型——酶约束网络代谢模型

基因组规模代谢模型(genome-scale metabolic model, GEM)是一种系统性描述细胞代谢过程的

数学模型。利用计算机辅助的建模方法进行所谓的“干实验”, 能够指导并部分替代那些耗时耗力或难以实施的“湿实验”, 从而显著提升研究效率, 已广泛应用于生物制造、生命健康等多个领域。但GEM模型存在缺乏细胞生理信息约束的短板, 为了整合各类酶信息, 酶约束模型(enzyme-constrained model, ecModel)应运而生^[88]。Wang等^[89]基于酶约束模型和原有的嗜热菌丝体酶(*Myceliophthora thermophila*)基因组规模代谢模型, 对原模型进行了更新和修正, 利用机器学习的 k_{cat} 数据在ECMpy框架内构建了一个酶约束的基因组规模代谢模型(ecGEM), 提高了嗜热菌丝体酶突变体代谢工程修改目标的预测准确性, 对准确修改其蛋白质结构和功能具有指导意义。

5 总结

目前, 工业化体系对工程酶的要求越来越高, 既要耐高温高压, 又要高效表达产物、无毒无害, 天然酶很难满足商品化需求, 于是对酶分子进行设计、改造及定向进化的研究发展迅速。随着人工智能技术的不断进步, 其在酶工程领域特别是在酶分子定向进化中起到了关键作用。机器学习和深度学习算法已经被广泛应用于酶的设计和优化中, 显著提高了酶分子改造的效率和精确度。同时, 各种软件算法、数据库平台的构建和共享使得定向进化的计算量和流程大大缩短, 具有广泛的应用前景。

定向进化技术为酶的改造提供了一种有效的途径, 但由于成本高昂、耗时长, 限制了其在工业领域的广泛应用。基于大数据和算法的优势, 人工智能的应用为解决这些问题提供了新的视角。AI辅助的酶设计不仅能够减少实验工作量, 还能够通过预测模型指导突变体的筛选, 通过多种算法计算为筛选提供相关的参考标准, 同时也能在一定程度上缩小误差, 加速获得理想特性的酶变体。其中最重要的两个技术是分子描述符和算法。分子描述符是连接蛋白质序列与功能特性的桥梁, 对于提高预测模型的准确性至关重要。综合运用多种算法, 包括但不限于高斯过程、贝叶斯优化、深度神经网络等, 这些算法在提高酶的热稳定性、催化活性、底物特异性等方面展现出巨大潜力。

人工智能在酶工程中的应用仍存在一些挑战。例如,预测模型的准确性受到训练集质量和采样策略的影响,在蛋白质上位性学习等方面还存在瓶颈。在未来的研究中,AI辅助的酶工程需要集中于开发更高效的算法、构建更全面的数据库,并提高模型的泛化能力,探索酶分子与生物系统相互作用的复杂性,以实现更深层次的酶功能改造,在工业生物催化和生物技术的发展中发挥更加广泛而深远的作用。

作者贡献声明:

孙 萌:参与论文选题和设计,搜集资料,起草及修改文章;

王曼璐:数据收集与分析,起草及修改文章;

李雨泽:参与论文选题和设计;

袁 波:图表制作,修改文章;

陈晨曦:论文编辑与校对;

王梁华:对文章的知识性内容作批评性审阅;

陈 涵:文章修改与校对;

孙铭娟:对文章的知识性内容作批评性审阅。

利益冲突声明: 本文不存在任何利益冲突。

参考文献

- [1] 张睿钦. 新型AA9LPMOs的性质研究、蛋白质工程改造及应用[D]. 济南: 山东大学, 2020
- [2] Hall BG. Changes in the substrate specificities of an enzyme during directed evolution of new functions. *Biochemistry*, 1981, 20(14): 4042-4049
- [3] Liao H, McKenzie T, Hageman R. Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. *Proc Natl Acad Sci USA*, 1986, 83(3): 576-580
- [4] Stemmer WPC. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, 1994, 370(6488): 389-391
- [5] Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. *Nat Methods*, 2019, 16(8): 687-694
- [6] 陈珏, 黄佳敏, 燕天鹤, 等. 随机突变文库构建与筛选研究进展. *生物工程学报*, 2021, 37(1): 163-177
- [7] Hess GT, Frésard L, Han K, et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat Methods*, 2016, 13(12): 1036-1042
- [8] Butt H, Zaidi SSA, Hassan N, et al. CRISPR-based directed evolution for crop improvement. *Trends Biotechnol*, 2020, 38(3): 236-240
- [9] 金帆, 李奉庭, 夏霖. CRISPR/Cas在定向进化技术中的应用. *集成技术*, 2021, 10(4): 33-49
- [10] Hess GT, Tycko J, Yao D, et al. Methods and applications of CRISPR-mediated base editing in eukaryotic genomes. *Mol Cell*, 2017, 68(1): 26-43
- [11] 崔金明, 刘陈立. 合成生物学中的高通量筛选与测量技术. *中国细胞生物学学报*, 2019, 41(11): 2083-2090
- [12] Zeng W, Guo L, Xu S, et al. High-throughput screening technology in industrial biotechnology. *Trends Biotechnol*, 2020, 38(8): 888-906
- [13] 马富强, 杨广宇. 基于液滴微流控技术的超高通量筛选体系及其在合成生物学中的应用. *生物技术通报*, 2017, 33(1): 83-92
- [14] 江正强, 杨绍青. 食品酶学与酶工程原理[M]. 北京: 中国轻工业出版社, 2018
- [15] 王国桢, 谢志鹏, 张建国. 定向进化和半理性设计改造顺式环氧琥珀酸水解酶. *微生物学通报*, 2023, 50(5): 1741-1755
- [16] 蒋迎迎, 曲戈, 孙周通. 机器学习助力酶定向进化. *生物学杂志*, 2020, 37(4): 1-11
- [17] Xu Y, Verma D, Sheridan RP, et al. Deep dive into machine learning models for protein engineering. *J Chem Inf Model*, 2020, 60(6): 2773-2790
- [18] Hu R, Fu L, Chen Y, et al. Protein engineering via Bayesian optimization-guided evolutionary algorithm and robotic experiments. *Brief BioInf*, 2023, 24(1): bbac570
- [19] Greenhalgh JC, Fahlberg SA, Pfleger BF, et al. Machine learning-guided acyl-ACP reductase engineering for improved *in vivo* fatty alcohol production. *Nat Commun*, 2021, 12(1): 5825
- [20] Vasina M, Vanacek P, Hon J, et al. Advanced database mining of efficient haloalkane dehalogenases by sequence and structure bioinformatics and microfluidics. *Chem Catal*, 2022, 2(10): 2704-2725
- [21] Lu H, Diaz DJ, Czarnecki NJ, et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature*, 2022, 604(7907): 662-667
- [22] Xia S, Chen E, Zhang Y. Integrated molecular modeling and machine learning for drug design. *J Chem Theor Comput*, 2023, 19(21): 7478-7495
- [23] Vázquez Torres S, Leung PJY, Venkatesh P, et al. *De novo* design of high-affinity binders of bioactive helical peptides. *Nature*, 626(7998): 435-442
- [24] Chen K, Xu R, Hu X, et al. Recent advances in the development of DprE1 inhibitors using AI/CADD approaches. *Drug Discov Today*, 2024, 29(6): 103987
- [25] Zhang Z, Li Z, Yang M, et al. Machine learning-guided multi-site combinatorial mutagenesis enhances the thermostability of pectin lyase. *Int J Biol Macromol*, 2024, 277: 134530
- [26] Adhikari N, Ayyannan SR. Development and validation of machine learning models for the prediction of SH-2

- containing protein tyrosine phosphatase 2 inhibitors. *Mol Divers*, 2024, 28(4): 1889-1905
- [27] 卞佳豪, 杨广宇. 人工智能辅助的蛋白质工程. *合成生物学*, 2022(3): 429-444
- [28] Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436-444
- [29] 曲戈, 朱彤, 蒋迎迎, 等. 蛋白质工程: 从定向进化到计算设计. *生物工程学报*, 2019, 35(10): 1843-1856
- [30] Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression). *Comput Stat*, 2010, 2(1): 97-106
- [31] Cortes C, Vapnik V. Support-vector networks. *Mach Learn*, 1995, 20(3): 273-297
- [32] Quinlan JR. Induction of decision trees. *Machine Learn*, 1986, 1(1): 81-106
- [33] Heckerman D. A tutorial on learning with Bayesian networks[M]. Berlin, Heidelberg: Springer, 2008: 33-82
- [34] Ismail HD, Saigo H, Kc DB. RF-NR: random forest based approach for improved classification of nuclear receptors. *IEEE ACM Trans Comput Biol Bioinf*, 2018, 15(6): 1844-1852
- [35] Hou J, Adhikari B, Cheng J, et al. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 2018, 34(8): 1295-1303
- [36] Zacharaki EI. Prediction of protein function using a deep convolutional neural network ensemble. *Peer J Comput Sci*, 2017, 3: e124
- [37] White C, Ismail HD, Saigo H, et al. CNN-BLPred: a convolutional neural network based predictor for β -lactamases (BL) and their classes. *BMC Bioinf*, 2017, 18(S16): 577
- [38] Ragoza M, Hochuli J, Idrobo E, et al. Protein-ligand scoring with convolutional neural networks. *J Chem Inf Model*, 2017, 57(4): 942-957
- [39] Han X, Ning W, Ma X, et al. Improving protein solubility and activity by introducing small peptide tags designed with machine learning models. *Metab Eng Commun*, 2020, 11: e00138
- [40] Yang KK, Wu Z, Bedbrook CN, et al. Learned protein embeddings for machine learning. *Bioinformatics*, 2018, 34(23): 4138
- [41] Sandberg M, Eriksson L, Jonsson J, et al. New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. *J Med Chem*, 1998, 41(14): 2481-2491
- [42] Mei H, Liao ZH, Zhou Y, et al. A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers*, 2005, 80(6): 775-786
- [43] Biou V, Gibrat JF, Levin JM, et al. Secondary structure prediction: combination of three different methods. *Protein Eng Sel*, 1988, 2(3): 185-191
- [44] Ofer D, Linial M. ProFET: feature engineering captures high-level protein functions. *Bioinformatics*, 2015, 31(21): 3429-3436
- [45] Tian F, Zhou P, Li Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *J Mol Structure*, 2007, 830(1-3): 106-115
- [46] Yang L, Shu M, Ma K, et al. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids*, 2010, 38(3): 805-816
- [47] Liang G, Li Z. Factor analysis scale of generalized amino acid information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides. *QSAR Comb Sci*, 2007, 26(6): 754-763
- [48] Zaliani A, Gancia E. MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. *J Chem Inf Comput Sci*, 1999, 39(3): 525-533
- [49] van Westen GJ, Swier RF, Wegner JK, et al. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *J Cheminform*, 2013, 5(1): 41
- [50] Georgiev AG. Interpretable numerical descriptors of amino acid Space. *J Comput Biol*, 2009, 16(5): 703-723
- [51] Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 1976, 9(6): 945-950
- [52] Asgari E, Mofrad MRK, Kobeissy FH. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*, 2015, 10(11): e0141287
- [53] Goan E, Fookes E. Bayesian neural networks: an introduction and survey. case studies in applied bayesian data science. 2020, arXiv: 2006.12024.
- [54] Zhang J, Ding Q, Li B, et al. Bidirectional k-nearest neighbor spatial crowdsourcing allocation protocol based on edge computing. *Peer J Comput Sci*, 2023, 9: e1244
- [55] Cui Y, Shi J, Wang Z. Development of quantum local potential function networks based on quantum assimilation and subspace division. *IEEE Trans Neural Netw Learn Syst*, 2018, 29(1): 63-73
- [56] Rougier NP, Detorakis GI. Randomized self-organizing map. *Neural Computation*, 2021, 33(8): 2241-2273
- [57] Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*, 2008, 9: 2579-2605
- [58] Yang X, Wang Y, Byrne R, et al. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev*, 2019, 119(18): 10520-10594
- [59] Borstelmann SM. Machine learning principles for radiology investigators. *Academic Radiol*, 2020, 27(1): 13-25

- [60] Kingman A, Shlens J, Jaitly N, et al. Auto-encoding variational bayes. 2013, arXiv, [1312.6114](#)
- [61] Makhzani A, Shlens J, Jaitly N, et al. Adversarial autoencoders. arXiv, 2015: 1511.05644
- [62] 池燕飞, 李春, 冯旭东. 机器学习在蛋白质功能预测领域的研究进展. *生物工程学报*, 2023, 39(6): 2141-2157
- [63] Varadi M, Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*, 2023, 23(17): e2200128
- [64] Metropolis N, Rosenbluth AW, Rosenbluth MN, et al. Equation of state calculations by fast computing machines. *J Chem Phys*, 1953, 21(6): 1087-1092
- [65] M Sobol' I. On the distribution of points in a cube and the approximate evaluation of integrals. *Russ Math Surv*, 1967, 7(4): 86-112
- [66] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970, 57(1): 97-109
- [67] Gao H, Wang ST, Wang H, et al. Characterization and exploration of an artifact in the reducing capillary electrophoresis-sodium dodecyl sulfate analysis of the 'me-too' drug zuberitamab related to rituximab. *J Pharm BioMed Anal*, 2023, 228: 115347
- [68] Kouskoumvekaki I, Shublaq N, Brunak S. Facilitating the use of large-scale biological data and tools in the era of translational bioinformatics. *Brief Bioinf*, 2014, 15(6): 942-952
- [69] Sayers EW, Beck J, Bolton EE, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 2021, 49(D1): D10-D17
- [70] Ribeiro AJM, Riziotis IG, Borkakoti N, et al. Enzyme function and evolution through the lens of bioinformatics. *Biochem J*, 2023, 480(22): 1845-1863
- [71] Meng L, Chan WS, Huang L, et al. Mini-review: recent advances in post-translational modification site prediction based on deep learning. *Comput Struct Biotechnol J*, 2022, 20: 3522-3532
- [72] Li ZR, Lin HH, Han LY, et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*, 2006, 34(Web Server): W32-W37
- [73] Liu B, Liu F, Wang X, et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*, 2015, 43(W1): W65-W71
- [74] Wang J, Yang B, Revote J, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics*, 2017, 33(17): 2756-2758
- [75] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596(7873): 583-589
- [76] Hou Q, Kwasigroch JM, Rooman M, et al. SOLart: A structure-based method to predict protein solubility and aggregation. *Bioinformatics*, 2020, 36(5): 1445-1452
- [77] Cao DS, Xu QS, Liang YZ. Propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, 2013, 29(7): 960-962
- [78] Xiao N, Cao DS, Zhu MF, et al. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, 2015, 31(11): 1857-1859
- [79] Cao DS, Xiao N, Xu QS, et al. Rcp: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*, 2015, 31(2): 279-281
- [80] Kashani-Amin E, Tabatabaei-Malazy O, Sakhteman A, et al. A systematic review on popularity, application and characteristics of protein secondary structure prediction tools. *Curr Drug Discov Technol*, 2019, 16(2): 159-172
- [81] Janson G, Zhang C, Prado MG, et al. PyMod 2.0: improvements in protein sequence-structure analysis and homology modeling within PyMOL. *Bioinformatics*, 2017, 33(3): 444-446
- [82] 康里奇, 谈攀, 洪亮. 人工智能时代下的酶工程. *合成生物学*, 2023, 4(3): 524-534
- [83] 周恒, 巴庆芳, 袁哲明, 等. 基于逐步非线性回归的血管紧张素转化酶抑制肽QSAR建模. *化学通报*, 2022, 85(6): 736-745
- [84] 余俊霖, 李国波. 计算机辅助金属酶靶向药物发现的研究进展. *中国现代应用药学*, 2022, 39(21): 2828-2833
- [85] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P, et al. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics*, 2018, 34(21): 3666-3674
- [86] Jiménez J, Škalič M, Martínez-Rosell G, et al. K_{DEEP} : protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model*, 2018, 58(2): 287-296
- [87] Jiang D, Hsieh CY, Wu Z, et al. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions. *J Med Chem*, 2021, 64(24): 18209-18232
- [88] Chen Y, Gustafsson J, Tafur Rangel A, et al. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. *Nat Protoc*, 2024, 19(3): 629-667
- [89] Wang Y, Mao Z, Dong J, et al. Construction of an enzyme-constrained metabolic network model for *Myceliophthora thermophila* using machine learning-based kcat data. *Microb Cell Fact*, 2024, 23(1): 138