特约评述

DOI: 10.12211/2096-8280.2021-032

人工智能辅助的蛋白质工程

卞佳豪，杨广宇
（上海交通大学 生命科学技术学院，微生物代谢国家重点实验室，上海 200240）

摘要：蛋白质工程是合成生物学领域的重要研究方向之一。但目前人类对于蛋白质折叠、酶天然进化机制等基础生物学问题的理解仍很有限，因此基于理性设计方法进行蛋白质的功能从头设计（*de novo* design）仍然是一个难题。定向进化（directed evolution）通过在实验室模拟自然进化的原理，可以在不依赖结构和机制信息的基础上对蛋白质的功能进行有效优化。但是定向进化高度依赖高通量筛选方法，也限制了其对缺少高通量筛选方法的蛋白质进行改造的能力。近年来，人工智能辅助的蛋白质工程逐渐发展成为一种高效的蛋白质分子设计新策略，在蛋白质的结构预测、功能预测、溶解度预测和指导智能文库设计等多个方面显现出独特的优势，成为理性设计和定向进化之后的又一次技术发展的浪潮。本文综述了近年来人工智能辅助的蛋白质工程的应用进展，对其中的代表性工作进行了重点阐述。在简单介绍了人工智能蛋白质工程策略的原理和流程之后，对数据、分子描述符和人工智能算法等三个影响预测模型性能的关键点进行了分析，总结了该策略中的主要数据库、分子描述符和算法的主流工具包及平台，介绍了它们的功能、用途和网址。我们还对人工智能策略目前仍面临的不足进行了探讨，如高质量数据不足、实验数据存在偏差、缺少通用模型等。随着自动基因功能注释技术、超高通量筛选技术和人工智能算法的不断发展，将会给人工智能辅助的蛋白质工程提供足够的高质量数据和更准确的算法，从而不断提升人工智能辅助的蛋白质工程预测准确度，为合成生物学研究提供更大的助力。

关键词：蛋白质工程；合成生物学；人工智能；预测模型；数据库；分子描述符

中图分类号：Q816　　**文献标志码：**A

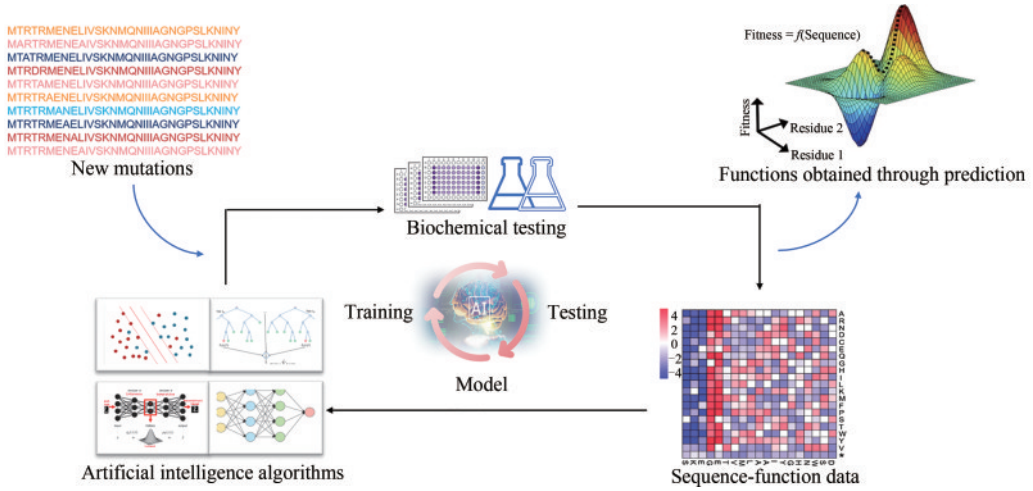
Artificial intelligence-assisted protein engineering

BIAN Jiahao，YANG Guangyu
（*State Key Laboratory of Microbial Metabolism*，*School of Life Sciences and Biotechnology*，*Shanghai Jiao Tong University*，*Shanghai 200240*，*China*）

Abstract: Protein engineering is one of the important research fields of synthetic biology. However, *de novo* design of protein functions based on rational design is still challenging, because of the limited understanding on biological fundamentals such as protein folding and the natural evolution mechanism of enzymes. Directed evolution is capable of

收稿日期：2021-03-16 修回日期：2021-05-24
基金项目：国家自然科学基金（32030063，21627812）
引用本文：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
Citation：BIAN Jiahao，YANG Guangyu. Artificial intelligence-assisted protein engineering[J]. Synthetic Biology Journal，2022，3(3)：429-444

optimizing protein functions effectively by mimicking the principle of natural evolution in the laboratory without relying on structure and mechanism information. However, directed evolution is highly dependent on high-throughput screening methods, which also limits its applications on proteins which lack high-throughput screening methods. In recent years, artificial intelligence has been developed very rapidly for integrating into multidisciplinary fields. In synthetic biology, artificial intelligence-assisted protein engineering has become an efficient strategy for protein engineering besides rational design and directed evolution, which has shown unique advantages in predicting the structure, function, solubility of proteins and enzymes. Artificial intelligence models can learn the internal properties and relationships from given sequence-function data sets to make predictions on properties for virtual sequences. In this article, we review the application of artificial intelligence-assisted protein engineering. With the basic and process of the strategy introduced, three key points that affect the performance of the predictive model are analyzed: data, molecular descriptors and artificial intelligence algorithms. In order to provide useful tools for researchers who want to take advantage of this strategy, we summarize the main public database, diverse toolkits and web servers of the common molecular descriptors and artificial intelligence algorithms. We also comment on the functions, applications and websites of several artificial intelligence-assisted protein engineering platforms, through which a complete prediction task including protein sequences representation, feature analysis, model construction and output can be completed easily. Finally, we analyze some challenges that need to be solved in the artificial intelligence-assisted protein engineering, such as the lack of high-quality data, deviation in data sets and lacking of the universal models. However, with the development of automated gene annotations, ultra-high-throughput screening technologies and artificial intelligence algorithms, sufficient high-quality data and appropriate algorithms will be developed, which can enhance the performance of artificial intelligence-assisted protein engineering and thus facilitate the development of synthetic biology techniques.



Keywords: protein engineering; synthetic biology; artificial intelligence; predictive model; database; molecular descriptor

合成生物学是一个广泛的研究领域，通过将生物学和工程学相结合来设计和创建具有新颖功能的生物系统[1-2]。这一过程需要功能各异、形式多样并且能够良好实现预期功能的生物元件，特别是蛋白质功能元件（催化酶、转录因子、转运蛋白、蛋白支架等）[3]。但是，天然来源蛋白质元件大部分都不能满足人工生物系统的需要，实际应用中往往表现出折叠错误、细胞毒性、功能不

适宜等缺陷[4-5]。蛋白质从头设计或对天然蛋白质进行分子改造，成为解决这一问题的重要途径。对于蛋白质或酶的分子改造，已经成为合成生物学的重要研究领域[6-9]。

在天然蛋白质分子改造方面，主要包括定向进化（directed evolution）和理性设计（rational design）两种策略[10-12]，见图1。前者通过模拟自然选择过程，对目标基因进行多轮突变和筛选实验，直至获得所需水平的优良变体，但是该技术受限制于较低的筛选速率和序列空间中庞大的变体数量[13]。后者依据序列和结构信息，选择较少的关键位点进行精准改造，从而构建较小的突变文库，但是需要对结构功能信息有深入了解，并且需要巨大的计算资源[14]。

人工智能辅助的蛋白质工程策略是一种由数据驱动的新策略[15]。该策略通过学习已有数据中的信息，建立起输入属性（如序列）到输出属性（如功能）的映射关系，不需要详细的物理或生物层面的基础信息[16]。一旦得到足够准确的映射关系（或者说预测模型），就能够通过实验中容易得到的输入值来预测输出值，从而免除大量的重复性实验。目前，该策略已经成功应用在蛋白质工程的很多方面，包括蛋白分子结构预测[17-18]、蛋白分子功能预测[19-20]、蛋白分子溶解度预测[21-22]和指导设计智能组合文库[23-26]等。

目前已有多篇综述详细介绍了有关机器学习的基础概念[27-31]。这些文章多从数据和算法的角度来对人工智能的主要进展进行了介绍，但是对于非生物信息学背景的研究人员而言，这类综述读起来较为深奥。为了使更多实验生物学背景的人员理解人工智能蛋白设计的进展，本文将主要介绍人工智能辅助蛋白分子设计的应用实例、已开

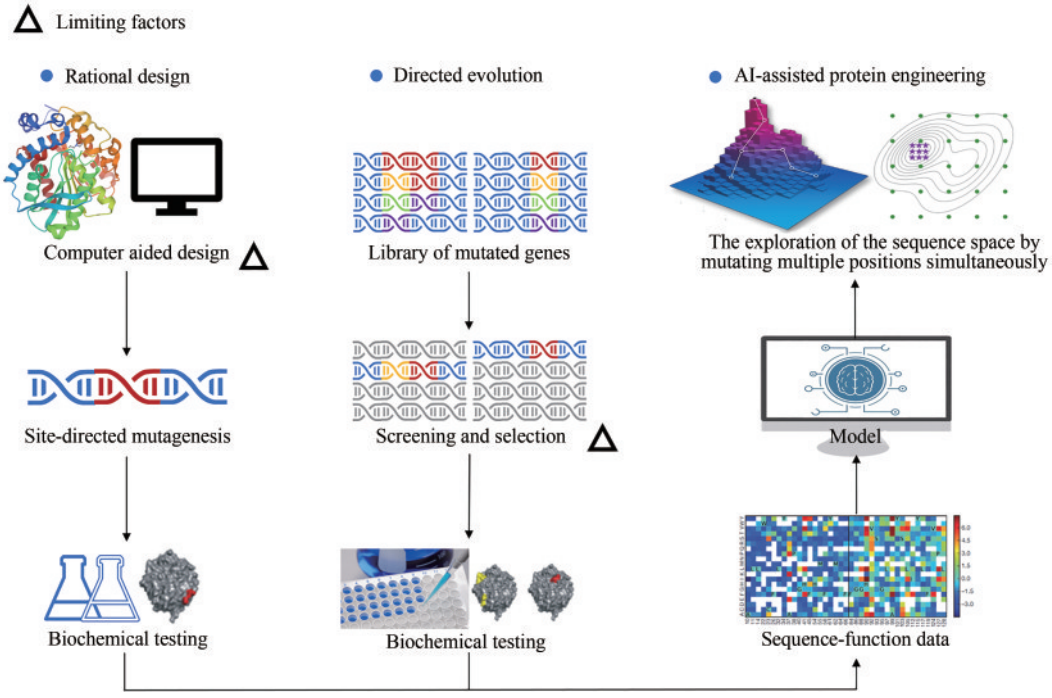


图1 理性设计，定向进化和人工智能辅助的蛋白质工程策略示意图

（理性设计依赖序列和结构信息，精准设计突变体文库，但难以应用于缺少结构功能信息的蛋白质。定向进化中对目标基因进行多轮突变和筛选实验，不受结构功能信息限制，但是需要进行高通量的筛选方法。人工智能辅助的蛋白质工程则需要大量的序列-功能数据，可以来源于实验、计算和数据库等多方面，通过构建的预测模型，能够更有效地探索蛋白质突变体序列空间）

Fig. 1 Schematic diagram for rational design, directed evolution and artificial intelligence-assisted protein engineering

(Rational design relies on sequence and structural information to design mutant libraries accurately. However, it is difficult for being applied to proteins lacking structural and functional information. In the directed evolution strategy, multiple rounds of mutation and screening experiments are performed on target genes, which are not limited by structural and functional information, but high-throughput screening methods are required. Artificial intelligence-assisted protein engineering requires a large amount of sequence-function data, which can be derived from experiments, calculations, and databases. Through the predictive model, the sequence space of protein mutants can be explored more effectively)

发的数据库和平台工具等几个方面，为希望进入人工智能蛋白质工程领域的入门者提供帮助。

1 人工智能辅助的蛋白质工程应用实例

人工智能算法由于准确度高、计算速度快、不受蛋白质结构功能信息限制等优点，近年来被大量应用于蛋白质工程领域，包括蛋白质的结构、功能、热稳定性、对映体选择性、光敏性及指导设计智能组合文库等多个方面。其中除了经典的机器学习算法（决策树、支持向量机和高斯过程回归等）外，多种深度学习算法和基于深度学习的自然语言处理技术也获得了成功的应用。在下文中，我们重点集中于近几年在蛋白质结构预测、功能预测、溶解度预测和指导设计智能组合文库四个方面的成功案例，系统地分析人工智能算法在蛋白质工程中应用的优势。

1.1 蛋白质结构预测

截至2018年，蛋白质数据库中发布了超过145 000个蛋白质结构，但与目前已知的超过2亿条蛋白质序列相比，仍仅占很小的比例[32]，因此蛋白质结构预测是生物学中经久不衰的热点问题。早在1992年，机器学习算法就被用于预测蛋白质二级结构[33]。近几年，利用深度学习算法和蛋白质序列的三维结构预测模型取得了不小的进展[34]。首先是在2018年第13届全球蛋白质结构预测竞赛（CASP）上，AlphaFold模型结合深度残差卷积神经网络和快速Rosetta模型，获得了预测43种蛋白中的25种蛋白结构的最高分，实现了预测成功率的突破[17]。2019年底，David Baker团队发表了trRosetta方案，综合了深度学习和Rosetta的优势和进展，具有良好预测精度的同时，能够在本地电脑上就可以完成计算，使得预测蛋白结构的门槛大大降低[18]。在2020年的CASP14中，AlphaFold 2再次获得冠军。根据DeepMind官方的信息，AlphaFold 2在无模板的自由建模任务中，拿到了87.0的GDT_TS分数（global distance test[35]），在常规项目中拿到了92.4分，这意味着该系统预测的均方根偏差（即预测数据与实验数据在原子位置上的偏差）大约为0.16 nm，已经达到了常规蛋白质晶体结构的实验精度。尽管AlphaFold目前最好的成绩是针对单链蛋白质分子，但这种成绩本身就足以证明人工智能算法在蛋白质结构预测中的巨大潜力，例如减少繁琐的结晶条件探索工作，以及提供以常规实验方法难以获得的蛋白质结构等。

1.2 蛋白质功能预测

天然蛋白的功能表征实验需要大量工作，其速度远远低于新蛋白序列的获取速度[36-37]。借助人工智能算法来预测蛋白质的功能是另外一个研究热点。2018年，研究者通过收集来自拟南芥的54种GT1家族糖基转移酶的序列信息和它们91种底物的物理化学特性（如疏水常数lg*P*、分子表面积）和结构信息（如官能团拷贝数、框架类型），建立了初始的数据集，并以多种基于决策树的算法来构建酶功能的预测模型（图2）[19]。在不需要进行任何实验的条件下，该预测模型利用酶序列，就能够准确地预测其他植物中（苜蓿和燕麦）GT1糖基转移酶的活性，对来自细菌的GT1酶活性的预测准确率也在70%以上。这表明能够利用高通量数据进行学习的人工智能算法在底物混杂、已解析结构少的酶的功能注释中具有巨大潜力。此外，人工智能算法也被应用于预测酶的EC编号（enzyme commission number），帮助对酶分子进行分类。先后发展出的PRIAM[38]、CatFam[39]、EFICAz2.5[40]、SVM-prot[41]、COFACTOR[42]、DEEPre[36]、DETECT v2[43]、ECPred[44]和DeepEC[20]等多种预测工具，在计算时间、计算精度和覆盖范围等预测性能方面逐渐改进，简要内容见表1。其中，DeepEC方法包括三个独立的卷积神经网络，利用氨基酸序列，就能对氨基酸序列是否为酶分子、酶分子EC编号的三位和四位数值进行预测。与CatFam、DETECT v2、ECPred、EFICAz2.5和PRIAM五种代表性的酶EC编号预测工具相比，在Swiss-Prot数据库中选取的201个酶进行验证时，DeepEC表现最佳，准确率

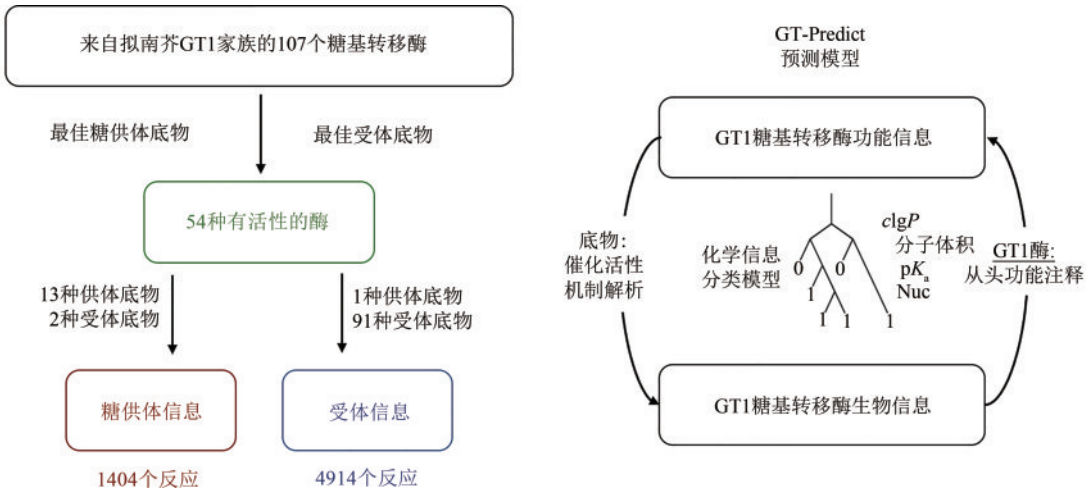


图2 GT1家族糖基转移酶预测模型（GT-Predict）的工作流程[19]

（基于功能的算法学习方法GT-Predict，使用来源于酶、亲电试剂和亲核试剂的多种训练集来创建基于物理化学和局部序列的分类器，从而预测GT1糖基转移酶的催化活性和功能信息。Nuc表示亲核基团的数量/类型）

Fig. 2 Workflow for predicting the GT1 glycosyltransferase model (GT-Predict)[19]

(The function-based algorithmic learning approach, GT-Predict, uses a diverse training set of enzymes, electrophiles, and nucleophiles to create a physicochemical and local-sequence-based classifier for predicting the novel transformations and functional annotation of GT group-transfer enzymes.)

表1 EC编号预测工具汇总表

Tab. 1 Forecast tools for EC numbers

名称	发表日期	分子描述符	程序/算法	软件/工具包地址
PRIAM	2003年11月15日	序列比对构建的同源模块	PSI-BLAST序列比对程序	http://genopole.toulouse.inra.fr/bioinfo/priam/
CatFam	2008年07月17日	序列比对和分层聚类	ClustalW和PSI-BLAST序列比对程序	http://www.bhsai.org/downloads/catfam.tar.gz
EFICAz2.5	2012年08月24日	序列比对和分层聚类	支持向量机(SVM)和分类树(classification trees)	http://cssb.biology.gatech.edu/EFICAz2.5
SVM-Prot	2016年08月15日	多种分子描述符（多种氨基酸残基特性描述符和整体描述符）	支持向量机(SVM)，K最近邻(KNN)和概率神经网络(sPNN)	http://bidd2.nus.edu.sg/cgi-bin/svmprot/svmprot.cgi
COFACTOR	2017年05月02日	来自BioLiP文库的结构信息	基于TM分数的蛋白质结构比对算法	http://zhanglab.ccmb.med.umich.edu/COFACTOR/。
DEEPre	2017年10月23日	序列独热编码，位置特异性得分矩阵，溶剂可及性，二级结构独热编码和功能结构域	卷积神经网络(CNN)和循环神经网络(RNN)	http://www.cbrc.kaust.edu.sa/DEEPre
DETECT v2	2018年05月02日	酶EC编号的正负密度分布图	贝叶斯框架(Bayesian framework)	https://github.com/ParkinsonLab/DETECT-v2
ECPred	2018年09月21日	三种基于子序列，序列相似性和氨基酸物理化学特征的分子描述符：SPMap，BLAST-kNN和Pepstats-SVM	二进制分类算法	https://ecpred.kansil.org/
DeepEC	2019年06月20日	独热编码	卷积神经网络(CNN)	https://bitbucket.org/kaistsystemsbiology/deepec

（accuracy）和召回率（recall）分别为0.920和0.455。即45.5%的阳性样本能被预测模型准确识别，这其中92.0%样本的预测值与真实值是一致的。

1.3 蛋白质溶解度预测

蛋白质的溶解度对于其行使功能起到重要作

用。溶解度过低是蛋白质大规模生产中常见的主要瓶颈[45-46]，而溶解度的测量费时费力，因此非常需要能够准确对蛋白质溶解度进行预测的生物信息学工具。新加坡国立大学的Han等[21]测试了逻辑回归、决策树、支持向量机、朴素贝叶斯、条件随机森林、XGboost和人工神经网络等七种算法构建基于序列的溶解度预测模型，其中支持向量机算法构建的模型在此预测任务中显示出最高的准确性。在预测结果为代表“可溶”和“不溶”的二分值“1”和“0”时，该模型的预测准确率为0.7628。除此之外，该模型还可以预测蛋白质连续的溶解度值（离心后上清液的蛋白质质量与总蛋白质质量之比）。但这种情况下，模型预测的准确性有所降低，决定系数为0.41。最近，中山大学的Chen Jianwen等[22]利用蛋白质接触图（contact map）和图神经网络算法（GCN）开发了一种新的利用氨基酸序列预测蛋白质溶解的模型GraphSol，在同样利用eSOL数据库中的蛋白质溶解度数据进行验证时，进一步提升了预测模型的性能，其决定系数为0.48。在蛋白质工程中，输出结果为简单的二分值时，重要的氨基酸突变对溶解度的贡献无法分析。例如，“不溶”和“可溶”的群体中，不同突变对蛋白质溶解度的贡献无法分辨。并且，当存在大量“可溶”的预测变体时，无法从中选出表现最佳的少数变体进行实验验证[21]。因此，能够预测蛋白质连续的溶解度的模型更适用于辅助蛋白质工程。随着可用数据集的扩大和算法框架的优化，基于序列的蛋白质溶解度预测模型将能够有越来越高的准确率。

1.4 指导设计智能组合文库

人工智能策略在酶定向进化中也具有重要的应用潜力。依靠人工智能算法，可以基于已有的序列/结构信息，直接建立起序列/结构-功能的映射关系，因此理论上可以极大减少筛选工作量，并且更加有效地探索整个组合突变体的序列空间[26，47]。例如，在指导绿色荧光蛋白向黄色荧光蛋白进化的研究中，研究者们对选定的四个关键位点构建了单点饱和突变库和随机诱变库，共包含218个变体。但将所有变体筛选之后，没有发现比参考黄色荧光蛋白性能更好的突变体。随后，他们选择其中的155个变体的序列-功能数据作为初始数据集，以高斯过程回归算法来构建预测模型。通过预测模型，遍历了整个四点组合序列空间中的近16万个变体，并对其性能打分。在仅仅对预测突变体文库中排名靠前的78个变体进行验证的情况下，就找到了12个黄色荧光强度高于参考蛋白的突变体[23]。

此外，在Frances H. Arnold团队[24]的研究中，他们从对*S*-对映体有76%ee一氧化氮双加氧酶出发，利用455个突变体来构建从序列预测功能的模型。通过该模型对涵盖了七个位置（两个区域）的组合序列空间中约168 000个变体的性能进行预测，再进行两轮筛选，共验证了360个变体后，就获得了对*S*-对映体有93%ee和对*R*-对映体有79%ee的两种优良变体。

在2018年，Manfred T. Reetz团队[25]利用一种innov’SAR的人工智能方法来指导在环氧水解酶的对映体选择性的进化过程中组合突变文库的设计，在仅使用了38个突变体的序列-功能数据的情况下，预测模型对九个位点上共512种突变体的功能进行了预测，经过简单验证后就找到了多个优于经随机突变文库筛选得到的最佳突变体的酶分子。

2019年，为了解决视紫红质通道蛋白筛选通量太低，并且要同时保留其多种特性的问题，Frances H. Arnold团队[26]使用了人工智能辅助的蛋白质工程策略（图3）。其方法为首先利用实验表征的和文献报道得到的183个序列-功能数据，构建一个分类模型，从而有效排除重组文库120 000条序列中绝大多数的非功能序列。然后根据已经表征的视紫红质通道蛋白的特性信息，针对不同的目标属性来建立不同的回归模型，例如电流强度、关闭动力学（即曝光后通道关闭所需的时间）和激活的波长敏感度等，对所有具有功能的序列进行特性的得分的预测。最后从预测库中选择少部分排名靠前的突变体（28个）进行实验验证，并得到了目标属性都优于现有的视紫红质通道蛋白的三个变体ChRger1、ChRger2和ChRger3。

2 人工智能辅助的蛋白分子设计策略概述

在人工智能辅助的蛋白分子设计策略中，本质是基于已有的数据，引入不同的机器学习算法来进行“输入特征-输出特征”的映射关系的构建。根据训练数据是否拥有标记信息（即规定的输出值），机器学习大致可划分为监督学习（supervised learning）和无监督学习（unsupervised learning）。由于在蛋白质工程中，最终目的是获得或者优化目标蛋白的一个或多个属性，因此至少会有一个属性值作为标记信息，属于监督学习[48]。

图4描述了监督学习的工作流程，主要可以分为三个步骤[27]。步骤1：通过数据库、实验和文

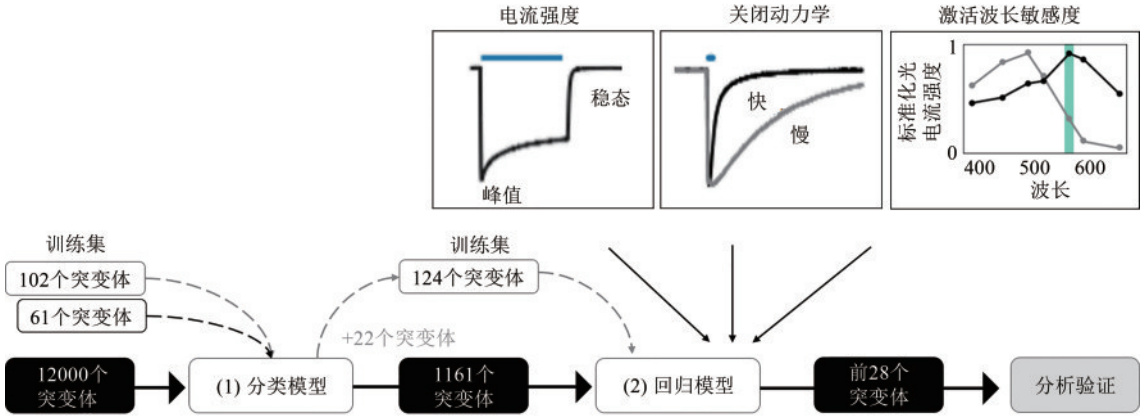


图3 人工智能辅助的视紫红质通道蛋白改造的工作流程[26]

[在重组文库中表征的102种ChR蛋白和文献中报道的61种变体，共同构成了（1）分类模型的训练集。然后，使用经过训练的分类模型来预测12000个未表征的ChR序列变体是否具有功能。接下来，构建了三个（2）回归模型，分别针对不同的ChR光电流特性：光电流强度，关闭动力学和光电流的波长敏感性]

Fig. 3 Workflow for machine learning-guided channelrhodopsin engineering[26]

[102 ChR proteins characterized in the recombinant library, together with 61 variants reported in the literature, constitute the training set of the classification model (1). Then the trained classification model was used to predict whether 12000 uncharacterized ChR sequence variants are functional, and three regression models (2) were trained, one for each of the ChR photocurrent properties of interest: photocurrent strength, off-kinetics and wavelength sensitivity of the photocurrents.]

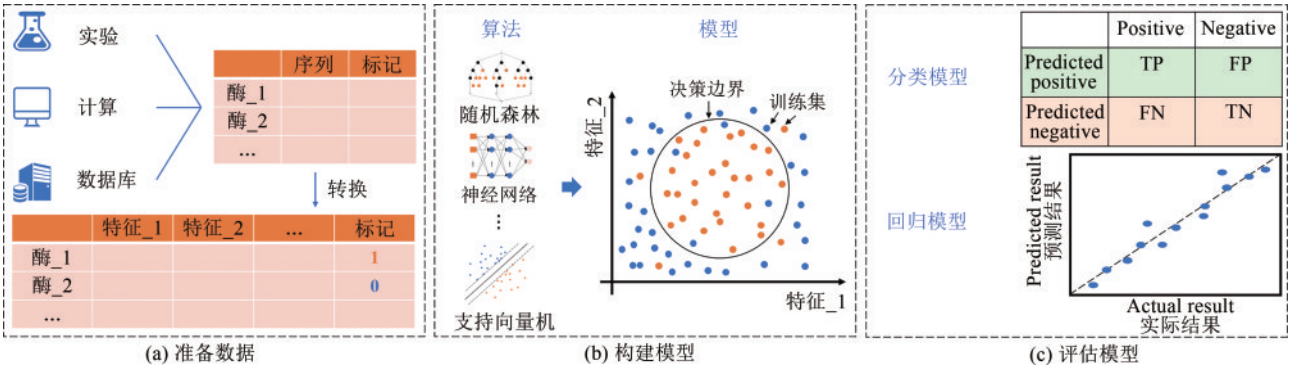


图4 监督学习的流程示意图[27]

（a）准备数据：来源于实验，计算或数据库的数据通常会转换成计算机可以识别的格式，并拆分为训练集和测试集；（b）构建预测模型：利用训练集训练不同的算法以找到决策边界，构建预测模型，例如随机森林，神经网络和支持向量机；（c）验证模型：对于分类问题或者回归问题，应选择合适的评估方法

Fig. 4 Schematic diagram of the supervised learning process[27]

Step (a): Preparing data. The data from experiments, calculations or databases are usually converted to a format that the computer can recognize and split into the training and test parts. Step(b): Constructing a predictive model. Using the training set to train different algorithms to find decision boundaries, such as random forests, neural networks and support vector machines, so as to build predictive models.
Step (c): Validating the model. An appropriate evaluation method should be selected for tasks with classification or regression.

献等方式收集初始数据，将序列作为输入特征，将蛋白质的功能信息（如对某种目标底物的活性）作为标记信息（如1代表该序列对底物有活性，0代表该序列无活性），转为计算机能够识别的数字格式，并拆分为训练集和测试集。步骤2：选用合适的算法，利用训练集进行预测模型的训练，建立起“序列-活性”的映射关系。步骤3：利用训练的模型，输入测试集的序列，得到预测值（0或1），通过比较测试集中的真实值和预测值之间的差异，评估预测模型的性能。在整个流程中，有两个关键点对预测模型的性能至关重要：数据、分子描述符和算法，人工智能方法的开发重点也是集中于这两个方面。

2.1 数据

由于人工智能算法严重依赖数据，初始数据的数量和质量决定了训练得到的模型的泛化性能[49-50]。数据集的数量不足或者质量过低会导致模型出现过拟合或者欠拟合的问题，往往会进行交叉验证来检测模型中是否存在该问题，例如k折交叉验证（即将整个数据集平均拆分为k份，每一份轮流作为测试集，其余作为训练集，如图5），因此数据收集是重要且耗时的步骤。一般来说，人工智能辅助策略很适合与其他蛋白质改造策略联用，利用在随机突变或（半）理性设计后生成的数据作为初始数据[51]。但是，一方面，就来自单轮实验的数据而言，数据集通常仅包括数十种到数百种变体，这在人工智能算法框架中属于较小的样本量[52]。另一方面，从实验中以及部分数据库中的数据是存在一定偏差的，特别是针对蛋白质某项属性进行改造时，表现不好的突变体通常直接被丢弃掉，因此导致初始数据集中数据不均匀。因此，如果采用人工智能辅助的蛋白质工程策略，应当注意收集阴性数据来保证数据的无偏性。针对训练数据的数量偏少的问题，一方面许多数据库一直在收集、整理来源于文献或实验的数据，涵盖蛋白质的序列、结构、功能和溶解度等多个属性，可以为人工智能算法提供许多优质的数据；另一方面，随着超高通量筛选和二代测序等高通量生物学实验技术的逐渐成熟，可以相信在不远的未来可用数据的数量和质量都会得到大幅度的提升，为更精准的人工智能算法提供充足的资源。

2.2 分子描述符

分子描述符（molecular descriptors），就是将分子的化学信息（例如结构特征）转换成有用的数字形式的工具。算法，即学习算法（learning algorithm），是机器学习中用于帮助计算机系统从数据中产生模型（model）、总结“经验”的方法[53]。但计算机系统仅能理解数字向量，所以算法不能直接作用于蛋白质序列[16]。因此，在获得序列之后，一般还需要利用合适的分子描述符将氨基酸序列处理为计算机能够识别的格式。以最简单的独热编码描述符为例，对于N个长度为L的多个蛋白质突变体序列，它们若在某一相同位点上包含S种不同的氨基酸（$S \leqslant N$，$S \leqslant 20$），则该位

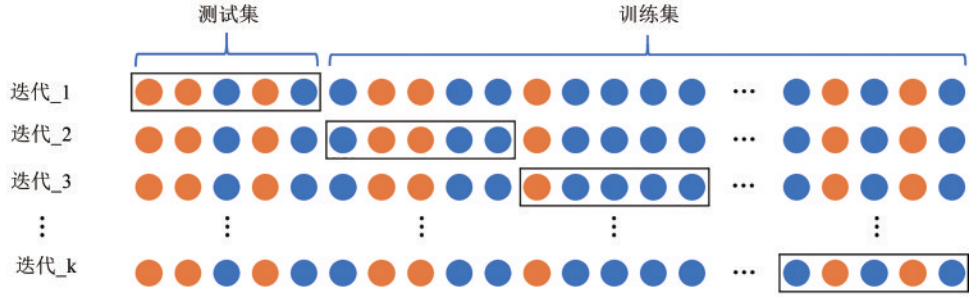


图5 k折交叉验证示意图

（将训练数据进一步细分为k个子集，并且将训练工作流程重复k次，同时保留k个子集中的一个用于评估，其余k-1个子集用于训练）

Fig. 5 Schematic diagram for k-fold cross-validation

(The training data is further subsplit into k subsets, and the training workflow is repeated k times with each of the k subsets holding for evaluation and the remaining k−1 subsets used for training)

置的所有氨基酸都可以用一个S维向量表示，每一个S维向量都包括S−1个0和一个1，其中1的位置表明该氨基酸的身份，如图6。氨基酸序列也可以根据物理性质进行编码，每种氨基酸可以由其电荷、体积或疏水性等特性或者这些特性的组合来表示，如AAindex[54]中就包含了大量类似的描述符。目前常用到的描述符有4种类型，包括基于氨基酸序列特征的描述符、结构信息描述符、嵌入式表示描述符以及突变指示描述符，在综述［16，30，55］中均有详细描述，本文不再赘述。

2.3 算法

除此之外，人工智能领域也已经提出了大量算法。从模型复杂度角度，机器学习分为经典机器学习和深度学习[56]。前者中的偏最小二乘回归[57]、支持向量机[58]、决策树/随机森林[59]和贝叶斯网络[60]等常见算法以及后者中的变分自编码器[61]、卷积神经网络[62]和循环神经网络[63]等都已用于辅助蛋白分子设计。

经典机器学习和深度学习二者的不同在于，经典机器学习算法强烈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[64]。而深度学习是通过深度神经网络，将数据进行分层抽象处理，能有效排除噪声、发现隐藏信息，因此非常适用于从高维数据发现复杂结构[56]。各个算法的入门介绍可以参考综述［16，28，31，55］。

在选择算法时，一般会以线性模型作为基线。如果线性模型的准确性不足，并且初始数据集中数据小于10 000时，偏最小二乘回归、随机森林和支持向量机都可能构建出最佳的预测模型，而神经网络则通常在更大的数据集上表现出最佳性能[16]。在计算速度方面，由于复杂程度和所需训练集大小等因素影响，深度学习往往也需要花费更多时间[55]。因此，如何选择合适的算法，需要研究者在具体的预测任务中仔细衡量准确率、计算速度和实现难度等因素。

在人工智能辅助的酶定向进化策略中，选择合适的分子描述符和机器学习算法对构建准确的预测模型而言至关重要。没有一种分子描述符和算法能够满足所有的学习任务[65]，研究人员必须结合专业知识或者同时构建多个模型进行比较。Frances H. Arnold团队使用高斯过程算法，嵌入式表示、蛋白质指数和独热编码等氨基酸编码方式进行了未知功能蛋白的功能预测，结果发现，使用嵌入式表示描述符训练的模型预测能力与其他模型的预测能力相当，甚至超过它们[66]；而在Jennifer M. Johnston等人的研究中，使用多种描述符和卷积神经网络模型构建了蛋白质序列/活性关系预测模型，结果发现，基于序列的氨基酸特性相关描述符的卷积神经网络模型表现较好，而嵌入式表示描述符表现不佳[55]。这恰恰证明了没有

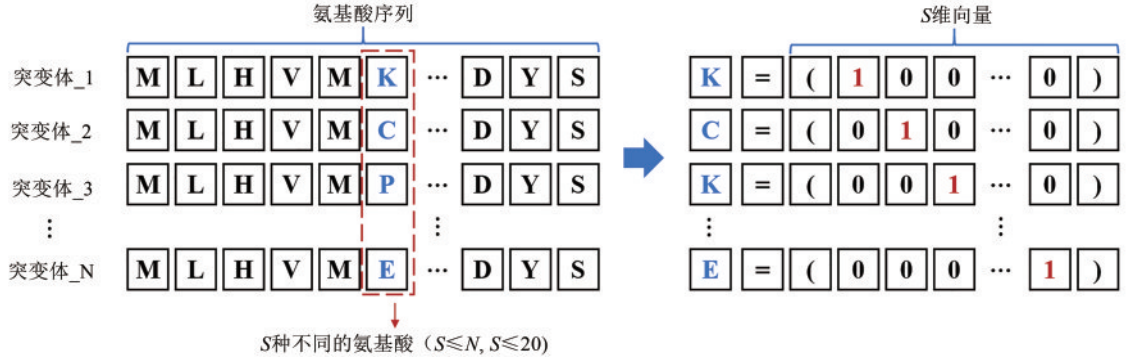


图6 独热编码示意图

（N个蛋白质突变体序列中L个氨基酸中某一相同位置包含S种不同的氨基酸，独热编码将这S个氨基酸都表示为包括S−1个0和一个1的S维向量，其中1的位置表示该位置的氨基酸的种类）

Fig. 6 Schematic diagram for one-hot encoding

(A certain position of the L amino acids in the N protein mutant sequence contains S different amino acids. The one-hot encoding represents all S amino acids as an S-dimensional vector including S−1 zeros and one 1. The position of 1 indicates the type of amino acid at that position.)

一种分子描述符和算法能够满足所有的学习任务。

3 相关的数据库和线上平台

3.1 数据库

除了与其他分子改造策略联用之外，随着高通量筛选和二代测序技术的不断发展，越来越多的蛋白质信息被挖掘，目前已经有许多优秀的数据库收集并整理了多种可作为该策略初始数据的信息，是优良的数据来源。即便数据库中大量蛋白质序列信息没有功能注释，也可以用于构建预测模型，即通过人工智能算法从这些序列中学习、提取特征，然后作为下一步从“已知特征”到“目的属性”的顶层预测模型的输入数据。例如，在2019年George M. Church团队利用了大约2400万条蛋白质序列训练递归神经网络算法，构建了一个UniRep模型[67]。该模型能够预测氨基酸序列中下一个氨基酸是什么，以此来提取氨基酸序列中不可见的特征。这些特征可以作为其他算法（如随机森林、稀疏线性回归等）的输入信息，来构建顶层特征（图7）。在应用方面，基于UniRep模型的预测模型在预测蛋白质稳定性和荧光蛋白序列优化任务中，性能都明显优于Frances H. Arnold团队曾报道的Doc2Vec模型[66]。该研究说明人工智能算法能够深度挖掘蛋白质序列中隐藏信息，为提高蛋白质工程的效率、解决蛋白质表征实验费时费力问题提供了一个全新的方法。

除了最常见的蛋白质序列和结构数据库外，越来越多的数据库在自动或手动收集整理蛋白质突变稳定性、溶解度等信息，表2对部分比较常见的数据库的类型、大小和特点进行了介绍。

3.2 线上平台

事实上，学者们已经开发了许多线上平台或者工具包来帮助人们获得蛋白质序列中的特征信息以及使用人工智能算法的工具，汇总信息见表3。大多数工具包和线上平台都只关注于完成整个生物序列分析任务的一部分，例如，大多数工具都只能利用不同类型的分子描述符从序列中生成特征。但是其中BioSeq-Analysis2.0和

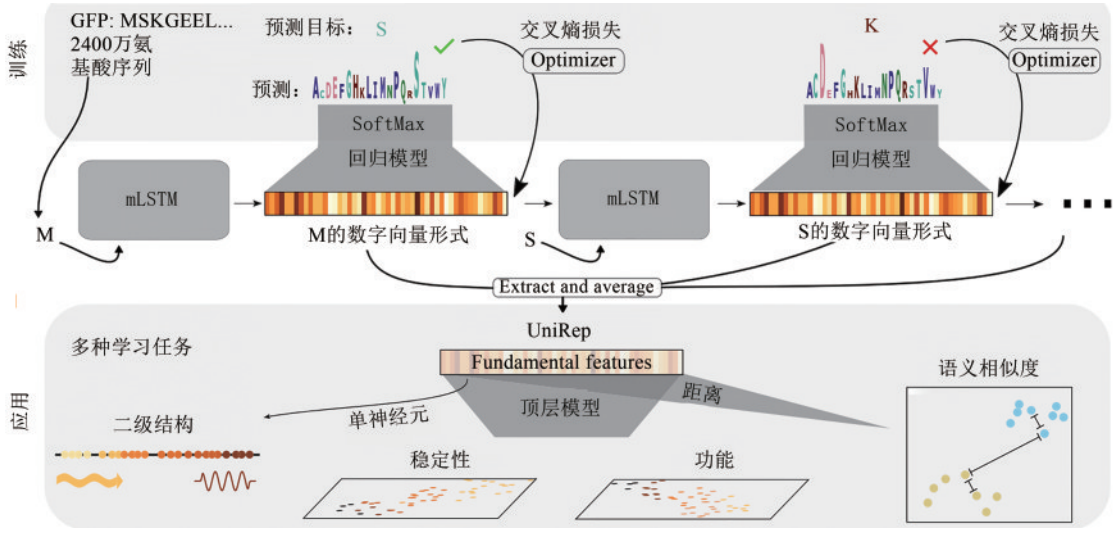


图7 UniRep模型的工作流程[67]

[在训练部分，UniRep模型使用了2400万个氨基酸序列作为训练集。然后使用训练好的模型来预测下一个氨基酸（使交叉熵损失最小化），从而学会如何正确表示氨基酸。在应用部分中，训练后的模型通过提取和平均各个氨基酸的数字向量，从而生成输入序列的单个固定长度矢量表示。这些向量可以用于训练顶级模型，从而应用于多种序列-功能预测任务]

Fig. 7 Workflow for the UniRep model[67]

[In the training part, 24 million amino acid sequences are used to train the UniRep model. Then the trained model is used to predict the next amino acid (minimizing the cross-entropy loss), so as to learn how to correctly represent the amino acid. In the application part, by extracting and assessing the numerical vector associated with the amino acid, the trained model is used to generate a single fixed-length vector representing the input sequence. Next, these vectors can be used to train top models, which can be applied to various sequence-function prediction tasks.]

表2 常见数据库汇总表

Tab. 2 Commonly used database

名称	类型	数目/大小	参考文献	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序列，UniProtKB/TrEMBL则包括了2亿多条自动注释的蛋白序列	[68]	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[66-67,69]
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	[70]	该数据库为蛋白质结构预测模型的构建提供了大量的初始数据
ProThermDB	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多条数据	[71]	突变体数据中突变类型包括野生型、单点突变和多点突变
FireProtDB	蛋白质稳定性数据的数据库	242个蛋白质的6715个变体数据	[72]	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数据库	917条突变数据	[73-78]	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	[79]	手动输入，不仅仅储存各种类型的突变信息，还提供完整的序列信息

表3 基于蛋白质序列的特征生成工具汇总表

Tab. 3 Feature generation tools based on protein sequences

名称	主要功能	类型	参考文献
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pI值6种特征	网页平台	[80]
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	[81]
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	[82]
PseAAC-General	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种基于PseAAC的分子描述符	Linux/Windows软件	[83]
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	[84]
Rcpi	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	[85]
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	[86]
ProFET	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，不支持PSSM矩阵和GO注释等非基于序列的特征	Python工具包	[87]
PseKRAAC	从蛋白质序列生成多种基于PseAAC的特征，并且利用氨基酸簇的概念，降低了特征向量的维度	网页平台	[88]
POSSUM	从蛋白质序列生成21种基于PSSM矩阵的特征	网页平台	[89]
iFeature	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18种分子描述符，并且提供12种常用的特征聚类，选择和降维算法	Python工具包/网页平台	[90]
BioSeq-Analysis2.0	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；提供多种人工智能算法构建预测模型；提供特征选择算法和模型验证方法	Windows/Linux/Unix软件/网页平台	[91]
iLearn	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；提供多种人工智能算法构建预测模型；提供特征选择算法和模型验证方法	Python工具包/网页平台	[92]
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，不能选择算法和评估方式	网页平台	[93]

iLearn两个平台可以自动执行整个蛋白序列分析任务的步骤，SOLart平台则额外引入了结构信息来预测目标蛋白质溶解度，下面进行详细阐述。

3.2.1 Protein-Analysis2.0

Protein-Analysis2.0是服务器BioSeq-Analysis2.0的线上蛋白质服务器，可以通过三个主要步骤完成蛋白序列分析任务：特征提取，预测模型构建

以及性能评估[91]。其中在特征提取方面，包括13种基于氨基酸残基水平的分子描述符和39种基于氨基酸序列水平的分子描述符。为了避免某些分子描述符导致编码后向量维度爆炸的情况，该平台还添加了两种特征选择方法。在人工智能算法方面，该平台仅整合两种分类算法（支持向量机和随机森林）和一种序列标记算法（条件随机场）。在性能评估方面，该平台支持5折交叉验证或独立数据集两种方式。同时，作者利用文献［94］中的数据作为基准数据，预测蛋白质的无序区域，其中，其中条件随机场-One-hot（6-bit）预测模型表现最佳，与文献中的方法高度可比，证明了该平台的实用性。平台地址为：

http：//bliulab.net/BioSeq-Analysis2.0/home/

3.2.2 iLearn

iLearn线上平台与BioSeq-Analysis2.0类似，不同之处在于：① iLearn平台中包含更多种分子描述；②拥有更丰富的特征分析功能，支持聚类、特征向量归一化、降维和5种特征选择方法；③支持更多的机器学习算法和更多的评估指标；④选择一种或多种机器学习算法进行提交，可以返回具有最佳性能的模型等[93]。在应用方面，作者从文献［95］中收集初始数据集和独立测试数据集，利用BLOSUM62、CKSAAP、Binary、Z-scales、AAindex、AAC和EAAC其中不同的分子描述符来进行蛋白质丙二酰化位点预测模型的构建，最终EAAC编码模型的AUC值为0.73，与原始工作中报告的AUC值为0.739相当，表明iLearn可以作为一种方便有效的工具来构建相关的预测模型。平台地址为：

https：//ilearn.erc.monash.edu/

3.2.3 SOLart

SOLart线上平台要求的输入信息仅仅是蛋白质结构，该结构可以由用户手动上传，也可以从Protein Data Bank自动上传，无需其他额外操作。其原理是在基于序列的特征（如蛋白长度和氨基酸组成）之外，引入了溶解度依赖距离电位、溶剂可及表面积和二级结构等结构特征，并以此训练随机森林算法构建预测模型。在交叉验证中，实验和预测的溶解度值之间的皮尔森相关系数几乎达到0.7，表现出了较好的预测能力[93]。平台地址为：

http：//babylone.ulb.ac.be/SOLART/index.php

4 总结

目前人工智能策略在蛋白质工程领域的应用范围主要包括蛋白质结构预测、酶功能预测、蛋白质溶解度预测以及指导智能组合文库设计等。在短短数年中，人工智能策略已经在蛋白质工程领域展现了显而易见的应用潜力和价值。要进一步挖掘人工智能在蛋白质工程领域的潜能，提升预测模型的性能，还需解决许多问题。首先，目前数据库中自动注释的蛋白质的信息质量难以让人信服，手动管理的高质量数据库中数据量的大小又远不如前者，缺少大量可用于训练和验证的标准化的数据。在后续工作中，应该构建更加高质量的基础性蛋白质序列-结构-功能数据库，有助于更加高效地构建人工智能预测模型。其数据集应该是相关的、有代表性的、非冗余的，并且包含通过实验确定的阳性和阴性数据，具有统一的标准格式等[50]。其次，在早期的实验中，更容易被表征或者具有更好表型的蛋白质往往会在后续工作中进行表征和确认，而表现不佳的蛋白质则会被丢弃，导致数据出现偏差，模型的预测性能下降[96]。此外，人工智能辅助的蛋白质工程策略还处于早期阶段，大多数例子中的预测模型可能无法直接推广应用到其他学习任务中，需要重新进行训练和验证。最后，随着越来越多的复杂的人工智能算法被用于蛋白质工程，难以对预测模型的原理进行解释等等。

随着相关研究的逐渐深入，最近已经有一些针对这些问题的研究。如今，基因功能注释领域中的自动功能预测（automatic function prediction，AFP）飞速发展，虽然还不足以解决上面提到的新蛋白质序列表征的问题，但是也已经提出一些类似于CASP竞赛性质的比赛，如CAFA[97]、EFI[98]和COMBREX[99]等。相信在未来，会出现具有足够精度的人工智能算法能准确预测新蛋白质序列的功能，为人工智能辅助的蛋白质工程提供大量优质的数据。除此之外，随着微流控筛选、荧光激活的细胞分选、噬菌体辅助连续进化等超高通量筛选技术的突破与二代测序技术的成熟，二者联用产生的蛋白质深度突变扫描技术应运而生[100-102]，应用它们来获得大量更全面、更均匀的

实验数据是未来重要的发展方向之一。并且，近几年人工智能算法仍在飞速发展，迁移学习模型取得了一些进展，除了 Frances H. Arnold 团队和 George M. Church 团队所采用的自然语言算法模型外，自动编码器和变分自编码器神经网络算法也可以从输入的蛋白质序列中生成、提取深层的特征，从而基于序列就可以执行多种预测任务。例如 Debora S. Marks 团队开发的 DeepSequence 仅基于序列就可以预测突变带来的影响[103]。最后，人工智能算法的可解释性也是重要研究方向。相信在未来，能够清晰明了地解析预测模型内部原理。随着数据和人工智能算法的不断发展，性能更好的人工智能预测模型将会成为蛋白质工程的强大工具。

参 考 文 献

[1] WAY J C, COLLINS J J, KEASLING J D, et al. Integrating biological redesign: Where synthetic biology came from and where it needs to go[J]. Cell, 2014, 157(1): 151-161.

[2] XIE M Q, HAELLMAN V, FUSSENEGGER M. Synthetic biology—application-oriented cell engineering[J]. Current Opinion in Biotechnology, 2016, 40: 139-148.

[3] BOYLE P M, SILVER P A. Parts plus pipes: Synthetic biology approaches to metabolic engineering[J]. Metabolic Engineering, 2012, 14(3): 223-232.

[4] FOO J L, CHING C B, CHANG M W, et al. The imminent role of protein engineering in synthetic biology[J]. Biotechnology Advances, 2012, 30(3): 541-549.

[5] ERB T J, JONES P R, BAR-EVEN A. Synthetic metabolism: Metabolic engineering meets enzyme design[J]. Current Opinion in Chemical Biology, 2017, 37: 56-62.

[6] PLEISS J. Protein design in metabolic engineering and synthetic biology[J]. Current Opinion in Biotechnology, 2011, 22(5): 611-617.

[7] CHEN R P, GAYNOR A S, CHEN W. Synthetic biology approaches for targeted protein degradation[J]. Biotechnology Advances, 2019, 37(8): 107446.

[8] GAINZA-CIRAUQUI P, CORREIA B E. Computational protein design—the next generation tool to expand synthetic biology applications[J]. Current Opinion in Biotechnology, 2018, 52: 145-152.

[9] BADENHORST C P S, BORNSCHEUER U T. Getting momentum: From biocatalysis to advanced synthetic biology[J]. Trends in Biochemical Sciences, 2018, 43(3): 180-198.

[10] EASON M G, DAMRY A M, CHICA R A. Structure-guided rational design of red fluorescent proteins: Towards designer genetically-encoded fluorophores[J]. Current Opinion in Structural Biology, 2017, 45: 91-99.

[11] ZEYMER C, HILVERT D. Directed evolution of protein catalysts[J]. Annual Review of Biochemistry, 2018, 87: 131-157.

[12] RIBEIRO L F, AMARELLE V, ALVES L F, et al. Genetically engineered proteins to improve biomass conversion: New advances and challenges for tailoring biocatalysts[J]. Molecules, 2019, 24(16): 2879.

[13] MARKEL U, ESSANI K D, BESIRLIOGLU V, et al. Advances in ultrahigh-throughput screening for directed enzyme evolution[J]. Chemical Society Reviews, 2020, 49(1): 233-262.

[14] LIU Q, XUN G H, FENG Y. The state-of-the-art strategies of protein engineering for enzyme stabilization[J]. Biotechnology Advances, 2019, 37(4): 530-537.

[15] LIAO J, WARMUTH M K, GOVINDARAJAN S, et al. Engineering proteinase K using machine learning and synthetic genes[J]. BMC Biotechnology, 2007, 7: 16.

[16] YANG K K, WU Z, ARNOLD F H. Machine-learning-guided directed evolution for protein engineering[J]. Nature Methods, 2019, 16(8): 687-694.

[17] SENIOR A W, EVANS R, JUMPER J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.

[18] YANG J Y, ANISHCHENKO I, PARK H, et al. Improved protein structure prediction using predicted interresidue orientations[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(3): 1496-1503.

[19] YANG M, FEHL C, LEES K V, et al. Functional and informatics analysis enables glycosyltransferase activity prediction[J]. Nature Chemical Biology, 2018, 14(12): 1109-1117.

[20] 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.

[21] HAN X, WANG X N, ZHOU K. Develop machine learning-based regression predictive models for engineering protein solubility[J]. Bioinformatics, 2019, 35(22): 4640-4646.

[22] 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. Journal of Cheminformatics, 2021, 13: 7.

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

[24] WU Z, KAN S B J, 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.

[25] CADET F, FONTAINE N, LI G Y, et al. 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.

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

[27] MAZURENKO S, PROKOP Z, DAMBORSKY J. Machine learning in enzyme engineering[J]. ACS Catalysis, 2020, 10 (2): 1210-1223.

[28] YANG X, WANG Y F, BYRNE R, et al. Concepts of artificial intelligence for computer-assisted drug discovery[J]. Chemical Reviews, 2019, 119(18): 10520-10594.

[29] BADILLO S, BANFAI B, BIRZELE F, et al. An introduction to machine learning[J]. Clinical Pharmacology and Therapeutics, 2020, 107(4): 871-885.

[30] 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 Y Y, QU G, SUN Z T. Machine learning-assisted enzyme directed evolution[J]. Journal of Biology, 2020, 37(4): 1-11.

[31] 胡如云, 张嵩亚, 蒙海林, 等 . 面向合成生物学的机器学习方法及应用[J]. 科学通报, 2021, 66(3): 284-299.
HU R Y, ZHANG S Y, MENG H L, et al. Machine learning for synthetic biology: Methods and applications[J]. Chinese Science Bulletin, 2021, 66(3): 284-299.

[32] CONSORTIUM T U. UniProt: a worldwide hub of protein knowledge[J]. Nucleic Acids Research, 2018, 47(D1): D506-D515.

[33] 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.

[34] ALQURAISHI M. AlphaFold at CASP13[J]. Bioinformatics, 2019, 35(22): 4862-4865.

[35] KINCH L N, SHI S Y, CHENG H, et al. CASP9 target classification[J]. Proteins: Structure, Function, and Bioinformatics, 2011, 79(S10): 21-36.

[36] LI Y, WANG S, UMAROV R, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning[J]. Bioinformatics, 2017, 34(5): 760-769.

[37] BOUTET E, LIEBERHERR D, TOGNOLLI M, et al. UniProtKB/Swiss-prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view[J]. Methods in Molecular Biology, 2016, 1374: 23-54.

[38] CLAUDEL-RENARD C, CHEVALET C, FARAUT T, et al. Enzyme-specific profiles for genome annotation: PRIAM[J]. Nucleic Acids Research, 2003, 31(22): 6633-6639.

[39] YU C G, ZAVALJEVSKI N, DESAI V, et al. Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases[J]. Proteins, 2009, 74(2): 449-460.

[40] KUMAR N, SKOLNICK J. EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes[J]. Bioinformatics, 2012, 28(20): 2687-2688.

[41] LI Y H, XU J Y, TAO L, et al. SVM-prot 2016: A web-server for machine learning prediction of protein functional families from sequence irrespective of similarity[J]. PLoS One, 2016, 11 (8): e0155290.

[42] ZHANG C X, FREDDOLINO P L, ZHANG Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information[J]. Nucleic Acids Research, 2017, 45(W1): W291-W299.

[43] NURSIMULU N, XU L L, WASMUTH J D, et al. Improved enzyme annotation with EC-specific cutoffs using DETECT v2 [J]. Bioinformatics, 2018, 34(19): 3393-3395.

[44] DALKIRAN A, RIFAIOGLU A S, MARTIN M J, et al. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature[J]. BMC Bioinformatics, 2018, 19(1): 334.

[45] HOU Q Z, BOURGEAS R, PUCCI F, et al. Computational analysis of the amino acid interactions that promote or decrease protein solubility[J]. Scientific Reports, 2018, 8: 14661.

[46] BHANDARI B K, GARDNER P P, LIM C S. Solubility-Weighted Index: Fast and accurate prediction of protein solubility[J]. Bioinformatics, 2020, 36(18): 4691-4698.

[47] LI G Y, DONG Y J, REETZ M T. Can machine learning revolutionize directed evolution of selective enzymes? [J]. Advanced Synthesis and Catalysis, 2019, 361(11): 2377-2386.

[48] SONG H, BREMER B J, HINDS E C, et al. Inferring protein sequence-function relationships with large-scale positive-unlabeled learning[J]. Cell Systems, 2021, 12(1): 92-101.e8.

[49] YU L A, WANG S Y, LAI K K. An integrated data preparation scheme for neural network data analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2): 217-230.

[50] YANG Y, UROLAGIN S, NIROULA A, et al. PON-tstab: Protein variant stability predictor. importance of training data quality[J]. International Journal of Molecular Sciences, 2018, 19 (4): 1009.

[51] SIEDHOFF N E, SCHWANEBERG U, DAVARI M D. Machine learning-assisted enzyme engineering[J]. Methods in Enzymology, 2020, 643: 281-315.

[52] BASTIAN F B, CHIBUCOS M C, GAUDET P, et al. The Confidence Information Ontology: A step towards a standard for asserting confidence in annotations[J]. Database, 2015, 2015: bav043.

[53] 周志华 . 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University

Press, 2016.

[54] KAWASHIMA S, POKAROWSKI P, POKAROWSKA M, et al. AAindex: amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2007, 36(suppl 1): D202-D205.

[55] XU Y T, VERMA D, SHERIDAN R P, et al. Deep dive into machine learning models for protein engineering[J]. Journal of Chemical Information and Modeling, 2020, 60(6): 2773-2790.

[56] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[57] ABDI H. Partial least squares regression and projection on latent structure regression (PLS Regression)[J]. WIREs Computational Statistics, 2010, 2(1): 97-106.

[58] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[59] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[60] HECKERMAN D. A tutorial on learning with Bayesian networks [M]//HOLMES D E, JAIN L C. Innovations in Bayesian networks: theory and applications. Berlin, Heidelberg; Springer , 2008: 33-82.

[61] SINAI S, KELSIC E, CHURCH G M, et al. Variational autoencoding of protein sequences[EB/OL]. 2017: arXiv: 1712.03346[q-bio.QM]. https://arxiv.org/abs/1712.03346

[62] LECUN Y, BENGIO Y. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10): 1995.

[63] LOHMANN R, SCHNEIDER G, BEHRENS D, et al. A neural network model for the prediction of membrane-spanning amino acid sequences[J]. Protein Science, 1994, 3(9): 1597-1601.

[64] 曲戈, 朱彤, 蒋迎迎, 等. 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 Y Y, et al. Protein engineering: From directed evolution to computational design[J]. Chinese Journal of Biotechnology, 2019, 35(10): 1843-1856.

[65] WOLPERT D H. The lack of *A priori* distinctions between learning algorithms[J]. Neural Computation, 1996, 8(7): 1341-1390.

[66] YANG K K, WU Z, BEDBROOK C N, et al. Learned protein embeddings for machine learning[J]. Bioinformatics, 2018, 34 (15): 2642-2648.

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

[68] CONSORTIUM T U, BATEMAN A, MARTIN M J, et al. UniProt: the universal protein knowledgebase in 2021[J]. Nucleic Acids Research, 2020, 49(D1): D480-D489.

[69] RIVES A, MEIER J, SERCU T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[EB/OL]. bioRxiv, 2020, DOI: 10.1101/622803.

[70] BURLEY S K, BHIKADIYA C, BI C X, et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences[J]. Nucleic Acids Research, 2020, 49(D1): D437-D451.

[71] NIKAM R, KULANDAISAMY A, HARINI K, et al. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years[J]. Nucleic Acids Research, 2020, 49(D1): D420-D424.

[72] STOURAC J, DUBRAVA J, MUSIL M, et al. FireProtDB: database of manually curated protein stability data[J]. Nucleic Acids Research, 2020, 49(D1): D319-D324.

[73] TIAN Y, DEUTSCH C, KRISHNAMOORTHY B. Scoring function to predict solubility mutagenesis[J]. Algorithms for Molecular Biology: AMB, 2010, 5: 33.

[74] SORMANNI P, APRILE F A, VENDRUSCOLO M. The CamSol method of rational design of protein mutants with enhanced solubility[J]. Journal of Molecular Biology, 2015, 427 (2): 478-490.

[75] ZAMBRANO R, JAMROZ M, SZCZASIUK A, et al. AGGRESCAN3D (A3D): Server for prediction of aggregation properties of protein structures[J]. Nucleic Acids Research, 2015, 43 (W1): W306-W313.

[76] YANG Y, NIROULA A, SHEN B R, et al. PON-Sol: Prediction of effects of amino acid substitutions on protein solubility[J]. Bioinformatics, 2016, 32(13): 2032-2034.

[77] PALADIN L, PIOVESAN D, TOSATTO S C E. SODA: prediction of protein solubility from disorder and aggregation propensity[J]. Nucleic Acids Research, 2017, 45(W1): W236-W240.

[78] MAZURENKO S. Predicting protein stability and solubility changes upon mutations: data perspective[J]. Chem Cat Chem, 2020, 12,5590-5598.

[79] WANG C Y, CHANG P M, ARY M L, et al. ProtaBank: A repository for protein design and engineering data[J]. Protein Science: a Publication of the Protein Society, 2018, 27(6): 1113-1124.

[80] SHEN H B, CHOU K C. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition[J]. Analytical Biochemistry, 2008, 373(2): 386-388.

[81] RAO H B, ZHU F, YANG G B, et al. Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence[J]. Nucleic Acids Research, 2011, 39(suppl_2): W385-W390.

[82] CAO D S, XU Q S, LIANG Y Z. Propy: a tool to generate various modes of Chou's PseAAC[J]. Bioinformatics, 2013, 29(7): 960-962.

[83] DU P F, GU S W, JIAO Y S. PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid

composition for large-scale protein datasets[J]. International Journal of Molecular Sciences, 2014, 15(3): 3495-3506.

[84] XIAO N, CAO D S, ZHU M F, et al. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences[J]. Bioinformatics, 2015, 31(11): 1857-1859.

[85] CAO D S, XIAO N, XU Q S, et al. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions[J]. Bioinformatics, 2014, 31(2): 279-281.

[86] LIU B, LIU F L, WANG X L, et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences[J]. Nucleic Acids Research, 2015, 43(W1): W65-W71.

[87] OFER D, LINIAL M. ProFET: Feature engineering captures high-level protein functions[J]. Bioinformatics, 2015, 31(21): 3429-3436.

[88] ZUO Y C, LI Y, CHEN Y L, et al. PseKRAAC: a flexible web server for generating pseudo *K*-tuple reduced amino acids composition[J]. Bioinformatics, 2016, 33(1): 122-124.

[89] WANG J W, YANG B J, REVOTE J, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles[J]. Bioinformatics, 2017, 33 (17): 2756-2758.

[90] CHEN Z, ZHAO P, LI F Y, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences[J]. Bioinformatics, 2018, 34(14): 2499-2502.

[91] LIU B, GAO X, ZHANG H Y. BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches[J]. Nucleic Acids Research, 2019, 47(20): e127.

[92] CHEN Z, ZHAO P, LI F Y, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data [J]. Briefings in Bioinformatics, 2019, 21(3): 1047-1057.

[93] HOU Q Z, KWASIGROCH J M, ROOMAN M, et al. SOLart: a structure-based method to predict protein solubility and aggregation[J]. Bioinformatics, 2019, 36(5): 1445-1452.

[94] LIU Y M, WANG X L, LIU B. IDP-CRF: intrinsically disordered protein/region identification based on conditional random fields[J]. International Journal of Molecular Sciences, 2018, 19(9): 2483.

[95] CHEN Z, HE N N, HUANG Y, et al. Integration of A deep learning classifier with A random forest approach for predicting malonylation sites[J]. Genomics, Proteomics & Bioinformatics, 2018, 16(6): 451-459.

[96] SCHNOES A M, REAM D C, THORMAN A W, et al. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space[J]. PLoS Computational Biology, 2013, 9(5): e1003063.

[97] ZHOU N H, JIANG Y X, BERGQUIST T R, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens[J]. Genome Biology, 2019, 20(1): 244.

[98] GERLT J A, ALLEN K N, ALMO S C, et al. The enzyme function initiative[J]. Biochemistry, 2011, 50(46): 9950-9962.

[99] ROBERTS R J, CHANG Y C, HU Z J, et al. COMBREX: a project to accelerate the functional annotation of prokaryotic genomes[J]. Nucleic Acids Research, 2010, 39(S1): D11-D14.

[100] FOWLER D M, FIELDS S. Deep mutational scanning: A new style of protein science[J]. Nature Methods, 2014, 11(8): 801-807.

[101] FOWLER D M, STEPHANY J J, FIELDS S. Measuring the activity of protein variants on a large scale using deep mutational scanning[J]. Nature Protocols, 2014, 9(9): 2267-2284.

[102] NEWBERRY R W, LEONG J T, CHOW E D, et al. Deep mutational scanning reveals the structural basis for α-synuclein activity[J]. Nature Chemical Biology, 2020, 16(6): 653-659.

[103] Riesselman A J, Ingraham J B, Marks D S. Deep generative models of genetic variation capture the effects of mutations[J]. Nature Methods, 2018, 15(10): 816-822.

通讯作者：杨广宇（1980—），男，研究员，博士生导师。研究方向为酶定向进化、酶高通量筛选、酶技术应用、体外合成生物学等。

E-mail：yanggy@sjtu.edu.cn

第一作者：卞佳豪（1997—），男，硕士研究生。研究方向为人工智能辅助定向进化的组合方法。

E-mail：bjh2170@sjtu.edu.cn