DOI: 10.13345/j.cjb.230748
Jun. 25, 2024, 40(6): 1728-1741

·综 述·

人工智能辅助的酶分子改造应用进展

徐沛[1]，汪卫华[1]，宁洪伟[3]，曹瑞芬[3]，刘胜[1]，范培锋[4]，宋小平[2*]

1 安徽大学物质科学与信息技术研究院，安徽 合肥 230061
2 安徽医学高等专科学校药学院，安徽 合肥 230061
3 安徽大学计算机科学与技术学院，安徽 合肥 230061
4 安徽大学物理与光电工程学院，安徽 合肥 230061

徐沛, 汪卫华, 宁洪伟, 曹瑞芬, 刘胜, 范培锋, 宋小平. 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 Pei, WANG Weihua, NING Hongwei, CAO Ruifen, LIU Sheng, FAN Peifeng, SONG Xiaoping. Progress in the application of artificial intelligence-assisted molecular modification of enzymes[J]. Chinese Journal of Biotechnology, 2024, 40(6): 1728-1741.

摘 要：天然酶在活性、对映体选择性或热稳定性等方面经常难以满足应用与研究的需求，探索高效的酶分子改造技术改善该类酶的某些特性是酶工程的重要任务。酶分子改造技术主要包括理性设计、定向进化和人工智能辅助设计等。定向进化和理性设计是由实验驱动的酶分子改造策略，已经成功地应用于酶工程，但由于蛋白质序列空间的尺寸巨大以及实验数据少，现行的酶分子改造方法仍然面临着重大挑战。随着新一代测序、高通量筛选方法、蛋白质数据库和人工智能技术的发展，数据驱动的酶工程有望应对这些挑战。其中，采用人工智能辅助的统计学习方法，通过数据驱动方式构建序列/结构-酶性能的预测模型，依据预测模型挑选优良突变酶，大大提高了酶分子改造效率。基于酶分子改造的应用需求，本文综述了人工智能辅助酶分子改造的数据采集方法以及人工智能辅助酶分子改造的应用实例等，重点叙述了采用卷积神经网络预测蛋白质热稳定性的方法，以期为该领域的研究人员提供参考。

关键词：人工智能；定向进化；酶分子改造；卷积神经网络

资助项目：国家自然科学基金(62373001)；安徽省自然科学基金(1808085MC86)；安徽高校教师自然科学研究重点项目(2022AH052316，2023AH050089)
This work was supported by the National Natural Science Foundation of China (62373001), the Natural Science Foundation of Anhui Province (1808085MC86), and the Key Research Project of Natural Science for University Teachers in Anhui Province (2022AH052316, 2023AH050089).
*Corresponding author. E-mail: sxp20081012@sina.com
Received: 2023-10-30; Accepted: 2024-03-15; Published online: 2024-06-07

Progress in the application of artificial intelligence-assisted molecular modification of enzymes

XU Pei[1], WANG Weihua[1], NING Hongwei[3], CAO Ruifen[3], LIU Sheng[1], FAN Peifeng[4], SONG Xiaoping[2*]

1 Institutes of Physical Science and Information Technology, Anhui University, Hefei 230061, Anhui, China
2 School of Pharmacy, Anhui Medical College, Hefei 230061, Anhui, China
3 School of Computer Science and Technology, Anhui University, Hefei 230061, Anhui, China
4 School of Physics and Optoelectronic Engineering, Anhui University, Hefei 230061, Anhui, China

Abstract: Natural enzymes are often difficult to meet the needs of application and research in terms of activity, enantiomer selectivity or thermal stability. Therefore, it is an important task of enzyme engineering to explore efficient molecular modification technologies to improve the properties of such enzymes. The molecular modification technologies of enzymes mainly include rational design, directed evolution, and artificial intelligence-assisted design. Directed evolution and rational design are experiment-driven molecular modification approaches of enzymes and have been successfully applied to enzyme engineering. However, due to the huge space sizes of protein sequences and the lack of experimental data, the current modification methods still face major challenges. With the development of next-generation sequencing, high-throughput screening, protein databases, and artificial intelligence (AI), data-driven enzyme engineering is emerging as a promising solution to these challenges. The AI-assisted statistical learning method has been used to establish a model for predicting the sequence/structure-properties of enzymes in a data-driven manner. Excellent mutant enzymes can be selected according to the prediction results, which greatly improve the efficiency of molecular modification. Considering the application requirements of molecular modification of enzymes, this paper reviews the data acquisition methods and application examples of AI-assisted molecular modification of enzymes, with focuses on the convolutional neural network method for predicting protein thermostability, aiming to provide reference for researchers in this field.

Keywords: artificial intelligence; directed evolution; molecular modification of enzymes; convolutional neural network

酶是由活细胞产生的、对其底物具有高度特异性和高度催化效能的蛋白质或 RNA，大多数酶化学本质是蛋白质，故也称酶蛋白[1]。大量研究发现，天然酶在稳定性、选择性、耐受性等方面常常难以满足工业或实验室的需求，需要对天然酶进行分子改造，来提高其特定性能以满足实际应用需求，这也是当前酶工程的一项重要任务[2-4]。

根据技术原理的不同，现有酶分子改造技术可分为基于蛋白质工程技术的传统酶分子改造与人工智能辅助的数据驱动酶分子改造两种方法[5]。其中，传统酶分子改造方法主要包括定向进化与理性设计[6-7]。利用定向进化与理性设计技术，可在实验室模拟天然酶的自然进化

☏：010-64807509 ✉：cjb@im.ac.cn

过程，通过对目的基因进行多轮突变、表达与筛选，迭代累积有益突变体以分离出期望性能的酶突变体[8]。虽然定向进化和理性设计在酶工程中均取得了显著成效，但在生物催化剂分子改造过程中，这些技术都需要进行大量的计算或开展实验筛选工作[9]。因此，想要实现高效的酶分子改造，提高酶分子进化高度，还需引入更为先进的技术来指导改造过程。

面对突破传统酶分子改造效率低的技术瓶颈，人工智能(artificial intelligence, AI)为酶工程提供了新的技术手段，并在该领域得到了有效利用与发展，且逐渐成为酶工程领域的研究热点[9,10-12]。人工智能辅助酶工程有两大任务：酶功能预测和酶分子改造。目前，学者们基于机器学习并采用公开数据集，已成功解决蛋白质的溶解度预测、功能预测、稳定性预测和结构预测等预测类问题[13-16]。迄今为止，公开的蛋白质数据库中已经收集了数百万条蛋白质序列、数十万个蛋白质结构、数千万个生物物理值以及数百个带注释的催化机制，而且数据还在不断丰富，为人工智能指导的酶工程中机器学习算法奠定了坚实的数据基础[5]。

尽管采用公开数据集，机器学习方法已成功解决酶功能预测类问题，然而这些研究并不涉及酶分子改造。两者之间的共同点是依赖数据构建预测模型，获得蛋白质编码结果与功能(或改造特性)之间的映射关系(即预测模型)，主要区别在于功能预测只需要依据预测模型给出预测结果，而酶分子改造需要依据预测模型挑选优良突变体作为下一步的实验验证对象[5,9,17]。基于机器学习的监督学习，研究人员可在无生物学或物理学等先验知识的情况下，仅通过氨基酸序列特征提取、预测模型构建与验证对象筛选 3 个步骤便可以实现酶分子改造[18]。然而，机器学习指导酶分子改造的根本困难是可利用样本少，导致从实验中收集到的数据量少。因此，在实验数据样本少的情况下，构建泛化效果好的预测模型便成为酶分子改造的一个关键环节，直接影响改造效率。基于酶分子改造的应用需求，本文将围绕传统酶分子改造中的技术瓶颈，就人工智能辅助的数据驱动预测模型构建及其应用进行综述，以期为利用人工智能开展酶分子改造的研究人员提供帮助。

1 定向进化与理性设计

蛋白质一级序列中离散的氨基酸具有高度的进化相关性，因此是基于传统酶分子改造方法的主要编辑改造对象[9]。定向进化通过模拟自然选择过程，利用易错 PCR (error-prone PCR)、交错延伸(staggered extension process)和 DNA 改组(DNA shuffling)等技术对目标基因进行多轮突变，每轮获得的最优突变序列都将作为下一轮改造的模板，从而不断累积有益突变，直至筛选到优良的突变体[3,19-20] (图 1)。定向进化策略因不需要了解目标酶蛋白的结构信息与催化机理就可实现改造酶的目的，广泛用于酶分子改造并取得了很大的成功。Arango 等[21]采用定向进化技术降低了葡萄糖氧化酶对氧气约 97%的依赖并提高了 5.7 倍酶活。Chen 等[22]利用易错 PCR 技术对枯草杆菌蛋白酶 E (subtilisin E)经过 3 轮诱变和筛选，最终得到了比野生酶活性提高了 256 倍的 6 位点突变体，成功实现了枯草杆菌蛋白酶 E 的进化。Stemmer[23]采用定向进化技术对 β-内酰胺酶(β-lactamase)进行了 3 轮改组(shuffling)和两轮回交(backcrossing)，得到的突变体对头孢霉素的抑制浓度比野生型提高了 16 000 倍。定向进化通过对酶分子的改造，极大地促进了代谢工程、酶工程以及医药等领域的发展，是改善蛋白质性能最有前途的方法之一[8]。然而，该策略应用成功的前提是先进

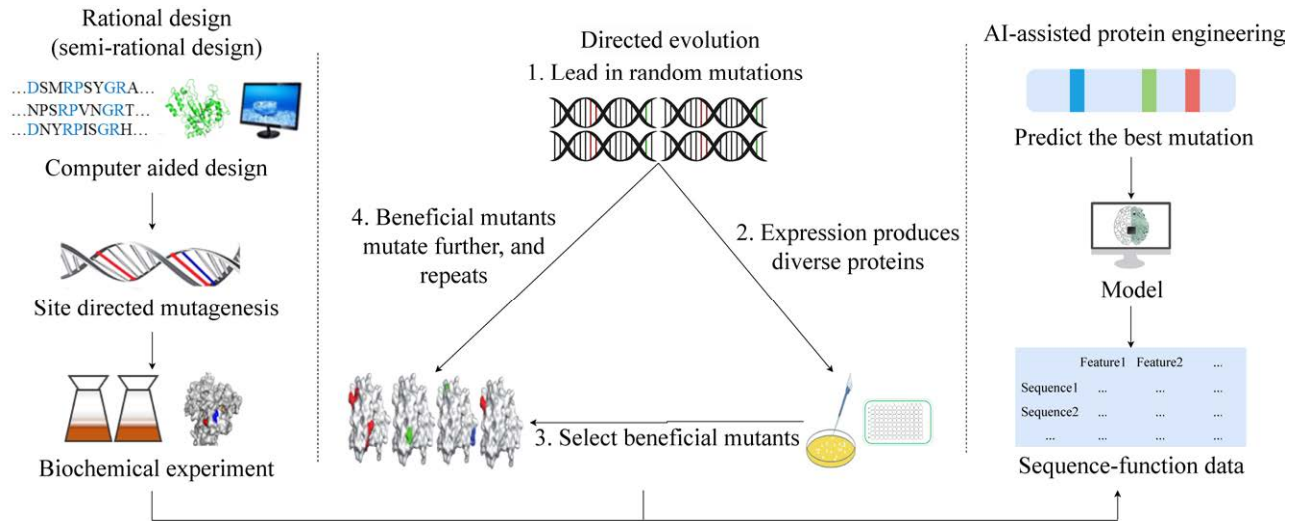

图 1　定向进化、理性设计和人工智能辅助的酶工程策略示意图

Figure 1　Schematic diagram of enzyme engineering strategies assisted by directed evolution, rational design and artificial intelligence.

的定向进化技术和易行通用的高通量筛选方法，而对于酶分子改造来说，易行通用的高通量筛选实验的设计仍然是一个挑战，而且多轮突变和迭代筛选会导致过长的实验周期，这对于分子生物学实验来说是沉重的负担。因此，定向进化仍需要更为先进的技术指导实验过程，以降低酶分子改造的实验成本，提高酶分子改造效率[5,8]。

与定向进化不同，理性设计是基于蛋白质空间结构及其功能区域信息，在计算机的辅助下，运用分子对接、分子动力学模拟、第一性原理计算等方法，计算单个或者多个突变给蛋白质结构和功能带来的具体影响[9,24]。基于计算结果，研究人员通过预测并评估突变体在结构、自由能、底物结合能等方面的变化，选择较少的关键位点进行精准改造，就可以构建小而精的突变文库，从而大幅度降低了定向进化策略中筛选突变体的实验工作量，且筛选过程不再需要高通量筛选方法[24](图 1)。例如采用理性设计对谷氨酰胺转氨酶的酶分子活性位点和末端进行改造，最终得到的突变体比酶活提高了 50%[25]。采用理性设计对吸水链霉菌来源的谷氨酰胺转氨酶的前导肽 C 端、成熟酶的 N 端做了改造，最终突变体相比野生型在 50 ℃和 60 ℃的半衰期分别提高了 3.0 倍和 2.1 倍[26]。通过理性设计对甲酸脱氢酶结构进行模拟、计算和预测，在 I 239 位引入半胱氨酸突变与 C262 构成二硫键，获得的突变酶在 60 ℃处理 3 min，相对酶活仍可保持 60%以上[27-28]。尽管理性设计在酶分子改造中已经取得了一定效果，但该策略依然面临着难以获取高精度的酶分子空间构象、缺乏对结构-功能关系与催化机理的理性认知以及过分依赖计算机资源等缺点[21,28]。

半理性设计是一种介于定向进化与理性设计之间的蛋白质改造策略，该策略克服了两者的缺点，降低了技术需求[5,29]。半理性设计策略主要借助生物信息学方法，基于同源蛋白序列比对、三维结构或已有知识，识别潜在有益替代氨基酸位点作为改造靶点，结合单密码子饱和突变(single code saturation mutagenesis, SCSM)、双密

码子饱和突变(double code saturation mutagenesis, DCSM)和三密码子饱和突变(triple code saturation mutagenesis, TCSM)等半理性设计新方法，对多种生物催化剂的立体/区域选择性及催化活力等多参数的改造[30-31]。半理性设计通过对特定靶点进行组合突变，构建“小而精”的高质量突变体文库，不仅提高了虚拟筛选优良突变体的效率，还大大缩小了突变空间的规模，更有利于在采用该策略完成有益位点的选择后用数据驱动的方式指导其后续实验过程，以降低酶分子改造对已有知识的依赖以及改造过程中的成本投入[5,31]。

2 人工智能辅助的酶分子改造应用

随着算法、算力(计算能力)与算料(数据)的发展，人工智能方法在酶分子改造中的应用越来越广泛。目前已有基于序列、结构等多种特征，使用经典机器学习和深度学习对酶分子改造的实例，减少了实验室筛选的工作量，避免了有益突变的损失，加速了定向进化进程。本节就人工智能所依赖的数据采集、经典机器学习辅助酶分子改造及深度学习辅助酶分子改造结合具体应用案例进行综述。

2.1 人工智能技术与数据采集方法

人工智能技术对数据有着高度依赖，初始数据的数量和质量决定了模型的精准性和泛化能力[32-33]。数据集样本数量不足或者质量过低将会导致模型出现欠拟合，即模型无法很好地拟合训练数据的真实模式和规律[34]。反之则会出现过拟合，即模型在训练数据上表现得很好，但在测试集中表现较差[35-36]。因此，初始数据的采集是重要且耗时的一步。

目前，学者们基于人工智能技术，采用UniProt、PDB 等公开数据库已成功解决了 GT1 家族糖基转移酶(glycosyltransferase, GTase)、P450 酶、聚对苯二甲酸乙二酯水解酶(poly-ethylene terephthalate hydrolase, PETase)、肌酸酶、谷胱甘肽转移酶、环氧水解酶等多种酶功能预测问题[37-42]。然而，不同于酶功能预测问题，酶分子改造公开的数据较少，因此酶分子改造无法从公开数据集中收集到训练数据，只能从定向进化或理性设计后对酶的改造特性实验中获取的数据作为人工智能策略的初始数据[13,15]。在实际的酶分子改造中，有限的实验资源导致可利用的样本数目非常少，缺少大量的可用于训练和验证的标准化数据；实验中性能不佳的蛋白质突变体数据常被丢弃，缺乏阴性数据，因而导致初始数据不均匀[16]。因此，小样本问题是基于人工智能的酶分子改造不得不面临的一大挑战。

为了解决数据不足问题，研究人员提出了各种方法生成新的数据。2002 年，Chawla 等[43]提出了一种经典的过采样方法(synthetic minority over-sampling technique, SMOTE)，该方法主要是通过合成少数类样本来增加其在数据集中的数量，以达到样本平衡；通过 SMOTE 过采样，可以使得模型更好地学习到少数类别的特征，从而提高模型的泛化能力和准确性。2004 年，Zhou 和 Jiang[44]提出了基于神经集成的 C4.5 决策树方法，该方法首先训练神经网络，然后利用它生成新的训练集。Li 和 Lin[45]于 2006 年提出了一种基于内化核密度估计的虚拟样本生成方法，该方法首先确定样本的概率密度函数，然后使用该函数生成训练样本。然而，这些方法无法利用样本的固有特征，导致训练模型的局限性[45-46]。近年来，随着新一代人工智能和大数据的快速发展，基于深度神经网络(deep neural network, DNN)的生成式对抗网络(generative adversarial network, GAN)的使用为

创建解决数据问题的新方法提供了机会，这也使得深度学习在小样本情况下的应用成为可能[47-48]。GAN 是 Goodfellow 等[48-50]在 2014 年提出的一种功能强大的生成模型，它可以用来生成与真实数据具有相同分布的合成样本，以解决标注数据不足的问题。GAN 由生成器和鉴别器 2 个深层架构函数组成，在学习过程中，生成器捕获真实数据的潜在分布并生成合成样本，而鉴别器则尽可能准确地区分真实样本和合成样本[41,45]。从训练数据不足的监督学习角度来看，GAN 是一种过采样方法，有研究表明，在提高数据质量方面，GAN 的数据增强比 SMOTE 过采样方法更有效[44,50]。

2.2 经典机器学习辅助酶分子改造

机器学习因其相对简单的模型结构、较快的运行速度以及对较小数据库的包容性，已经成功地应用于新酶的分类、酶或其底物属性的预测、反应最佳微环境的预测等[49-53] (表 1)。机器学习高度依赖于人工提取的特征，一般与基于氨基酸特征或序列整体特征的分子描述符配套使用，但可能会受限于定义好的特征值，而忽略数据中隐藏的信息[3]。在选择算法时，一般会选择线性回归模型作为比较基准，当数据量小于 10^4 时，可以选择机器学习算法，如偏最小二乘回归、支持向量机、决策树/随机森林和贝叶斯网络等常见算法[67-72]。

基于机器学习的酶分子改造过程如图 2 所示，其中预测模型构建和优化是最重要过程，将酶的序列或结构信息作为输入信息，突变酶与天然酶的改造特性实验值之比作为突变体的量化输出，可获得酶编码结果与改造特性之间的映射关系。但需要指出的是，机器学习指导酶分子改造的根本困难是可利用的样本少。因此，构建泛化效果好的预测模型便成为酶分子改造中的一个关键(图 2)。现有研究倾向于采用统计学习算法构建预测模型，如采用线性模型构建了 K 蛋白质酶的活性预测模文献[73]；基于支持向量机构建了环氧水解酶的立体选择性预

表 1 人工智能在酶工程的应用

Table 1 Application of artificial intelligence in enzyme engineering

Models	Algorithms	Input types	Applications	References
CWLy-SVM	SVM	Sequence	Identification of cell wall catalytic enzymes	[36]
SVR	SVM	Sequence	Enhance enzyme activity and solubility	[54]
Innov'SAR	PLSR	Sequence	Finding the best combination of mutations to increase activity	[55-57]
ProSAR	PLSR	Sequence	Enhancement of halohydrin dehalogenase activity	[57]
–	GP	Sequence	Modifying the fluorescence of green fluorescent protein	[58]
TOME	RF	Sequence	Predicting the optimal temperature	[59-60]
GOLabeler	LTR	Sequence	Predicting the function of unknown proteins	[61]
DeepGOPlus	CNN	Sequence	Predicting protein function using protein sequence and structure information	[32]
GAT-GO	GAT	Sequence	Predicting protein function from PPI networks	[62]
ESM-1v	Transformer	Sequence	Predicting protein function under unsupervised learning	[63]
DeepFRI	GCN	Sequence and structure	Predicting protein function by learning the features of protein sequences and contact maps	[64]
DeepGraphGO	GCN	PPI network	Using protein data from all species in a single model to predict protein function	[65]
–	CNN	Sequence	Prediction of protein thermostability	[66]

SVM: Support vector machine; PLSR: Partial least square regression; GP: Gaussian progress; RF: Random forest; LTR: Learning to rank; CNN: Convolutional neural network; GAT: Graph attention network; GCN: Graph convolutional network.

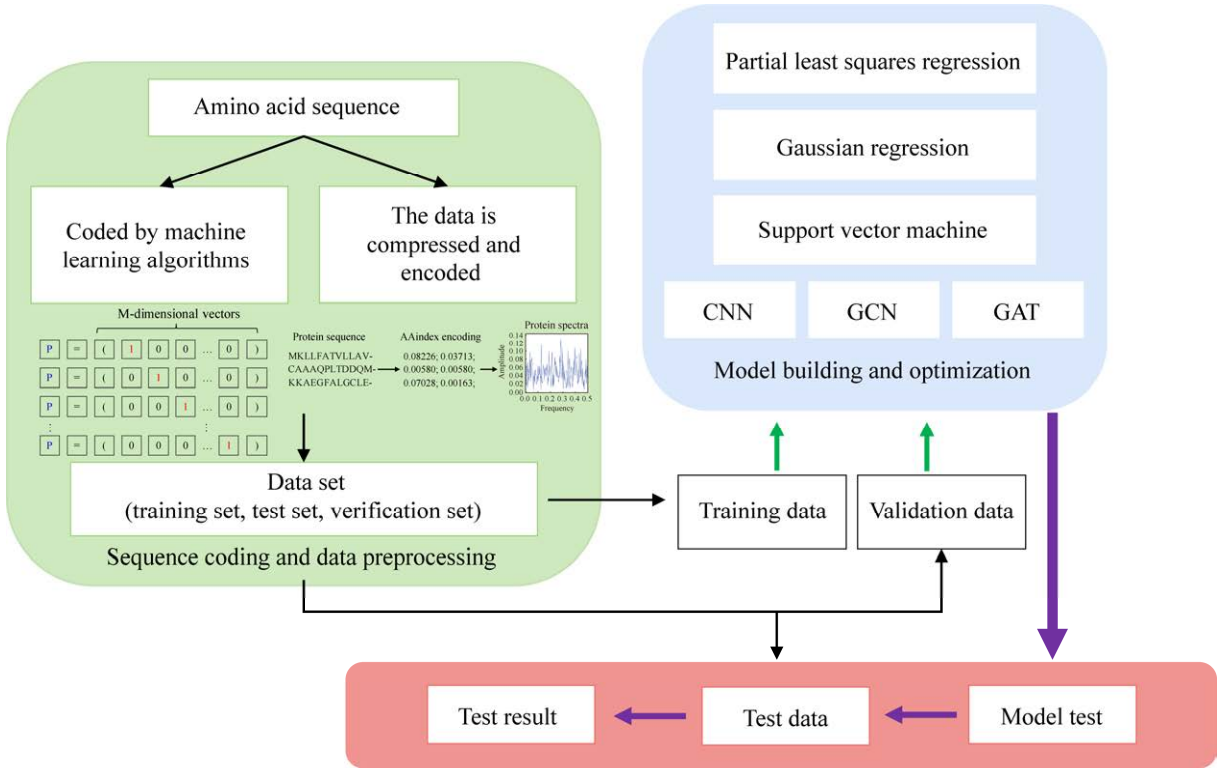

图 2　基于机器学习的酶分子改造过程

Figure 2　Enzyme molecular modification process based on machine learning.

测模型[42]；使用高斯过程回归构建了 P450 酶的热稳定性预测模型[38]以及绿色荧光蛋白质的荧光性预测模型[58]。其中 Romero 等[38]使用高斯过程回归算法对 P450 酶进行热稳定性改造，对机器学习算法选择的最优突变体进行实验测试，最终其热稳定性较野生型提高了 8.7 ℃。Cadet 等[56]使用偏最小二乘回归来寻找具有改进立体选择性的环氧水解酶，仅对 38 个序列-标签数据进行建模，然后对 9 个位点上共 512 种突变体进行预测，并选择了几个预测性能较好的突变体进行实验测量，结果明显优于野生型。上述应用实例展示了机器学习方法的实用性，特别适用于在定向进化实验中昂贵或难以筛选的情况。

虽然经典机器学习已经大大提高了酶定向进化速度，但是仍存在一定的局限性。大部分蛋白质的突变序列较少，而有些蛋白质的序列长度又很大，当序列长度大于样本量时，常规的线性回归可能因为共线性而无法求解出唯一解，进而影响模型的稳定性和泛化能力，大大限制了可以选择的经典机器学习算法种类[5]。此外，经典机器学习一般只适用于单一酶蛋白的功能/结构预测或性能改造，对于不同种蛋白质则需要重新构建模型，故模型的迁移性较差，计算资源浪费较多。因此，机器学习辅助酶分子改造成功案例尚待研究积累。

2.3　深度学习辅助酶分子改造

针对经典机器学习在酶分子改造中的局限性，深度学习显示出其较强的应用能力，尤其是在酶功能预测方面有不少成功的实例，而在酶分子改造方面还仍然处于探索阶段。表 1 列出了几种深度学习模型在酶工程的应用，如

DeepGOPlus 仅通过蛋白质序列就可以预测蛋白质的功能[32]；GOLabeler[61]同时考虑了基序、蛋白质家族、序列特征等信息，通过对不同的特征进行训练进而预测未知蛋白功能；而 DeepFRI 则基于蛋白质接触图和序列，通过提取局部序列和全局结构特征，对蛋白质功能进行预测，由于考虑了结构信息，所以 DeepFRI 可以获得比仅基于序列信息的模型具备更高的准确率[64]。深度学习可以直接学习开放数据库中的数据，不需要通过实验搜集数据，就可以预测未知蛋白质的功能。尤其是当样本数量足够多时，深度学习的准确性比传统机器学习更高，同时模型可以自动学习复杂的数据特征，适用于氨基酸序列的高维数据。

近年来，不少研究者积极探索深度学习用于指导酶分子改造，以提高其特定性能。目前，最成功的例子是Fang等[66,74]通过使用卷积神经网络(convolutional neural network, CNN)构建了蛋白质热稳定性预测器，模型的准确率与 RIF 策略(Rosetta ddG、I Mutant 3.0 和 FoldX，RIF)相比高出了近 10%。该模型与其他模型的不同之处在于采用了蛋白质突变信息三维编码(图 3)，而常规蛋白质编码方法为一维向量，即将蛋白质突变位置和邻域信息都编码为一维向量，这忽略了蛋白质序列的特征，影响了预测的准确性和敏感性。三维编码方法不仅有效地考虑了蛋白质序列的特征，还编码了氨基酸特征、突变部位的邻域特征，甚至空间特征。其中横轴表示特征通道，即每个二维矩阵是氨基酸某一特性的正交编码，如突变信息、分子量或疏水性；纵轴代表蛋白质序列，竖轴代表 20 个氨基酸。对于突变位点用−1 表示野生型氨基酸，1 表示突变后的氨基酸，即对应到突变位点有 1 个−1、1 个 1 和 18 个 0，而突变位点两边的点有 1 个 1 和 19 个 0，氨基酸的性质采用 AAindex 编码[75]。这种三维编码方法类似彩色数字图像有红绿蓝 3 个关于颜色的通道，同样该方法结合多个通道可以捕捉更多有关蛋白质的突变信息。最终输入数据被编码为 $X^{20\times L\times N}$，其中 L 表示突变点邻域大小，N 表示特征数量。

模型的结构如图 4 所示。模型包含 1 个卷积层、1 个池化层、1 个 dropout 层和 1 个全连接层，最后输出预测结果。将三维编码信息输入模型后，首先通过卷积层，经过多尺度卷积学习蛋白质序列的特征。卷积学习本质上是线性变化，为了更好地学习序列特征，在卷积层输出后加入非线性激活函数。因为卷积学习具有一定的冗余性，所以在卷积层后接入最大池化层以减少冗余，同时降低特征的维度。Dropout 层通过随机剔除神经元以达到防止过拟合的目的，同时可以提高模型的泛化能力。最终各卷积核学习到的特征被拼接成一维向量，输入给全连接层并通过一个普通的神经网络输出预测结果，0 表示稳定性未增强，1 表示稳定性增强。Fang 等[66,76]使用来自 ProTherm 数据库的数据(单点突变)进行训练，最终结果显示在多个指标上与传统方法不相上下，阳性样本(SEN+、SPE+)和马修斯相关系数(matthews correlation coefficient, MCC)的检测效果最好，表明该方法在阳性样本检测和稳定突变识别方面具有显著优势；随后作者使用 RIF 策略选择了 24 个突变，并通过实验测量，然后用 CNN 模型对这 24 个变体进行预测，预测准确率为 75.2%，比使用 RIF 策略预测的准确率提高了 7.3%。证明了卷积神经网络预测蛋白质热稳定性方法的可行性。

3 总结与展望

多年来，定向进化、理性设计和人工智能辅助设计等在酶工程领域取得较大进展。定向

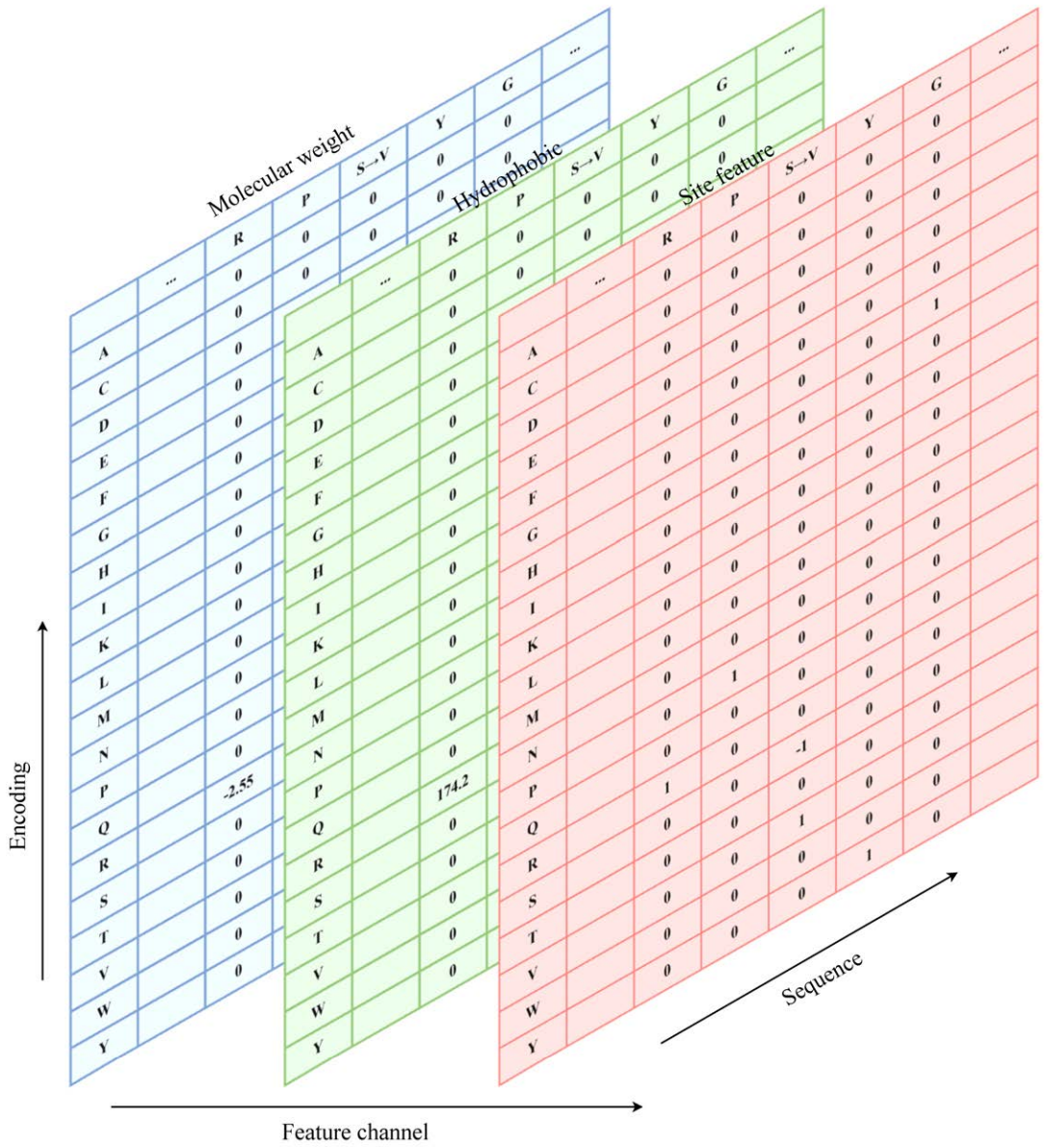

图 3　蛋白质序列及性质编码

Figure 3　Coding of protein sequence and properties.

进化不需要研究者详细了解酶结构、催化机理和氨基酸突变对酶的影响等方面的知识，就可实现对酶的改造，突破了酶工程对先验知识要求的壁垒，但定向进化方案存在筛选成本过大、筛选工作量大和实验周期过长的问题。理性设计建立在研究者对结构以及催化机理深入了解的基础上，可以改造酶的特性，构建具有期望功能的酶，以满足实际应用需求。利用酶蛋白在定向进化或理性设计后生成的数据作为人工智能辅助策略的初始数据，通过 SMOTE 过采样、虚拟样本生成、GAN 生成以及 GAN 与 DNN 相结合等方法增强数据，以解决小样本量带来的酶分子改造问题[14,33]。

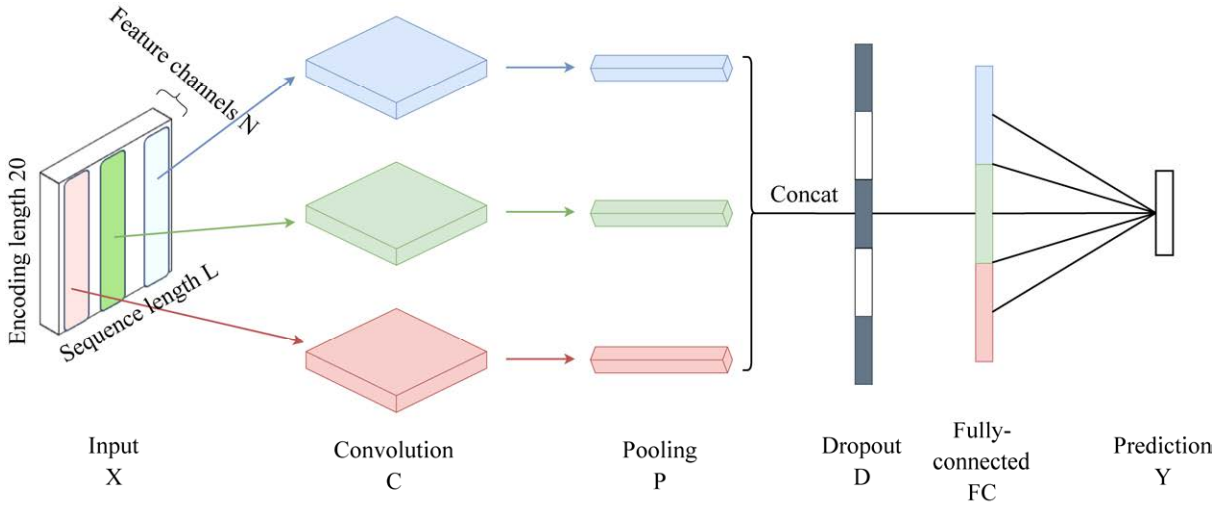

图 4　模型结构

Figure 4　Structure of the model.

在合适的采样方法引导下，CNN 可以在酶的活性、热稳定性甚至是选择性等性能改造方面给出置信度较高的建议，但作为一种数据驱动的计算方法，其预测结果往往受到训练集的限制。因此更高质量的数据库构建是酶分子改造工作的重点，需要构架更为全面的基础性蛋白质序列/结构→性能数据库，充分利用高通量技术的发展和蛋白质深度突变扫描技术获得更全面、更均衡的数据集。其次，人工智能对酶性能预测还处在早期阶段，大部分模型对数据有高度依赖，无法直接将模型迁移到新的任务上，还需要对模型进行训练和微调以提升模型预测的准确性。近几年发展的基于深度学习的酶性能预测模型策略有利于解决上述问题：如 DeepGOPlus 模型通过将神经网络预测与基于序列相似性的方法相结合来捕捉同源和间接相互作用信息，解决了特征缺失的问题；GAT-GO 模型结合了序列特征、蛋白质嵌入和残基接触图，从蛋白质相互作用力网中预测蛋白质功能，而不依赖于高质量的数据集；ESM-1v 模型依靠公共数据库，而不需要实验室突变实验数据，大大降低了生物实验所形成的人力物力负担，也解决了监督学习对于只有几十条注释的小样本量标签预测准确率低的问题。这些策略为提升预测模型性能指明了方向。

随着计算能力和实验技术的进步，人工智能在酶分子改造方面的应用会越来越完善，如果上述问题得到解决，可以获得具有较高准确度的模型算法来准确预测突变酶的性能，为数据库提供优质数据，并且随着序列-性能对数据的不断增长，具备高泛化能力和准确预测性能的神经网络模型将是降低酶分子进化成本、缩短实验周期、提高酶分子进化高度的重要手段。

REFERENCES

[1] 王镜岩. 生物化学. 上册[M]. 北京: 高等教育出版社, 2002.
WANG JY. Biochemistry. Volume I[M]. Beijing: Higher Education Press, 2002 (in Chinese).

[2] REETZ MT, CARBALLEIRA JD, VOGEL A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability[J]. Angewandte Chemie International Edition, 2006, 45(46): 7745-7751.

[3] 曲戈, 朱彤, 蒋迎迎, 吴边, 孙周通. 蛋白质工程:

从定向进化到计算设计[J]. 生物工程学报, 2019, 35(10): 1843-1856.
QU G, ZHU T, JIANG YY, WU B, SUN ZT. Protein engineering: from directed evolution to computational design[J]. Chinese Journal of Biotechnology, 2019, 35(10): 1843-1856 (in Chinese).

[4] 张锟, 曲戈, 刘卫东, 孙周通. 工业酶结构与功能的构效关系[J]. 生物工程学报, 2019, 35(10): 1806-1818.
ZHANG K, QU G, LIU WD, SUN ZT. Structure-function relationships of industrial enzymes[J]. Chinese Journal of Biotechnology, 2019, 35(10): 1806-1818 (in Chinese).

[5] 赵永耀. 基于机器学习的酶改造方法研究[D]. 江苏: 南京邮电大学硕士学位论文, 2023.
ZHAO YY. Research of enzyme modification methods *via* machine learning[D]. Jiangsu: Master's Thesis of Nanjing University of Posts and Telecommunications, 2023 (in Chinese).

[6] ARNOLD FH. The nature of chemical innovation: new enzymes by evolution[J]. Quarterly Reviews of Biophysics, 2015, 48(4): 404-410.

[7] MUSIL M, KONEGGER H, HON J, BEDNAR D, DAMBORSKY J. Computational design of stable and soluble biocatalysts[J]. ACS Catalysis, 2019, 9(2): 1033-1054.

[8] 蒋迎迎, 曲戈, 孙周通. 机器学习助力酶定向进化[J]. 生物学杂志, 2020, 37(4): 1-11.
JIANG YY, QU G, SUN ZT. Machine learning-assisted enzyme directed evolution[J]. Journal of Biology, 2020, 37(4): 1-11 (in Chinese).

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

[10] SIKANDER R, WANG YP, GHULAM A, WU XJ. Identification of enzymes-specific protein domain based on DDE, and convolutional neural network[J]. Frontiers in Genetics, 2021, 12: 759384.

[11] JING XY, LI FM. Predicting cell wall lytic enzymes using combined features[J]. Frontiers in Bioengineering and Biotechnology, 2021, 8: 627335.

[12] WAN ZY, WANG QD, LIU DC, LIANG JH. Accelerating the optimization of enzyme-catalyzed synthesis conditions *via* machine learning and reactivity descriptors[J]. Organic & Biomolecular Chemistry, 2021, 19(28): 6267-6273.

[13] SAITO Y, OIKAWA M, SATO T, NAKAZAWA H, ITO T, KAMEDA T, TSUDA K, UMETSU M. Machine-learning-guided library design cycle for directed evolution of enzymes: the effects of training data composition on sequence space exploration[J]. ACS Catalysis, 2021, 11(23): 14615-14624.

[14] Del RIO-CHANONA EA, FIORELLI F, ZHANG DD, AHMED NR, JING KJ, SHAH N. An efficient model construction strategy to simulate microalgal lutein photo-production dynamic process[J]. Biotechnology and Bioengineering, 2017, 114(11): 2518-2527.

[15] GADO JE, HARRISON BE, SANDGREN M, STÅHLBERG J, BECKHAM GT, PAYNE CM. Machine learning reveals sequence-function relationships in family 7 glycoside hydrolases[J]. The Journal of Biological Chemistry, 2021, 297(2): 100931.

[16] SIEDHOFF NE, SCHWANEBERG U, DAVARI MD. Machine learning-assisted enzyme engineering[J]. Methods in Enzymology, 2020, 643: 281-315.

[17] 王慕锪, 陈琦, 马薇, 李春秀, 欧阳鹏飞, 许建和. 机器学习方法在酶定向进化中的应用进展[J]. 生物技术通报, 2023, 39(4): 38-48.
WANG MQ, CHEN Q, MA W, LI CX, OUYANG PF, XU JH. Advances in the application of machine learning methods for directed evolution of enzymes[J]. Biotechnology Bulletin, 2023, 39(4): 38-48 (in Chinese).

[18] YANG KK, WU Z, ARNOLD FH. Machine-learning-guided directed evolution for protein engineering[J]. Nature Methods, 2019, 16: 687-694.

[19] BUETTNER K, HERTEL TC, PIETZSCH M. Increased thermostability of microbial transglutaminase by combination of several hot spots evolved by random and saturation mutagenesis[J]. Amino Acids, 2012, 42(2): 987-996.

[20] BÖHME B, MORITZ B, WENDLER J, HERTEL TC, IHLING C, BRANDT W, PIETZSCH M. Enzymatic activity and thermoresistance of improved microbial transglutaminase variants[J]. Amino Acids, 2020, 52(2): 313-326.

[21] ARANGO GUTIERREZ E, MUNDHADA H, MEIER T, DUEFEL H, BOCOLA M, SCHWANEBERG U. Reengineered glucose oxidase for amperometric glucose determination in diabetes analytics[J]. Biosensors & Bioelectronics, 2013, 50: 84-90.

[22] CHEN K, ARNOLD FH. Tuning the activity of an enzyme for unusual environments: sequential random

mutagenesis of subtilisin E for catalysis in dimethylformamide[J]. Proceedings of the National Academy of Sciences of the United States of America, 1993, 90(12): 5618-5622.

[23] STEMMER WPC. Rapid evolution of a protein *in vitro* by DNA shuffling[J]. Nature, 1994, 370: 389-391.

[24] ROMERO-RIVERA A, GARCIA-BORRÀS M, OSUNA S. Computational tools for the evaluation of laboratory-engineered biocatalysts[J]. Chemical Communications, 2016, 53(2): 284-297.

[25] YOKOYAMA K, UTSUMI H, NAKAMURA T, OGAYA D, SHIMBA N, SUZUKI E, TAGUCHI S. Screening for improved activity of a transglutaminase from *Streptomyces mobaraensis* created by a novel rational mutagenesis and random mutagenesis[J]. Applied Microbiology and Biotechnology, 2010, 87(6): 2087-2096.

[26] 陈康康. 分子改造强化 *Streptomyces hygroscopicus* 谷氨酰胺转胺酶催化性能研究[D]. 无锡: 江南大学博士学位论文, 2013.
CHEN KK. The study on the molecular modification of *Streptomyces hygroscopicus* transglutaminase for enhanced catalytic properties[D]. Wuxi: Doctoral Dissertation of Jiangnan University, 2013 (in Chinese).

[27] 倪晗朦, 胡孟凯, 张恒维, 张显, 潘学玮, 饶志明, 周楠迪. 半理性设计提高甲酸脱氢酶(CbFDH)活力及热稳定性[J]. 食品与生物技术学报, 2023, 42(10): 1-8.
NI HM, HU MK, ZHANG HW, ZHANG X, PAN XW, RAO ZM, ZHOU ND. Enhanced activity and thermal stability of *Formate dehydrogenase* (CbFDH) *via* semi-rational design[J]. Journal of Food Science and Biotechnology, 2023, 42(10): 1-8 (in Chinese).

[28] ARABNEJAD H, dal LAGO M, JEKEL PA, FLOOR RJ, THUNNISSEN AM WH, TERWISSCHA van SCHELTINGA AC, WIJMA HJ, JANSSEN DB. A robust cosolvent-compatible halohydrin dehalogenase by computational library design[J]. Protein Engineering, Design & Selection: PEDS, 2017, 30(3): 173-187.

[29] SUN ZT, LIU Q, QU G, FENG Y, REETZ MT. Utility of B-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability[J]. Chemical Reviews, 2019, 119(3): 1626-1665.

[30] QU G, LI AT, ACEVEDO-ROCHA CG, SUN ZT, REETZ MT. The crucial role of methodology development in directed evolution of selective enzymes[J]. Angewandte Chemie International Edition, 2020, 59(32): 13204-13231.

[31] SUN ZT, LONSDALE R, ILIE A, LI GY, ZHOU JH, REETZ MT. Catalytic asymmetric reduction of difficult-to-reduce ketones: triple-code saturation mutagenesis of an alcohol dehydrogenase[J]. ACS Catalysis, 2016, 6(3): 1598-1605.

[32] KULMANOV M, HOEHNDORF R. DeepGOPlus: improved protein function prediction from sequence[J]. Bioinformatics, 2020, 36(2): 422-429.

[33] 卞佳豪, 杨广宇. 人工智能辅助的蛋白质工程[J]. 合成生物学, 2022, 3(3): 429-444.
BIAN JH, YANG GY. Artificial intelligence-assisted protein engineering[J]. Synthetic Biology Journal, 2022, 3(3): 429-444 (in Chinese).

[34] 李怡凡, 王怡, 张凯丽, 李帅. 深度突变扫描技术在蛋白研究中的应用[J]. 生物工程学报, 2023, 39(9): 3710-3723.
LI YF, WANG Y, ZHANG KL, LI S. Application of deep mutational scanning technology in protein research[J]. Chinese Journal of Biotechnology, 2023, 39(9): 3710-3723 (in Chinese).

[35] WIJMA HJ, FLOOR RJ, JEKEL PA, BAKER D, MARRINK SJ, JANSSEN DB. Computationally designed libraries for rapid enzyme stabilization[J]. Protein Engineering, Design and Selection, 2014, 27(2): 49-58.

[36] MENG CL, GUO F, ZOU Q. CWLy-SVM: a support vector machine-based tool for identifying cell wall lytic enzymes[J]. Computational Biology and Chemistry, 2020, 87: 107304.

[37] YANG M, FEHL C, LEES KV, LIM EK, OFFEN WA, DAVIES GJ, BOWLES DJ, DAVIDSON MG, ROBERTS SJ, DAVIS BG. Functional and informatics analysis enables glycosyltransferase activity prediction[J]. Nature Chemical Biology, 2018, 14: 1109-1117.

[38] ROMERO PA, KRAUSE A, ARNOLD FH. 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.

[39] LU HY, DIAZ DJ, CZARNECKI NJ, ZHU CZ, KIM W, SHROFF R, ACOSTA DJ, ALEXANDER BR, COLE HO, ZHANG Y, LYND NA, ELLINGTON AD, ALPER HS. Machine learning-aided engineering of hydrolases for PET depolymerization[J]. Nature, 2022, 604: 662-667.

[40] 陆宸叶. 机器学习辅助的肌酸酶的突变进化分析[D]. 上海: 华东师范大学硕士学位论文, 2022.
LU CY. The analysis of machine learning-assisted mutation evolution of creatinase[D]. Shanghai: Master's Thesis of East China Normal University, 2022 (in Chinese).
[41] MUSDAL Y, GOVINDARAJAN S, MANNERVIK B. Exploring sequence-function space of a poplar glutathione transferase using designed information-rich gene variants[J]. Protein Engineering, Design and Selection, 2017, 30(8): 543-549.
[42] ZAUGG J, GUMULYA Y, MALDE AK, BODÉN M. Learning epistatic interactions from sequence-activity data to predict enantioselectivity[J]. Journal of Computer-Aided Molecular Design, 2017, 31(12): 1085-1096.
[43] CHAWLA NV, BOWYER KW, HALL LO, KEGELMEYER WP. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
[44] ZHOU ZH, JIANG Y. NeC4.5: neural ensemble based C4.5[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(6): 770-773.
[45] LI DC, LIN YS. Using virtual sample generation to build up management knowledge in the early manufacturing stages[J]. European Journal of Operational Research, 2006, 175(1): 413-434.
[46] LIU YF, ZHOU Y, LIU X, DONG F, WANG C, WANG ZH. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology[J]. Engineering, 2019, 5(1): 156-163.
[47] WANG KF, GOU C, DUAN YJ, LIN YL, ZHENG XH, WANG FY. Generative adversarial networks: introduction and outlook[J]. IEEE/CAA Journal of Automatica Sinica, 2017, 4(4): 588-598.
[48] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, XU B, WARDE-FARLEY D, OZAIR S, COURVILLE AC, BENGIO Y. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.
[49] CRESWELL A, WHITE T, DUMOULIN V, ARULKUMARAN K, SENGUPTA B, BHARATH AA. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.
[50] SHARMA A, SINGH P, CHANDRA R. SMOTified-GAN for class imbalanced pattern classification problems[J]. IEEE Access, 2022, 10: 30655.
[51] REMMERT M, BIEGERT A, HAUSER A, SÖDING J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment[J]. Nature Methods, 2012, 9: 173-175.
[52] HEFFERNAN R, YANG YD, PALIWAL K, ZHOU YQ. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility[J]. Bioinformatics, 2017, 33(18): 2842-2849.
[53] GREENHALGH JC, FAHLBERG SA, PFLEGER BF, ROMERO PA. Machine learning-guided acyl-ACP reductase engineering for improved *in vivo* fatty alcohol production[J]. Nature Communications, 2021, 12: 5825.
[54] HAN X, NING WB, MA XQ, WANG XN, ZHOU K. Improving protein solubility and activity by introducing small peptide tags designed with machine learning models[J]. Metabolic Engineering Communications, 2020, 11: e00138.
[55] OSTAFE R, FONTAINE N, FRANK D, CHONG MNF, PRODANOVIC R, PANDJAITAN R, OFFMANN B, CADET F, FISCHER R. One-shot optimization of multiple enzyme parameters: tailoring glucose oxidase for pH and electron mediators[J]. Biotechnology and Bioengineering, 2020, 117(1): 17-29.
[56] CADET F, FONTAINE N, LI GY, SANCHIS J, NG FUK CHONG M, PANDJAITAN R, VETRIVEL I, OFFMANN B, REETZ MT. A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes[J]. Scientific Reports, 2018, 8: 16757.
[57] FOX RJ, DAVIS SC, MUNDORFF EC, NEWMAN LM, GAVRILOVIC V, MA SK, CHUNG LM, CHING C, TAM S, MULEY S, GRATE J, GRUBER J, WHITMAN JC, SHELDON RA, HUISMAN GW. Improving catalytic function by ProSAR-driven enzyme evolution[J]. Nature Biotechnology, 2007, 25: 338-344.
[58] SAITO Y, OIKAWA M, NAKAZAWA H, NIIDE T, KAMEDA T, TSUDA K, UMETSU M. Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins[J]. ACS Synthetic Biology, 2018, 7(9): 2014-2022.
[59] LI GY, DONG YJ, REETZ MT. Can machine learning revolutionize directed evolution of selective

enzymes?[J]. Advanced Synthesis & Catalysis, 2019, 361(11): 2377-2386.
[60] LI G, RABE KS, NIELSEN J, ENGQVIST MKM. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima[J]. ACS Synthetic Biology, 2019, 8(6): 1411-1420.
[61] YOU RH, ZHANG ZH, XIONG Y, SUN FZ, MAMITSUKA H, ZHU SF. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank[J]. Bioinformatics, 2018, 34(14): 2465-2473.
[62] LAI BQ, XU JB. Accurate protein function prediction *via* graph attention networks with predicted structure information[J]. Briefings in Bioinformatics, 2022, 23(1): bbab502.
[63] MEIER J, RAO R, VERKUIL R, LIU J, SERCU T, RIVES A. Language models enable zero-shot prediction of the effects of mutations on protein function[J]. Advances in Neural Information Processing Systems, 2021, 34: 29287-29303.
[64] GLIGORIJEVIĆ V, RENFREW PD, KOSCIOLEK T, LEMAN JK, BERENBERG D, VATANEN T, CHANDLER C, TAYLOR BC, FISK IM, VLAMAKIS H, XAVIER RJ, KNIGHT R, CHO K, BONNEAU R. Structure-based protein function prediction using graph convolutional networks[J]. Nature Communications, 2021, 12: 3168.
[65] YOU RH, YAO SW, MAMITSUKA H, ZHU SF. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction[J]. Bioinformatics, 2021, 37(supplement_1): i262-i271.
[66] FANG XR, HUANG JS, ZHANG R, WANG F, ZHANG QY, LI GL, YAN JY, ZHANG HJ, YAN YJ, XU L. Convolution neural network-based prediction of protein thermostability[J]. Journal of Chemical Information and Modeling, 2019, 59(11): 4833-4843.
[67] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.
[68] QUINLAN JR. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.
[69] JENSEN FV. An Introduction to Bayesian Networks[M]. London: UCL Press, 1996.
[70] GELADI P, KOWALSKI BR. Partial least-squares regression: a tutorial[J]. Analytica Chimica Acta, 1986, 185: 1-17.
[71] LI Y, SONG K, ZHANG J, LU S. A computational method to predict effects of residue mutations on the catalytic efficiency of hydrolases[J]. Catalysts, 2021, 11(2): 286.
[72] ABDI H, WILLIAMS LJ. Principal component analysis[J]. WIREs Computational Statistics, 2010, 2(4): 433-459.
[73] LIAO J, WARMUTH MK, GOVINDARAJAN S, NESS JE, WANG RP, GUSTAFSSON C, MINSHULL J. Engineering proteinase K using machine learning and synthetic genes[J]. BMC Biotechnology, 2007(7): 1-19.
[74] LI GL, FANG XR, SU F, CHEN Y, XU L, YAN YJ. Enhancing the thermostability of *Rhizomucor miehei* lipase with a limited screening library by rational-design point mutations and disulfide bonds[J]. Applied and Environmental Microbiology, 2018, 84(2): e02129-e02146.
[75] KAWASHIMA S, OGTA H, KANEHISA M. AAindex: amino acid index database[J]. Nucleic Acids Research, 1999, 27(1): 368-369.
[76] NIKAM R, KULANDAISAMY A, HARINI K, SHARMA D, GROMIHA MM. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years[J]. Nucleic Acids Research, 2021, 49(D1): D420-D424.

(本文责编 郝丽芳)