基于人工智能的蛋白质属性预测的潜能与应用

李金甲[1]，陈都鑫[2]，柴人杰[3*]，虞文武[2**]

（1. 东南大学-蒙纳士大学联合研究生院，江苏 苏州 215123；2. 东南大学数学学院，江苏 南京 210096；3. 东南大学生命科学与技术学院，江苏 南京 210096）

[摘要] 蛋白质作为生命系统的核心构成要素，其结构和功能的精准理解对揭示生物学过程及药物开发至关重要。随着人工智能技术的飞速发展，其在蛋白质工程领域的应用，特别是在蛋白质属性预测方面，已成为研究的热点。综述简介了生物医学工程领域的最新动态，突出了人工智能技术在蛋白质工程中的重要作用，并着重讨论了人工智能在蛋白质属性预测方面取得的创新性成果，及其在此领域的应用潜力与面临挑战。

[关键词] 人工智能；深度学习；蛋白质工程；蛋白质属性预测

[中图分类号] TP18 [文献标志码] A [文章编号] 1001-5094（2023）10-0733-08

DOI: 10.20053/j.issn1001-5094.2023.10.003

Potential and Application of Artificial Intelligence-Based Protein Attribute Prediction

LI Jinjia[1], CHEN Duxin[2], CHAI Renjie[3], YU Wenwu[2]

(1. Southeast University-Monash University Joint Graduate School, Suzhou 215123, China; 2. School of Mathematics, Southeast University, Nanjing 210096, China; 3. School of Life Science and Technology, Southeast University, Nanjing 210096, China)

[Abstract] Proteins are the core building blocks of living systems, and a precise understanding of their structure and function is essential for revealing biological processes and drug development. With the rapid development of artificial intelligence technology, its application in the field of protein engineering, especially in the prediction of protein attribute, has become a hot research topic. This article reviews the latest developments in the field of biomedical engineering, with focus on the important role of artificial intelligence technology in protein engineering, and highlights the innovative achievements of artificial intelligence algorithms in predicting protein attribute, with deep discussion on its potential application and challenges in the field.

[Key words] artificial intelligence; deep learning; protein engineering; protein attribute prediction

生物医学工程是一个多学科交叉的领域，其主要特点是运用工程学和应用科学的知识与技术解决生物学和医学领域的科学问题，进而充分研究生命系统及其行为，并开发相关的生物医学系统，最终提高人类健康水平[1]。生物医学工程这个领域的范围非常广泛，涉及生物学、医学、工程学和计算科学等多个领域的交叉融合。该领域的重要性在于它为医学领域带来了革命性的创新，对于改善疾病诊断和治疗、提高手术和医疗设备的效率、开发新型药物和生物疗法，以及推动生物学研究的进展具有巨大的潜力[2]。

近年来，人工智能（artificial intelligence，AI）在生物医学工程中的作用日益显著，特别是在蛋白质属性预测方面。蛋白质属性预测作为蛋白质工程的一个重要领域，其重要性体现在能够提供对蛋白质的深入理解，从而指导药物设计、疾病治疗等方面。准确预测蛋白质的结构和功能等属性对于揭示其在生物学过程中的作用和发挥其潜在应用价值至关重要。蛋白质属性预测存在的难点包括蛋白质复杂空间结构的理解、蛋白质功能机制的多样性和复杂性及蛋白质间复杂的相互作用模式等。这些挑战使得传统的生物学方法和实验技术在蛋白质属性预测精度和效率上受到限制。随着 AI 技术的进步，

接受日期：2023-10-08

项目资助：国家自然科学基金（No.62203004，No. 62273090，No. 62073076）

* 通信作者：柴人杰，教授，博士生导师；

研究方向：神经干细胞和内耳干细胞的转录调控机制，听觉神经元和毛细胞再生；

E-mail：renjiec@seu.edu.cn

** 通信作者：虞文武，教授，博士生导师；

研究方向：科学与人工智能交叉领域；

Tel：025-52090590；E-mail：wwyu@seu.edu.cn

特别是深度学习和机器学习的发展，以上难点逐渐得以解决。AI 技术的应用，如利用卷积神经网络（convolutional neural network，CNN）进行蛋白质结构预测和利用图神经网络分析蛋白质相互作用，已显著提高了蛋白质属性预测的准确性和效率。AI 不仅加速了蛋白质属性的预测过程，也为新药发现和生物学研究提供了新的工具。本文将围绕应用 AI 技术进行蛋白质属性预测这一主题，深入探讨 AI 在蛋白质结构预测、功能预测这 2 种属性预测方面的应用。

1 人工智能促进生物医学工程研究

1.1 人工智能方法演进

机器学习是人工智能 AI 领域的核心组成部分，它涵盖了一系列算法和技术，使计算机系统能够从数据中学习并不断改进模型的性能。机器学习这一领域的发展得益于计算能力的提高、大规模数据集的可用性以及算法的不断创新。机器学习在生物医药领域的应用范围广泛，包括但不限于基因组学、药物发现、疾病诊断和治疗优化等。

近 30 年来，机器学习领域经历了显著的发展，孕育出众多创新的算法和模型，如主成分分析（principal component analysis，PCA）、支持向量机（support vector machines，SVM）、随机森林和谱聚类方法等。这些方法在数据降维、分类和回归等任务中取得了显著的成效 [3]。PCA 通过线性变换找到数据中的主要特征，有效地减少数据维度和复杂性 [4]。SVM 通过寻找最优的超平面对数据进行分割，提高了数据分类的准确性和效率 [5]。随机森林作为一种集成学习方法，通过构建多个决策树并综合其结果，提升了对数据的分类和回归预测能力 [6]。谱聚类作为一种基于图论的强大聚类方法，通过分析数据的相似性矩阵，能在复杂数据集中识别出固有的群组结构 [7]。这些方法为机器学习的发展提供了坚实的基础。

深度学习是机器学习领域的一个分支，近年来取得了巨大的突破和成果，其发展历程可追溯到 20 世纪 80～90 年代的神经网络研究。然而，由于计算资源和数据集的限制，神经网络在那个时期并没有得到广泛应用。随着计算机计算能力的提升和大规模数据集的可用性，深度学习在 2006 年以后迅速发展起来。其中，深度学习的一个重要里程碑是 Hinton 等 [8] 在 2006 年提出了深度信念网络（deep belief network，DBN）；DBN 是一种多层次的神经网络模型，通过无监督学习逐层训练，可以学习到更抽象和更高级的数据特征表示。

1.2 深度学习的核心研究领域

计算机视觉和自然语言处理为 AI 领域的 2 个核心分支，一直以来都备受关注。计算机视觉致力于使计算机系统能够理解和解释图像、视频以及其他视觉数据，从而模拟人类视觉系统的功能。计算机视觉领域的研究涵盖了图像识别、物体检测、图像生成等众多任务，其应用包括自动驾驶 [9]、医疗影像分析 [10] 等众多领域，具有巨大的社会和经济价值。

自然语言处理旨在使计算机能够理解、分析和生成人类语言的文本数据。该领域包括了文本分类、情感分析、机器翻译、自动问答等任务，其应用广泛，涵盖了虚拟助手、智能搜索、智能客服等领域。自然语言处理的研究不仅涉及语言的语法和语义分析，还包括处理多语言数据、非结构化文本数据等复杂问题 [10]。

深度学习在计算机视觉和自然语言处理领域拥有诸多常见的算法，它们具有极强的通用性。在计算机视觉领域，CNN 是最重要的模型之一，通过卷积层、池化层和全连接层来提取图像特征并进行分类、检测和分割 [11]。此外，残差网络（residual network，ResNet）通过引入残差连接解决了深层网络的退化问题 [12]；Inception 网络引入了 Inception 模块和瓶颈层以提高计算效率和性能 [13]。在自然语言处理领域，递归神经网络（recurrent neural network，RNN）和长短期记忆网络（long short term memory network，LSTM）是常用的序列建模工具，RNN 通过循环连接处理序列数据，LSTM 通过引入门控机制解决了传统 RNN 在处理长序列数据时遇到的梯度消失问题 [14]。此外，注意力机制（attention mechanism）在自然语言处理任务中得到广泛应用，它能够提取关键信息，改善模型性能 [15]。

1.3 人工智能在生物医学工程领域的应用

深度学习技术在生物医学工程领域的应用引起

了广泛关注。生物医学工程领域数据量巨大、特征复杂，深度学习强大的表示学习和模式识别能力使其非常适合处理这些数据。深度学习在生物医学工程领域的应用范围包括基因组学、蛋白质研究、医学图像分析、药物发现和个性化医疗[16]。例如，在医学图像分析中，深度学习模型可以自动识别和定位病灶、分割器官结构，协助医生制定诊断和治疗计划[17]。此外，深度学习还在药物发现和个性化医疗中发挥着重要作用，加速新药的开发和治疗方案的优化[18-19]。

总的来说，在 AI 领域，机器学习作为一个关键技术，对生物医学工程领域产生了深远的影响。AI 及机器学习不仅加速了科学研究的进展，也为疾病的诊断和治疗提供了更加精准、个性化的解决方案。随着技术的不断发展和创新，可以预见，AI 在生物医学工程领域的作用将不断增强，为人类健康和医疗做出更大的贡献[16]。

2 蛋白质工程与人工智能

2.1 蛋白质工程概述

蛋白质工程旨在创建具有特定功能的蛋白质，如改善生物体的特征、增强酶的催化性能和提高抗体的效力[20]。该领域对药物发现、酶开发、生物传感器、诊断学以及其他生物技术的进步产生了深远影响，同时也为理解蛋白质结构与功能之间的关系提供了基础原理。此外，蛋白质工程还对可持续性和环保产生了积极影响。例如，通过设计和优化工业用酶，可以实现更环保的化学反应过程，减少有害废物的产生。蛋白质工程领域有望持续推动创新，为未来生活的改进提供可能性。

在蛋白质工程领域，主要采用了 2 种传统方法，分别是定向进化[21]和理性设计[22-23]。定向进化是一种用于创建具有改进或新功能的蛋白质或酶的过程[24]。定向进化方法涉及将突变引入目标蛋白质的遗传密码，然后筛选所得的变体以改善其功能。这个过程被称为“定向”，因为它受到期望结果的指导，例如提高活性、稳定性、特异性、结合亲和力和适应性。另一方面，理性设计则利用对蛋白质结构和功能的了解，有针对性地对蛋白质序列或结构进行特定的修改[23,25]。这 2 种方法均需要进行实验筛选，但考虑到蛋白质中氨基酸残基的多样性，这是昂贵、耗时且复杂的过程[26]。因此，即使使用最先进的高通量筛选技术，也只能对蛋白质中的一小部分突变空间进行实验探索。

2.2 机器学习辅助蛋白质工程

近年来，数据驱动的机器学习为定向进化和蛋白质工程方法[27-28]提供了新的解决方案。机器学习辅助蛋白质工程是指应用机器学习模型和技术，以提高蛋白质工程的效率和效力。该方法不仅能够降低成本并加速蛋白质工程的进展，还能够优化蛋白质的筛选和变体选择[29]，从而提高了工作效率和生产率。具体而言，通过机器学习分析和预测突变对蛋白质功能的影响，研究人员可以快速生成和测试大量变体，从而建立蛋白质的适应度映射关系（即适应度景观），然后采集实验数据[30-31]。这种方法极大地加速了蛋白质工程的进程。

数据驱动的机器学习辅助蛋白质工程的过程通常包括多个要素，如数据采集和预处理、模型设计、特征提取和选择、算法选择和设计、模型训练与验证、实验验证以及模型优化的反复迭代。电化学生物传感器和微流控技术的进步在高通量测序和筛选技术方面发挥着重要作用，积累了大量的蛋白质序列、结构和功能的通用实验数据集[32-33]。这些数据集以及专门用于蛋白质工程的深度突变扫描库[34]，为机器学习的训练和验证提供了宝贵的资源。

数据表示和特征提取是机器学习模型设计的关键步骤，因其有助于简化生物数据的复杂性，实现更有效的模型训练和预测。有多种典型的特征类型表示方法，包括基于序列、基于结构[35-36]、基于物理[37-38]和混合方法[39]。其中，基于序列的表示一直占据主导地位，因其成功利用了自然语言处理（natural language processing，NLP）方法，如 LSTM[40]、自动编码器[41]和 Transformers[42]，允许在大规模序列数据上进行无监督的预训练。基于结构的嵌入则依赖于现有蛋白质三维结构数据库[43]和高级结构预测，例如 AlphaFold2[43]；这些方法进一步利用先进的数学工具，如拓扑数据分析[44-45]、微分几何[46-47]或图形方法[48]。基于物理的方法使用物理模型，如密度泛函理论[49]、分子力学[50]、泊松玻

尔兹曼模型[51]等；虽然这些方法具有高度可解释性，但通常性能取决于模型的参数设置。混合方法可以选择多种特征类型的组合。

机器学习辅助蛋白质工程算法的设计和选择是受数据可用性和实验效率影响的。在实际应用中，常见情况是小规模标记训练数据集的场景，对于这种情况，通常使用简单的机器学习算法，如支持向量机和集成方法；而对于大规模训练数据集，深度神经网络更为适用。除了回归模型，还可以考虑使用无监督零样本学习方法来应对标记数据有限的情况[52-53]。实验和模型之间的迭代作用通过反复筛选和新数据的引入，构成了机器学习辅助蛋白质工程的另一个重要组成部分。因此，选择适当的模型需要考虑实验频率和实验成本等因素的影响。这个迭代细化的过程使机器学习辅助蛋白质工程能够提供优化的蛋白质工程成果。

3 基于人工智能方法的蛋白质属性预测

蛋白质属性预测领域聚焦于运用计算方法（如AI和机器学习技术）来预测蛋白质的结构、功能和相互作用等关键属性。蛋白质属性预测领域是蛋白质工程的核心组成之一。这些预测工具对于深入理解蛋白质的生物学特性以及设计和改良具有特定功能的蛋白质至关重要。借助这些预测结果，研究人员能够更加有效地开发出具有预定功能的蛋白质。

3.1 蛋白质属性预测的重要性

蛋白质是生命过程中不可或缺的分子实体，其展现的多样性和复杂性对生物学和医学研究至关重要。根据Koehler Leman等[54]的研究，蛋白质属性包括氨基酸序列、三维空间结构、生物学功能及其与其他分子的相互作用，尤其重要的是，蛋白质的三维结构和功能在其生物学角色中具有决定性作用；蛋白质精细调整的三维结构是由其氨基酸序列所决定的，这是其执行多样化分子功能的关键。因此，深入理解氨基酸序列与蛋白质结构之间的关系，对于推动生物学的理解和医学的应用具有重大意义。

首先，蛋白质的三维结构是其功能的决定性因素。蛋白质分子通过其独特的空间构型与其他分子相互作用，实现包括催化生化反应和信号传递等多种生物学功能。准确预测蛋白质的三维结构对理解其功能机制至关重要。近年来，随着计算技术特别是AI在蛋白质结构预测中的应用不断发展，从序列到结构的解析时间已显著缩短，预测精度也得到了提升。Kuhlman等[55]研究指出，如何通过这些技术进步来推动蛋白质结构预测和设计的前沿研究。

其次，蛋白质功能的预测对于生物医学领域的研究至关重要。蛋白质在生物体内承担的功能（如酶的催化活性和受体的信号传导）是其在生命过程中发挥关键作用的基础。尽管高通量测序技术的发展能迅速获取大量蛋白质序列，但对这些序列的功能理解仍不充分。Jeffery等[56]研究认为，开发有效的计算方法预测蛋白质的潜在功能对于新药物的开发和疾病机制的研究等极为关键。

最后，蛋白质间的相互作用构成了一项关键属性。蛋白质-蛋白质相互作用（protein-protein interactions，PPIs）是细胞内多种生物过程的基本组成部分，对信号传导和代谢途径等具有显著影响。传统的实验方法虽能提供关于PPIs的数据，但通常耗时费力且易产生假阳性结果。Durham等[57]研究认为，计算方法在预测PPIs方面的作用日益凸显，其目的是更高效地识别和验证蛋白质间的相互作用，从而推动生物学和医学领域的研究和应用。

3.2 人工智能在蛋白质结构预测方面的应用

在蛋白质结构预测领域，AI技术的突破性进展尤为显著，特别是DeepMind公司开发的AlphaFold2和Baker实验室的RoseTTAFold在该领域的应用成果备受瞩目。

AlphaFold2在关键评估蛋白质结构预测竞赛（critical assessment of protein structure prediction，CASP）中以其创新性算法大放异彩，实现了令人印象深刻的92.4的中位数得分，远超90分的高准确性标准，这一成绩意味着其预测的结构与实验确定的结构高度吻合，大幅超越了传统预测方法[58]。与此同时，RoseTTAFold同样运用深度学习技术，仅依赖一块RTX2080显卡，便可在大约10 min内完成不超过400个氨基酸残基的蛋白质结构预测[59]。这些先进的AI工具不仅大幅提升了蛋白质结构预测的精准度，还显著缩减了从蛋白质序列解析到结构

预测的时间，对于揭示生物分子的功能机制、促进相关疾病的研究和治疗具有重大意义。

从 AI 技术的视角看，AlphaFold 通过运用深度学习技术学习蛋白质氨基酸序列与其三维结构之间的复杂关系。AlphaFold 的创新性在于将多序列比对（multiple sequence alignment，MSA）数据与物理生物学信息相结合，预测氨基酸序列的距离和角度[60]。AlphaFold2 的主要创新在于其深度学习架构，其利用了自注意力机制的 Transformer 架构和一个名为“Evoformer”的模块来有效地整合蛋白质序列和结构信息，从而提高了预测精度。这一架构特别擅长捕捉蛋白质序列中的模式，并结合进化信息来预测蛋白质的三维结构[58]。另一方面，RoseTTAFold 则采用了一种三轨神经网络，其可以兼顾蛋白质序列的模式、氨基酸如何相互作用以及蛋白质可能的三维结构，其多轨神经网络架构能够同时处理不同维度的信息，从而有效学习预测蛋白质结构[59]。上述模型不仅展示了深度学习在生物学领域的巨大潜力，也为生物医学研究和药物开发提供了新的可能性。

3.3 人工智能在蛋白质功能预测方面的应用

在蛋白质功能预测领域，AI 技术也显现出其巨大潜力。通过综合分析蛋白质的氨基酸序列和结构信息，AI 算法能够预测蛋白质的功能类别、活性位点及其潜在的作用对象。例如，DeepFRI 融合了自监督语言模型和图卷积网络的先进方法，能够利用从蛋白质序列的自监督模型中提炼出的序列特征及蛋白质结构来预测其功能[61]。DeepFRI 在性能方面超越了现有的先进方法（如 DeepGO 和 FunFams），其设计允许扩展序列数据库的规模。此外，DeepFRI 通过使用同源建模来增加训练样本的数量，显著增加了可预测的蛋白质功能的数量，减少了训练数据中正负例之间的不平衡。值得强调的是，即使在使用由计算方法生成的蛋白质结构来代替实验获得的蛋白质结构时，DeepFRI 的预测性能仅略有降低，表明 DeepFRI 具有一定的去噪能力。另一方面，Bileschi 等[62]利用深度学习模型对未经比对的氨基酸序列进行功能注释的准确预测，这些模型不仅推断出已知的进化替代模式，还学会了准确聚类未见家族的序列；该方法通过分析和比对蛋白质序列，扩大了 Pfam 蛋白家族数据库的覆盖范围，即预测了 360 种之前未在 Pfam 数据库中详细注释的蛋白质的功能。此外，Hakala 等[63]开发了一个综合系统，结合了随机森林和神经网络分类器，对输入的蛋白质序列进行基因本体论（GO）术语的预测；在 CAFA3 评估中，该模型展现出了竞争性的性能，在超过 100 个提交系统中排名前列。

在蛋白质功能预测领域中，AI 技术的核心方法涵盖了多种先进的计算模型和分析工具。ProteinBERT 使用基于双向编码器表示变换器（bidirectional encoder representations from transformers，BERT）的深度学习模型来学习蛋白质氨基酸序列和自然语言之间的相似性，从而有效地编码蛋白质序列并捕捉其生物学性质，以预测蛋白质的结构和功能[64]。InterProScan 被用于识别蛋白质家族和功能域，结合了多个数据库和预测工具，如 Pfam、PROSITE、SUPERFAMILY 等，为全面分析蛋白质功能域提供支持[65]。BLAST 和 HMMER 等工具利用启发式算法和隐藏马尔可夫模型，分别快速识别序列间的局部相似性和更精确地识别序列的同源性，从而有助于功能预测和探究蛋白质与核酸序列的进化关系。DALI 和 TM-align 专注于蛋白质三维结构的精确比较与对齐。DALI 通过对已知结构数据库进行查询，利用预计算的结构相似性进行分层分类，实现结构的比较。TM-align 结合 TM-score 旋转矩阵和动态规划，提供比现有方法更快速、准确的蛋白质结构相似性度量方法。DALI 和 TM-align 这 2 个工具在生物信息学和结构生物学领域中，对于理解蛋白质的功能预测和进化关系具有重要作用。上述方法的综合应用，在生物医学工程领域对蛋白质功能预测提供了强有力的支持。

4 结语与展望

AI 技术，尤其是深度学习，已在蛋白质结构和功能预测方面取得显著进展，这对于加速药物发现和疾病理解至关重要。蛋白质间的相互作用预测同样重要，AI 方法可用于揭示细胞内的信号传导和代谢途径。然而，AI 模型通常被视为“黑盒”，其预测结果难以解释，这是未来研究的一个重点。未来研究还应关

注以下方向：1）更多的数据收集和处理方法的开发，以应对噪声和错误；2）跨学科合作的促进，加速生物学、计算机科学和化学等领域的创新；3）开发可解释性强的 AI 模型，以帮助理解蛋白质属性预测的基础。相信 AI 技术在生物医学工程领域的作用将不断增强，为人类健康和医疗做出更大的贡献。

[参考文献]

[1] Houssein A, Lefor A K, Veloso A, *et al*. BMC Biomedical Engineering: a home for all biomedical engineering research[J]. *BMC Biomed Eng*, 2019, 1: 1–4. DOI: 10.1186/s42490-019-0004-1.

[2] Lantada A D, Morgado P L. Rapid prototyping for biomedical engineering: current capabilities and challenges[J]. *Annu Rev Biomed Eng*, 2012, 14: 73–96. DOI: 10.1146/annurev-bioeng-071811-150112.

[3] Dong S, Wang P, Abbas K. A survey on deep learning and its applications[J]. *Comput Sci Rev*, 2021, 40: 100379. DOI: 10.1016/j.cosrev.2021.100379.

[4] Lever J, Krzywinski M, Altman N. Points of significance: principal component analysis[J]. *Nat Methods*, 2017, 14(7): 641–643.

[5] Cortes C, Vapnik V. Support-vector networks[J]. *Mach Learn*, 1995, 20(3): 273–297.

[6] Breiman L. Random forests[J]. *Mach Learn*, 2001, 45(1): 5–32.

[7] Qin X, Dai W, Jiao P, *et al*. A multi-similarity spectral clustering method for community detection in dynamic networks[J]. *Sci Rep*, 2016, 6(1): 31454. DOI: 10.1038/srep31454.

[8] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. *Neural Comput*, 2006, 18(7): 1527–1554.

[9] 杜明宇 . 自动驾驶综述 [J]. 中国科技纵横 , 2018(6): 215–216.

[10] 倪杭建 , 何必仕 , 徐哲 , 等 . 区域医疗影像重复检查分析及关联挖掘 [J]. 中国医疗设备 , 2016, 31(10): 154–158.

[11] Liu S, Deng W. Very deep convolutional neural network based image classification using small training sample size[C]//2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). New York: IEEE, 2015: 730–734.

[12] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). New York: IEEE, 2016: 770–778.

[13] Szegedy C, Liu W, Jia Y, *et al*. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). New York: IEEE, 2015: 1–9.

[14] Chai J, Li A. Deep learning in natural language processing: a state-of-the-art survey[C]//2019 International Conference on Machine Learning and Cybernetics (ICMLC). New York: IEEE, 2019: 1–6.

[15] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[EB/OL]. (2023-08-02)[2023-10-01]. https://arxiv.org/abs/1706.03762v3.

[16] Gao W, Mahajan S P, Sulam J, *et al*. Deep learning in protein structural modeling and design[J]. *Patterns*, 2020, 1(9): 100142. DOI: 10.1016/j.patter.2020.100142.

[17] Fourcade A, Khonsari R H. Deep learning in medical image analysis: a third eye for doctors[J]. *J Stomatol Oral Maxillofac Surg*, 2019, 120(4): 279–288.

[18] Nabirotchkin S, Peluffo A E, Rinaudo P, *et al*. Next-generation drug repurposing using human genetics and network biology[J]. *Cur Opin Pharmacol*, 2020, 51: 78–92. DOI: 10.1016/j.coph.2019.12.004.

[19] MacEachern S J, Forkert N D. Machine learning for precision medicine[J]. *Genome*, 2021, 64(4): 416–425.

[20] Narayanan H, Dingfelder F, Butté A, *et al*. Machine learning for biologics: opportunities for protein engineering, developability, and formulation[J]. *Trends Pharmacol Sci*, 2021, 42(3): 151–165.

[21] Arnold F H. Design by directed evolution[J]. *Acc Chem Res*, 1998, 31(3): 125–131.

[22] Karplus M, Kuriyan J. Molecular dynamics and protein function[J]. *Proc Natl Acad Sci U S A*, 2005, 102(19): 6679–6685.

[23] Boyken S E, Chen Z, Groves B, *et al*. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity[J]. *Science*, 2016, 352(6286): 680–687.

[24] Romero P A, Arnold F H. Exploring protein fitness landscapes by directed evolution[J]. *Nat Rev Mol Cell Biol*, 2009, 10(12): 866–876.

[25] Bhardwaj G, Mulligan V K, Bahl C D, *et al*. Accurate de novo design of hyperstable constrained peptides[J]. *Nature*, 2016, 538(7625): 329–335.

[26] Pierce N A, Winfree E. Protein design is NP-hard[J]. *Protein Eng*, 2002, 15(10): 779–782.

[27] Siedhoff N E, Schwaneberg U, Davari M D. Machine learning assisted enzyme engineering[J]. *Method Enzymol*, 2020, 643: 281–315. DOI: 10.1016/bs.mie.2020.05.005.

[28] Mazurenko S, Prokop Z, Damborsky J. Machine learning in enzyme engineering[J]. *ACS Catal*, 2019, 10(2): 1210−1223.

[29] Diaz D J, Kulikova A V, Ellington A D, *et al*. Using machine learning to predict the effects and consequences of mutations in proteins[J]. *Curr Opin Struct Biol*, 2023, 78: 102518. DOI: 10.1016/j.sbi.2022.102518.

[30] Wittmann B J, Johnston K E, Wu Z, *et al*. Advances in machine learning for directed evolution[J]. *Curr Opin Struct Biol*, 2021, 69: 11−18. DOI: 10.1016/j.sbi.2021.01.008.

[31] Yang K K, Wu Z, Arnold F H. Machine-learning-guided directed evolution for protein engineering[J]. *Nat Methods*, 2019, 16(8): 687−694.

[32] Berman H M, Westbrook J, Feng Z, *et al*. The protein data bank[J]. *Nucleic Acids Res*, 2000, 28(1): 235−242.

[33] Apweiler R, Bairoch A, Wu C H, *et al*. UniProt: the universal protein knowledgebase[J]. *Nucleic Acids Res*, 2004, 32(Suppl 1): D115−D119.

[34] Notin P, Dias M, Frazer J, *et al*. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval[EB/OL]. (2023-08-02)[2023-10-01]. https://arxiv.org/abs/2205.13760.

[35] Cang Z, Wei G W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction[J]. *Int J Numer Method Biomed Eng*, 2018, 34(2): e2914. DOI: 10.1002/cnm.2914.

[36] Wang M, Cang Z, Wei G W. A topology-based network tree for the prediction of protein-protein binding affinity changes following mutation[J]. *Nat Mach Intell*, 2020, 2(2): 116−123.

[37] Schymkowitz J, Borg J, Stricher F, *et al*. The FoldX web server: an online force field[J]. *Nucleic Acids Res*, 2005, 33(Suppl 2): W382−W388.

[38] Leman J K, Weitzner B D, Lewis S M, *et al*. Macromolecular modeling and design in Rosetta: recent methods and frameworks[J]. *Nat Methods*, 2020, 17(7): 665−680.

[39] Qiu Y, Wei G W. Persistent spectral theory-guided protein engineering[J]. *Nat Comput Sci*, 2023, 3(2): 149−163.

[40] Alley E C, Khimulya G, Biswas S, *et al*. Unified rational protein engineering with sequence-based deep representation learning[J]. *Nat Methods*, 2019, 16(12): 1315−1322.

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

[42] Rives A, Meier J, Sercu T, *et al*. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. *Proc Natl Acad Sci U S A*, 2021, 118(15): e2016239118. DOI: 10.1073/pnas.2016239118.

[43] Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021, 596(7873): 583−539.

[44] Chazal F, Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists[J]. *Front Artif Intell*, 2021, 4: 667963. DOI: 10.3389/frai.2021.667963.

[45] Zomorodian A, Carlsson G. Computing persistent homology[J]. *Discrete Comput Geom*, 2005, 33(2): 249−274.

[46] Nguyen D D, Wei G W. DG-GL: differential geometry-based geometric learning of molecular datasets[J]. *Int J Numer Method Biomed Eng*, 2019, 35(3): e3179. DOI: 10.1002/cnm.3179.

[47] Wee J J, Xia K. Ollivier persistent Ricci curvature-based machine learning for the protein-ligand binding affinity prediction[J]. *J Comput Chem*, 2020, 41(21): 1924−1936.

[48] Nguyen D D, Wei G W. AGL-Score: algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening[J]. *J Chem Inf Model*, 2019, 59(7): 3291−3304.

[49] Ryczko K, Strubbe D A, Tamblyn I. Deep learning and density-functional theory[J]. *Phys Rev A*, 2019, 100(2): 022512. DOI: 10.1103/PhysRevA.100.022512.

[50] Butler K T, Davies D W, Cartwright H, *et al*. Machine learning for molecular and materials science[J]. *Nature*, 2018, 559(7715): 547−555.

[51] Chen J, Geng W, Wei G W. MLIMC: machine learning-based implicit-solvent Monte Carlo[J]. *Chin J Chem Phys*, 2021, 34(6): 683−694.

[52] Hsu C, Nisonoff H, Fannjiang C, *et al*. Learning protein fitness models from evolutionary and assay-labeled data[J]. *Nat Biotechnol*, 2022, 40(7): 1114−1122.

[53] Wittmann B J, Yue Y, Arnold F H. Informed training set design enables efficient machine learning-assisted directed protein evolution[J]. *Cell Syst*, 2021, 12(11): 1026−1045.

[54] Koehler Leman J, Szczerbiak P, Renfrew P D, *et al*. Sequence-structure-function relationships in the microbial protein universe[J]. *Nat Commun*, 2023, 14(1): 2351. DOI: 10.1038/s41467-023-37896-w.

[55] Kuhlman B, Bradley P. Advances in protein structure prediction and design[J]. *Nat Rev Mol Cell Biol*, 2019, 20(11): 681−697.

[56] Jeffery C J. Current successes and remaining challenges in protein function prediction[J]. *Front Bioinform*, 2023, 3: 1222182. DOI: 10.3389/fbinf.2023.1222182.

[57] Durham J, Zhang J, Humphreys I R, *et al*. Recent advances in predicting and modeling protein-protein interactions[J]. *Trends Biochem Sci*, 2023, 48(6): 527–538.

[58] Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021, 596(7873): 583–589.

[59] Baek M, DiMaio F, Anishchenko I, *et al*. Accurate prediction of protein structures and interactions using a three-track neural network[J]. *Science*, 2021, 373(6557): 871–876.

[60] 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.

[61] Gligorijević V, Renfrew P D, Kosciolek T, *et al*. Structure-based protein function prediction using graph convolutional networks[J]. *Nat Commun*, 2021, 12(1): 3168. DOI: 10.1038/s41467-021-23303-9.

[62] Bileschi M L, Belanger D, Bryant D H, *et al*. Using deep learning to annotate the protein universe[J]. *Nat Biotechnol*, 2022, 40(6): 932–937.

[63] Hakala K, Kaewphan S, Björne J, *et al*. Neural network and random forest models in protein function prediction[J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2020, 19(3): 1772–1781.

[64] Brandes N, Ofer D, Peleg Y, *et al*. ProteinBERT: a universal deep-learning model of protein sequence and function[J]. *Bioinformatics*, 2022, 38(8): 2102–2110.

[65] Blum M, Chang H Y, Chuguransky S, *et al*. The InterPro protein families and domains database: 20 years on[J]. *Nucleic Acids Res*, 2021, 49(D1): D344–D354.

[专家介绍] 虞文武：东南大学首席教授（二级）、博士生导师、数学学院院长。曾入选国家级高层次人才（2020）、国家级青年高层次人才 2 项（2014、2016），并于 2013 年获批国家优秀青年科学基金项目资助。作为重点研发计划项目的首席科学家，担任东南大学校学术委员会委员、江苏省网络群体智能重点实验室常务副主任、复杂工程系统测量与控制教育部重点实验室副主任、江苏国家应用数学（东南大学）中心常务副主任、网络通信与安全紫金山实验室数理基础研究中心课题负责人、华为-东南大学网络群体智能联合创新实验室主任。

主要研究方向涵盖系统科学与人工智能的交叉领域，包括分析、控制、优化、学习等。出版合编书 1 部、专著 2 部、教材 1 章节，发表 100 余篇 IEEE 汇刊文章。研究成果在 Google 和 SCI 上的引用超过 2 万次，SCI H 指数为 60，有 30 篇 ESI 高被引论文（学科前 1%）。

研究成果获得国家自然科学二等奖 1 项，江苏省科学技术奖 / 自然科学奖一等奖 2 项及国家一级学会科学技术奖一等奖 1 项等。担任 *IEEE Trans Circuits Syst II*，*IEEE Trans Ind Cyber-Phys Syst*，*IEEE Trans Ind Inform*，*IEEE Trans Syst Man*，*Cybern Syst*，《中国科学信息科学》《中国科学技术科学》《自动化学报》《系统科学与数学》《智能科学与技术》等杂志的编委。2014—2022 年连续 9 次入选科睿唯安（原汤森路透）全球高被引科学家（工程学）名单。

[专家介绍] 柴人杰：东南大学首席教授、二级教授，东南大学生命健康高等研究院执行院长、东南大学附属中大医院耳鼻喉科双聘教授；教育部长江学者特聘教授，国家重点研发计划首席科学家，青年千人，国家优青。现任中国生物物理学会听觉分会会长；中国生理学会干细胞生理专委会副主委、候任主委；中国细胞生物学学会发育生物学分会副会长；中国生物医学工程学会干细胞工程技术分会副会长；中国听基会基础研究专委会主委；ESCI 期刊 *Am J Stem Cells* 执行主编，*Neurosci Bull*，*Front Cell Dev Biol*，《实用医院临床杂志》副主编，《药学进展》编委等。

柴人杰教授长期致力于神经元和内耳毛细胞的再生和保护研究，信号通路及药物递送对内耳干细胞调控机制的研究。近 5 年以通信作者发表 SCI 论文 130 篇（总影响因子 1 246.38，平均影响因子 9.33，其中 IF > 30 分有 4 篇，IF 10~30 分有 37 篇），其中 4 篇论文被 Faculty1 000 列为推荐文章，1 篇论文入选 *Cell* 正刊年度最佳论文，13 篇论文入选了 ESI 高被引论文，2 篇论文入选热点论文，总他引 6 000 余次，h-index 48。先后获中国干细胞研究创新科学家奖，第三届全国创新争先奖，2022 年度中国科协十大强国青年科学家，教育部霍英东教育基金会高等院校青年科学奖一等奖，中国青年科技奖，教育部自然科学奖一等奖，中华医学科技奖二等奖，江苏省医学科技奖一等奖，上海市技术发明奖一等奖，湖北省科学技术奖二等奖，树兰医学青年奖，江苏省青年科技奖和江苏省青年十大科技之星等奖励。