重点选题“人工智能与医学”·特约专栏

人工智能在基础医学数据分析中的应用▲

李　璞[1,2]　秦剑秋[1,2]　邱　洪[3]　胡艳玲[3*]

（1 南宁市疾病预防控制中心，广西南宁市　530023；2 广西重点传染病防控监测检测重点实验室，广西南宁市　530023；3 广西医科大学生命科学研究院，广西南宁市　530021）

胡艳玲，广西医科大学二级教授、博士研究生导师，入选广西高层次人才。毕业于上海交通大学生物医学工程专业，主攻生物信息学及病原学方向，曾在美国威斯康辛大学学习深造，目前的主要研究方向为生物医学大数据的分析与挖掘、病原微生物的检测与分析、人工智能诊断。近年来主持国家自然科学基金、科技部重点研发计划子项目和广西各类重点项目20多项。以通信作者或第一作者在*Nature*、*Cell*、*Nature Microbiology*、*Nature Communications*等顶级刊物发表SCI论文70多篇，总引用频次达5 000多次。获专利或软件著作权16部，获中华医学科技奖二等奖1项、省级科学技术进步奖二等奖3项。现为中国生物信息学学会生物数据资源专业委员会委员、中国生物工程学会噬菌体分会委员、亚太科学与工程学会主席、中华医学生物免疫学会基础免疫学分会常务委员、中华医学生物免疫学会理事、广西预防医学会病原监测与生物信息专业委员会主任委员、广西人工智能学会生物信息学分会副主任委员、广西性病学会HIV基层专业委员会副主任委员、*iMeta*编委、*iCell*主编。

【提要】 近年来，人工智能技术迅猛发展，在基础医学数据分析中的应用日益广泛。人工智能经历从专家规则系统到数据驱动深度学习和多模态模型的快速演进，为基因组学、蛋白质结构预测、单细胞转录组研究、生物序列分析及微生物组学等领域的研究带来突破性进展。然而，这些技术也面临诸如数据隐私、资源限制和可解释性不足等瓶颈。本文主要探讨人工智能技术在基础医学数据分析中的应用现状、关键方法、面临的挑战及发展趋势。

【关键词】 人工智能；基础医学；数据分析；模型

【中图分类号】 R 319　【文献标识码】 A　【文章编号】 0253-4304(2025)08-1072-10

DOI：10.11675/j.issn.0253-4304.2025.08.02

Application of artificial intelligence in basic medical data analysis

LI Pu[1,2], QIN Jianqiu[1,2], QIU Hong[3], HU Yanling[3]

(1 Nanning Center for Disease Control and Prevention, Nanning 530023, Guangxi, China; 2 Guangxi Key Laboratory of Key Infectious Disease Prevention and Control, Surveillance, and Testing, Nanning 530023, Guangxi, China; 3 Life Sciences Institute, Guangxi Medical University, Nanning 530021, Guangxi, China)

【**Abstract**】 In recent years, the rapid advancement of artificial intelligence technology has led to its increasingly widespread application in basic medical data analysis. Artificial intelligence has undergone a swift evolution from expert rule-based systems to data-driven deep learning and multi-modal models, bringing groundbreaking progress to research fields such as genomics, protein structure prediction, single-cell transcriptomics, biological sequence analysis, and microbiology. However, these technologies also face bottlenecks, including data privacy concerns, resource limitations, and insufficient

▲基金项目：国家重点研发计划(2023YFC2605400)

第一作者简介：李璞，硕士，研究方向为宏基因组学和病毒基因组。

*胡艳玲为通信作者及本期专栏主持人。

interpretability. This paper primarily explores the current applications and key methodologies of artificial intelligence in basic medical data analysis, along with the challenges and future development trends.

【Key words】 Artificial intelligence, Basic medicine, Data analysis, Model

基础医学作为医学科学的基石，主要研究人体的结构、功能及其在健康与疾病状态下的变化，涵盖解剖学、生理学、生物化学、病理学、生物信息学等多个核心学科。基础医学不仅为临床实践提供了坚实的理论依据，也为医疗从业者掌握实践技能构建了必要框架[1]。近年来，随着高通量测序（基因组学、转录组学、蛋白质组学、代谢组学等）、表型组学、单细胞测序及高分辨率生物医学成像（显微镜图像、组织切片、细胞成像等）等数据获取技术的飞速发展，基础医学领域正经历一场前所未有的数据获取变革。虽然这些数据都是基础研究的宝贵资源，但其巨大的体量和内在的复杂性也给传统的数据分析方法带来了严峻挑战[2]。

在这一背景下，人工智能（artificial intelligence，AI）技术，特别是深度学习、自然语言处理、图神经网络等方法，以强大的数据挖掘和模式识别能力，正逐渐成为基础医学数据分析的核心工具。AI可以从复杂的基因组数据中识别疾病相关变异，从高分辨率显微镜图像中自动提取细胞特征，从复杂医学数据中挖掘潜在规律，辅助疾病机制解析、生物标志物发现及药物靶点预测，整合多维度医学数据构建疾病预测模型，极大加速了基础研究到临床转化的进程，成为解锁基础医学大数据价值的关键技术[3-4]。AI在基础医学中的应用正从单一任务向系统生物学方向发展。通过迁移学习和元学习等技术，AI模型可以在不同实验平台和疾病类型间实现知识迁移。这种通用型AI框架将显著降低医学研究的计算成本，促进跨学科合作。AI与基础医学的融合正在重塑生物医学研究的范式，不仅提高了研究效率，更为精准医疗、个体化治疗策略的制订提供了技术支持。本文深入探讨AI在基础医学数据分析中的应用现状、技术挑战及未来发展方向，旨在为相关领域的研究提供参考与启示。

1 用于基础医学数据分析的基本AI技术

AI是一种旨在模拟人类智能以完成复杂任务的技术。早期的AI较为依赖专家规则，典型例子是斯坦福大学在1972年开发的MYCIN系统，其主要用于血液感染类型判断和治疗方案推荐[5]。AI发展的一个重要里程碑是2012年AlexNet的提出[6]。作为一种基于卷积神经网络（convolutional neural network，CNN）的构架，AlexNet在医学影像识别和分类任务中展现出突破性能力，同时促进基于AlexNet的医学模型迅速发展，为数据驱动的生物医学研究开辟了新途径[7]。当前，AI的主要发展方向已经转向基于数据驱动进行自动学习与性能优化的机器学习、深度学习及大型语言模型（large language model，LLM）。随着AI的迅速发展，AI在医学数据分析领域的应用已愈加广泛。

机器学习作为AI的一个重要子集，在基础医学数据分析中发挥着关键作用。机器学习可不依赖明确的专家规则，而是在输入数据中进行学习，并且随着特定学习任务而提高性能[8]。根据学习方式的不同，机器学习主要分为监督学习、无监督学习和强化学习。监督学习算法利用标注数据进行训练，一般情况下算法会创建一个数学函数，将输入特征与预期的输出值联系起来，并在分类或回归中预测未标注数据的结果。常用的监督学习方法包括逻辑回归、支持向量机、随机森林、极端梯度提升（extreme gradient boosting，XGBoost）等，在疾病风险评估、医学诊断、生物标志物等特征分类和识别任务中表现良好。无监督学习算法则通过数据驱动的方式在未标注数据上进行训练，在学习过程中揭示数据的内在模式和相似性关系，常用于数据聚类、特征提取和降维。常用的无监督学习算法包括k均值聚类、主成分分析、主坐标分析、t-分布邻域嵌入算法等。无监督学习在基因表达量、单细胞分析及高维度医学数据降维分析中应用广泛。强化学习通过环境驱动，以与预定目标的比较结果获取奖励或惩罚，在医学领域中强化学习被应用于AI辅助外科手术操作、动态优化抗癌方案及个性化给药方案设计等复杂决策问题。这些机器学习方法经常被结合应用于临床相关AI模型的构建。机器学习在医学领域的应用日益广泛，且实际应用趋于成熟，不同的学习方法各具特色[3,9]。其中，监督学习因其在分类、诊断和预后预测任务中的高准确性，已成为医疗领域的核心方法之一。无监督学习在缺乏标注数据时展现出独特优势，通过聚类、异常检测和特征提取技术，帮助鉴别患者亚群、识别罕见病例或降低医疗数据的维度。强化学习的

独特优势在于能够应对高度复杂和动态变化的医疗环境，通过与环境的持续交互和反馈，改善决策方案，实现最优甚至创新性的临床决策路径，这对于提升医疗服务的个体化和智能化水平具有重要意义。

深度学习是机器学习中越来越受重视的分支，它使用具有多个处理层的人工神经网络，采用监督学习、无监督学习和强化学习等机器学习方法进行训练，在处理涉及大量高维度数据的复杂任务方面表现出色。深度学习的成功更依赖于网络深度带来的层次化特征学习能力，而非单纯增加参数数量。这种结构更接近人脑的分层信息处理机制，也是其解决复杂问题的核心优势[10-11]，在实际应用中需要结合训练数据、模型要求和计算资源的限制来共同决定。基于监督学习的深度网络包括CNN、循环神经网络（recurrent neural network，RNN）、长短期记忆网络（long short-term memory，LSTM）、Transformer构架、多层感知器等。其中，CNN在医学影像分析（X射线、CT、MRI等）中表现出色；RNN及LSTM在时间序列和医疗数据分析方面应用广泛；Transformer构架具有自注意力机制，能够捕捉长距离依赖关系，这对于分析医学数据中的复杂关系和依赖性尤其有用。同时，Transformer构架具有更强的可扩展性，使模型参数可扩展到数十亿甚至数千亿，而模型参数的增多和能力的增强使Transformer构架能够处理多模态和跨模态数据[4,12]。

近年来，ChatGPT、DeepSeek等LLM受到广泛关注。而LLM亦是基于海量数据训练的大规模深度学习模型，其核心能力在于通过预训练实现对复杂语义的深度理解与生成。作为基础模型，LLM不仅能够精准捕捉上下文关联性、生成符合语境的连贯文本，还展现出跨领域迁移学习的优势，在生物信息学领域实现了突破性进展。根据模型构架的不同，LLM可大致分为编码器模型、解码器模型和编码器-解码器模型。其中，编码器模型主要侧重于分析和理解输入文本的数据，这类模型的主要任务是将文本转换为一种易于处理的数字表示，捕捉输入的语义和上下文信息。基于双向编码器表示（bidirectional encoder representations from transformers，BERT）构架的生物序列模型如DNABERT、DNABERT2，能够将生物序列转换为结构、功能、表达量等特征，从而实现基于这些特征的任务，但缺乏自回归解码机制使其在生成类的下游任务中受到较大的限制[13-14]。解码器模型主要用于生成，使用较为宽松的自回归算法，具有单向的注意力机制，这意味着它们始终基于先前生成的信息来生成输出。在预训练阶段，该模型一般采用无监督的训练方式，这极大优化了训练数据的获取能力，而海量的数据使模型更容易获得对上下文的理解能力。该模型特别适合序列生成、结构预测和功能注释，因此其在需要从头合成序列和预测建模的生物信息学研究中具有很高的应用价值。编码器-解码器模型可实现从输入到输出的端到端模式，特别适合不同模式内容之间的转换，如RoseTTAFold[15]和ESMFold[16]可用于蛋白结构的预测。然而，它们的性能高度依赖于大规模特定领域的数据集，对用于训练和推理的计算资源也有较高的要求，这些因素限制了编码器-解码器模型的发展。

AI在医学数据分析中的发展核心依赖于模型构架、训练方式和数据驱动3大要素。模型构架从早期的专家规则系统逐步演变至复杂的深度神经网络，其中CNN在医学影像分析中表现突出，RNN及LSTM擅长时序数据处理，而Transformer构架凭借自注意力机制和可扩展性成为多模态分析的主流构架，LLM（DNABERT和ESMFold等）进一步拓展了生物序列的理解与生成能力。在训练方式上，监督学习（逻辑回归、XGBoost等）在疾病诊断和预后预测中占据主导地位，无监督学习（聚类、降维等）可挖掘未标注数据的价值，强化学习则能够优化决策过程，深度学习的层次化特征提取结合预训练-微调范式显著提升了模型性能。数据驱动是AI发展的基石，医学影像、基因组学、电子病历等多模态数据为模型训练提供丰富素材，数据规模的扩大直接推动模型突破，但数据隐私、标注成本和领域特异性仍是其应用过程中的挑战。未来，模型构架的创新、训练方式的优化及高质量数据的积累，将进一步推动AI在疾病预测、精准治疗和生物医学研究中的深度应用，实现从辅助诊断到个性化医疗的跨越。

2 基础医学数据分析场景和前沿模型

2.1 *蛋白质结构的分析、预测与蛋白质模型* 蛋白质结构的分析与预测一直是生命科学研究的重要领域。传统的实验室分析方法如X射线晶体分析、生物核磁共振波谱和冷冻电镜，被用于解析蛋白质的三维结构，但这些方法耗时长且价格昂贵，一个蛋白质结构的解析可能需要数月甚至数年。近年来，AI的发展为蛋白质结构的分析与预测提供了新的方向。由Google公司旗下DeepMind公司开发的AI程序AlphaFold2[17]，在2020年度蛋白质结构预测大赛（CASP14）中取得优

异成绩，其准确性可与实验室水平相媲美，表明基于AI的蛋白质结构预测研究迈出坚实的一步[17]。AlphaFold2首先构建多序列比对（multiple sequence alignment，MSA），同时搜索结构数据库，通过多层Evoformer神经网络块，实现在MSA表示和残基对表示之间的双向信息传递。Evoformer之后的网络主干是“结构模块”，该模块以蛋白质中每个残基的旋转和平移形式引入明确的3D结构，并迅速优化为准确的蛋白质结构，从而对整个网络进行“循环利用”的迭代优化[18]。作为后续升级版本，AlphaFold3引入了Pairformer模块，用“扩散模块”取代了上一代中非常重要的“结构模块”，从而简化了模型构架，并支持蛋白质、DNA、RNA、配体、金属离子等混合输入，能够构建广泛的生物分子复合物，因此其在各种场景中的准确性得到大幅提升，这些场景包括蛋白质-小分子、蛋白质-核酸及抗体-抗原相互作用等[19]。这些创新和优化使AlphaFold成为更通用、更高效的生物分子建模工具。Google公司在2024年11月11日对AlphaFold3进行了模型开源，允许用户申请和下载模型。2024年诺贝尔化学奖授予该团队成员中的Hassabis D和Jumper JM，以表彰他们在蛋白质结构预测方面的贡献。

ESM是由Facebook研究团队构建的蛋白质系列模型[20-21]，包含结构预测、功能预测、突变效应预测等众多模型和功能。ESM Cambrian（ESM-C）和ESM3是ESM系列的前沿模型，目前仅开源了部分小参数的模型。ESM-C专注于构建蛋白质基础生物学特性的表征，模型大小为300M/6B，用于替代ESM2。ESM3采用多模态深度学习构架，可同步处理蛋白质的氨基酸序列（一级结构）、折叠构象（三级结构）与生物功能之间的复杂关系，更专注于蛋白质的可控生成，其最大参数数量达到了惊人的980亿。该模型支持条件式生成控制，研究者可通过输入结构约束或功能描述，引导模型生成满足特定需求的蛋白质变体。

2.2 RNA/单细胞RNA测序数据的分析与RNA模型 RNA在遗传信息传递过程中发挥重要的桥梁作用，可实现遗传信息在蛋白质上的表达。RNA不仅能够直接参与蛋白质的合成，还在基因表达调控、细胞功能及疾病发展中扮演核心角色。随着RNA测序（RNA sequencing，RNA-seq）技术的不断发展，生物学数据呈现指数增长，研究者可从公共数据库获取训练相关数据，促使基于RNA的AI迅猛发展[22]。基因表达是一个复杂的过程，而RNA的转录和翻译则是决定蛋白质丰度的关键环节。基因表达数据是AI模型训练中的重要特征之一，这些表达数据主要来源于RNA-seq、染色质免疫沉淀测序、CAGE测序和核糖体印记测序（ribosome profiling sequencing，Ribo-seq）等技术的检测结果。RNA结构是维持RNA稳定性、功能和分子间相互作用的核心，RNA结构信息也在解析基因表达调控和其他生物学功能中起着重要作用。此外，一些调控分子对于多种生物过程至关重要，故RNA分子与其他调控分子的结合数据是AI模型训练的另一类重要数据，包括RNA与转录因子、miRNA及RNA结合蛋白之间的作用等。这些表达、结构、结合数据通过先进的编码方法整合到AI模型中，使研究者能够更好地解析RNA的功能与调控网络，并为RNA相关的疾病治疗和新药研发提供了强有力的工具支持。

Basenji和ExPecto都是基于CNN的RNA模型，两者均可基于多种转录调控数据进行训练和推理，可预测基因组中序列表达量，且预测结果与表达数量性状位点分析结果较为吻合[23-24]。相比而言，Xpresso没有利用表观遗传学数据进行训练，而是严格依赖基因组序列，其在不同细胞系甚至跨物种分析中都有较为理想的结果，表现出强大的泛化能力[25]。mRNA与蛋白的表达具有直接关系，利用mRNA的表达量可预测蛋白质的表达量。但是，越来越多的研究者分析两者之间的差异后发现，人类细胞系mRNA与蛋白质表达量的Pearson相关系数在0.39~0.79之间，平均仅约为0.6[26-27]，因此在疾病的基因研究中使用mRNA表达量代替蛋白质表达量可能并不准确[28]。Translatomer是一个基于Transformer构架的多模态深度学习框架，其预训练模型使用了成对的RNA-seq和Ribo-seq表达数据，输出为核糖体表达量的信号，因序列嵌入数据中包含了RNA二级结构和其调控信息，基于Pearson相关系数衡量的准确度达到了0.784[28]。

单细胞RNA测序（single-cell RNA sequencing，scRNA-seq）已成为揭示细胞多样性和组织复杂性的强大工具。传统的RNA-seq仅捕捉细胞群体的平均转录水平，而scRNA-seq则提供单个细胞水平上的测序结果[29]。scRNA-seq通过分析成千上万的细胞，能深入揭示不同类型细胞在健康状态和疾病状态中的作用，有助于在单体细胞水平上更深入地了解疾病机制。scRNA-seq数据具有高维度的特点，不同细胞类型差异巨大且容易产生批次效应，这些问题导致细胞的聚类分群和注释具有挑战性[29-30]。随着AI技术的发展，深度学习已成为分析和解释这些数据的强大工具。scBERT是一种基于深度学习的scRNA-seq数

据分析工具，旨在优化细胞类型注释的准确性和可靠性[31]。scBERT采用了预训练BERT的方法，首先在大规模的无标签scRNA-seq数据上进行预训练，以学习基因间的潜在关联模式，随后通过监督微调和适配不同数据集的细胞类型标注任务[31]。该模型在细胞类型注释准确性、未知细胞亚群发现能力、跨数据集泛化性能及生物学可解释性等关键指标上均优于现有方法[31]，为单细胞组学研究提供了新的计算模式[31]。由于Transformer构架在多个领域的成功，scGPT将其应用于单细胞研究领域。该模型在包含3 300万个细胞的庞大数据库中进行训练，可以完成包括细胞类型注释、去除批次效应、遗传扰动预测和基因网络推断等下游任务[32]。scGPT作为一个基础RNA模型，展现了提取有价值生物学见解的能力，并且可以通过迁移学习以进一步优化，从而能够应用于多种下游任务[32]。而scTab则采用专门针对表格数据的TabNet构架，并结合特征注意机制，从而优化了训练效率，能够精准识别对细胞类型标注最为重要的基因[33]。这样的设计在处理scRNA-seq得到的基因表达矩阵时具有显著优势，提升了模型可解释性，且避免了对不相关特征过多计算所产生的负担[33]。

基于RNA-seq或scRNA-seq数据的RNA模型可以系统性地构建RNA序列、结构、功能之间的联系，在单个细胞水平上为细胞生物学和疾病机制的探索提供更多的视角，在一些难以完成的实验任务中展现出巨大的潜力，其可以通过准确的推理构建连接疾病风险、遗传变异与翻译调控机制的桥梁，并提供疾病特异性和组织特异性的新见解。这种计算驱动的工具将显著加速功能基因组学研究和疾病基因的精准识别，为未来个性化医学的发展奠定基础。随着RNA数据与AI的深度融合及多模态模型的整合，研究者得以更准确地将疾病风险、遗传变异与翻译调控等分子机制相联系，为功能基因组学、疾病靶点筛选、个性化医学的未来发展奠定坚实基础。

2.3 DNA的分析、生成与DNA模型 DNA、RNA和蛋白质是生命的3大重要生物大分子。虽然它们都至关重要，但由于DNA作为遗传信息的主要载体，是构成细胞生命基因组的基本单位，可转录为RNA并翻译为蛋白质，结构的复杂性和功能调控的多层次性使得DNA模型比RNA模型和蛋白质模型在深度学习中更具挑战性[34-35]。深度学习技术的突破性进展，尤其是生成式AI的进步，使得AI模型能够完成对基因组的阅读、理解和生成等全方位的任务，基于LLM的DNA模型已成为连接计算生物学与AI领域的重要桥梁。DNABERT是一种基于BERT构架改进的DNA序列分析模型，其采用单编码器的Transformer构架，专为处理生物序列数据设计[13]。DNABERT将自然语言处理中的Transformer技术应用于DNA序列，通过K-mer分词和预训练策略来捕捉生物序列中的复杂模式。它可以被微调用于许多序列分析任务，例如启动子区域和转录因子结合位点的预测、可变剪切位点的识别、变异位点的检测等任务[13]。

作为DNA语言模型，Omni-DNA[36]采用Transformer自回归构架，为适应不同计算资源的需求，Omni-DNA的参数从2 000万至10亿不等。通常生物序列模型需要针对单一任务进行微调，而Omni-DNA可针对多任务进行微调，在Nucleotide Transformer[37]和GB benchmarks[38]两个基准数据集的测试中，Omni-DNA模型在26个基因组任务中表现出色，并在其中18个任务上取得了当前最先进（SOTA）的性能[36]。同时，研究者还设计了DNA2Function（将DNA序列映射至功能文本描述）和Needle-in-DNA（将DNA序列映射至图像）两个复杂的基因组任务，进一步验证Omni-DNA在跨模态领域的适用性[36]。Evo是基于StripedHyena构架的长文本（最长可达131 kb）的DNA基础模型，能够在小型基因组上实现预测和生成任务。Evo拥有70亿个参数，采用单核苷酸分辨率的编码方式，使用含有3 000亿个核苷酸的原核生物线粒体基因组作为预训练数据[35]。Evo主要围绕生物学两大核心特征——中心法则的多模态性和进化过程的多尺度性，完成从分子到基因组尺度的预测和生成任务，其不仅能够跨越DNA、RNA和蛋白质模态进行零样本功能预测，还可以生成功能性的CRISPR-Cas分子复合物和IS200/IS605转座子系统[35]。在Evo的基础上，Evo2进一步提升了模型的训练数据量和参数规模，其使用除真核生物病毒外的9.3万亿个核苷酸进行训练，有两个版本的参数规模（70亿和400亿），上下文长度可达100万kb[39]。超长的上下文使得该模型可进行长距离序列的分析任务和真核生物完整基因的生成任务[39]。

总之，DNA基础模型，如DNABERT、Omni-DNA和Evo系列等，通过使用序列进行大量的训练，能够理解序列并掌握其潜在的功能和结构，并可通过微调进一步提高其在特定任务中的表现。DNA模型的飞跃式发展包含了对RNA与蛋白质序列、结构和功能的全面理解，在变异解读、疾病预测、个性化医疗、药物研发、合成生物学方面有着巨大的应用前景，正迅速发展成为生物学研究的强大工具。此外，DNA模型的开放性和可解释性也日益受到重视，许多研

究团队致力于开发开源工具和框架，使更多研究人员能够使用和改进现有模型。未来，随着计算能力的提升和数据收集能力的增强，生物序列语言模型有望在精准医疗、创新药物开发和生物工程等方面发挥更大的作用。这一领域的发展不仅推动了生物信息学和计算生物学的前沿研究，还为研究者们理解生命的基本运行机制提供了新的视角和工具。

2.4 微生物感染性疾病的标志物分析与预测模型 近年来，微生物群落与宿主健康之间复杂而深远的关系备受医学界关注。随着AI技术的快速发展，研究者们能够以更高效和精准的方式分析微生物数据，揭示其与疾病之间的潜在联系。这些技术克服了传统统计方法在处理微生物数据时遇到的诸多挑战，例如数据的高维稀疏性和结构复杂性等。同时，这些技术还可通过预测宿主的疾病状态与识别关键生物标志物，帮助研究者更深入地理解微生物与疾病之间的关系。PopPhy-CNN是一个开创性的预测模型，通过纳入微生物的系统发育树来预测宿主的表型特征[40]。该模型将微生物分类学中的系统发育结构嵌入到机器学习框架中，将元基因组数据处理为二维矩阵输入，以捕获微生物群落数据中的深层级联关系，这不仅能够实现宿主表型的准确预测，还能识别与疾病密切相关的微生物标志物[40]。另一种先进模型GDmicro则通过结合GCN和深度适配网络，以应对疾病状态分类中的两大难题：有限的标注数据及模型对跨研究数据集的泛化能力下降[41]。GDmicro采用半监督学习和领域适配技术，从受限的数据中建立更加强大的预测模型，因此，其在跨研究数据集中的分类能力显著提升[41]。例如，在炎症性肠病的分类中，GDmicro将曲线下面积（area under the curve，AUC）从0.783提高到0.949。此外，该模型还能精准识别影响疾病发生和发展的关键微生物标志物，并进一步阐明其对宿主健康的影响机制[41]。DeepMicro则利用深度表示学习方法，专注于减小微生物数据高维度和稀疏性带来的限制，通过自动编码器压缩高维数据到低维空间，从而在保证特征完整性的同时增强了分析的效率，使得后续的机器学习分类更为高效[42]。相较于传统分析方法，DeepMicro在多个数据集上表现出更卓越的疾病预测能力，同时缩短了模型训练和超参数优化所需的时间，显著加速了建模过程[42]。该方法展现了深度学习技术在微生物数据分析中的重要潜力，为健康状态与疾病预测提供了新的视野。

AI技术正在推动微生物与疾病的相关性研究进入全新阶段。这些前沿方法不仅完成了对宿主疾病状态的高效分类，还精确捕捉与疾病相关的微生物标志物，为理解疾病的潜在机制及开发个性化治疗方案提供了重要线索。此外，它们在技术层面的突破显著提升了模型的跨数据集应用能力，并加快了模型开发和训练效率，扩大了其在临床研究和实践中的应用场景。随着算法的持续迭代及微生物数据资源的不断丰富，这一领域的研究将进一步揭示微生物群落与宿主健康复杂而精细的网络结构，为疾病的预防、诊断和干预开辟更多新途径。

2.5 传染病防控与病毒预测模型和疫苗（抗体）设计模型 传染病的全球性暴发对公共卫生构成巨大挑战，病毒变异和抗原进化的预测则成为应对传染病的核心环节，AI技术的引入显著提升了传染病防控和疫苗研发的速度与精准性。MLAEP是基于深度学习预测病毒抗原进化的模型，其通过结合结构建模、对抗学习和遗传算法，构建病毒适应性预测构架，并利用计算驱动的演化分析探索潜在的抗原变化[43]。研究表明，MLAEP不仅可以准确预测病毒抗原进化轨迹，还对病毒免疫逃逸突变具有较好的预测能力，在新型冠状病毒演化研究中具有突破性意义，例如该模型可成功识别新型冠状病毒新突变毒株（XBB.1.5）及免疫抑制患者所感染病毒携带的新型突变，为疫苗设计提供宝贵参考[43]。E2VD则着眼于病毒变异驱动因子的预测，以统一框架应对多种病毒的复杂演化特性[44]。E2VD利用病毒进化特性先验知识，通过无结构限制的演化驱动设计，预测病毒高风险突变位点，洞察病毒演化趋势，捕捉病毒变异规律[44]。香港科技大学的研究人员开发出一种针对甲型流感病毒H3N2亚型的季节性预测框架，模拟世界卫生组织疫苗株选择流程，利用HA1基因序列和相关元数据作为输入，通过AdaBoost集成学习算法预测H3N2亚型抗原性变化[45]。该模型通过数据驱动方法解释了抗原变化的非线性特征，在缺乏数据的情况下仍然具有较高的预测能力[45]，从而可深化研究者对流感病毒抗原漂移机制的理解，强化流感疫苗设计的科学依据。该方法可用于辅助病毒红细胞凝集抑制试验、流感疫苗株推荐、全球流感监测及公共卫生管理[45]。

在治疗性抗体开发中，筛选和工程化具有高亲和力及其他药物特性的分子是必不可少的。传统治疗性抗体的优化通常需要先针对在哺乳动物细胞中表达的全长抗体进行低通量筛选，这一过程的效率较低且资源消耗大。通过引入AI技术进行预测和筛选，可以显著提高筛选速度和效率、减少资源消耗，

极大地增加了新抗体序列空间的探索度，并加速高效、药物型抗体的开发[46-47]。PALM-H3是一个LLM，为从头开始生成人工抗体重链互补决定区3提供了新的方法，显著减少了对天然抗体的依赖，与传统方法相比，其避开了从血清中分离特异性抗体这一耗时、耗资源的过程[48]。研究者还构建了高精度模型抗原-抗体结合预测器（A2binder），用于预测抗原表位序列与抗体序列的结合特异性和亲和力[48]。研究表明，PALM-H3生成的抗体对SARS-CoV-2抗原（包括新兴的XBB变体）表现出高结合亲和力和强中和能力，为抗体设计的基本原理提供了关键技术支持[48]。

2.6 生物影像分析 生物影像分析涵盖了多种成像模态，包括光学显微镜、电子显微镜、分子成像技术（荧光、生物发光、正电子发射断层扫描等）、结构成像技术（X射线、CT、MRI、超声波等）及各种光谱技术。多样化的成像技术产生了具有不同特征的生物图像数据，对分析方法提出了独特挑战。AI尤其是深度学习，为复杂生物图像的处理和信息提取提供了强大工具。随着计算能力的提升和算法的进步，AI在生物影像分析中的应用呈现出多元化和普及化的特点。近期，多模态生物成像，即整合多种成像技术以克服单一方法局限性的方法，被认为是值得关注的发展趋势。多模态整合虽然实施复杂，但在空间生物学理解方面具有革命性潜力。

在模型方面，传统的CNN和新兴的Transformer构架共同推动着生物影像分析的发展。以乳腺超声图像分类为例，研究表明，基于Transformer构架（ViT）[49-50]的ViT-B/32模型能够获得较高的准确率（86.7%）和AUC（0.95），优于多种基于CNN的模型（ResNet50的准确率、AUC分别为85.3%、0.94；VGG16的准确率、AUC分别为82%、0.92）[51]。此外，基于混合构架如U-Net Transformer和CNN-Transformer的混合模型，可通过结合CNN的局部特征提取能力和Transformer构架的全局注意力机制，在3D医学图像分割等任务中展现出卓越性能[51-53]。

基于生物图像数据的深度学习模型正逐步朝着平台化与整合化的发展方向迈进。Piximi作为一种无需编程的、基于网络的解决方案迅速流行，其可提供分类器、标注工具、分割工具和测量功能，涵盖了完整的生物影像分析工作流程[54]。Piximi的客户端构架能确保数据隐私，适用于处理敏感的医学图像或网络连接受限的环境[54]。其他工具如Cellpose、DeepImageJ及CellProfiler等，也通过整合深度学习能力，提高生物影像分析的可访问性和易用性，使得原本复杂的医学影像识别、分割和分析变得更加简便且高效，为生物影像分析注入了新的活力[54]。

3 AI在基础医学数据分析中的发展趋势和挑战

3.1 现有AI技术的局限性 医疗数据的敏感性体现在其包含了个人基因信息、疾病历史、用药记录等多种隐私信息，任何数据共享过程都可能成为隐私泄露的风险点[55]。医疗机构必须确保患者信息安全，防止非法买卖和泄露。这些因素共同导致医疗数据共享的现实困境，限制了AI模型的训练数据来源和应用范围[56-57]。医疗数据的获取与使用受到严格的伦理和隐私保护法规的制约，这对AI模型的开发与应用构成了显著障碍。高质量且规模足够的数据是深度学习模型构建成功的前提，但生物医学数据的独特性使得其在数据质量和样本规模方面具有明显短板[58]。标注数据的稀缺限制了模型的泛化能力，降低其适应性。现有研究数据通常集中于特定物种或疾病，忽略了其他生物学背景，导致预测结果欠缺广泛适用性。基础医学研究中样本数量少但特征维度高的问题较为普遍，这种不平衡可能导致深度学习模型过拟合[59]。领域内缺乏标准化的基准评测数据集，不同模型采用的比较方法也有较大差异，阻碍了不同AI方法间性能的一致性比较。

深度学习模型复杂性极高，这表现在模型预测结果的可解释性不足，而可解释性在医学领域至关重要。例如，模型能够发现某些基因与疾病之间的关联，但无法解释其因果关系。模型在诊断和治疗领域的可信度需要基于透明的解释性，这直接关系到临床应用的安全性与实用性。尽管已有方法可以提高模型解释性[60]，但仍然难以解决深度学习模型的复杂机制和因果关系问题。同时，语言模型容易生成看似合理但非真实的内容。这种现象引发了研究者对可靠性的担忧[61]。深度学习模型在处理基因组数据长序列时，其效率和硬件容量均面临瓶颈，基因组序列长度常达数千至数万碱基对，这导致模型处理长序列时显存需求激增[62]。训练大型模型需要密集的计算资源，这使得某些研究机构难以负担。同时，随着模型参数的增加，推理成本也明显增加，对硬件的要求愈发严苛，这使模型的部署和广泛应用受到巨大的限制。

3.2 未来发展趋势 尽管当前AI在医学领域的应用面临诸多技术性挑战，其发展潜力依然十分广阔。未来的技术突破将围绕数据标准化、多模态整合、新型构架设计和智能体技术等方面展开，从而推动AI与医学研究的深度融合。

数据标准化和开放性平台的建设将成为促进AI技术深入应用的重要基础。医学研究的多样性和复杂性要求数据处理必须具备高度的规范性。从数据的收集到预处理，再到分析的每一个环节，建立统一的标准将为不同研究团队间的数据共享与技术协作提供保障。同时，开放性平台的构建也将成为推动AI发展的重要驱动力。越来越多的医学数据已被托管在如 Hugging Face 和 ModelScope 等开放平台上，这种共享机制有助于加快生物医学模型的开发和优化过程，让研究者能够以更低的门槛参与其中，从而推动整个领域的技术发展和应用普及。在基础医学研究领域，AI带来的风险要求加强AI伦理基础教育，并进行社区规范加强和约束[57]。

多模态整合在未来的研究中也将占据重要位置。单一数据源的局限性已逐渐显现，而通过AI模型对基因组学、蛋白质组学、临床记录及医学影像等不同数据模态的联合分析，则有望揭示复杂生物系统的深层规律。为实现这一目标，未来的AI构架需要解决数据模态间的异质性问题，例如探索类似跨模态 Transformer 的技术，将不同类型的数据进行无缝连接。此外，自监督学习的技术进步也将进一步推动多模态数据的整合，使AI能够在多维数据中挖掘潜在规律，实现从早期诊断到治疗方案设计的全方位优化。

在技术构架层面，新型模型构架设计将打破现有的技术瓶颈，成为基础医学领域创新的重要推动力。针对生物学中常见的长序列数据分析需求，适配长序列处理的新型模型构架也将得到更多的研发投入。未来的设计还可能进一步简化推理方式，从繁琐的多步骤流程转向端到端的处理模式，从而提升下游任务的效率。同时，模型将更加注重轻量化的结构和高效的训练，例如 LoRA、QLoRA 及低精度模型。这种优化不仅可以使大型模型在资源有限的环境中运行，还为AI开辟了更多细分场景的应用可能，从而成为降低计算成本的潜在解决方案。

智能体技术正在成为AI推动基础医学领域创新的下一步重点，这类技术将AI的应用由工具性扩展到主动参与科学探索[63]。未来的智能体技术不仅能够自动进行数据分析，还可以主动提出科学假设，设计实验方案，从而加速科研发现的过程[64-65]。此外，通过模拟跨学科专家间的合作，智能体技术能够为复杂课题提供综合性解决方案，并优化问题解决的整体流程。不仅如此，随着知识不断更新，新型智能体技术还将具备动态学习和自我演化的能力，从而保持在生物医学研究中的前沿性与适应性。人机协作也将是未来研究形式的重要特点，AI与研究者之间的高度协作将结合人类的直觉与AI的强大计算能力，在提升诊断和预测准确性的同时，为结果的解释性增添更多科学深度。

4 总 结

AI技术在基础医学数据分析中呈现出巨大的潜力，为医学研究和临床应用带来了深远的影响。从早期的专家规则系统到如今以深度学习主导的跨模态模型，AI的发展有效推动了基础医学的前沿探索。监督学习、无监督学习、深度神经网络及LLM等技术，在疾病诊断、基因组分析、蛋白质结构预测、单细胞转录组研究、微生物组学等领域展现出了卓越能力，尤其是生物序列语言模型、多模态整合和智能体技术的出现，为复杂生物医学问题的解决开辟了新路径。尽管如此，AI在基础医学数据分析中的应用仍面临多方面挑战，包括数据隐私保护、获取高质量样本的限制、模型可解释性不足及计算资源需求增长等。这些问题限制了AI技术在医学领域的全面推广，需要持续迭代与优化，为精准医疗、医学研究和健康管理提供新的动力与视野，从而助力基础医学迈向新的发展阶段。

参 考 文 献

[1] DeFranco DB, Sowa G. The importance of basic science and research training for the next generation of physicians and physician scientists[J]. Mol Endocrinol, 2014, 28(12): 1919-1921.

[2] Leung CK. Biomedical informatics: state of the art, challenges, and opportunities[J]. Bio Med Informatics, 2024, 4(1): 89-97.

[3] An Q, Rahman S, Zhou JW, et al. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges[J]. Sensors (Basel), 2023, 23(9): 4178.

[4] Khan N, Das A. Deep learning models for tumor detection and segmentation in medical image analysis: a comprehensive review of ResNet, U-Net, DETR, and Inception Variants[J]. IP J Diagn Pathol Oncol, 2024, 9(4): 195-206.

[5] van Melle W. MYCIN: a knowledge-based consultation program for infectious disease diagnosis[J]. Int J Man Mach Stud, 1978, 10(3): 313-322.

[6] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks[J]. Commun ACM, 2017, 60(6): 84-90.
[7] Tang W, Sun J, Wang S, et al. Review of AlexNet for medical image classification[DB/OL]. (2023-11-15)[2025-04-24]. https://arxiv.org/abs/2311.08655.
[8] Bishop CM. Pattern recognition and machine learning[M]. New York: Springer, 2007.
[9] Jiang Y, Luo J, Huang D, et al. Machine learning advances in microbiology: a review of methods and applications[J]. Front Microbiol, 2022, 13: 925454.
[10] Montúfar G, Pascanu R, Cho K, et al. On the number of linear regions of deep neural networks[C]//Ghahramani Z, Welling M, Cortes C, et al. Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. Cambridge: MIT Press, 2014: 2924-2932.
[11] Raghu M, Poole B, Kleinberg J, et al. On the expressive power of deep neural networks[C]//Precup d, Teh YW. Proceedings of the 34th International Conference on Machine Learning-Volume 70. Cambridge: MIT Press, 2017: 2847-2854.
[12] Min S, Lee B, Yoon S. Deep learning in bioinformatics[J]. Brief Bioinform, 2017, 18(5): 851-869.
[13] Ji YR, Zhou ZH, Liu H, et al. DNABERT: pre-trained bidirectional encoder representations from Transformers model for DNA-language in genome[J]. Bioinformatics, 2021, 37(15): 2112-2120.
[14] Zhou Z, Ji Y, Li W, et al. DNABERT-2: efficient foundation model and benchmark for multi-species genome[C/OL]// ICLR. Proceedings of ICLR 2024. Vienna: ICLR, 2024: 3857. (2024-01-16)[2025-04-24]. https://openreview.net/pdf?id=oMLQB4EZE1.
[15] 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.
[16] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.
[17] Jumper J, Evans R, Pritzel A, et al. Applying and improving AlphaFold at CASP14[J]. Proteins, 2021, 89(12): 1711-1721.
[18] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.
[19] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024, 630(8016): 493-500.
[20] Hayes T, Rao R, Akin H, et al. Simulating 500 million years of evolution with a language model[J]. Science, 2025, 387(6736): 850-858.
[21] Lin Z, Akin H, Rao R, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction[DB/OL]. (2022-11-21)[2025-04-24]. https://www.biorxiv.org/content/10.1101/2022.07.20.500902v3.
[22] Hwang H, Jeon H, Yeo N, et al. Big data and deep learning for RNA biology[J]. Exp Mol Med, 2024, 56(6): 1293-1321.
[23] Kelley DR, Reshef YA, Bileschi M, et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks[J]. Genome Res, 2018, 28(5): 739-750.
[24] Zhou J, Theesfeld CL, Yao K, et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk[J]. Nat Genet, 2018, 50(8): 1171-1179.
[25] Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks[J]. Cell Rep, 2020, 31(7): 107663.
[26] Edfors F, Danielsson F, Hallström BM, et al. Gene-specific correlation of RNA and protein levels in human cells and tissues[J]. Mol Syst Biol, 2016, 12(10): 883.
[27] Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues[J]. PLoS Comput Biol, 2017, 13(5): e1005535.
[28] He J, Xiong L, Shi S, et al. Deep learning prediction of ribosome profiling with Translatomer reveals translational regulation and interprets disease variants[J]. Nat Mach Intell, 2024, 6(11): 1314-1329.
[29] Wang S, Sun ST, Zhang XY, et al. The evolution of single-cell RNA sequencing technology and application: progress and perspectives[J]. Int J Mol Sci, 2023, 24(3): 2943.
[30] Bhattachan P, Jeschke MG. Single-cell transcriptome analysis in health and disease[J]. Shock, 2024, 61(1): 19-27.
[31] Yang F, Wang WC, Wang F, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data[J]. Nat Mach Intell, 2022, 4(10): 852-866.
[32] Cui HT, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI[J]. Nat Methods, 2024, 21(8): 1470-1480.
[33] Fischer F, Fischer DS, Mukhin R, et al. scTab: scaling cross-tissue single-cell annotation models[J]. Nat Commun, 2024, 15(1): 6611.
[34] Benegas G, Ye C, Albors C, et al. Genomic language models: opportunities and challenges[J]. Trends Genet, 2025, 41(4): 286-302.
[35] Nguyen E, Poli M, Durrant MG, et al. Sequence modeling and design from molecular to genome scale with Evo[J]. Science (1979), 2024, 386(6723): eado9336.
[36] Li Z, Subasri V, Shen Y, et al. Omni-DNA: a unified genomic foundation model for cross-modal and multi-task learning[DB/OL]. (2025-02-05)[2025-04-24]. https://arxiv.org/pdf/2502.03499.
[37] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics[J]. Nat Methods, 2025, 22(2): 287-297.

[38] Grešová K, Martinek V, Čechák D, et al. Genomic benchmarks: a collection of datasets for genomic sequence classification[J]. BMC Genomic Data, 2023, 24(1): 25.

[39] Brixi G, Durrant MG, Ku J, et al. Genome modeling and design across all domains of life with Evo2[DB/OL]. (2025-02-21)[2025-04-24]. https://www.biorxiv.org/content/10.1101/2025.02.18.638918v1.

[40] Reiman D, Metwally AA, Sun J, et al. PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data[J]. IEEE J Biomed Health Inform, 2020, 24(10): 2993-3001.

[41] Liao HR, Shang JY, Sun YN. GDmicro: classifying host disease status with GCN and deep adaptation network based on the human gut microbiome data[J]. Bioinformatics, 2023, 39(12): btad747.

[42] Oh M, Zhang LQ. DeepMicro: deep representation learning for disease prediction based on microbiome data[J]. Sci Rep, 2020, 10(1): 6026.

[43] Han WK, Chen NN, Xu XZ, et al. Predicting the antigenic evolution of SARS-COV-2 with deep learning[J]. Nat Commun, 2023, 14(1): 3478.

[44] Nie Z, Liu X, Chen J, et al. E2VD: a unified evolution-driven framework for virus variation drivers prediction[J/OL].(2024-07-17)[2025-04-24]. https://www.biorxiv.org/content/10.1101/2023.11.27.568815v3.

[45] Shah SAW, Palomar DP, Barr I, et al. Seasonal antigenic prediction of influenza A H3N2 using machine learning[J]. Nat Commun, 2024, 15(1): 3833.

[46] Makowski EK, Kinnunen PC, Huang J, et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space[J]. Nat Commun, 2022, 13(1): 3788.

[47] Mason DM, Friedensohn S, Weber CR, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence *via* deep learning[J]. Nat Biomed Eng, 2021, 5(6): 600-612.

[48] He HH, He B, Guan L, et al. *De novo* generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model[J]. Nat Commun, 2024, 15(1): 6867.

[49] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[DB/OL].(2020-10-22)[2025-04-24].https://arxiv.org/pdf/2010.11929v2.

[50] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[DB/OL]. (2021-03-24)[2025-04-24]. https://arxiv.org/pdf/2103.13413.

[51] Saad M, Ullah M, Afridi H, et al. BreastUS: vision transformer for breast cancer classification using breast ultrasound images[C]//IEEE. Proceeding of 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). Dijon: IEEE, 2022: 246-253.

[52] Hatamizadeh A, Yang D, Roth H, et al. UNETR: transformers for 3D medical image segmentation[DB/OL].(2021-03-18)[2025-04-24]. https://arxiv.org/pdf/2103.10504v1.

[53] Shaker A, Maaz M, Rasheed H, et al. UNETR++: delving into efficient and accurate 3D medical image segmentation[J]. IEEE Trans Med Imaging, 2024, 43(9): 3377-3390.

[54] Moser LM, Gogoberidze N, Papaleo A, et al. Piximi-an images to discovery web tool for bioimages and beyond[DB/OL]. (2024-06-04)[2025-04-24]. https://www.biorxiv.org/content/10.1101/2024.06.03.597232v2.

[55] Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges[J]. J Med Internet Res, 2024, 26: e53008.

[56] Wang C, Zhang J, Lassi N, et al. Privacy protection in using artificial intelligence for healthcare: Chinese regulation in comparative perspective[J]. Healthcare (Basel), 2022, 10(10): 1878.

[57] Prunkl C. AI meets biology: a call for community governance[J]. Nat Methods, 2024, 21(8): 1407-1408.

[58] Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine[J]. J R Soc Interface, 2018, 15(141): 20170387.

[59] Chen C, Weiss ST, Liu YY.Graph convolutional network-based feature selection for high-dimensional and low-sample size data[J]. Bioinformatics, 2023, 39(4): btad135.

[60] Salh AM, Raisi-Estabragh Z, Galazzo IB, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME[DB/OL].(2024-06-27)[2025-04-24]. https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202400304.

[61] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[DB/OL]. (2023-11-09)[2025-04-24]. https://arxiv.org/pdf/2311.05232v2.

[62] Chen R, Han W, Zhang H, ET AL. An embarrassingly simple approach to enhance transformer performance in genomic selection for crop breeding[C]//Larson K. Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. Jeju: IJCAI, 2024: 7215-7223.

[63] Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare[J]. Nat Biomed Eng, 2025, 9(4): 432-438.

[64] Gao SH, Fang A, Huang YP, et al. Empowering biomedical discovery with AI agents[J]. Cell, 2024, 187(22): 6125-6151.

[65] Zou J, Topol EJ. The rise of agentic AI teammates in medicine[J]. Lancet, 2025, 405(10477): 457.

（收稿日期：2025-03-07　修回日期：2025-05-11）