



评述

深度学习在生物医学领域中的应用简介

王瑞松^{1*}, 王胜男², 石铁流^{2*}

1. 湖南文理学院医学院, 常德 415000;

2. 华东师范大学生命科学学院, 上海 200241

* 联系人, E-mail: ruisong.wang@huas.edu.cn; tieliushi@yahoo.com

收稿日期: 2024-05-09; 接受日期: 2024-07-01; 网络版发表日期: 2024-07-31

湖南省自然科学基金(批准号: 2023JJ50050)、湖南文理学院科学研究项目重点项目(批准号: 23ZZ02)资助

摘要 深度学习是人工智能领域中的一种新兴技术, 已在生物医学科学领域展现出巨大的潜力. 本文总结了深度学习的算法、原理和应用. 深度学习模型包含多个层次, 能够从数据中进行分层特征学习. 与传统的机器学习相比, 深度学习在自动特征提取、抗噪声能力、泛化能力和多功能性方面具有优势. 深度学习已成功应用于基因组学、转录组学、蛋白质组学、药物发现和疾病生物学等多个生物医学领域的问题中. 然而, 深度学习模型的可解释性还不足, 并在多模态数据集成方面面临挑战. 解决模型可解释性并将生物学知识集成的限制将加速深度学习技术转化为临床实践. 总的来说, 深度学习正在通过从大型生物医学数据中获取新的见解, 并改变生物医学研究和转化领域. 网络架构、可解释性和训练方法的持续进步将进一步释放深度学习推进人类健康发展的巨大潜力.

关键词 生物医学, 人工智能的应用, 深度学习的应用, 疾病的理解与分析

随着基因组学、医学成像、电子健康记录和移动健康等大规模高通量生物技术迅速发展导致了与人类健康和疾病相关的复杂、多维数据的爆炸式增长^[1]. 基因、转录组、蛋白质、代谢物、细胞、组织、患者和人群的大规模数据集, 为通过数据驱动的转化生物医学提供了前所未有的机会^[2]. 然而, 大规模生物医学数据的规模、噪声、异质性、不完整性和复杂性对传统计算方法构成了重大挑战^[3]. 因此, 我们需要新的计算方法以解决这些问题, 以充分发掘这些大规模生物医学数据的潜力. 近年来, 人工智能、机器学习和深度学习等先进技术在处理大规模数据方面表现出强大的能力, 为生物医学研究带来新的解决方案. 人工智能是广义上指使机器具备智能的技术, 涵盖但不限于机器

学习. 机器学习作为其重要分支, 通过相关算法, 使计算机系统的性能随着经验自我优化. 深度学习则是机器学习的分支, 它主要使用人工神经网络(artificial neural network, ANN)作为学习模型. 这三者构成包含关系, 非并列或对立(图1).

2006年起, 深度学习快速发展, 在多个领域展示了其卓越性能^[4]. 深度学习模型由输入层、输出层及多个隐藏层组成, 这些隐藏层是分层架构的非线性信息处理单元, 用于特征学习和模式分类^[5]. 卷积神经网络(convolutional neural networks, CNNs)和循环神经网络(recurrent neural networks, RNNs)是两种广泛应用的深度学习模型. CNNs主要处理网格拓扑结构的数据, 如图像和视频; RNNs则主要用于处理序列数据, 如文本

引用格式: 王瑞松, 王胜男, 石铁流. 深度学习在生物医学领域中的应用简介. 中国科学: 生命科学, 2025, 55: 1268–1287

Wang R, Wang S, Shi T. An introduction to the application of deep learning in the biomedical field (in Chinese). Sci Sin Vitae, 2025, 55: 1268–1287, doi: 10.1360/SSV-2023-0297

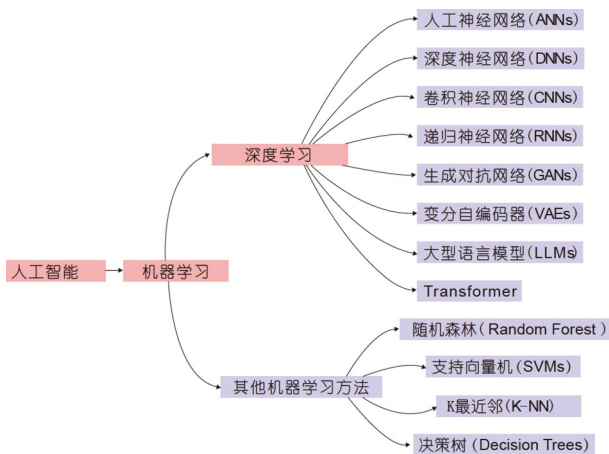


图 1 人工智能、机器学习和深度学习三者间的关系

Figure 1 The interrelationship between artificial intelligence (AI), machine learning (ML), and deep learning (DL)

和语音。因此，CNNs与RNNs并非相互对立，而是针对不同数据类型的深度学习模型。

深度学习技术在自然语言处理(natural language processing, NLP)领域取得了显著成就。NLP作为人工智能的一个重要分支，旨在让计算机理解、生成及处理人类语言。传统的NLP方法依赖于规则和统计模型；相比之下，深度学习通过基于RNN的语言模型和基于Transformer的预训练模型(如双向编码器表示，即BERT(bidirectional encoder representations from transformers))，在众多NLP任务上展示了前所未有的性能。因此，深度学习技术为NLP注入了新活力。深度学习与NLP之间存在相辅相成的关系。

深度学习技术，特别是CNNs^[6-8]和RNNs^[7,9]，已广泛应用于生物学数据分析^[10]。在基因组学研究中，CNNs与RNNs已利用基因序列的分类和功能预测，助力科学家探索新的基因与疾病关联。此外，这些技术已应用于蛋白质分类^[11]、预测^[12-14]和相互作用分析^[10,15-17]，从而揭示了新药靶点和疾病机制^[18-20]。深度学习技术不仅应用于基因组学和蛋白质组学，还应用于其他生物学数据分析领域。例如，结合NLP和深度学习技术，可以从非结构化的文本数据(如病例报告、医学文献和社交媒体数据)提取价值的信息，以更全面理解患者临床记录与医生诊断，进而优化疾病的诊断和治疗方案。代谢组学利用深度学习算法，从大量的代谢物谱数据中识别与疾病风险相关的生物标志物，助力疾病的早期诊断和个性化治疗。深度学习几乎

可以解决生物学领域中的诸多问题。然而，深度学习的成功仍然依赖于特定领域的知识和数据的质量。应用深度学习于生物学领域时，必须考虑数据的特点、任务要求和模型的可解释性等因素，以保证模型的有效性和可靠性。

总体而言，人工智能、机器学习以及深度学习等先进技术，已经为生物学数据分析引入了革命性变革。利用这些技术，我们能够深入理解及有效利用大规模生物学数据的潜力，进而为转化生物学研究和医疗服务带来更多机会和改进方案。

1 简介

深度学习前沿的人工智能技术，已在从庞大复杂数据集中提取有意义的模式和关系上取得领先^[5]。深度学习涉及具有多个层的人工神经网络，这些层级结构使能够直接从原始数据中学习层次特征表示^[21]。传统的机器学习通常依赖大量人力进行特征提取(例如编码或手动筛选)以训练模型^[8]。相比之下，深度学习通过其连续的神经网络层自动进行特征提取^[22]，每一层都将输入转换为更高级、更抽象地表示，从而能够建模极其复杂的函数^[23]。训练完成后，深度神经网络(deep neural networks, DNNs)能通过层级中的非线性变换快速处理新输入进行预测，其灵活结构适用于图像、序列、图形、和文本等多种生物学应用的定制^[24]。在大多数情况下，深度学习减少了手动特征工程的需求，使模型能从原始数据中直接学习复杂表示。

1.1 深度学习在生物学领域的应用现状与未来发展重点

深度学习的高级功能适用于应对大规模生物学数据分析挑战。深度学习广泛应用于医学图像疾病诊断^[5,25,26]、药物发现^[5,7]、基因组目标识别^[5,27,28]、患者风险预测^[29-31]以及医院运营^[32]等领域。例如，Gulshan等人^[25]显示，深度学习能够通过视网膜基底照片检测糖尿病视网膜病变，准确率与专家眼科医生相当。DeepChem通过深度学习预测小分子的性质，从而加速药物开发^[33]。DeepSEA利用卷积神经网络预测人类基因组中非编码变异的效应^[27]。Rajkomar等人^[29]开发了一种基于电子健康记录的深度学习模型，用于预测患者的多项医疗事件。

随着分子、细胞、组织、人群及时间尺度上生物医学数据集规模和复杂性的日益增大, 开发更强大、更通用的深度学习技术迫切必要。生物医学深度学习发展的关键优先事项涉及多模态数据整合^[2,34]、模型可解释性^[34,35]及对分布变化的鲁棒性^[3]。整合不同类型的数据可以更全面地理解生物医学现象, 而可解释的模型可以阐明预测背后的机制。此外, 提高对分布变化的鲁棒性能够增强模型在实际应用的泛化能力和临床影响。

1.2 深度学习相较于传统方法的优势

相较于传统方法, 深度学习在特征学习、鲁棒性、泛化能力及通用性方面具有显著优势。深度学习能够自动提取数据的特征, 无需人工设计, 从而大幅降低了特征工程的复杂性和对专业知识的依赖^[5]。深度学习模型在面对输入数据的噪声、缺失或者其他形式的不完整性时, 展现出卓越的鲁棒性, 能够自动识别并忽略次要变化, 增强对异常数据和攻击的抵抗力。这一点由Szegedy等人^[36]的研究充分证实。在大规模和复杂的数据集上, 深度学习模型展现出卓越的泛化能力, 能够准确预测训练数据未覆盖的新情形^[37]。Zhang等人^[38]在光学设计中使用模拟监督的深度学习神经网络进行空间变化的反卷积, 实现了在传统显微镜的景深的10倍以上的改进, 并且展现了良好的泛化能力。此外, 深度学习模型显示出强大的通用性, 能够应用于图像识别、语音识别、自然语言处理等多种任务。其泛化能力尤其在处理来自不同设备和采集协议的数据时显著, 展现了在多样化应用场景中的优异性能^[39]。特别是在处理不同设备和采集协议产生的数据时的泛化性能, 这体现了深度学习的通用性, 即它能够在各种不同的应用场景和数据类型中展现出良好的性能。Bashyam等人^[26]研究表明, 即使在处理来自不同设备和采集协议的医学图像数据时, 这些模型的预测性能仍然保持稳定, 展示了其在各种环境下的强大适应能力。

1.3 深度学习的主要模型

深度学习已经成为当前的人工智能研究领域一种主流的技术手段, 尤其在图像识别、语音识别、自然语言处理等领域取得显著成就。深度学习核心涉及多种神经网络模型, 如DNNs、CNNs、RNNs和生成对

抗网络(generative adversarial networks, GANs)、变分自编码器(variational autoencoders, VAEs)、大型语言模型(large language models, LLMs)和Transformer等。图2A这些主要深度学习模型的相互关系。

DNNs是一种多层的神经网络, 设计灵感源于人脑的神经网络结构(图2B)。DNNs的工作原理基于多个非线性变换层, 这些层逐层对输入数据进行抽象和表示学习, 每层接收前一层的输出, 并通过激活函数实现非线性处理。随着网络深度的增加, DNNs能够捕捉到数据的高级抽象特征。DNNs的主要特征包括多层结构、前馈传播、使用反向传播算法进行训练, 以及能够处理高度非线性的复杂模式。在生物医学领域, DNNs广泛应用于疾病诊断和药物发现等任务, 展示了卓越性能。

CNNs是专为处理具有网格结构的数据(如图像)设计的深度学习模型(图2C)。通过卷积层、池化层和全连接层的组合, CNNs有效地提取图像的局部特征和全局特征^[37]。卷积层主要通过一组可学习的卷积核, 对输入图像执行滑动窗口操作, 从而提取各种局部特征。池化层通过降低特征维度来增强模型的鲁棒性和计算效率。全连接层将提取的特征映射至输出空间, 用于执行图像分类、目标检测等任务。CNNs的核心特征包括局部连接、权重共享、池化和多层次特征提取等。在医学图像分析中, CNNs取得显著成功, 特别是在辅助放射科医生进行肿瘤诊断等应用中。

RNNs是专为处理序列数据设计的神经网络, 设计灵感来源于人脑处理时间序列信息方式(图2D)。RNNs的关键特性在于其循环结构, 这使得网络可以储存并利用历史信息, 以更有效地理解和预测序列数据。具体来说, RNNs通过结合当前输入与前一时刻的隐藏状态, 能够捕捉序列中的长期依赖性。RNNs的主要特征包括循环结构、参数共享、处理变长序列的能力, 以及梯度消失或爆炸问题。因此, RNNs尤其适用于语音识别和文本生成等任务。在生物医学领域, RNNs广泛应用于分析基因表达数据和电子健康记录等时序数据, 用以识别与疾病相关的模式和规律。

GANs由生成器和判别器两个神经网络组成, 通过对抗学习, 生成器能产生接近真实数据分布的样本(图2E)。GANs的核心原理是将生成器和判别器置于一个双玩家最小化-最大化博弈中, 生成器努力生成逼真样本以欺骗判别器, 而判别器力图区分真实与生成样本。

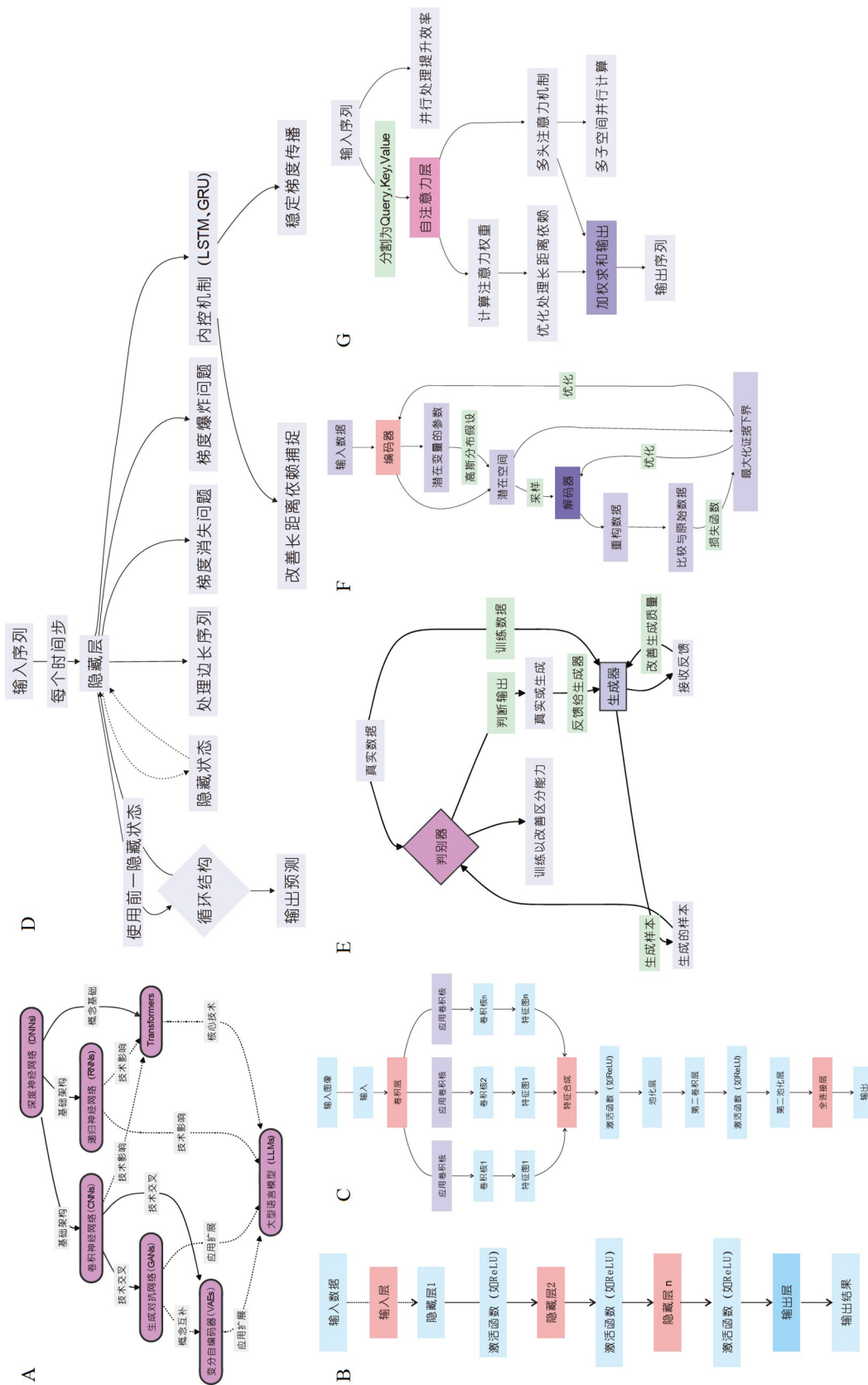


图 2 几种深度学习的主要模型之间的关系以及示意图 A: 主要模型间的相互关系; B: DNNs, 线性整流函数(rectified linear unit, ReLU); C: CNNs; D: RNNs, GRU, 长短期记忆网络(long short-term memory, LSTM); E: 门控循环单元(gate recurrent unit); F: Generative Adversarial Networks; G: Variational Autoencoders; H: Transformers

通过持续的对抗学习, 生成器学习真实数据的分布. GANs能生成高质量的图像和音频等内容^[40]. GANs的主要特征包括无监督学习、生成器与判别器的对抗学习, 以及训练和评估的复杂性. 在生物医学领域, GANs被用于生成合成医学图像以缓解标注数据不足问题, 并应用于药物分子设计, 以加速新药的开发.

变分自编码器(VAEs)是一种基于概率图的生成模型, 包含编码器和解码器两个部分(图2F). VAEs通过最大化变分下界(variation lower bound, ELBO)来近似数据的边缘对数似然, 目的是学习数据的压缩表示形式. 编码器将输入数据映射到一个低维潜在空间, 假定潜在变量遵循特定的先验分布(例如高斯分布); 解码器从这一潜在空间采样, 重构原始数据. VAEs的主要特征包括无监督学习、概率建模、学习数据的低维表示以及生成新样本的能力. 在生物医学领域, VAEs被用于分析基因表达数据和发现疾病相关的生物标志物, 同时也用于生成合成医学图像以辅助研究.

Transformer是一种基于自注意力机制的神经网络架构, 已成为自然语言处理领域的主流模型(图2G). Transformer的核心原理是利用自注意力机制来计算序列元素间的依赖关系, 有效捕捉长距离依赖性. 具体而言, Transformer通过查询(query)、键(key)和值(value)三个向量来计算注意力权重, 并通过加权求和来生成输出. 多头注意力机制允许模型在多个子空间内同时计算注意力, 从而增强了模型的表达力. Transformer的主要特征包括完全依赖于注意力机制、并行计算能力、处理长序列的能力, 以及适应迁移学习的优势. Transformer在机器翻译和文本摘要等任务中取得显著成就, 并已扩展应用至计算机视觉和语音识别等领域. 在生物医学领域, Transformer被用于生成医患对话和自动编写临床报告等任务, 从而提高了医疗服务的效率和质量. 自2020年以来, Transformer架构已成为自然语言处理的核心, 同时扩展至计算机视觉和音频处理等领域, 催生了多种创新模型如iDNA-ABF^[41]、GPT(generative pre-trained transformer)^[42]系列等.

大语言模型(LLM)是一种基于深度学习技术的人工智能模型, 专门在大规模文本数据上进行训练. 大语言模型通常基于Transformer架构, 能够从海量数据中自学语言规律和知识, 展现出卓越的语言理解和生成能力. 大语言模型适用于多种自然语言处理任务, 如问答、翻译、摘要和文本生成等, 不仅代表了人工智

能领域的重大突破, 也为机器理解和处理语言的方式提供了新的视角.

总体而言, 这些模型在结构设计和应用场景上各具特色, 均通过数据的多层次非线性变换, 实现了复杂模式的自动学习和表示. 随着深度学习技术的进一步发展, 这些模型在生物医学领域的应用正变得日益广泛, 预计将在疾病诊断、药物开发和个性化医疗等领域带来革命性变革.

2 深度学习在多种组学数据分析中的应用

在大数据时代, 将生物医学大数据转化为有价值的知识构成了生物信息学领域的主要挑战之一. 为了从生物信息学的大数据中提取知识, 机器学习已成为从生物信息学大数据中提取知识的一种广泛使用并成功的方法. 这些方法中, 支持向量机(support vector machine, SVM)、随机森林(random forest, RF)、隐马尔可夫模型(hidden Markov model, HMM)、贝叶斯网络等已广泛应用于基因组学、蛋白质组学、系统生物学等领域. 然而, 传统机器学习算法的性能高度依赖于所谓的特征表示. 近年来, 深度学习的兴起克服了传统方法的许多局限性, 并在生物信息学领域取得显著进展^[22]. 下文将从多个角度介绍深度学习在生物医学领域的应用(具体研究内容见表1), 并分析及展望其未来的应用前景(图3).

2.1 基因调控网络分析中的应用

基因是生物体的遗传物质, 通过其相互作用及表达调控来控制生物体的性状和功能. 深度学习在基因组学数据挖掘中的关键应用之一包括预测基因与调控因子之间的相互作用. 基因表达受转录因子等调控因子的精确调控, 深度学习则提供了预测这些调控关系的有效工具. 例如, DeepSEA^[27]和DeepBind^[43]利用深度学习从DNA序列和基因组数据中预测转录因子与DNA的结合位点, 从而揭示基因表达调控的关键线索. 同样, Keilwagen等人^[44]开发的深度学习方法能准确预测不同细胞类型中转录因子的结合位点, 助于理解细胞特异性的基因表达调控机制. Avsec等人^[45]开发了BPNet模型, 利用卷积神经网络从DNA序列直接预测多个转录因子结合位点, 实现了高分辨率预测. BPNet通过多任务学习等策略优于传统方法, 并可解释模型

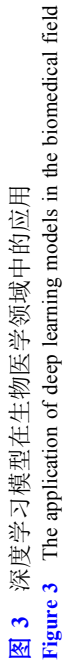


表 1 深度学习模型在生物医学领域中的应用

Table 1 Applications of deep learning models in the biomedical field

任务	模型架构	期刊	发表时间	参考文献
从蛋白质序列预测蛋白质功能	CNN	<i>Bioinformatics</i>	2017年	[11]
预测细胞类型特异性转录因子结合	深度学习	<i>Genome Biology</i>	2019年	[44]
从基因表达数据中鲁棒地预测表型的 基于生物网络的正则化人工神经网络 模型	人工神经网络, 正则化	<i>BMC Bioinformatics</i>	2017年	[47]
基因网络嵌入的深度学习框架	GNE	<i>BMC Systems Biology</i>	2019年	[46]
DNA甲基化预测	BERT	<i>Genome Biology</i>	2022年	[41]
推断细胞间通讯	基于图神经网络(GNN)的节点中心的表达 建模	<i>Nature Biotechnology</i>	2023年	[50]
蛋白质结构预测	深度学习, 注意力机制, AlphaFold	<i>Nature</i>	2021年	[54]
蛋白质序列和功能预测	基于传统的Transformer/BERT的ProteinBERT	<i>Bioinformatics</i>	2022年	[56]
蛋白质结构预测	语言模型, Transformer	<i>Science</i>	2023年	[57]
蛋白质结构预测	RGN	<i>Nature Biotechnology</i>	2022年	[55]
RBP结合位点预测	多任务基于的DNN 模型Deep RNA-Binding Protein binding preference (DeepRiPe)	<i>Genome Research</i>	2020年	[49]
lncRNA-疾病关联预测	基于拓扑特征提取与Transformer编码器的 LDAformer	<i>Briefings in Bioinformatics</i>	2022年	[90]
利用质谱数据进行精准医学	卷积神经网络(VGG16)	<i>Briefings in Bioinformatics</i>	2022年	[92]
糖尿病视网膜病变新生血管渗漏检测	集成三个CNN的深度学习算法	<i>Scientific Reports</i>	2023年	[94]
非功能性胰腺神经内分泌肿瘤淋巴结 转移预测	基于影像组学和深度学习的预测模型	<i>eClinicalMedicine</i>	2023年	[97]
单细胞水平癌症药物反应预测	基于深度迁移学习框架的scDEAL	<i>Nature Communications</i>	2022年	[80]
阿尔茨海默病结构性MRI异常检测和 病理进展模式分析	3D卷积神经网络 (Ensemble 3DCNN)	<i>Advanced Science</i>	2023年	[95]
甲状腺上皮肿瘤风险分类预测	利用Resnet (CNN)提取影像组学特征	<i>European Journal of Radiology</i>	2023年	[96]
癌症预后预测	多模态深度学习	<i>Cancer Cell</i>	2022年	[102]
分子指纹预测	与深度神经网络(DNN)的结合	<i>Bioinformatics</i>	2022年	[105]
人类生命轨迹追踪	迁移学习的microDELTA	<i>Briefings in Bioinformatics</i>	2023年	[101]
肠道菌群健康预测	利用对中机器学习分类器进行PCA统计建模	<i>Microbiome</i>	2023年	[99]
肽段识别	基于深度学习模型的MSBooster与Percolator 以及MSFragger结合	<i>Nature Communications</i>	2022年	[106]
空间基因组学	CNN和RNN	<i>Nature Biotechnology</i>	2023年	[98]
基因表达预测	CNN和RNN	<i>Briefings in Bioinformatics</i>	2023年	[113]
电子病历癌症队列构建	半监督的机器学习算法结合自然语言处理	<i>JAMA Network Open</i>	2021年	[103]
微卫星不稳定性检测	图注意力网络(GAT) 和图卷积网络(GCN)	<i>Briefings in Bioinformatics</i>	2023年	[111]
癌症转移预测	卷积变分自编码器(CVAE)和深度神经网络 (DNN)	<i>Computational and Structural Biotechnology Journal</i>	2021年	[109]
表观基因组、染色质组织和转录组 预测	EPCOT(基于预训练和微调的深度学习框架)	<i>Nucleic Acids Research</i>	2023年	[112]
GC-MS数据解析	U-Net、k-CNN、U-Net for elution Region detection、Deep learning-assisted Multivariate Curve Resolution	<i>Talanta</i>	2022年	[107]
药物相似度整合	基于注意力机制的多视图自编码器(multi-view graph autoencoders , GAE)的GraphCNN	<i>IJCAI</i>	2018年	[78]

(表1续1)

任务	模型架构	期刊	发表时间	参考文献
全基因组反卷积揭示阿尔茨海默氏症弹性表观遗传特征	Cellformer, 基于深度学习的将批量染色质科技型测序(ATAC-seq)解卷积为整个基因组上的特定细胞类型表达的算法	<i>Nature Communications</i>	2023年	[110]
预测药物-靶标结合亲和力	基于图神经网络的模型GraphDTA	<i>Bioinformatics</i>	2020年	[79]
深度挖掘异构网络以预测新的药物-靶标关联	基于Linked Tripartite Network (LTN)架构的DeepWalk算法	<i>Bioinformatics</i>	2017年	[16]
预测DNA和RNA结合蛋白的序列特异性	深度卷积神经网络DeepBind	<i>Nature Biotechnology</i>	2015年	[43]
预测非编码变异的功能效应	深度卷积神经网络DeepSEA	<i>Nature Methods</i>	2015年	[27]
预测胶质瘤的IDH突变状态	深度学习和影像学混合模型, 包含两个CNN	<i>Neuro-Oncology</i>	2021年	[30]
预测乳腺癌的空间基因表达图谱	基于CNN的ST-Net	<i>Nature Biomedical Engineering</i>	2020年	[51]
转录组可变剪接分析	DNN	<i>Nature Methods</i>	2019年	[48]
预测氨基酸变异的有害性	随机森林机器学习	<i>Nucleic Acids Research</i>	2017年	[64]
预测人类基因组中所有可能的错义突变的致病性	基于AlphaFold蛋白质结构预测模型, 在人类和灵长类错义突变频率数据上微调得到的变体致病性预测模型AlphaMissense	<i>Science</i>	2023年	[4]
预测人类基因组中所有可能的编码变异的影响	基于ESM1b这个包含6.5亿参数、在2.5亿条蛋白质序列上预训练的蛋白质语言模型	<i>Nature Genetics</i>	2023年	[68]
预测lncRNA-疾病关联	DeepWalk和LTN	<i>Mathematical Biosciences</i>	2019年	[91]
泛癌分析APOBEC诱变特征及其在膀胱癌预后和免疫治疗中的预测价值	多组学分析, 机器学习(随机森林等)	<i>Theranostics</i>	2022年	[87]
使用不完整的多组学数据进行疾病预测和生物标志物识别	不完整多组学变分神经网络	<i>Briefings in Bioinformatics</i>	2023年	[100]
使用深度学习预测蛋白质点突变引起的稳定性变化	共享残基对神经网络	<i>Journal of Chemical Information and Modeling</i>	2019年	[66]
基于基因共表达网络和深度学习的肺癌风险分层模型	CNN	<i>BioMed Research International</i>	2018年	[93]
激酶抑制剂活性预测	多尺度感知Transformer	<i>Briefings in Bioinformatics</i>	2023年	[82]
预测氨基酸突变对蛋白质-蛋白质结合亲和力的影响	几何编码器(graph attention network)+梯度提升树(GBT)	<i>PLOS Computational Biology</i>	2021年	[67]
蛋白质-蛋白质相互作用预测	双视角分层图学习模型, 包含蛋白图卷积网络(BGNN)和PPI网络图同构网络(TGNN)	<i>Nature Communications</i>	2023年	[62]
预测突变对蛋白质稳定性和结合亲和力的影响	集成机器学习方法, 利用同源建模和结构特征	<i>Bioinformatics</i>	2016年	[65]
结合体内交联质谱信息和深度学习预测蛋白质结构	修改后的AlphaFold2算法AlphaLink, 将实验得到的距离约束信息整合进网络架构中, 在蛋白质结构预测中引入实验数据	<i>Nature Biotechnology</i>	2023年	[53]
从头蛋白质序列和结构设计	利用trRosetta结构预测网络, 通过蒙特卡罗序列采样优化序列与背景分布的KL散度, 生成新的折叠蛋白序列	<i>Nature</i>	2021年	[59]
从头设计具有特定功能的蛋白质序列	称为ProteinGAN的专门化GAN变体, 通过随机噪声向量生成多样化的蛋白质序列, 并使用判别器评估优化	<i>Nature Machine Intelligence</i>	2021年	[60]
给定骨架结构设计氨基酸序列	基于图神经网络的ProteinMPNN, 灵活的解码顺序, 支持多链/对称性设计, 通过实验广泛验证	<i>Science</i>	2022年	[58]
转录组可变剪接分析	深度神经网络, 贝叶斯假设检验	<i>Nature Methods</i>	2019年	[48]
表观基因组学个性化预测	基于注意力机制的深度学习模型, 迁移学习	<i>Nature Communications</i>	2023年	[108]

(表1续2)				
任务	模型架构	期刊	发表时间	参考文献
基于转录组数据预测药物疗效	使用SMILES编码和神经网络预测化合物引起的转录组变化(CTPs), 然后用基因集富集分析(GSEA)评估对疾病的疗效	<i>Nature Biotechnology</i>	2021年	[81]
预测蛋白质复合物的结构	直接用AlphaFold2对蛋白质异源二聚体进行建模,通过优化多序列比对提高准确性	<i>Nature Communications</i>	2022年	[61]
预测人类错义变异的致病性	PrimateAI-3D, 一个在蛋白质3D结构上运行的半监督3D卷积神经网络	<i>Science</i>	2023年	[69]
从DNA序列预测多个转录因子(Oct4、Sox2、Nanog和Klf4)在基因组每个碱基位置的结合概率	多任务卷积神经网络	<i>Nature Genetics</i>	2021年	[45]
单细胞蛋白质组数据分析, 包括细胞聚类、批次校正、细胞类型注释、临床分析和空间蛋白质组数据探索	基于深度图对比学习的通用框架, 包含多任务异方差回归和图对比学习两个阶段	<i>Nature Methods</i>	2024年	[76]
使用生成式AI构建单细胞多组学的基础模型	基于生成式对抗网络(GAN)的scGPT	<i>Nature Methods</i>	2024年	[70]
区分目标单细胞RNA-seq数据集中与背景数据集相比富集的潜在结构和变异	对比变分推断(Contrastive Variational Inference)	<i>Nature Methods</i>	2023年	[73]
从单细胞和空间转录组数据中量化神经元激活	深度学习模型	<i>Nature Communications</i>	2023年	[74]
预测单细胞DNA甲基化状态	深度神经网络	<i>Genome Biology</i>	2017年	[75]
单细胞RNA测序数据的细胞类型注释	基于大规模预训练语言模型scBERT	<i>Nature Machine Intelligence</i>	2022年	[71]
单细胞RNA-seq分析中的细胞类型注释	使用GPT-4进行评估	<i>Nature Methods</i>	2024年	[72]

学到的特征, 揭示motif间相互作用和新的生物学见解. 这彰显了深度学习从海量数据中学习复杂调控规律的巨大潜力.

另一主要应用方向是使用深度学习构建基因相互作用网络. 基因在复杂的生物学网络中不是孤立存在的, 它们共同作用影响表型^[30]. Kc等人^[46]提出的基因网络嵌入(gene network embedding, GNE)深度学习框架展现了优良性能, 能学习嵌入式表示并在缺少基因网络结构信息的情况下整合已知的基因相互作用网络和表达数据, 用于预测基因相互作用. Kang等人^[47]开发的正则化的人工神经网络能够将基因及其调控因子的相互依赖性编码进分类器, 专用于利用基因表达数据预测表型, 如对治疗的反应. 将生物网络中的节点转化为神经网络节点的方法为分析生物网络中的调控关系提供了新的研究方向.

2.2 转录组学数据分析中的应用

转录组学是一门研究生物体在特定条件下所有转

录本的总和的学科. 转录本是从基因组DNA被转录成RNA的产物, 这些RNA中的一部分可以转译成蛋白质, 而其他的则在细胞中执行各种功能. 深度学习算法在转录组学数据分析中被广泛应用.

首先, 深度学习算法可以分析特定条件下基因的表达量变化, 并转录过程中的调控机制和调控因子. 例如, Keilwagen等人^[44]开发的一个深度学习方法可以准确地预测不同细胞类型中转录因子的结合位点, 进而揭示细胞特异性的基因调控机制. 多尺度深度学习框架iDNA-ABF能从序列数据中预测和解释DNA甲基化^[41].

其次, 深度学习可以研究基因在转录过程中的剪接方式和剪接过程. 深度学习开算法, 如DARTS^[48]和Ghanbari及Ohler的预测模型^[49], 即便在测序深度较低的样本中也显著提升了RNA测序(RNA-seq)定量差异剪接的准确度. 这些技术使研究人员能预测多个可能的转录本和编码的蛋白质, 加速对RNA结合蛋白(RNA-binding proteins, RBP)研究, 并揭示它们在相关

疾病中的作用。

最后, 深度学习模型在基因表达谱分析中已被广泛应用。全基因组表达谱分析用于描述细胞在各种条件下的活动, 如癌细胞对不同治疗条件的反应。组织RNA表达谱的差异可以反映疾病的异质性, 分析不同区域的RNA表达量有助于识别细胞类型和调控网络的差异, 为精准治疗提供依据。使用创新的图神经网络方法NCEM(node-centric expression modeling)^[50], 增强了分析的精度。然而, 并非所有的研究组都有足够的资源进行大规模空间转录组测序。人体全基因组含有约60000个基因, 其中约20000个基因编码蛋白质, 绝大多数基因的表达之间高度相关。He等人^[51]发现, 深度学习方法可以在未进行空间转录组测序的情况下, 识别表达异质性和基因表达差异。通过基于基因表达谱的分子标记, 研究人员不仅可以区分细胞的表型, 还可识别各种疾病的生物标志物, 从而有助于疾病的诊断和治疗。

2.3 蛋白质组学数据中的应用

深度学习模型在蛋白质组学数据的应用展现了巨大潜力, 并已在蛋白质研究和生物医学领域带来了重要的进展。蛋白质组学研究生物体内蛋白质组成、结构、功能以及相互作用, 对于揭示生物学过程、疾病机制以及药物研发至关重要。以下是深度学习在蛋白质组学数据中的主要应用领域:

深度学习模型已经被应用于预测蛋白质的三维结构, 即其折叠形态。预测蛋白质的三维结构在生物医学领域极为重要, 它有助于解析蛋白质功能、相互作用及疾病机制, 为药物设计和治疗研究提供关键依据。将深度学习模型应用于蛋白质序列和结构数据分析显著提高了蛋白质结构预测的准确性。AlphaFold2利用深度学习技术, 在蛋白质结构预测方面实现了显著的性能提升^[52]。AlphaLink^[53]整合交联质谱数据, 进一步增强了AlphaFold在预测低同源序列蛋白及其实验条件下构象的表现^[15,54]。正如AlphaFold2可以经过数据整合成为更好性能的AlphaLink一样, 近期在蛋白质语言模型和蛋白质设计领域的进展为利用成熟算法进行蛋白质组学研究提供了新的思路。Chowdhury等人^[55]和Brandes等人^[56]分别开发了基于BERT的蛋白质语言模型AminoBERT和ProteinBERT, 它们通过对蛋白质序列进行预训练, 为蛋白质结构预测和功能注释等下

游任务提供了关键的上下文信息。这启示我们可以将GPT-4等先进的语言模型引入到AlphaFold等蛋白质结构预测算法中, 以提高预测的准确性和可解释性^[57]。

最近的研究利用AlphaFold2这类成熟的蛋白质结构预测模型, 配合其他深度学习技术, 设计并优化具有特定功能的新蛋白质。例如, Dauparas等人^[58]利用图神经网络ProteinMPNN优化由AlphaFold2产生的不稳定设计, 成功制造了能正确折叠且含有大量突变的蛋白质; Anishchenko等人^[59]通过Monte Carlo序列优化技术, 反向工程AlphaFold, 创造了实验验证的新蛋白质设计。

此外, 深度学习在蛋白质设计领域已取得显著进展。例如, ProteinGAN^[60]和基于AlphaFold2的技术^[54,61]展示了利用生成模型和深度学习设计新型稳定蛋白质的潜力, 开拓了治疗学和生物材料学等领域的新途径。

深度学习方法被广泛应用于预测蛋白质之间的相互作用。蛋白质之间的相互作用在生物学中至关重要, 因为它们涉及许多生物过程和细胞功能。预测蛋白质间的相互作用有助于理解它们的功能、信号传导路径和与疾病相关的机制。方法如DeepGO^[11]和HIGH-PPI^[62]利用深度学习整合蛋白质交互作用(protein-protein interaction, PPI)和Gene Ontology(GO)数据, 创建蛋白质的嵌入表示以预测蛋白质间相互作用, 为生物学研究开辟了新途径。

遗传变异已被证明与多种人体疾病密切相关, 但目前尚无突变致病性预测的金标准方法。蛋白质是代谢、免疫反应、信号传递等生命活动的直接执行者, 蛋白质序列突变的致病性预测因此可以直接应用于复杂疾病(如癌症)的基因筛查和诊断。深度学习技术已被应用于预测蛋白质突变对功能和稳定性的影响, 这对理解疾病机制和药物设计至关重要。该领域内多种算法不断涌现, 如ENCoM^[63]、DEOGEN2^[64]和ELASPIC^[65]这些都是基于氨基酸突变数据分析其对蛋白质稳定性影响以预测突变效应的工具。例如, DeepDDG^[66]、GeoPPI^[67]、AlphaMissense^[4]、ESM1b^[68]和PrimateAI-3D^[69]等利用深度学习整合序列和结构信息, 显著提高了突变效应预测的准确性。Cheng等人^[4]对不同方法的评估显示, AlphaMissense在区分致病和良性变异方面表现最佳, 可用于解释病人中的不明确遗传变异并发现候选疾病基因。

总的来说, 这些研究表明深度学习极大地扩展了

在蛋白质组学的研究范围和可扩展性, 并相较于传统方法更有效地揭示了序列、结构与功能之间的关系. 主要挑战在于提升模型的准确性和处理序列多样性, 以及更灵活地定义结合相互作用之外的其他属性. 持续的技术进步预计将在治疗学、生物材料学和合成生物学等领域的带来革命性变革.

2.4 单细胞测序技术中的应用

深度学习与单细胞测序技术的结合, 尤其是单细胞测序转录组学, 已成为近年来生命科学领域的研究热点. 在这一背景下, 众多创新性的深度学习模型相继涌现, 为单细胞数据分析提供了新的视角和有力工具. 例如, scGPT通过大规模预训练学习基因和细胞的生物学信息, 从而捕捉基因表达模式、细胞类型特异性等关键特征, 并应用于多种下游任务, 如细胞聚类、细胞类型注释和基因表达预测等^[70]. scBERT模型则将Transformer架构引入单细胞转录测序(single-cell RNA sequencing, scRNA-seq)分析, 能够有效地学习基因之间的复杂关系, 并在跨数据集泛化方面展现出优势^[71]. CelltypeGPT 软件包(R语言) 利用GPT-4(generative pre-trained transformer 4)的强大能力, 能够根据scRNA-seq数据自动识别和注释细胞类型, 展现出在scRNA-seq分析中细胞类型注释方面巨大的应用潜力^[72].

除了分析单细胞转录组数据, 深度学习模型也被应用于其他类型的单细胞数据分析. contrastiveVI模型旨在区分目标scRNA-seq数据集中与背景数据集相比富集的变异, 可以用于识别特定细胞状态或亚群相关的基因^[73]. NEUROeSTIMator模型利用深度学习技术从转录组数据中量化神经元激活, 为研究神经元功能和疾病提供了新的工具^[74]. DeepCpG模型可以预测单细胞DNA甲基化状态, 有助于研究表观遗传调控机制^[75]. scPROTEIN模型通过深度图对比学习解决了单细胞蛋白质组数据的肽段定量不确定性、数据缺失和批次效应等难题, 推动了单细胞蛋白质组学的发展^[76]. 尽管这些模型取得了显著的进步, 但仍面临一些挑战, 例如模型的可解释性、对批次效应的敏感性以及对多组学数据整合的局限性等. 未来, 开发更具鲁棒性、可解释性和多组学整合能力的深度学习模型将是单细胞数据分析领域的关键方向, 这将有助于更全面地解析细胞异质性, 深入理解复杂的生物过程, 并最终推动

疾病诊断、治疗和药物开发的进展.

2.5 药物发现中的应用

深度学习技术在药物发现领域的应用已经广泛展开, 特别是在药物与靶标关联研究中, 通过评估蛋白质与药物的亲和力, 这类研究对发现新靶标具有重要价值. 此外, 研究药物间的关联性在临床医学中也扮演关键角色, 它有助于识别新药的潜在副作用、适应证以及药物间的相互作用^[7,77]. 例如, Zong等人^[16]开发的基于深度学习的药物靶标预测方法利用DeepWalk算法在已知药物-靶点网络上计算特征向量和节点相似性, 进而通过药物基于相似性的药物(DBSI)和靶点(TBSI)推断来探索它们之间的关联. 另一方面, Ma等人^[78]使用多视图图形自动编码器(graph autoencoder, GAE)来学习药物的相似性. 在他们的研究中, 药物被模型化为关联网络中的节点, 使用扩展的图卷积网络(graph convolutional network, GraphCNN)获得多视图节点和边缘嵌入, 进而预测药物间的相似性. 类似的, GraphDTA是一种基于图神经网络的深度学习算法, 专门用于预测药物-靶标亲和力^[79]. 该算法将药物和蛋白质靶标的拓扑结构信息编码为图形, 利用图神经网络来分析这些复杂结构. GraphDTA的方法使其能直接从药物和蛋白质的结构中学习, 而无须依赖预定义的特征, 从而提高了预测的准确性和效率. 利用迁移学习和大规模药物基因组数据集训练而成的scDEAL能够预测来自单细胞RNA测序中的药物反应和基因特征, 这有望促进更有效的个性化治疗策略^[80]. 同样, Xie团队^[81]发出的DLEPS是基于深度学习的工具, 能够综合多种复杂的基因指标以预测候选药物分子在疾病治疗中的应用. KinScan是一个先进的计算工具, 用于预测化学化合物与激酶靶标结合亲和力, 这种方法采用多尺度上下文感知变换器框架, 在391种蛋白激酶的虚拟分析模型中表现出强大的预测性能, 对于激酶全谱分析和药物选择性研究具有显著的实用价值^[82].

总的来说, 深度学习技术在药物发现领域的应用, 尤其是在药物-靶标关联和药物间关联性的研究中, 已经取得了显著的进展. 这些研究不仅为新靶标的发现和新药的副作用预测提供了强大的工具, 还促进了药物相互作用的预测, 从而加速了药物发现的进程. 然而, 值得注意的是, 尽管大多数对于药物反应的注释通常分为敏感或者抵抗(sensitive/resistant)两种, 实际

上, 药物敏感性更适合被视为一个连续的数值变量, 例如半数抑制浓度(IC50, half maximal inhibitory concentration)^[83-85]或曲线下面积(area under the curve, AUC)值^[86-88], 而不是简单的二分类变量^[89]. 连续的药物敏感性值能够更准确地反映药物的疗效, 对制定个性化用药方案和预测疗效具有至关重要的重要意义^[3]. 因此, 在构建预测模型时, 应优先使用定量的药物敏感性数据, 而非仅依赖于二元的药物反应注释.

2.6 疾病多组学中的应用

深度学习算法的应用在生物医学领域已经取得显著成就, 广泛应用于疾病诊断、分类、患者预后分析以及生物标记物的发现. 这些技术通过分析生物网络数据结构来探索疾病的全局和发病机制, 并发现与疾病相关生物分子. 四川大学团队开发的LDAformer工具, 基于拓扑特征提取和Transformer编码器, 整合长链非编码RNA(long non-coding RNA, lncRNA)、疾病和微小RNA(miRNA, microRNA)的关联, 构建异质网络, 并以lncRNA-disease-miRNA加权邻接矩阵形式表示内部类相似性和跨类关联, 用于预测lncRNA与疾病之间的关联^[90]. Zhang等人^[91]开发了一种基于相似性的方法, 利用miRNA疾病网络的拓扑结构预测miRNA与疾病的潜在关联. deepPseudoMSI将原始液相色谱-质谱联用技术(liquid chromatography-mass spectrometry, LC-MS)代谢组学数据转换为伪质谱图像, 并应用深度学习图像分析技术进行疾病诊断^[92]. 与传统的靶点分析方法相比, 这种方法更灵活且精度更高.

2.6.1 影像学诊断

由于其非侵入性和获得结果便利性, 基于影像的生物标志物在推动个性化医疗方面展现出广阔的前景. 深度学习技术, 如CNNs和RNNs, 已广泛应用于医学影像分析. 例如, CNNs已用于构建肺癌系统级风险分层模型^[93], 超宽视野荧光素血管造影图像分析以检测糖尿病视网膜病变^[94], 以及利用Ensemble 3DCNN分析阿尔茨海默病患者的全脑结构磁共振成像(magnetic resonance imaging, MRI)变化^[95].

影像基因组学(radiogenomics)融合影像组学与基因组学, 通过医学影像揭示个体的基因信息, 代表着从解剖到基因水平的重要演进. Zhou等人开发并验证了一种基于深度学习的影像组学模型, 用于预测胸腺上

皮肿瘤(thymic epithelial tumors, TETs)的组织学风险分类, 其研究表明, 结合深度学习与临床因素可显著提高预测准确性^[96]. Gu等人^[97]开发并验证了利用增强CT图像的放射深度学习特征(radiomic deep learning features, RDPs), 用以预测非功能性(non-functional pancreatic neuroendocrine tumors, NF-PanNETs)患者的淋巴结转移(lymph node metastasis, LNM). RDPs在预测淋巴结转移和无病生存(disease-free survival, DFS)方面展现了高准确性, 显示出其在指导术前决策和制定精准治疗策略中的潜在价值.

尽管影像基因组学结合深度学习展示了巨大潜力, 确保结果的可重复性和标准化、处理大数据集依赖以及在不同队列和成像技术中的严格验证仍然是该领域面临的挑战. 这些挑战凸显了影像基因组学作为跨学科领域的复杂性. 总体来看, 影像基因组学, 尤其是其与深度学习的结合, 为精准医疗的显著进展铺设了基础. 随着该领域的不断发展, 预期将开发出更加个性化和有效的治疗策略, 尤其是在肿瘤学领域, 有望弥合影像表型遗传基础之间的研究深度的差异.

2.6.2 肠道微生物组学

肠道微生物组与人类健康之间的紧密联系已经得到广泛证实, 深度学习技术在这一领域的应用也在日益增多. 例如, 空间测序技术SHM-seq(spatial host-microbiome sequencing)用于分析肠道内宿主与微生物的生态位^[98], 而“hiPCA”则是一种统计监测方法, 用于量化肠道微生物组的健康状态并预测疾病^[99]. 在炎症性肠病研究中, Hu等人^[100]使用变分自编码器和特定的分布模型整合不完整的多组学数据, 开发了IMOVNN算法. 这一算法在准确性上超过了以往的方法, 并能高效地识别相关生物标志物. Zhang等人^[101]等人开发了一种名为“microDELTA”的迁移学习技术, 使用肠道微生物组数据来追踪人类的寿命轨迹, 揭示了不同生命周期的肠道微生物组变化. 相较于传统的纵向分析方法, microDELTA在上下文感知、数据不足和批处理效应方面实现了显著改进. 总之, 这些研究展示了深度学习在利用肠道微生物组数据进行健康监测和人类寿命的疾病预测方面的创新应用.

2.6.3 医疗信息管理

在癌症治疗领域, 多模态数据融合技术正在改变

制定治疗方案的过程。电子健康记录(electronic health record, EHR)、分子数据、数字病理学和影像学图片的整合,为癌症患者的诊断和治疗提供了全面的信息,这种综合性的方法能够提高对患者进行个性化治疗的能力。例如,基于注意力机制的多实例学习和自我标准化网络成功整合了病理图像和分子数据^[102],而NLP算法从非结构化临床数据EHR中提取临床信息^[103],提高了对癌症患者进行个性化治疗的能力。

深度学习技术、算法在生物医学领域的应用正不断拓展,为个性化医学提供了强大的动力。这些技术、算法已成功应用于疾病的诊断、分类、预后分析以及生物标记物的发现的研究中,它们通过分析和整合多组学数据,包括基因组学、转录组学和影像学数据,增强了对疾病机制的理解,并推动了精准治疗策略的发展。尽管存在样本规模和数据限制的挑战,深度学习模型的预训练权重和交互式门户正加速各研究领域的创新和研究。

2.6.4 非编码RNA与代谢组学

深度学习在非编码RNA和代谢组学研究中展示了广泛的应用潜力。四川大学团队开发了一种创新的基于拓扑特征提取和Transformer编码器的深度学习工具LDAformer^[90],通过整合lncRNA、疾病和miRNA之间的关联,构建了一个异质网络,其中LDAformer模型使用加权邻接矩阵表达网络内部的类似性和跨类关联,有效预测lncRNA与疾病之间的潜在关联。同样,Zhang等人^[91]利用miRNA疾病网络的拓扑结构进行相似性预测,揭示了miRNA与疾病之间的关系。

在代谢组学研究中,一系列创新性工作进已显著推动了质谱数据的深度学习应用。例如deepPseudoMSI^[92]将原始的LC-MS代谢组学数据转换成伪质谱图像,并使用深度学习算法进行分析,这种方法比传统方法更灵活,并在疾病诊断和预测中显示出明显优势。PEAKS^[104]用于从串联质谱中高效测序新肽段;同时,Dührkop^[105]开发了一种深度核学习方法,将SVM和DNNs整合在一起,用于从串联质谱中预测分子指纹,优于当前的黄金标准。MSBooster^[106]算法结合保留时间、离子迁移以及卷积和循环网络的光谱预测,已被集成到FragPipe平台中,用以提升肽的鉴定率。MetFID^[107]则是一个基于ANN对化合物分子指纹预测的

算法,可用于代谢组研究中的代谢物注释。

这些研究展示了深度学习在整合和挖掘多组学数据方面的巨大潜力,为深入理解疾病的发生和发展机制以及制定新的诊断和治疗策略提供了新见解。随着深度学习技术的不断发展和完善,它在疾病研究中的作用日益重要。

2.7 表观遗传学

在表观遗传学领域,深度学习技术已能推断出健康与疾病过程中的复杂调控机制,成为推动个性化医学发展的关键技术。然而,我们面临的一大挑战是精确测定特定细胞类型中动态变化的表观修饰相关的基因组表达模式^[108]。虽然目前创建个体特异性表观基因组图谱尚有困难,计算生物学方法已能利用现有数据来弥补这一空白,助力我们深入理解表观基因组的复杂性,并预测精准医学中的个体差异^[108]。

研究者们已经开发出了多种工具,如跨物种DNA甲基化预测工具iDNA-ABF^[41],以及卷积变分自编码器(convolutional variational autoencoder, CVAE)和DNN算法组合,后者已经在宫颈癌数据中展现了高准确性和AUC值^[109]。这进一步证实了整合DNA甲基化和基因表达数据在癌症表观遗传学精准医疗中的巨大潜力。

这些研究利用包括注意力机制、卷积网络、图网络在内的深度学习模型,整合表观基因组学、转录组学和染色质构象数据,从而预测和解释功能基因组学。例如,Berson等人^[110]开发的Cellformer模型能够将批量染色质可及性测序(assay for transposase-accessible chromatin using sequencing, ATAC-seq)数据分解为与神经系统疾病相关的细胞类型特异性配置文件,而MSI-XGNN模型则能识别出用于患者分层的关键甲基化和表达生物标志物^[111]。虽然存在样本规模和数据限制,这些模型展现了在跨组织、跨疾病乃至跨物种的研究中的推断潜力。

EPCOT^[112]和TRAMHap^[113]等模型展现出优于先前方法的预测能力,并提供了将远端增强子甲基化与基因表达调控之间关系的解释性洞见。深度学习和表观遗传学数据的融合,为研究非编码区域的疾病变异机制,开展患者分层分析,以及指导治疗开发等方面提供了可行且高效的策略。这种跨学科的结合,为精准医疗的发展带来了新的契机。当前方法在应对人类

疾病的预训练和模型优化方面存在局限, 且在处理罕见细胞子集的性能有待提高. 然而, 共享统一的模型框架、预训练权重和交互式平台将无疑为各研究团队的探索工作提供助力, 加速推进相关研究.

3 现存的挑战和未来展望

识别当前深度学习方法存在的缺陷有助于改进深度学习算法.

第一, 上述研究中使用的训练数据集的样本量较小, 可能导致过拟合或泛化性弱. 预测准确性受训练数据的影响较大, 新数据集表现可能有所不同. 为了提高深度学习算法的通用性和鲁棒性, 未来的工作需要扩大训练数据集的规模(如空间转录组数据等^[102])采用多组学策略整合更多类型的数据^[114]. 在蛋白质预测中, 增加PPI和蛋白质特性的信息, 可使预测更贴近实验结果. ELASPIC2通过引入超过2000个蛋白质复合物的大型数据、新特征如蛋白质的进化、热力学和动力学信息, 以及使用新Transformer架构来处理长距离依赖性, 提高了预测精度和适用性, 例如在预测蛋白质与小分子的相互作用方面^[115]. 基于Transformer架构的MolFPG模型, 通过集成多级指纹的图, 显著提高了药物毒性预测的准确性和鲁棒性, 并且增强了模型的解释性^[31]. 然而, 训练数据集中的偏差、噪声或不全可能会影响ELASPIC2和MolFPG的性能. 要全面验证这些模型的通用性, 需在更大且多样化的数据集上进一步评估.

第二, 尽管现有的AI工具和算法能够帮助研究者处理测序、质谱及互作数据, 未来仍需要更高效的计算生物学工具. 这些工具应利用预训练或大型模型的优势, 以提升其在下游任务的表现, 进而增强计算效率和准确性. 例如, Hiranuma等人^[116]开发的DeepAccNet框架, 可以估计蛋白质结构的精确度和残基间误差, 这些预测有助于指导Rosetta蛋白质结构优化. 另一例是AlphaFold2, 该模型运用预训练的Transformer来分析蛋白质序列, 并融合进化信息与物理约束, 在CASP14比赛中大幅提升性能^[54,61]. 这些案例说明, 利用预训练或大型模型优化并结合专业知识, 可以开发出更高效、精准的计算生物学工具. 这些新工具将帮助研究者从大量预测中提取关键信息, 促进技术迭代, 并加深对蛋白质复合物等复杂结构的理解. 随着AI技

术的发展, 这些工具预期将在生物医学研究中扮演更重要角色, 为人类健康和生命科学的进步作出显著贡献.

第三, 目前用于跨物种转移学习的数据集在物种数量和样本量上有限, 可能会影响模型的泛化性和鲁棒性. 为了更全面地理解的基因调控网络, 需要为更多物种收集大规模转录组学数据. 这有助于发现物种间共有的基因表达调控模式, 并验证预测模型的泛化性. 我们需要开发算法显式处理同源基因在不同物种间表达差异. 这些算法应引入进化信息或域适应技术, 以提升预测精度. 因此, 算法需要能够鉴别和校正这些差异, 以提高跨物种预测的准确性. 这可以通过引入进化信息或使用域适应等人工智能技术来实现. 除了数据和算法, 超参数优化^[117](如模型学习率、隐藏层数等)对模型性能至关重要. 利用贝叶斯优化和强化学习等技术可以高效地找到最佳超参数组合. 同时, 需注意控制模型复杂度以避免过拟合.

第四, 理解疾病状态的基因调控网络对于揭示疾病机制以及确定干预靶点至关重要. 然而, 对于罕见疾病和临床上不可触及的组织, 这类研究通常会受到患者样本数量的限制^[118]. 为克服样本数量限制, Theodoris等人^[118]使用了接近3000万个来自各种组织的单细胞的转录组数据开发了自监督深度学习模型Geneformer, 从而能够深入理解基因调控网络的动态. 针对特定场景, 通过微调只需使用少量的单细胞数据进行, Geneformer就能有效预测驱动疾病的关键网络, 并确定有高治疗潜力的靶点^[118]. 另外, LLMs在多个领域的小样本学习中显示出巨大潜力, 如CancerGPT即在样本很少或甚至无样本的情况下, 也能达到高准确性^[119].

深度学习模型因其庞大的参数和复杂的网络结构, 难以解析其预测或决策的内部机制, 导致缺乏系统的生物学解释^[80,102]. 在临床应用中, 解释模型的决策过程对于医务人员和患者至关重要^[92], 它不仅增强模型决策的可信度, 还有助于医务人员理解诊断结果, 从而赢得患者的信任并促进治疗合作. 为了提高模型的可解释性, 未来研究应探索更多融合生物学知识的深度学习方法. Transformer框架在增强模型解释性表现出潜力, 如MolFPG的注意力权重帮助研究人员识别活性团, 进而解释药物作用^[31]. Wang等人^[120]开发的端到端神经网络融合了12层预训练的ESM(evolution-

ary scale modeling) Transformer模型、多尺度卷积层和全局多头注意力网络, 有效提取序列特征并关注关键分类贡献。该模型结合了一个12层预训练的变体Transformer架构ESM蛋白质语言模型以提取有用的序列表示, 多尺度卷积层以捕获局部特征, 以及一个全局多头注意力网络以关注用于分类的逐位置贡献。该算法采用交叉熵损失和标签平滑进行训练, 优化策略包括网格搜索和交叉验证, 其在独立测试集的表现优于前者, 并通过可视化注意力分数增强了可解释性。另外, 一种基于Reactome通路层次结构的可视神经网络(visual neural network, VNN)并纳入了逐层相关传播(layer-wise relevance propagation, LRP)的DTox, 不仅匹敌传统模型的预测性能, 还大幅增强了模型可解释性^[121]。利用LLMs也是提高模型解释性的一个研究方向, CancerGPT则可以通过几个已知的科学事实来推断药物协同效应进行^[119]。

尽管深度学习模型解释方法还在发展中, 但其在临床应用中的前景非常值得期待。研究人员正致力于改进这些方法, 以提高模型在临床实践中可解释性。

随着技术进步, 深度学习预计将在临床实践中发挥更大作用, 改善医疗决策过程。

4 结论

总体而言, 深度学习在基因组学、转录组学、蛋白质组学、药物发现和疾病生物学等生物医学领域展现出了巨大的潜力并取得了显著的进步。相比于传统的机器学习方法, 深度学习在特征提取、抗噪声、泛化和多功能性方面更为先进。随着大型生物医学数据集的可用性和模型架构的进步, 深度学习已经成为主流技术, 广泛应用于预测蛋白质结构、基因调控、药物-靶点相互作用和疾病结果等任务。然而, 挑战如模型可解释性和多模态数据的集成方面需解决, 以促进其在临床应用中广泛使用。未来, 尽管需要谨慎考虑模型限制, 深度学习依然将通过解析大型生物医学数据提供新的洞见, 从而革新生物医学领域。网络设计和训练方法的持续进步将进一步发挥深度学习的潜力, 增益人类健康。

参考文献

- 1 Ritchie M D, Holzinger E R, Li R, et al. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*, 2015, 16: 85-97
- 2 Chen R, Mias G I, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 2012, 148: 1293-1307
- 3 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*, 2019, 25: 1337-1340
- 4 Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 2023, 381: eadg7492
- 5 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521: 436-444
- 6 Xu H, Jia P, Zhao Z. DeepVISP: deep learning for virus site integration prediction and motif discovery. *Adv Sci*, 2021, 8: 2004958
- 7 Jing Y, Bian Y, Hu Z, et al. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *AAPS J*, 2018, 20: 58
- 8 Shu X, Zhang L, Wang Z, et al. Deep neural networks with region-based pooling structures for mammographic image classification. *IEEE Trans Med Imag*, 2020, 39: 2246-2255
- 9 Auslander N, Gussow A B, Benler S, et al. Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Res*, 2020, 48: e121
- 10 Boone M, Ramasamy P, Zuallaert J, et al. Massively parallel interrogation of protein fragment secretability using SECRiFY reveals features influencing secretory system transit. *Nat Commun*, 2021, 12: 6414
- 11 Kulmanov M, Khan M A, Hoehndorf R, et al. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 2017, 34: 660-668
- 12 Ornes S. Researchers turn to deep learning to decode protein structures. *Proc Natl Acad Sci USA*, 2022, 119: e2202107119
- 13 Altae-Tran H, Kannan S, Suberski A J, et al. Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science*, 2023, 382: eadi1910

- 14 Zheng W, Wuyun Q, Zhou X, et al. LOMETS3: integrating deep learning and profile alignment for advanced protein template recognition and function annotation. *Nucleic Acids Res*, 2022, 50: W454–W464
- 15 Suchanek M, Radzikowska A, Thiele C. Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. *Nat Methods*, 2005, 2: 261–268
- 16 Zong N, Kim H, Ngo V, et al. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics*, 2017, 33: 2337–2344
- 17 Grønning A.G.B., Doktor T.K., Larsen S.J., et al. DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning. *Nucleic Acids Res*, 2020, 48: 7099–118
- 18 Zhu H. Big data and artificial intelligence modeling for drug discovery. *Annu Rev Pharmacol Toxicol*, 2020, 60: 573–589
- 19 Singh R, Sledzieski S, Bryson B, et al. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc Natl Acad Sci USA*, 2023, 120: e2220778120
- 20 Wang G, Liu X, Wang K, et al. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nat Med*, 2023, 29: 2007–2018
- 21 Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netws*, 2015, 61: 85–117
- 22 Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform*, 2016, bbw068
- 23 Mamoshina P, Vieira A, Putin E, et al. Applications of deep learning in biomedicine. *Mol Pharm*, 2016, 13: 1445–1454
- 24 Faust O, Hagiwara Y, Hong T J, et al. Deep learning for healthcare applications based on physiological signals: a review. *Comput Methods Programs Biomed*, 2018, 161: 1–13
- 25 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 2016, 316: 2402
- 26 Bashyam V M, Doshi J, Erus G, et al. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *Magn Reson Imag*, 2022, 55: 908–916
- 27 Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods*, 2015, 12: 931–934
- 28 Zhang L, Shi J, Ouyang J, et al. X-CNV: genome-wide prediction of the pathogenicity of copy number variations. *Genome Med*, 2021, 13: 132
- 29 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Med*, 2018, 1: 18
- 30 Choi Y S, Bae S, Chang J H, et al. Fully automated hybrid approach to predict the *IDH* mutation status of gliomas via deep learning and radiomics. *Neuro-Oncology*, 2021, 23: 304–313
- 31 Teng S, Yin C, Wang Y, et al. MolFPG: multi-level fingerprint-based Graph Transformer for accurate and robust drug toxicity prediction. *Comput Biol Med*, 2023, 164: 106904
- 32 The Lancet . Artificial intelligence in health care: within touching distance. *Lancet*, 2017, 390: 2739
- 33 Zhang J, Zhao J, Lin H, et al. High-speed chemical imaging by dense-net learning of femtosecond stimulated raman scattering. *J Phys Chem Lett*, 2020, 11: 8573–8578
- 34 Ching T, Himmelstein D S, Beaulieu-Jones B K, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*, 2018, 15: 20170387
- 35 Liu Q, Hu P. Extendable and explainable deep learning for pan-cancer radiogenomics research. *Curr Opin Chem Biol*, 2022, 66: 102111
- 36 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint, arXiv: 1312.6199, 2013
- 37 Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017, 542: 115–118
- 38 Zhang Y, Song X, Xie J, et al. Large depth-of-field ultra-compact microscope by progressive optimization and deep learning. *Nat Commun*, 2023, 14: 4118
- 39 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529: 484–489
- 40 Jiang H, Luo X, Yin J, et al. Orthogonal subspace representation for generative adversarial networks. *IEEE Trans Neural Netw Learn Syst*, 2024, 1–15
- 41 Jin J, Yu Y, Wang R, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. *Genome Biol*, 2022, 23: 219
- 42 Yang M, Huang L, Huang H, et al. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Res*, 2022, 50: e81

- 43 Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 2015, 33: 831–838
- 44 Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol*, 2019, 20: 9
- 45 Avsec Ž, Weilert M, Shrikumar A, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet*, 2021, 53: 354–366
- 46 Kc K, Li R, Cui F, et al. GNE: a deep learning framework for gene network inference by aggregating biological information. *BMC Syst Biol*, 2019, 13: 38
- 47 Kang T, Ding W, Zhang L, et al. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. *BMC Bioinf*, 2017, 18: 565
- 48 Zhang Z, Pan Z, Ying Y, et al. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat Methods*, 2019, 16: 307–310
- 49 Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res*, 2020, 30: 214–226
- 50 Fischer D S, Schaar A C, Theis F J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat Biotechnol*, 2023, 41: 332–336
- 51 He B, Bergensträhle L, Stenbeck L, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng*, 2020, 4: 827–834
- 52 Callaway E. ‘The entire protein universe’: AI predicts shape of nearly every known protein. *Nature*, 2022, 608: 15–16
- 53 Stahl K, Graziadei A, Dau T, et al. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nat Biotechnol*, 2023, 41: 1810–1819
- 54 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583–589
- 55 Chowdhury R, Bouatta N, Biswas S, et al. Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol*, 2022, 40: 1617–1623
- 56 Brandes N, Ofer D, Peleg Y, et al. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 2022, 38: 2102–2110
- 57 Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023, 379: 1123–1130
- 58 Dauparas J, Anishchenko I, Bennett N, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 2022, 378: 49–56
- 59 Anishchenko I, Pellock S J, Chidyausiku T M, et al. *De novo* protein design by deep network hallucination. *Nature*, 2021, 600: 547–552
- 60 Repecka D, Jauniskis V, Karpus L, et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nat Mach Intell*, 2021, 3: 324–333
- 61 Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun*, 2022, 13: 1265
- 62 Gao Z, Jiang C, Zhang J, et al. Hierarchical graph learning for protein–protein interaction. *Nat Commun*, 2023, 14: 1093
- 63 Frappier V, Chartier M, Najmanovich R J. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic Acids Res*, 2015, 43: W395–W400
- 64 Raimondi D, Tanyalcin I, Féré J, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res*, 2017, 45: W201–W206
- 65 Witvliet D K, Strokach A, Giraldo-Forero A F, et al. ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. *Bioinformatics*, 2016, 32: 1589–1591
- 66 Cao H, Wang J, He L, et al. DeepDDG: predicting the stability change of protein point mutations using neural networks. *J Chem Inf Model*, 2019, 59: 1508–1514
- 67 Liu X, Luo Y, Li P, et al. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS Comput Biol*, 2021, 17: e1009284
- 68 Brandes N, Goldman G, Wang C H, et al. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet*, 2023, 55: 1512–1522
- 69 Gao H, Hamp T, Ede J, et al. The landscape of tolerated genetic variation in humans and primates. *Science*, 2023, 380: eabn8153
- 70 Cui H, Wang C, Maan H, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods*, 2024,

doi: 10.1038/s41592-024-02201-0

- 71 Yang F, Wang W, Wang F, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell*, 2022, 4: 852–866
- 72 Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods*, 2024, doi: 10.1038/s41592-024-02235-4
- 73 Weinberger E, Lin C, Lee S I. Isolating salient variations of interest in single-cell data with contrastiveVI. *Nat Methods*, 2023, 20: 1336–1345
- 74 Bahl E, Chatterjee S, Mukherjee U, et al. Using deep learning to quantify neuronal activation from single-cell and spatial transcriptomic data. *Nat Commun*, 2024, 15: 779
- 75 Angermueller C, Lee H J, Reik W, et al. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol*, 2017, 18: 67
- 76 Li W, Yang F, Wang F, et al. scPROTEIN: a versatile deep graph contrastive learning framework for single-cell proteomics embedding. *Nat Methods*, 2024, 21: 623–634
- 77 Jin S, Zeng X, Xia F, et al. Application of deep learning methods in biological networks. *Brief BioInf*, 2020, 22: 1902–1917
- 78 Ma T, Xiao C, Zhou J, et al. Drug similarity integration through attentive multi-view graph auto-encoders. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press, 2018/ 3477–83
- 79 Nguyen T, Le H, Quinn T P, et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 2020, 37: 1140–1147
- 80 Chen J, Wang X, Ma A, et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat Commun*, 2022, 13: 6494
- 81 Zhu J, Wang J, Wang X, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nat Biotechnol*, 2021, 39: 1444–1452
- 82 Brahma R, Shin J M, Cho K H. KinScan: AI-based rapid profiling of activity across the kinome. *Brief BioInf*, 2023, 24: bbad396
- 83 Jaaks P, Coker E A, Vis D J, et al. Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, 2022, 603: 166–173
- 84 Robichaux J P, Le X, Vijayan R S K, et al. Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature*, 2021, 597: 732–737
- 85 Zhao Z, Qing Y, Dong L, et al. QKI shuttles internal m7G-modified transcripts into stress granules and modulates mRNA metabolism. *Cell*, 2023, 186: 3208–3226.e27
- 86 Pemovska T, Bigenzahn J W, Srndic I, et al. Metabolic drug survey highlights cancer cell dependencies and vulnerabilities. *Nat Commun*, 2021, 12: 7190
- 87 Shi R, Wang X, Wu Y, et al. APOBEC-mediated mutagenesis is a favorable predictor of prognosis and immunotherapy for bladder cancer patients: evidence from pan-cancer analysis and multiple databases. *Theranostics*, 2022, 12: 4181–4199
- 88 Woo X Y, Giordano J, Srivastava A, et al. Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. *Nat Genet*, 2021, 53: 86–99
- 89 Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 2012, 483: 603–607
- 90 Zhou Y, Wang X, Yao L, et al. LDAformer: Predicting lncRNA-disease associations based on topological feature extraction and Transformer encoder. *Brief BioInf*, 2022, 23: bbac370
- 91 Zhang H, Liang Y, Peng C, et al. Predicting lncRNA-disease associations using network topological similarity based on deep mining heterogeneous networks. *Math Biosci*, 2019, 315: 108229
- 92 Shen X, Shao W, Wang C, et al. Deep learning-based pseudo-mass spectrometry imaging analysis for precision medicine. *Brief BioInf*, 2022, 23: bbac331
- 93 Choi H, Na K J. A risk stratification model for lung cancer based on gene coexpression network and deep learning. *Biomed Res Int*, 2018, 2018: 1–11
- 94 Zhao P Y, Bommakanti N, Yu G, et al. Deep learning for automated detection of neovascular leakage on ultra-widefield fluorescein angiography in diabetic retinopathy. *Sci Rep*, 2023, 13: 9165
- 95 Pan D, Zeng A, Yang B, et al. Deep learning for brain MRI confirms patterned pathological progression in Alzheimer’s disease. *Adv Sci*, 2023, 10: e2204717
- 96 Zhou H, Bai H X, Jiao Z, et al. Deep learning-based radiomic nomogram to predict risk categorization of thymic epithelial tumors: a multicenter

- study. [Eur J Radiol](#), 2023, 168: 111136
- 97 Gu W, Chen Y, Zhu H, et al. Development and validation of CT-based radiomics deep learning signatures to predict lymph node metastasis in non-functional pancreatic neuroendocrine tumors: A multicohort study. [eClinicalMedicine](#), 2023, 65: 102269
- 98 Lötstedt B, Stražar M, Xavier R, et al. Spatial host–microbiome sequencing reveals niches in the mouse gut. [Nat Biotechnol](#), 2023, doi: [10.1038/s41587-023-01988-1](#)
- 99 Zhu J, Xie H, Yang Z, et al. Statistical modeling of gut microbiota for personalized health status monitoring. [Microbiome](#), 2023, 11: 184
- 100 Hu M, Zhu J, Peng G, et al. IMOVNN: Incomplete multi-omics data integration variational neural networks for gut microbiome disease prediction and biomarker identification. [Brief BioInf](#), 2023, 24: bbad394
- 101 Zhang H, Chong H, Yu Q, et al. Tracing human life trajectory using gut microbial communities by context-aware deep learning. [Brief BioInf](#), 2023, 24: bbac629
- 102 Chen R J, Lu M Y, Williamson D F K, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. [Canc Cell](#), 2022, 40: 865–878.e6
- 103 Yuan Q, Cai T, Hong C, et al. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. [JAMA Netw Open](#), 2021, 4: e2114723
- 104 Ma B, Zhang K, Hendrie C, et al. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. [Rapid Comm Mass Spectrometry](#), 2003, 17: 2337–2342
- 105 Dührkop K. Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra. [Bioinformatics](#), 2022, 38: i342–i349
- 106 Yang K L, Yu F, Teo G C, et al. MSBooster: improving peptide identification rates using deep learning-based features. [Nat Commun](#), 2023, 14: 4539
- 107 Fan Ziling, Alley Amber, Ghaffari Kian, Resson Habtom W., 等. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. [Metabolomics](#), 2020, 16: 10: 104
- 108 Hawkins-Hooker A, Visonà G, Narendra T, et al. Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning. [Nat Commun](#), 2023, 14: 4750
- 109 Albaradei S, Napolitano F, Thafar M A, et al. MetaCancer: a deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. [Comput Struct Biotechnol J](#), 2021, 19: 4404–4411
- 110 Berson E, Sreenivas A, Phongpreecha T, et al. Whole genome deconvolution unveils Alzheimer’s resilient epigenetic signature. [Nat Commun](#), 2023, 14: 4947
- 111 Cao Y, Wang D, Wu J, et al. MSI-XGNN: An explainable GNN computational framework integrating transcription- and methylation-level biomarkers for microsatellite instability detection. [Brief BioInf](#), 2023, 24: bbad362
- 112 Zhang Z, Feng F, Qiu Y, et al. A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. [Nucleic Acids Res](#), 2023, 51: 5931–5947
- 113 Gao S, Zhu H, Cai K, et al. TRAMHap: Accurate prediction of transcriptional activity from DNA methylation haplotypes in bisulfite-sequencing data. [Brief BioInf](#), 2023, 24: bbad214
- 114 Pammi M, Aghaepour N, Neu J. Multiomics, artificial intelligence, and precision medicine in perinatology. [Pediatr Res](#), 2023, 93: 308–315
- 115 Strokach A, Lu T Y, Kim P M. ELASPIC2 (EL2): combining contextualized language models and graph neural networks to predict effects of mutations. [J Mol Biol](#), 2021, 433: 166810
- 116 Hiranuma N, Park H, Baek M, et al. Improved protein structure refinement guided by deep learning based accuracy estimation. [Nat Commun](#), 2021, 12: 1340
- 117 Xiao X., Yan M., Basodi S., et al. Efficient hyperparameter optimization in deep learning using a variable length genetic algorithm. arXiv preprint, arXiv: [200612703](#), 2020
- 118 Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. [Nature](#), 2023, 618: 616–624
- 119 Li T, Shetty S, Kamath A, et al. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. [npj Digit Med](#), 2024, 7: 40
- 120 Wang L, Huang C, Wang M, et al. NeuroPred-PLM: An interpretable and robust model for neuropeptide prediction by protein language model. [Brief BioInf](#), 2023, 24: bbad077
- 121 Hao Y, Romano J D, Moore J H. Knowledge-guided deep learning models of drug toxicity improve interpretation. [Patterns](#), 2022, 3: 100565

An introduction to the application of deep learning in the biomedical field

WANG Ruisong^{1*}, WANG Shengnan² & SHI Tielu^{2*}

¹ *Hunan University of Arts and Science, College of Medicine, Changde 415000, China;*

² *School of Life Sciences, East China Normal University, Shanghai 200241, China*

** Corresponding authors, E-mail: ruisong.wang@huas.edu.cn; tieliushi@yahoo.com*

Deep learning, an emerging technology within the realm of artificial intelligence, has demonstrated significant promise in the field of biomedical sciences. This paper provides a comprehensive overview of the algorithms, principles, and applications associated with deep learning. Deep learning models are characterized by their multiple layers and ability to conduct hierarchical feature learning from data. In comparison to traditional machine learning methods, deep learning offers advantages such as automatic feature extraction, noise resistance, generalization capability, and versatility. Deep learning has demonstrated efficacy in addressing challenges within diverse biomedical domains, including genomics, transcriptomics, proteomics, drug discovery, and disease biology. Nevertheless, the interpretability of deep learning models remains a significant obstacle, particularly in the integration of multimodal datasets. Overcoming these challenges, specifically in enhancing model interpretability and integrating biological knowledge, will expedite the adoption of deep learning technology in clinical settings. Ultimately, the application of deep learning is revolutionizing the biomedical sciences by extracting novel insights from extensive biomedical datasets. Ongoing developments in network architecture, interpretability, and training techniques are poised to unlock the significant potential of deep learning in advancing human health.

biomedicine, application of artificial intelligence, application of deep learning, understanding and analysis of disease

doi: [10.1360/SSV-2023-0297](https://doi.org/10.1360/SSV-2023-0297)