*submit stencil*
http://www.scichina.com

《中国科学》杂志社
SCIENCE IN CHINA PRESS

# 人工智能时代的生物信息学展望：从基础解码到智能生物系统的跨领域应用

梁　丹　张　佳　梁雨朝　左永春*

内蒙古大学生命科学学院，呼和浩特 010000

收稿日期 2026-03-23　接受发表 2026-05-21

摘要　人工智能(Artificial Intelligence, AI)技术的快速发展为生物信息学提供了新的思路. 针对传统方法在动态生物过程解析、复杂高维数据处理以及多源异质信息整合等方面存在的局限，相关 AI 方法展现出明显优势. 本文系统回顾了人工智能在分子生物学、基因与细胞调控网络、生物医学数据库及跨学科场景中的关键应用进展，整理并介绍了 80 余种人工智能工具与平台和 50 个常用生物信息学数据库，分析了各类方法在实际任务中的表现、适用场景与不足. 随着算法的持续发展，人工智能在生物信息学中的应用正逐步从数据分析向系统层面拓展. 相关内容可为未来的研究提供重要的理论基础.

**Abstract:** The rapid advancement of artificial intelligence (AI) has opened new avenues for bioinformatics research. While traditional methods face inherent limitations in dissecting dynamic biological processes, handling complex high-dimensional data, and integrating multi-source heterogeneous information, AI-driven approaches offer distinct advantages in addressing these challenges. This review provides a systematic overview of the key progress in applying AI to molecular biology, gene and cellular regulatory networks, biomedical databases, and interdisciplinary applications. We collate and introduce over 80 AI-powered tools and platforms and 50 commonly used bioinformatics databases, and analyze their performance, applicable scenarios, and limitations in real-world tasks. As algorithms continue to evolve, the application of AI in bioinformatics is expanding beyond data analysis toward systematic, system-level biological modeling. The findings presented herein offer a solid theoretical foundation for future research in this field.

**关键词**　人工智能；生物信息学；分子生物学；调控网络；生物医学数据库

Key words: artificial intelligence; bioinformatics; molecular biology; regulatory network; biomedical database

生物信息学是生命科学与计算科学交叉发展的前沿领域，主要通过对复杂生物数据的计算分析，探索生命活动的内在机制. 传统分析手段在面对大规模、动态变化且来源多样的数据时，往往在处理效率

收稿日期: 2000—00—00; 接受日期: 2000—00—00
国家自然科学基金(批准号：62571279 和 62171241)、内蒙古自治区“英才兴蒙”工程创新团队项目(2025TEL25)的资助
† 共同第一作者
联系人, E—mail: yczuo@imu.edu.cn

与预测精度上存在显著局限性. 生物过程的高度动态特性、数据的高维度和复杂性特征，以及跨平台、跨组学数据间的异质性，都给多源信息整合和系统层面的分析带来了挑战[1].

随着生物信息学领域的发展，人工智能(Artificial Intelligence, AI)技术逐渐成为应对这些挑战的有力工具，AI 不仅可以显著提高生物数据的处理效率和分析精度，还能为生物系统动态建模和分子功能预测等核心任务提供新的研究思路，并在实际应用中展现出较高的技术价值[2].其中，最具代表性的是 DeepMind 团队研发的 AlphaFold 系列模型，其核心开发者凭借 2020 年首次推出的 AlphaFold2 模型在蛋白质结构预测领域的突出贡献，成功获得 2024 年的诺贝尔化学奖. AlphaFold2 模型有效提升了蛋白质三维结构预测的精准度，使其达到接近实验测定结果的水准，有效缓解了长期困扰生物学界的蛋白质折叠难题，也为后续药物研发设计、疾病发生机制探究等相关领域提供了重要的技术支撑[3]. 随着人工智能技术在生物领域的认可度不断提升，科研人员也开始进一步探索这项技术在更复杂生物建模任务中的应用空间，2025 年发布的 Evo 2[4]模型就是这类探索中的一项典型成果. Evo 2 依托多类生物基因组数据进行训练，不仅可以完成编码区与非编码区突变预测、自然基因组序列生成此类基础任务，还具备优异的零样本预测能力. 除此之外，Evo 2 还能实现表观基因组生成、染色质可及性调控等任务，支持精准的基因设计以及基因功能相关的生物特征识别，为基因组学研究和合成生物学的发展提供了关键的技术支撑.

为了更直观地体现人工智能在生物信息学领域的发展趋势，我们统计了近十年 PubMed 上的相关文献数量，结果如图 1 所示. 结果表明，相关研究持续增长，AI 已经成为推动生物信息学发展的重要力量.

本文系统梳理了近十年来人工智能在生物信息学中的多层次应用与研究进展，重点讨论了其在分子生物学、基因与细胞调控网络、跨学科应用等场景中的实际突破，旨在帮助研究者理解不同 AI 方法的适用场景. 此外，本文也关注现有模型在可解释性、泛化能力、数据偏差及伦理等方面存在的问题，展望人工智能与生命科学深度融合的未来方向，为生物信息学工具的进一步优化提供方向.

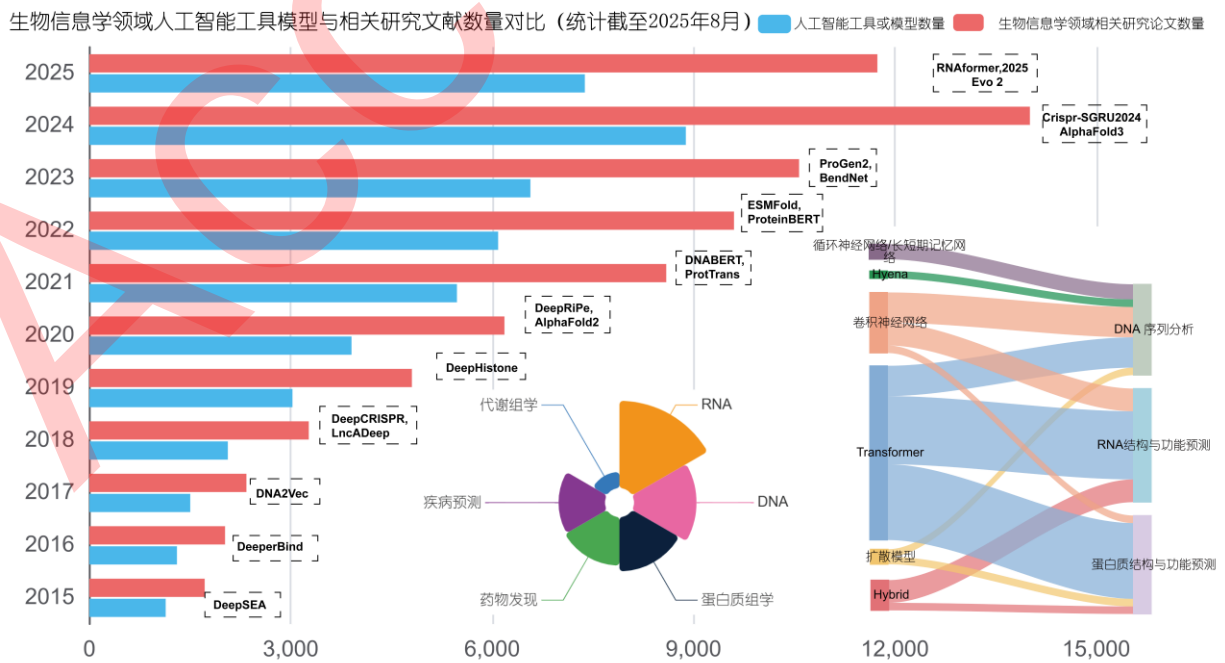


**图 1　人工智能在生物信息学中的发展趋势与应用结构.**

展示了 2015–2025 年间人工智能工具或模型数量（蓝色）与生物信息学领域相关研究论文数量（红色）的年度变化趋势，并以虚线框标注了具有代表性的里程碑模型（如 DeepSEA、DeepCRISPR、AlphaFold2 和 ESMFold）.

右下角的玫瑰图展示了人工智能方法在不同生物信息学研究领域（如 DNA、RNA、蛋白质、代谢组学及药物发现）中的应用分布，各扇区表示对应研究方向的相对比例. 同时，右侧的桑基图展示了不同人工智能模型架构，如变换器模型(Transformer)、卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)、长短期记忆网络(Long Short-Term Memory, LSTM)、扩散模型(Diffusion Model)及混合模型(Hybrid Model)与主要生物学研究任务（DNA、RNA 和蛋白质研究）之间的关联关系，其中连线宽度表示在本文纳入的代表性研究中对应关系的相对频率. 需要说明的是，由于部分研究涉及多种模型架构或研究任务，该图反映的是模型与任务关联的相对趋势，而非严格的文献计量统计结果. 图中统计数据基于 PubMed 数据库的关键词检索结果整理得到，用于反映整体趋势与相对分布.

**Figure 1　Development trends and application landscape of artificial intelligence in bioinformatics.**
This figure shows the annual trends in the number of AI tools or models (blue) and AI-related bioinformatics publications (red) from 2015 to 2025. Representative milestone models (e.g., DeepSEA, DeepCRISPR, AlphaFold2, and ESMFold) are highlighted with dashed boxes. The rose chart in the lower right illustrates the distribution of AI applications across major bioinformatics domains, including DNA, RNA, proteomics, metabolomics, and drug discovery, with each sector representing the relative proportion of studies in each area. Meanwhile, the Sankey diagram on the right depicts the relationships between different AI model architectures (e.g., Transformer, convolutional neural networks (CNN), recurrent neural networks/long short-term memory networks (RNN/LSTM), diffusion models, and hybrid models) and major biological research tasks (DNA, RNA, and protein studies). The width of each link indicates the relative frequency of these associations based on the representative studies included in this review. It should be noted that, as some studies involve multiple model architectures or research tasks, the diagram reflects general trends in model–task associations rather than a strict bibliometric analysis. The data presented in this figure were derived from keyword-based searches in the PubMed database and are intended to reflect overall trends and relative distributions.

## 1 AI 驱动的分子生物学研究

长期以来，序列分析与结构研究领域的传统方法一直依赖各类经典的计算手段，例如在基因注释和蛋白质家族识别等常规任务中发挥核心作用的基本局部比对搜索工具(Basic Local Alignment Search Tool, BLAST)等序列比对工具以及基于隐马尔可夫模型(Hidden Markov Model, HMM)的同源搜索方法[5, 6]，以及能够在原子尺度上刻画蛋白质构象演变及分子间的相互作用的分子动力学模拟方法[7].

面对海量的生物序列数据以及复杂的模式识别需求，近几年涌现的人工智能方法展现出了更为突出的性能. 以 AlphaFold 为例，它通过深度神经网络极大程度上提高了蛋白质结构预测的精度[3]；而引入了自注意力机制的 DNABERT 等模型，则擅长捕捉 DNA 序列中的长程依赖关系，在启动子预测等具体任务中，其性能通常优于早期的传统模型[8].

然而这并不意味着传统方法已经退出了历史舞台. 在涉及蛋白质构象的具体动态变化、配体结合路径等研究场景时，分子动力学模拟所能提供的时间维度信息和功能机制解释，依然是目前 AI 模型难以完全替代的，而在处理基因功能注释等依赖序列同源性的任务时，BLAST 和 HMM 也依然是常用方法. 因此当下的研究趋势绝非完全取代传统方法，而是要将人工智能的预测性能优势与传统生信方法的生物学解释力结合起来，从而实现进一步的效果提升.

当前，AI 正深刻变革分子生物学研究，在 DNA、RNA 和蛋白质等关键领域持续突破.从 DNA 的基因序列调控元件识别与基因预测[9]，到 RNA 的结构、功能及分子互作分析[10]，再到蛋白质的功能预测、结构解析[11]等，AI 借助深度学习、预训练模型等技术，为这些领域带来了更高效、精准的研究手段. 图 2 展现了 AI 在分子生物学各环节的技术工具与应用方向，共同推动了生物信息学从序列到功能研究的跨越式发展.

### 1.1 DNA 领域的 AI 应用

本节围绕序列层面及单一预测任务视角展开，包括 DNA 甲基化状态的预测工作，以及这类预测如何对基因表达产生影响、该过程如何建模等内容. 早期 DNA 分析受到算法与算力限制，大多依赖传统机器学习方法，研究人员通过手工提取的 k 长度子序列

(k-length subsequence, k-mer)频率、转录因子结合基序等特征，搭配支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest, RF)等简单模型完成分析. 但这类方法难以应对复杂基因组数据，泛化能力也较为有限[12].深度学习的出现提升了短序列分析的表现，比如 DeepD2V[13] 结合卷积神经网络(Convolutional Neural Networks, CNN)与双向长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)，并搭配词向量嵌入(Word2Vec)，优化了转录因子结合位点(Transcription factor binding sites, TFBS)的预测效果. 不过 CNN 在处理长序列时存在局限，难以捕捉序列间的远距离依赖，复杂场景下的分析能力仍受制约. 随着深度学习技术的不断发展，预训练模型有望成为突破传统方法瓶颈的切入点[14]. 研究者参考自然语言处理领域双向编码器表示(Bidirectional Encoder Representations from Transformers, BERT)、生成式预训练变换器(Generative Pre-trained Transformer, GPT)的思路，利用大规模无标注 DNA 序列完成预训练，构建了通用的 DNA 序列模型. 以 DNABERT-2[15]为例，它延续了 BERT 的预训练框架并优化网络结构，在超大规模无标注数据上完成预训练后，通过低秩适应(Low-Rank Adaptation, LoRA)实现高效微调，在多物种基因组分析、TFBS 预测、表观遗传标记预测等跨任务迁移场景中表现稳定.

在 DNA 甲基化建模方面，DeepMethyl[16]是一种基于堆叠去噪自编码器(Stacked Denoising Autoencoders, SdAs)的深度神经网络预测模型，该模型结合 DNA 序列信息与三维基因组拓扑信息，能够有效预测胞嘧啶-磷酸-鸟嘌呤二核苷酸(Cytosine-Phosphate-Guanine Dinucleotide, CpG)的甲基化状态，为后续探索癌症相关甲基化异常提供了计算工具基础. DeepPGD[17]进一步优化了网络结构，引入时序卷积网络(Temporal Convolutional Network, TCN)与 BiLSTM 机制，提升了对复杂序列上下文的建模能力，在不同物种数据集上均表现出稳定的高准确性. 此外，DeepMethyGene[18]模型基于残差网络，构建自适应递归卷积神经网络，能构建甲基化与基因表达的非线性映射关系，揭示了甲基化修饰如何影响转录活性，强化了表观遗传调控与功能基因表达之间的联系. DNA 领域还存在多个针对不同场景优化的 AI 工具，具体的核心特性、适用场景和关键指标对比详见表 1，可供研究人员作为模型选择参考.

综上所述，DNA 领域正在从早期的依赖人工特征工程方法，逐步转向能够从原始序列中直接学习特征表示的深度学习方法，因此能够更好地处理复杂调控元件识别这类任务.

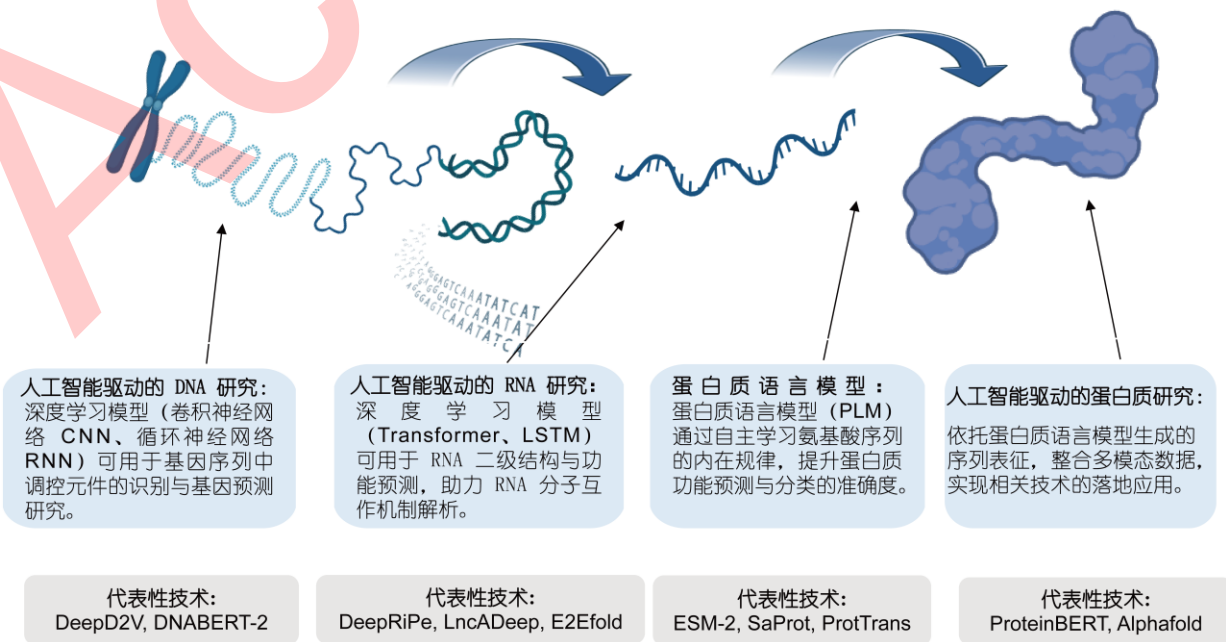

**图 2　人工智能在 DNA、RNA 和蛋白质序列分析中的应用流程.**

本图展示了人工智能在分子生物学不同环节中的应用，包括 DNA 序列分析、RNA 结构与功能预测以及蛋白质功能分析.基于 CNN、RNN、Transformer、LSTM 以及蛋白质语言模型(Protein Language Model, PLM)等深度学习模型的算法推动了基因调控元件识别、RNA 二级结构预测及蛋白质功能分类的发展.图中列出了各类代表性技术，体现了人工智能在序列生物学研究中的重要推动作用.

**Figure 2　AI application workflow in DNA, RNA, and protein sequence analysis.**

This figure illustrates the application of artificial intelligence across different stages of molecular biology, including DNA sequence analysis, RNA structure and function prediction, and protein function analysis. Deep learning approaches based on convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory networks (LSTM), Transformers, and protein language models (Protein Language Model, PLM) have facilitated the identification of gene regulatory elements, RNA secondary structure prediction, and protein function classification. Representative techniques are presented for each category, highlighting the important role of AI in sequence-based biological research.

**表 1 DNA 研究相关工具**

**Table1 Tools for DNA Research**

| 工具名称 | 年份 | 应用领域与算法特征 | 模型架构 | 优势 | 挑战与局限 |
|---|---|---|---|---|---|
| DANN[19] | 2015 | 基因变异致病性注释：多层 DNN(非线性、Dropout、动量) | 三层深度神经网络 + Sigmoid 输出层 | 显著提高非编码变异区分度(AUC +14% vs. SVM) | 特征解释性差；依赖 GPU/大样本 |
| DeepSEA[20] | 2015 | 非编码变异预测：深度学习用于染色质/多任务(转录因子结合、DNase I、组蛋白标记) | 深度卷积网络 | 单核苷酸敏感性；多任务学习增强预测 | 依赖于染色质数据；稀有变异预测的挑战 |
| DeeperBind[21] | 2016 | 转录因子-DNA 结合预测：LSTM-RNN + CNN | RNN + CNN | 模拟 DNA 位置信息；适应可变长度序列 | 在长范围依赖/多基序影响方面表现差 |
| DeepMethyl[16] | 2016 | DNA 甲基化预测：基因组(CpG)：堆叠去噪自编码器 + 基因组拓扑建模 | 堆叠去噪自编码器 + 基因组拓扑建模 | CpG 甲基化状态预测准确度高 | 依赖高质量染色质数据；解释性低 |
| DNA2Vec[22] | 2017 | DNA 序列嵌入：基于 Word2vec 的 k-mer 嵌入(余弦相似性) | 浅层双层神经网络 | 高效的 k-mer 表示；适用于长序列 | 在高维输入和慢训练时表现较差 |
| DNABERT[8] | 2021 | DNA 序列理解：预训练双向 Transformer(可微调) | 基于 Transformer 的 BERT 架构 | 高准确率(多物种)；对数据稀缺具有鲁棒性 | 预训练计算成本高；对特定任务适应性较弱 |
| DeepD2V[13] | 2021 | 转录因子结合位点识别：k-mer + word2vec + one-hot | CNN + BiLSTM | 整合四种序列表示(优于传统 one-hot 方法) | 高维输入需大量计算；对窗口大小敏感 |
| BendNet[23] | 2023 | DNA 序列柔性预测：多尺度 CNN + 胶囊网络 | 卷积神经网络+胶囊网络 | 可扩展的碱基分辨率 DNA 力学评估(覆盖 307 种物种) | 高维特征稀疏；泛化能力尚待验证 |
| HyenaDNA[24] | 2023 | 长序列基因组建模：长距离、隐式卷积 | Hyena 架构（解码器模型） | 处理长度达百万级 tokens 的序列；高效训练(低复杂度) | 准确率略低；依赖大规模上下文数据 |
| DNABERT-2[15] | 2023 | DNA 序列分析：Transformer(LoRA，快速注意力机制，Byte Pair Encoding) | Transformer 架构 | 高效分词(降低计算与内存开销) | 词汇表需调整；相较于 k-mer 模型，对超短序列(<70 bp)表现不佳 |
| DiscDiff[25] | 2024 | DNA 序列生成： LDM+ 吸收-逃逸后处理 | 潜在扩散模型 | 改进 DNA 生成质量；增强序列多样性与真实性 | 离散序列生成中存在舍入误差；计算成本高 |

| | | | | | |
|---|---|---|---|---|---|
| DeepPGD[17] | 2024 | DNA 甲基化分析：TCN + BiLSTM + 注意力机制(多尺度卷积) | BiLSTM + 时间卷积网络（复杂上下文建模） | 有效捕捉时间特征与长程 DNA 依赖关系 | 严重依赖高质量 DNA 甲基化数据；可解释性低 |
| dnaGrinder[26] | 2024 | 基因组建模：轻量级 Transformer 编码器(ME-BPE + 快速注意力机制 2 + ALiBi) | Transformer 编码器 + 多元素字节对编码 + 二代快速注意力机制 + ALiBi 位置编码 | 可处理 17k+ tokens(普通工作站) / 140k+ tokens(高性能 GPU)；轻量级(63.6M 参数)；在 30 项下游任务中达到 SOTA 性能 | ME-BPE 可能遗漏低频关键 token；部分任务预训练增益有限；超参数对小样本任务敏感 |
| Crispr-SGRU[27] | 2024 | sgRNA-DNA 脱靶预测：多尺度 Inception + 堆叠 BiGRU | Inception 模块 + 堆叠双向门控循环单元 | 精确预测脱靶活性(错配/插入缺失)；具有良好的鲁棒性 | 可解释性与可视化受限；对训练数据质量敏感 |
| Nucleotide Transformer[12] | 2025 | 基因组分析：Transformer 编码器(6-mer 分词，MLM 预训练于未注释的基因组数据) | Transformer 架构 | 支持高效微调、跨物种泛化与多任务学习 | 计算成本高；需大量训练数据 |
| DeepMethyGene[18] | 2025 | 基因表达预测：深度学习(DNA 甲基化数据输入) | 自适应残差递归卷积网络 | 预测性能高；自适应通道调整；缓解梯度消失；优化数据分布 | 在甲基化位点增多或远离 TSS 时预测优势减弱 |

注：表中总结了不同人工智能模型在 DNA 相关研究中的应用特点及性能表现. 由于部分研究涉及多种模型架构或任务类型，表中所列为主要方法及其典型应用.

缩写说明：DNN, 深度神经网络(Deep Neural Network); GRU, 门控循环单元(Gated Recurrent Unit); BiGRU, 双向门控循环单元(Bidirectional Gated Recurrent Unit); BPE, 字节对编码(Byte Pair Encoding); ME-BPE, 多嵌入字节对编码(Multi-embedding Byte Pair Encoding); MLM, 掩码语言模型(Masked Language Modeling); AUC, 曲线下面积(Area Under the Curve); DNase I, 脱氧核糖核酸酶 I (Deoxyribonuclease I); sgRNA, 单导 RNA(Single Guide RNA); ALiBi, 带线性偏置的注意力机制(Attention with Linear Biases); GPU, 图形处理单元(Graphics Processing Unit); SOTA，当前最优水平(State-of-the-Art).

### 1.2 RNA 领域的 AI 应用

RNA 分析领域的发展从早期主要依赖实验手段和基于规则的分析方法逐步转向与人工智能技术的深度结合. 这些新兴工具在 RNA 序列的结构预测、分子互作解析、非编码 RNA 功能注释以及 RNA 修饰等方向上发挥了关键作用.

人工智能使 RNA 结构预测效率与精度大幅提升，E2Efold[28]借助深度学习模型直接从 RNA 序列出发预测二级结构，显著提高了预测精度与效率. 在 RNA 分子互作解析的场景中，DeepRiPe[29]利用神经网络对交联免疫沉淀测序(Crosslinking Immunoprecipitation Sequencing, CLIP-seq)数据进行处理，能更准确地识别 RNA 结合蛋白的结合位点，帮助科研人员进一步理解 RNA 调控网络及其在疾病进程中的作用. 非编码核糖核酸(Non-coding Ribonucleic Acid, ncRNA)虽不参与蛋白质编码，但在基因表达调控、细胞分化、免疫应答以及疾病发生等多个生物过程中扮演重要角色[30]. 作为 ncRNA 家族的重要成员，微小核糖核酸(MicroRNA, miRNA)、长链非编码核糖核酸(Long Non-coding RNA, lncRNA)和环状核糖核酸(Circular RNA, circRNA)功能预测研究均受益于人工智能，如 LncADeep[31]可以通过序列特征预测 lncRNA 的编码潜力，再结合 lncRNA-蛋白互作预测和通路富集分析，间接完成功能推断.

RNA 分析离不开修饰研究，在 $N^6$-甲基腺苷($N^6$-methyladenosine, $m^6A$)修饰研究中，AI 展现出强大的应用潜力. m6Anet[32]是一种结合纳米孔信号特征与序列特征的深度学习模型，能捕捉 $m^6A$ 修饰的上下文依赖规律，进而揭示这类修饰在 mRNA 稳定性与翻译调控中的作用机制. 作为最早考虑组织间保守性的 AI 模型之一的 m6A-TCPred[33]则重点关注预测跨组织共享或组织特异性的 $m^6A$ 修饰位点，为 RNA 修饰的组织特异性研究开辟了新的思路. 此外，pum6a 模型[34]引入了正-未标记多实例学习(Positive-Unlabeled Multi-Instance Learning, PU-MIL)策略和注意力机制，在原始 RNA 测序数据中对 $m^6A$ 位点的检测准确性得到进一步提升，在低覆盖区域和异质性样本的分析中表现尤为突出.除文中提及的工具外，更多 RNA 深度学习工具的核心特性、适用场景及关键指标对比，详见表 2，可供研究人员参考.

从现有研究方法的发展来看，RNA 相关的各类

分析工作已逐渐脱离以往处理思路，从依靠既定规则或是常规统计模型，转向数据驱动的深度学习框架，这样的转变让 RNA 结构与功能的预测工作得以在更复杂的研究场景下达成更高的精准度，弥补了传统方法在复杂条件下适配性不足的问题.

**表 2 RNA 研究相关工具**

**Table 2　Tools for RNA Research**

| 工具名称 | 年份 | 应用领域与算法特征 | 模型架构 | 优势 | 挑战与局限 |
|---|---|---|---|---|---|
| CRISPRTarget[35] | 2013 | CRISPR-RNA 靶标预测：采用 BLASTn+PAM/seed 序列打分+交互式筛选模块 | 序列比对 + PAM/seed 打分模块 | 目标数据来源丰富；能识别与 CRISPR-RNA 相关的序列 | 需要大规模数据；在复杂环境中目标识别精度低 |
| DeepCRISPR[36] | 2018 | sgRNA 脱靶/靶标预测：采用去噪自编码器(预训练)+CNN 微调、数据增强、自助法(Bootstrap) | DCDNN + CNN | 融合表观遗传特征；目标预测框架统一；具有较强的泛化能力 | 训练过程复杂；特征可解释性有限 |
| LncADeep[31] | 2018 | LncRNA 识别/注释：采用 DBN(识别)+DNN(蛋白相互作用预测)+通路富集分析模块 | DBN + DNN +富集分析模块 | 能识别完整/部分转录本；支持跨物种预测；可注释蛋白相互作用 | 蛋白相互作用预测的可解释性有限；结果依赖注释质量 |
| E2Efold[28] | 2020 | RNA 二级结构预测：采用 Transformer(特征提取)+展开算法+后处理网络(结构约束) | Transformer +展开优化网络 | 能处理假结结构；端到端优化(无两阶段误差传播、无动态规划推理) | 结构约束设计复杂；在小数据集上拟合能力有限；可解释性需改进 |
| RNA-BERT[37] | 2020 | RNA 家族分类/结构预测：采用基于 BERT 的 RNA 预训练 | BERT 架构 | 通过有限数据完成 RNA 家族分类，适配不同长度序列，精度优、抗过拟合能力强 | 对大规模 RNA 数据处理速度慢；计算资源需求高 |
| DeepRiPe[29] | 2020 | RBP(RNA 结合蛋白)结合预测：采用多任务+多模态 CNN(序列/区域类型输入，结合序列位置上下文) | CNN + RNN (多任务学习) | 高效预测多种 RBP 结合位点；区域上下文提升性能；多任务学习共享特征以减少数据稀缺偏差 | 对低表达/低峰值数据敏感；CNN 第一层过滤部分基序；依赖高质量 CLIP-seq 数据 |
| RNA-FM[38] | 2022 | RNA 结构/功能建模：大型自监督 RNA 基础模型 | Transformer + 自监督学习 | 无标签数据预训练可提升 RNA 结构/功能预测性能 | 在功能性任务上的表现弱于结构性任务 |
| m6Anet[32] | 2022 | m6A 检测(直接 RNA 测序)：多实例学习框架 | 融合纳米孔（nanopore）信号与序列特征的深度网络 | 可适配纳米孔测序；可捕捉上下文依赖性 | 需要高质量的原始信号输入 |
| DRP-Score[39] | 2023 | RNA-蛋白复合物结构评估：基于深度学习的原生复合结构评分模型 | 基于深度学习的评分函数 | 高效、准确地预测 RNA-蛋白复合物结构；对多种数据集具有较强适应性 | 对结构解构变换的预测能力有限；需要大量物理模拟/实验数据 |
| DEMININ G[40] | 2024 | RNA 编辑/DNA 突变识别：基于深度学习与 DeepDDR(转录组数据分析) | DNN+迁移学习 | 能高效区分 A-to-I 编辑与 DNA 突变；适用于精确的转录组分析 | 需要跨物种训练集；对非人类数据集的适应性仍需改进 |
| RhoFold[10] | 2024 | RNA 三维结构预测：RNA 语言模型+多层自注意力+结构模块 | 端到端深度学习架构 | 可高效实现自动化的 3D 结构预测 | 大规模数据集需较高计算资源；模型对结构解构灵活性处理仍可改进 |
| RNA-former[41] | 2024 | RNA 二级结构预测：轻量级 Transformer 架构 | 轻量级 Transformer + RNA 特异性预训练 | 高效结构建模，直接模拟 RNA 配对矩阵，提升预测精度 | 模型深度受限于序列长度变化 |

| | | | | | |
|---|---|---|---|---|---|
| m6A-TCPred[33] | 2024 | 组织保守型 $m^6A$ 位点预测：基于机器学习 | 组织特异性建模 | 组织特异性建模；考虑组织间保守性，提高预测特异性 | 难以解释底层生物学机制 |
| NuFold[42] | 2025 | RNA 三级结构预测：基于深度神经网络的模型(核苷酸中心表示+糖环构象优化) | 端到端深度学习架构 | 端到端深度神经网络；通过多序列比对提升精度，适用于复杂 RNA 预测 | 训练数据集较小；对灵活末端和环区的建模仍有局限 |
| SAND-STORM + GARDN[43] | 2025 | RNA 分子设计/功能预测：结合 SANDSTORM(功能预测)与 GARDN(分子生成) | 神经网络 + GAN | 高效 RNA 序列/结构预测，支持多任务学习与实验设计 | 实验数据需求量高；难以处理极其复杂的序列结构 |
| pum6a[34] | 2025 | $m^6A$ 表观转录组图谱解码：基于直接 RNA 测序 | PU-MIL + 注意力机制 | 适用于低覆盖率/异质样本 | 依赖于正样本标注质量 |

注：表中总结了不同人工智能模型在 RNA 相关研究中的应用特点及性能表现. 由于部分研究涉及多种模型架构或任务类型，表中所列为主要方法及其典型应用.

缩写说明： DBN, 深度置信网络(Deep Belief Network); GAN, 生成对抗网络(Generative Adversarial Network); BLASTn，核苷酸序列比对工具(Basic Local Alignment Search Tool for Nucleotides); RBP, RNA 结合蛋白(RNA-Binding Protein, RBP); A-to-I, 腺苷到次黄嘌呤的 RNA 编辑(Adenosine-to-Inosine RNA Editing, A-to-I); PAM, 原间隔序列邻近基序(Protospacer Adjacent Motif); seed, 种子序列(Seed Sequence, seed).

### 1.3 蛋白质领域的 AI 应用

在进行蛋白质序列分析时，传统分析工具(如 BLAST)在应对长氨基酸序列以及大规模数据处理任务时，实际运行效率往往难以满足研究需求[44]. 传统的分析思路大多依托序列比对操作与保守区域靶向分析，不仅会占用大量计算资源，还会在面对结构复杂的蛋白质数据时大幅提升处理难度. 蛋白质语言模型(Protein Language Model, PLM)则可以通过自监督学习的方式，捕捉氨基酸序列在长期进化中形成的固有模式以及序列与功能之间的内在关联，构建通用的序列语义表征体系，为后续的蛋白质功能预测、空间结构分析提供扎实的序列层面支撑[45]. 例如 ESM-2[46]模型可以借助掩码语言建模的技术路径，精准抓取序列内部的保守特征与变异规律，自主完成氨基酸序列语法规则的学习过程，优化功能预测与序列分类任务的精准度，这种方式也能在一定程度上降低对前期实验验证结果的依赖. ProtTrans[47]模型将自然语言处理领域的成熟语言模型思路迁移至蛋白质序列分析场景，依托自监督学习模式完成模型训练，借助 Transformer 架构完成海量蛋白质序列数据的学习过程，从中提取有效的序列嵌入表示，进一步捕捉蛋白质疏水性、电荷分布这类基础物理化学性质，从而实现了无需多序列比对(Multiple Sequence Alignment, MSA)步骤的快速功能预测，可完成二级结构判定、亚细胞定位分析以及膜结合与水溶性分类等多项核心任务.

蛋白质三维结构预测一直是生物学研究领域的核心难点，这一方向的研究推进高度依赖 PDB(Protein Data Bank)数据库的数据[48]. 早期的结构预测工作，主要依靠 X 射线晶体学、核磁共振(Nuclear Magnetic Resonance, NMR)等实验技术[49]，这类方法虽然能获取原子级别的高精度结构数据，但实验成本偏高、测试周期偏长，直接限制了复杂蛋白质结构的深入研究进度. 后续随着硬件计算能力逐步提升，分子动力学模拟开始被应用于结构预测领域，这类方法依托辅助模型构建与能量优化力场(Assisted Model Building with Energy Refinement, AMBER)等经典力场模拟原子运动轨迹，进而推导蛋白质三维结构，在实验数据缺失或目标蛋白质结构过于复杂的场景下，能够发挥独特的作用[50]，只是仍然不足以支撑规模化批量预测场景.为了提升结构预测的整体效率，同源建模、折叠识别以及从头预测这类计算方法逐渐成为主流. 同源建模是发展最早、技术体系也相对成熟的路径，核心逻辑是依托序列相似性完成结构推断，不过这类方法仅适用于存在已知相似结构的蛋白质样本[51] 折叠识别，又称穿线法，通过与 PDB 数据库中的已知结构做比对，识别亲缘关系较远的同源蛋白，实际应用效果会受到数据库覆盖范围的直接限制[52]. 传统从头预测方法则结合了数据库现有知识与物理化学基本规则，拼接已知蛋白质片段，构建初始构象，再依靠统计打分函数完成近天然构象的筛选，这类方法更适合小分子蛋白质研究，但



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

是计算量庞大，最终的预测精准度也远低于实验测定结果[53].

21 世纪以后，蛋白质结构预测领域在 AI 技术的支持下取得了明显进步. 初代 AlphaFold[54]模型借助神经网络分析同源序列之间的协变信息，成功实现了蛋白质接触图的精准预测，在自由建模任务中表现优异. 升级后的 AlphaFold2[3]采用了基于 Transformer 优化的新型网络架构，把蛋白质结构预测转化为端到端的深度学习问题，利用氨基酸序列与 MSA 结果作为输入，完成蛋白质重原子三维坐标的预测，预测精度接近实验测定水平. 而最新的 AlphaFold3[55] 模型进一步引入扩散模型架构，实现了蛋白质、核酸以及小分子复合物结构的精准预测，重点优化了分子间相互作用的建模效果，同时通过改进多序列比对处理流程并降低计算成本.此外，RoseTTAFold[11]模型过整合序列信息、原子距离信息与空间结构信息，实现了多蛋白复合物结构预测的关键进展，进一步拓宽了人工智能技术在蛋白质结构建模领域的应用边界. 这类深度学习模型的落地应用，不仅支撑了蛋白质功能研究与靶向药物设计工作的推进，还为 PDB 数据库补充了大量高质量的预测结构数据，逐步形成实验验证与 AI 预测相互支撑、协同推进的研究模式，带动生物功能解析与疾病治疗方案开发领域的稳步发展.

在组蛋白修饰预测中，DeepHistone[56]采用密集连接卷积神经网络(Dense CNN)对修饰位点周围的序列上下文与染色质可及性数据进行建模，识别出不同类型的组蛋白标记所对应的染色质状态. 该方法可有效区分启动子、增强子等功能区域，并为染色质的开放性、活性调控和转录状态提供可解释的序列特征，推动了表观遗传图谱的构建与解读.更多蛋白质领域的相关工具和详细比较，详见表 3.

总体而言，蛋白质研究领域的相关方法同样经历了从早期依赖模板匹配或是基础物理模型到以深度学习为核心的端到端建模模式的转变，针对结构预测、功能分析以及蛋白质相互作用建模这类不同核心任务，各自有着明确的侧重方向，适用场景也存在明显区分.

**表 3　蛋白质研究相关工具**

**Table 3　Tools for Protein Research**

| 工具名称 | 年份 | 应用领域与算法特征 | 模型架构 | 优势 | 挑战与局限 |
|---|---|---|---|---|---|
| DeepHistone[56] | 2019 | 组蛋白修饰预测：基于深度学习 | 密集卷积神经网络(Dense CNN) | 高效捕捉复杂的 DNA 组学特征；整合多序列数据，提升组蛋白修饰预测准确度 | 依赖高质量 ChIP-seq/DNase-seq 数据；染色质可及性定量过程简单；跨表观基因组的预测性能低于单表观基因组模型 |
| AlphaFold2[3] | 2020 | 蛋白质结构预测：多模态数据(多序列比对 MSA、成对表示)+Transformer(Evoformer 模块、循环校正机制) | Transformer 架构，端到端三维坐标预测 | 实现单链蛋白原子级精度的结构预测 | 对高质量 MSA 数据的依赖性强；预测多蛋白复合物、与小分子/核酸结合的复杂结构以及蛋白质动态构象的能力有限 |
| ProtTrans[47] | 2021 | 蛋白质序列分析：基于自然语言处理架构(BERT、T5、XLNet)进行自监督预训练 | 包含基于经典自然语言处理架构(如 BERT、T5、XLNet)改造的蛋白语言模型变体 | 可从单一序列中提取蛋白质结构与功能特征；无需构建 MSA，显著降低计算负担；适用于无同源序列的蛋白质预测任务 | 需要高预训练资源(TPU / GPU 集群)；受位置编码与分段策略约束，对全长蛋白建模能力有限 |
| RoseTTAFold[11] | 2021 | 蛋白质结构与相互作用预测：采用三轨网络结构，整合多维序列–结构信息 | 多维序列–结构信息融合 | 能准确预测蛋白单体结构及相互作用；计算资源需求较低，适用范围更广 | 依赖高质量 MSA 数据；对于超高分子量蛋白复合物的预测精度与稳定性较低 |

| | | | | | |
|---|---|---|---|---|---|
| ESM-2[46] | 2022 | 蛋白质结构预测：大型语言模型(基于蛋白质序列)+下游 3D 结构预测模块 | Transformer 架构 | 单序列输入；结构预测精度高；速度快于 AlphaFold2；在低困惑度序列上表现出色 | 在复杂序列预测上仍面临挑战 |
| Protein-BERT[57] | 2022 | 蛋白质功能/序列理解：通过预训练(基因本体注释)+改进的 BERT(全局注意力+一维卷积) | 改进的 BERT 架构（全局注意力 + 一维卷积） | 有效处理长序列；适用于多种蛋白功能预测任务 | 在联合结构-功能预测方面仍有改进空间 |
| ProGen2[58] | 2023 | 蛋白质序列生成/适应性预测：自回归 Transformer 模型 | 自回归 Transformer 模型 | 能生成可自然折叠的蛋白质序列；具有较高的零样本适应度预测准确率 | 需要大规模计算资源；蛋白质适应性预测的部分难题仍未解决 |
| SaProt[59] | 2023 | 蛋白质语言建模/功能预测：基于结构感知词汇的 ESM-2 Transformer 模型 | Transformer 架构（基于 ESM-2 650M 骨干网络） | 引入结构信息增强的语言模型，有效提升蛋白质功能预测能力 | 训练数据依赖结构信息；在不同蛋白功能任务中泛化性受限 |
| ProstT5[60] | 2024 | 蛋白质序列-结构双模态建模：T5 编码-解码架构(Foldseek 3Di 结构编码) | T5 编码-解码架构 | 可生成结构兼容的蛋白序列(或反向预测结构)；优化蛋白质折叠任务 | 需要大量计算资源；依赖外部 3Di→3D 坐标转换工具；存在类别不平衡问题 |
| DeepLoc 2.1[61] | 2024 | 膜蛋白类型与亚细胞定位预测：Transformer + 注意力机制模型 | Transformer + 注意力机制 | 能准确预测膜蛋白类型，并可扩展到多种亚细胞定位场景 | 依赖高质量实验标注数据；对外周膜蛋白预测的敏感度较低 |
| DPLM-2[62] | 2024 | 序列-结构联合建模：扩散模型+ Transformer(基于氨基酸序列与 3D 结构数据的离散扩散建模) | 扩散模型 + Transformer 架构 | 同时生成结构与序列；优化结构-序列一致性 | 面临结构学习挑战；需要大规模数据与计算资源支持 |
| AlphaFold3[55] | 2024 | 蛋白质结构预测：扩散模型(原子坐标预测)+简化的 Pairformer | 扩散模型（蛋白-配体/核酸复合物建模） | 实现分子复合物的原子级联合结构预测；可减少对 MSA 的依赖 | 动态构象预测能力不足；高质量数据依赖问题尚未完全解决 |
| xTrimoP-GLM[63] | 2025 | 蛋白质结构/生成：Transformer + MLM/GLM(自回归与自编码预训练，用于理解/生成任务) | Transformer 架构 + MLM/GLM | 同时优化理解与生成任务；突破现有模型性能边界 | 需要大量计算资源；生成任务可控性有待提升 |

注：表中总结了人工智能模型在蛋白质中的典型应用及其性能特点. 由于部分方法涉及多种模型结构或任务类型，表中所列为其主要架构及代表性应用.

缩写说明： GLM, 广义语言模型(Generalized Language Model); T5, 文本到文本迁移模型(Text-to-Text Transfer Transformer); XLNet，自回归预训练语言模型; ChIP-seq，染色质免疫沉淀测序(Chromatin Immunoprecipitation Sequencing); DNase-seq，脱氧核糖核酸酶高通量测序(DNase I Hypersensitive Sites Sequencing); TPU, 张量处理单元(Tensor Processing Unit).

### 1.4 分子对接与药物设计领域的 AI 应用

依托分子结构与相关修饰信息，人工智能可以直接应用于药物靶点识别和靶向分子设计环节，缩短从基础分子机制研究到实际临床应用的转化周期. 传统的药物筛选流程，主要依赖大批量的实验验证，耗时长且成本高，限制了创新药物的整体开发速度[64].



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

近几年，生成式深度学习模型逐步融入药物研发环节，让药物筛选、靶点分子对接以及抗体设计相关工作的执行效率，都得到了明显提升，预测精准度也随之优化.

小分子药物设计相关研究中，DiffDock[65]作为依托扩散模型搭建的人工智能分子对接工具，能够在更宽泛的三维空间范围内，完成小分子与目标蛋白质结合构象的全面搜索，同时增强了自身对分子柔性特征以及多样分子构象的适配能力，这一特性使其能够延伸应用到复杂疾病对应靶点的小分子药物设计工作中. 而在抗体工程研究领域，免疫球蛋白语言模型IgLM[66]通过对海量抗体可变区序列的深度学习，实现了可控制物种类型和链型的抗体序列生成，生成的序列在可开发性与类人化特征上都表现更优，能够有效降低后续高通量实验筛选环节的资源损耗，在抗体库设计这类核心工作中，也体现出了可观的实用价值.

人工智能的融入让原本偏向试错的传统药物研发的模式逐步转向依靠模型驱动的精准预测路径，加快了候选药物分子的发现速度，同时为精准医疗和个性化用药方案的落地提供了可行支撑[67]. 但从实际应用情况来看，这类 AI 技术依旧存在不少亟待解决的问题，针对大尺度分子结构变化、高柔性分子体系的建模效果尚且不足，面对多靶点联动的复杂作用机制难以适配，对药物在真实生理环境下的吸收、分布、代谢与清除(Absorption, Distribution, Metabolism and Excretion, ADME)过程的模拟建模也存在局限，除此之外，目前还没有搭建起覆盖分子生成到实验验证的完整统一评估体系[68]. 针对这类短板，后续研究可以尝试搭建整合分子生成、药代动力学预测以及生物通路分析的一体化平台，以此推动人工智能技术在药物研发全流程中更全面的落地应用.

## 1.5 代表性基础模型的统一基准性能比较

本文在梳理现有研究的基础上，还挑选了各组学领域中具有代表性的基础模型进行结构化定量对比. 在确定筛选标准时，我们优先考虑了模型在所属领域的引用情况及其影响力，同时要求其在公开基准数据集上具备可复现的性能记录并兼顾技术路径的多样性.

考虑到不同组学任务在预测目标、标签定义及评测标准上存在显著差异，我们在子领域内部汇总模型在公开数据集上的主要性能表现，并同步补充了模型参数规模、上下文长度、推理效率以及训练资源消耗等关键特征. 需要指出的是，由于不同模型最初针对的任务设定并不完全一致，且在本地部署所有模型存在技术难度，本文选择引用原论文或官方报告，这些数值更多是用于反映模型效率的整体趋势，而非在严格控制变量条件下的实验结果.

在表 4 中，我们按照 DNA、RNA 和蛋白质的顺序，整理了部分代表性模型的核心指标. 由于不同领域的评价体系存在明显差异，所以该表数值主要是为了反映出性能区间和技术走势，无法进行跨领域的直接数值比较. 但参数规模和预训练数据的提升不仅能够显著提高了各领域的模型精度，还能呈现出优化推理效率、减轻对 MSA 依赖的趋势.

**表 4 代表性基础模型在公开基准数据集上的性能比较**

**Table 4 Performance of Representative Foundation Models on Public Benchmarks**

| 领域 | 模型 | 任务类型 | 参数规模 | 预训练数据规模 | 上下文长度 | 基准数据集 | 主要性能指标 | 推理时间 |
|---|---|---|---|---|---|---|---|---|
| DNA | DeepSEA[20] | 多标签分类(919 维染色质特征预测) | ≈30M(论文未直接报告) | 521.6 Mb genomic regions(≈2.6M 200 bp bins；输入为 1000 bp) | 1000 bp | ENCODE + Roadmap Epigenomics(919 chromatin features) | 中位 AUC: TF 0.958 / DHS 0.923 / HM 0.856 | 未报告 |
| | DNABERT-2[15] | 掩码语言模型预训练 + 下游基因组序列任务预测 | 117M | 32.49B nt(135 物种) | 700 bp(最大输入长度；有效上下文长度 600 bp，含两侧 padding) | GUE(36 个数据集，9 个任务) | GUE 平均得分 66.8；在 28 个数据集中取得 8 个最佳结果 | 未报告 |



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Hyena DNA[24] | 自回归预训练 + 下游序列分类任务 | 0.4M–6.6M(取决于模型/上下文长度) | 人类(*Homo sapiens*)参考基因组；最长上下文模型训练约 2T tokens(约 4 周) | 1k–1M tokens | Genomic 基准数据集；Nucleotide Transformer 基准数据集；DeepSEA；物种分类 | Genomic 基准数据集: 7/8 tasks SOTA (最高提升≈20%);<br>Nucleotide Transformer 基准数据集: 12/18 tasks SOTA;<br>物种分类: 99.5% (5-way, 1M context) | 1M tokens 序列推理时，较同硬件下 Transformer 模型快 160×(测试硬件：A100 GPU) |
| | Nucleotide Transformer[12] | 掩码语言模型预训练 + 多任务基因组功能预测 | 2.5B | ≈300B tokens；来源包括人类参考基因组、3,202 个人类基因组及 850 个物种基因组 | 6 kb (v1 版本)；12 kb (v2 版本，2025 年更新) | 18 个公开基因组基准任务 (GENCODE、ENCODE、EPD 等) | 基准 18 任务：平均 MCC≈0.755；剪接预测：PR-AUC=0.98；染色质特征：AUC 与 DeepSEA 性能相当 | 单 GPU 约 15 min 完成微调 |
| RNA | RNA-FM[38] | 自监督 RNA 语言模型 + 下游结构与功能预测 | 未报告总参数量(12 层 Transformer，hidden size=640，20 heads) | 23.7M RNA 序列 | 1024 tokens(最大输入长度) | Archive II 600；TS0(来自 bpRNA-1m 划分) | F1 (跨数据集)较 LinearFold 提升约 30%，较 SPOT-RNA 提升约 20%;<br>低冗余集较 UFold 提升约 4% | 未报告 |
| | NuFold[42] | RNA 三级结构预测 | 未公开具体规模 | 2860 条 PDB RNA 结构 + 11,101 条自蒸馏 RNA 序列(来自 bpRNA-1m) | 未报告 | RNA-Puzzles；独立 PDB RNA 结构测试集 | 平均 RMSD≈6.67 Å | 未报告 |
| | RhoFold[10] | 单链 RNA 3D 结构预测(端到端) | 未公开具体规模 | RNA-FM 预训练：23.7M RNA 序列；<br>3D 训练：PDB 结构 5,583 条(782 clusters)+ 自蒸馏数据 27,732 条 | 支持输入序列长度：16–256 nt(蒸馏训练数据匹配该范围) | RNA-Puzzles；CASP15 RNA targets；PDB 结构测试集 | RNA-Puzzles：平均 RMSD=4.02 Å；平均 TM=0.57;<br>CASP15：平均 RMSD 优于大多数参赛方法;<br>新 PDB 集：平均 RMSD=7.74 Å | 单样本 ≈0.14 s(无需 MSA 搜索) |
| Protein | AlphaFold2[3] | 蛋白质 3D 结构预测(MSA 依赖) | ≈93M(Evoformer trunk；完整模型参数量更大) | BFD(≈2.2B sequences)+ UniRef90 + MGnify + Uniclust30 + PDB70；PDB 结构监督 | 依赖 MSA(多序列比对输入) | CASP14 | CASP14 中位 GDT_TS = 92.4 | 推理时间主要受 MSA 搜索影响(分钟级至更长) |
| | AlphaFold3[55] | 生物分子复合体结构预测(蛋白/核酸/配体等) | 未公开具体规模 | 较 Alphafold2 新增 PDB 结构数据库(蛋白、核酸及配体复合体结构) | 支持多链、多分子复合体建模 | CASP15；RNA-Puzzles；蛋白-配体与蛋白-蛋白相互作用基准 | 蛋白-配体构象预测和蛋白-蛋白界面预测精度性能明显提升 | 未报告 (MSA 依赖性显著降低) |
| | ESMFold[46] | 单序列蛋白 3D 结构预测 | ESM-2(8M – 15B 参数；ESMFold 使用 15B 模型) | UniRef50(≈65M sequences) | 支持长序列输入(基于相对位置编码) | CAMEO；CASP14 FM | CAMEO 中位 TM≈0.83；CASP14 FM TM≈0.68 | 14.2 s / 384 aa(V100) |
| | Ro- | 蛋白质三维 | 未公开具体 | PDB 蛋白结构数 | 约 260 aa(训 | CASP14； | CASP14：平均 TM-score | RTX2080 |



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| seTTAFold[11] | 结构预测(序列→结构) | 规模 | 据库 + MSA 序列信息 | 练时使用裁剪片段) | CAMEO | 接近 AlphaFold2; CAMEO：在 medium/hard targets 中优于其他服务器方法 | GPU 约 10 min(<400 aa 蛋白) |

注：训练规模与模型开放性信息整理自原始论文及官方项目页面，不同实现环境下所需计算资源可能存在一定差异. 由于不同研究报告的硬件环境存在差异，部分工作未统一报告显存或内存占用指标，因此本文主要以推理时间作为效率参考指标.

缩写说明： PR-AUC, 精确率-召回率曲线下面积(Precision–Recall Area Under the Curve); MCC, 马修斯相关系数(Matthews Correlation Coefficient); RMSD, 均方根偏差(Root-Mean-Square Deviation); TM-score, 模板建模评分(Template Modeling Score); GDT_TS, 全局距离测试总分(Global Distance Test Total Score); nt，核苷酸(Nucleotide); ENCODE, DNA 元件百科全书项目(Encyclopedia of DNA Elements).

实际上，除了性能指标外，计算资源消耗以及实验的可重复性在当前的 AI 模型应用中同样是不容忽视的挑战. 近几年涌现的大量基础模型，往往要在训练阶段动用大规模的 GPU 或 TPU 集群，这种对算力的强依赖虽然换取了性能上的进展，却也让模型的复现与落地部署难度上升.

除此之外，代码的开源程度以及预训练权重是否能够直接获取，也直接决定了研究结果能否被同行有效复现. 为了应对这些问题，部分研究开始尝试通过提供 Docker 容器化环境或者标准化的依赖配置，来减轻实验复现时的技术负担. 考虑到这些实际因素，我们在表 5 中选择了部分代表模型，并对其训练规模、代码开放性以及权重的可用状态等信息进行了整理与对比.

**表 5 代表性基础模型的计算资源需求与可重复性比较**

**Table 5 Computational Resource Requirements and Reproducibility Comparison of Representative Foundation Models**

| 领域 | 模型 | 训练规模 | 代码开放性 | 预训练权重 |
|---|---|---|---|---|
| DNA | DNABERT-2[15] | 8×RTX2080Ti GPU，预训练约 14 天 | GitHub 开源 | 已公开(HuggingFace 提供预训练模型) |
| | Nucleotide Transformer[12] | Cambridge-1 超算：128×A100 GPU，训练约 28 天 | GitHub 开源 | 已公开(HuggingFace 提供模型权重) |
| RNA | RNA-FM[38] | 8×A100 80GB GPU，预训练约 1 个月 | GitHub 开源 | 已公开 |
| | RhoFold[10] | 未公开具体计算资源；训练数据：5583 条 PDB RNA 结构(782 clusters)+ 27,732 条自蒸馏数据 | GitHub 开源 | 已公开 |
| Protein | AlphaFold3[55] | TPU 集群训练(规模未公开) | GitHub 开源 | 已公开（需申请） |
| | ESMFold[46] | 大规模 GPU 集群训练（未公开具体 GPU 数量；基于 ESM-2 15B 参数模型微调） | GitHub 开源 | 已公开 |

注：训练规模信息来源于原论文. 部分模型报告计算资源规模(GPU/TPU)，部分模型仅报告训练数据规模. 部分模型还提供 Docker 或容器化运行环境，以进一步提高实验复现性和部署便利性.

## 2 AI 驱动的调控网络研究

### 2.1 AI 模型架构及其在生物序列与分子建模中的应用

随着深度学习方法的发展，生物序列分析领域出现了多种新的模型架构[69]. 其中，Transformer、图神经网络(Graph Neural Networks, GNN)以及扩散模型(Diffusion Models)是近年来应用较为广泛的几类代表性架构.

Transformer 模型是利用自注意力机制(self-attention)直接建立序列内部不同位置之间的关联. 相比于传统循环神经网络的逐步递归计算方式，Transformer 能够同步关注序列中的所有位置，这对于捕捉长距离依赖关系至关重要，因此这类模型近几年在基因调控区域识别、大规模蛋白质建模以及基因组预训练模型的构建中被广泛应用.

而针对分子结构或生物调控网络这种天然具备复杂空间关系的数据，图神经网络能够提供更为恰当



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

的建模方式. GNN 的核心原理是依靠节点之间的信息传递完成整个图结构特征表示的学习. 以消息传递神经网络(Message Passing Neural Networks, MPNN)为例，这类模型可以将多种图模型整合到统一的消息传递框架中，节点通过接收邻居节点传递的特征信息并不断更新自身状态，从而实现对整个图结构细节的逐步学习[70]. 图卷积网络(Graph Convolutional Network, GCN)则是通过对邻接矩阵归一化，使得各类特征能够在图拓扑上实现平滑传播[71]. 在目前的生信研究中，这类模型已广泛应用于分子性质的定量预测、蛋白质三维建模以及基因调控网络的深度分析等多项核心任务场景.

扩散模型是当前生成式人工智能领域的焦点，扩散模型可以划分为前向扩散与反向生成两个阶段，前向扩散过程对原始数据逐步添加噪声，使其转化为纯噪声，反向生成过程再逐步去除噪声，从而使模型具备从随机噪声中生成符合训练数据分布样本的能力[72]，其核心价值依赖于去噪函数的拟合精度，即如何从无序状态中找回原始的结构脉络. 正是由于这种机制在处理复杂数据分布时展现出的灵活性，扩散模型在蛋白质结构设计、小分子生成以及 RNA 预测等前沿任务中，正逐渐显现出其在分子设计领域的独特优势.

除了上述提到的各类模型架构外，近几年还有一些新兴的深度学习方法逐渐成为常用研究方法. 其中，大语言模型(Large Language Models, s)在医学文本的深度理解与知识提取方面展现出了强大的语义捕捉能力. LLM 依托大规模生物医学文献语料进行预训练，能够积累了丰富的领域语义信息. 例如 PubMedBERT[73]、BioGPT[74]、Galactica[75]和 BioMedLM[76]这类模型就是得益于在大规模 PubMed 语料库中的训练，从而在关系抽取、专业问答及文档分类等任务中取得了较为理想的表现.

大规模预训练模型在蛋白质结构研究中也展现出优异的性能. ESM[46]等蛋白质语言模型通过在海量蛋白质序列数据上进行训练，能够从序列本身信息中挖掘出那些隐含的结构与功能属性，从而实现对蛋白质结构的预测. AlphaFold 系列模型则将生物分子结构的预测推向新阶段，最新版本的 AlphaFold3 在前期版本的基础上对网络结构做了优化，并引入了扩散生成模块，能够对包含蛋白质、核酸、小分子配体以及离子在内的多种生物分子复合物实现预测，实现了较大的精度提升[55].这类模型已不再局限于单一序列建模，而是逐步整合序列、结构与功能，向多模态统一建模方式发展.

强化学习(Reinforcement Learning, RL)也正逐步应用于分子设计与优化的实际任务中. 强化学习方法将分子的生成或结构修饰过程看作是一个复杂的序列决策问题，从而让模型学会在庞大的化学空间里搜索具有特定性质的分子. 一些研究尝试利用循环神经网络与强化学习相结合，实现了基于简化分子线性输入规范表征的分子从头设计，并通过奖励函数来引导模型生成符合预期目标性质的候选分子[77]. 后续提出的 MolDQN 等方案则进一步利用深度 Q 网络对分子结构的修改路径进行策略学习，从而让分子性质的定向改进变得更加精准且可控[78].

除了单纯的预测与生成任务以外，人工智能方法也在逐步向实验设计以及主动学习等更具实操性的方向发展. 相关策略大多会参考模型的不确定性、候选样本能带来的信息增益或目标性质的优化空间进行筛选，优先选择最有信息价值的样本进入下一轮实验验证，进而在模型预测、实验筛选和数据回流之间构建可循环的迭代闭环. 通过这种循环，蛋白质工程、药物筛选以及合成生物学等领域的实验效率能够得到显著提升，降低试错成本[79, 80].

不同模型架构在生物信息学任务中已形成了比较清晰的分工. 这些模型架构很大程度上依赖于具体的任务类型而发挥优势. 例如 CNN 在转录因子结合位点识别这类捕捉局部序列模式的任务中表现尚可，但局部感受野的存在限制了 CNN 对长程依赖关系的建模. 因此 Transformer 模型被引入来解决此类问题，能够有效捕捉远距离调控关系，提高了在基因调控区域识别、蛋白质结构预测等相关任务中的性能[3, 8]. 扩散模型则在这一基础上进一步拓展至生成式任务，在蛋白质结构设计、小分子对接等任务中展现出优异的性能[65]. 各类算法在面对不同生物问题时有着明显的功能区分，彼此之间相互补足. 根据任务类型选择合适的算法，有利于充分发挥不同算法的优势，更好地解决各类生物学问题.

### 2.2 基因与表观遗传调控网络领域的 AI 应用

基于上述各类深度学习架构，人工智能方法在基



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

因调控与表观遗传调控网络的建模与解析中的应用日益广泛.与前文侧重单一分子层面的预测工作不同，本节会将关注点偏向基因与表观遗传调控网络的系统层面建模. 包括 DNA 甲基化在调控网络中所发挥的作用以及多组学数据的整合分析. 这类 AI 方法的应用，也为我们深入剖析复杂的生物调控逻辑，理解细胞功能与疾病的发生的内在机制，提供了更多切实可行的研究思路.

基因调控网络(Gene Regulatory Network, GRN)能够揭示转录因子、非编码 RNA 等与靶基因的相互作用，体现基因表达的调控规律，对细胞分化与稳态维持具有重要作用[81].传统的 GRN 推理方法大多依赖表达相关性和实验验证，推理效率较低，且难以捕捉其中复杂的非线性关系与细胞特异性[82].而人工智能技术凭借强大的建模能力，已成为 GRN 推理的重要工具，推动细胞功能与疾病机制研究[83].

基于图结构建模的相关方法会将基因抽象为图中的节点，将基因之间的调控关系抽象为边，以此搭建出一个复杂的拓扑图. 借助节点之间的信息传播机制，这类模型可以有效捕捉到网络中隐藏的拓扑特征与基因间的交互模式. DeepRIG[84]借助图自编码器整合单细胞 RNA 测序数据，在学习调控因子与基因表达模式的基础上，实现了细胞类型与关键调控网络的精准识别. 针对细胞特异性调控元件图谱的构建工作，SCRIPro[85]选取转录组与表观组数据作为分析基础，通过建模染色质可及性和基因表达之间的内在关联，让 GRN 推断的精准度与结果解释性都得到了一定提升. 在发育过程这类复杂研究场景中，时序动态建模方法的引入，能够捕捉到基因调控活动的变化趋势，也让这类场景下的分析思路更具针对性. TRENDY[86]模型通过双 Transformer 模块对协方差矩阵进行改进，不仅让 GRN 的准确性得到了提升，整体性能得到提升，同时还提高了生物机制层面的解释性与模型的泛化能力.DGRNS[87]则采用滑动窗口的方式提取时序相关特征，结合 GRU 与 CNN 两类模型联合优化 GRN 分类流程，这一设计思路有利于改善单细胞数据普遍存在的稀疏性问题.

表观遗传学揭示了基因表达的调控并非完全依赖基因序列本身，而是可以通过 DNA 甲基化、组蛋白修饰、染色质结构变化以及非编码 RNA 调控等多种常见机制实现，这类表观修饰对于细胞命运决定、各类疾病的发生发展过程都有着不可替代的作用[88]. 人工智能模型可以将输出的预测结果转化为表观遗传调控网络里的核心数据节点，为整合各类调控因子搭建了基础框架.例如，DeepCpG[89]融合卷积神经网络与 GRU 结构，实现了对单细胞甲基化状态的精准预判. CpGPT[90]模型依托生成式预训练 Transformer 架构做出了优化，让这类建模工作在不同样本与数据场景下的适配性得到了明显提升. 除了针对单一局部调控机制的分析，人工智能模型同样具备解析多类调控因子之间复杂关联的能力，ChromNet[91]以染色质免疫共沉淀测序(Chromatin Immunoprecipitation sequencing, ChIP-seq)测序数据为核心分析基础，结合条件依赖关系与分组图形模型开展运算，有效缓解了数据冗余问题，能够较为精准地完成染色质调控网络的重构工作，让相关分析结论更贴合实际的调控机制.

结合现有研究进展来看，AI 技术确实提升了基因及表观遗传调控网络的建模相关领域的研究进度，但是从实际应用与研究深度来看依旧存在若干不足之处. 最为突出的就是多模态表观标记的整合难题，这类数据本身带有高维稀疏的特性，常规处理手段很难实现高效整合，与此同时，现有模型大多难以捕捉调控过程中存在的非对称性特征，针对单细胞动态调控的分析尚且缺少成熟的高效研究框架，模型泛用性偏弱的问题也始终存在，再加上模型内部运行机制的可解释性不足，都在一定程度上限制了相关研究的进一步落地. 依赖多组学数据的不断积累和配套的人工智能分析方法优化改进，后续的调控网络建模工作将逐步达成高分辨率、强泛化能力的建模目标，也能进一步优化模型的生物可解释性，推动基础研究与临床应用实现更深层次的融合发展.

### 2.3 细胞通讯与因果推断领域的 AI 应用

细胞间相互作用(Cell–cell interactions, CCIs)是多细胞生物中细胞通过多种分子和膜结构，实现信号传递、黏附、代谢物交换等，激活信号通路、协调基因表达以推动细胞功能的过程[92].细胞间通讯(Cell–cell communication, CCC)特指 CCIs 中细胞间通过信号分子介导的信号传递过程，配体-受体相互作用(Ligand–receptor interactions, LRI)是其核心分子机制[93]. 传统 CCIs 研究受限于细胞异质性、空间信息



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

缺失及实验与计算结果难以匹配等问题. 近年来，结合单细胞空间组学与 AI 算法的新工具不断涌现，如 NICHES[94]通过构建单细胞分辨率配体-受体表达矩阵，保留细胞对间异质性，精准描绘局部组织通讯谱图；Tensor-cell2cell[95]利用张量分解对多场景通讯数据进行解析，揭示信号通路空间分布及动态变化；COMMOT[96]则结合集体最优传输理论与空间约束，实现物理意义上的通讯建模.这些 AI 方法通过图建模、张量分解等策略，显著提升了细胞通讯的空间分辨率和生物可解释性，但仍面临细胞注释误差和空间定位精度等挑战.

随着细胞通讯网络解析的深入，阐明信号传递对下游基因调控及细胞命运的影响成为关键，因果推断方法(Causal Inference)由此成为揭示其功能后果的关键工具.它旨在识别变量间的因果关系，超越相关性分析，在疾病机制和基因调控网络解析中意义重大. AI 方法为高维、多模态组学数据的因果推断带来突破，如 CIMLA[97]将 SHAP 特征归因与局部治疗效应结合，量化单细胞尺度的基因调控强度；FlowSig[98]通过图形因果建模与条件独立性检验，构建静态变量网络，实现细胞间信号流因果关系的高精度识别；Celcomen[99]通过将空间转录组数据建模为带有拉格朗日乘子约束的可因果识别的图神经网络，实现了细胞内与细胞间基因调控关系的因果解缠与空间反事实推断；NME[100]及其衍生方法基于“连续因果关系”框架，支持稳态与时序数据的有向因果推断，且具备捕捉非线性关联的能力.

综上所述，AI 在细胞通讯与因果推断中已展现重要价值，但仍存在局限，如细胞通讯分析受制于细胞注释误差与空间分辨率，因果推断难以充分融合动态时序与空间背景；部分模型过度依赖多标注数据，对信号下游调控及细胞命运机制的解析有待深化等. 未来，可以通过研发融合时空信息的 AI 模型，突破细胞注释、分辨率及因果推断时空融合局限，降低模型对标注数据依赖，深化信号调控与细胞命运机制解析，推动系统生物学与精准医学的发展.

### 2.4 单细胞与空间组学的人工智能解析方法

单细胞测序与空间组学技术的不断发展推动着生物信息学分析从过去的群体平均层面逐步深入到单细胞与空间分辨率的精细层面[101]. 传统分析方法如聚类、降维与统计模型，已在细胞异质性解析和组织空间结构研究中得到广泛应用. 然而面对高维稀疏数据、连续细胞状态转换以及复杂空间邻近关系等问题时，传统方法仍存在局限[102]. 近年来，单细胞与空间转录组数据分析开始大量引入在特征学习、非线性关系拟合与多模态信息的融合上拥有独特优势的 AI 方法. 在细胞类型注释、发育轨迹推断以及空间结构域识别等关键任务中，相关模型均展现出良好的应用潜力. 与此同时，人类细胞图谱(Human Cell Atlas, HCA)[103]、智人转录组图谱(Tabula Sapiens)[104]、人类生物分子图谱计划(Human Bio-Molecular Atlas Program, HuBMAP)[105]等规模公共数据库不断完善，也为人工智能算法的训练与评估提供了可靠的数据支撑.

单细胞转录组数据分析的两个关键任务分别是细胞类型注释与细胞轨迹推断. 传统注释手段大多依靠细胞聚类结果，结合已知标记基因进行人工判定，这类方法高度依赖实验人员的个人经验，且注释效果容易受参考图谱完善程度的影响. 深度学习与预训练模型则可以从大量表达谱数据中自动学习更泛化的细胞特征，减少人工标记带来的误差. 例如，自监督预训练模型 scFoundation[106]利用大量公共单细胞数据进行预训练，能够捕捉跨组织、跨数据集的共享表达特征，在跨样本、甚至跨物种的细胞类型注释任务中表现出更强的稳定性与泛化性. 此外，scANVI[107]和 CellTypist[108]等深度学习工具借助大规模参考图谱实现了自动化注释，在实际分析中大幅降低了对人工筛选标记基因的依赖.

细胞轨迹推断领域的传统伪时序分析方法，大多会先假定低维表达空间具有连续性，再根据细胞表达模式的相似性对细胞进行排序，进而描述细胞分化或状态转换的完整过程. 但面对复杂的发育进程或是疾病进展相关研究时，这种连续流形假设很难精准捕捉到多分支、非线性的细胞状态转变规律. 目前已有不少研究尝试借助图学习、动力学建模这类技术，去解析细胞状态之间复杂的非线性关联. RNA velocity 相关方法如 scVelo 通过对剪接与未剪接转录本的动态变化趋势进行建模，能够直接推断出细胞状态的转变方向，不仅可以精准重建多分支细胞分化轨迹，还可以定位发育与疾病进展研究中潜在的关键调控节点[109, 110].



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

空间转录组学分析更加依赖于将基因表达与空间位置信息进行联合建模. 在细胞分割与空间定位任务中, 传统的图像分割或规则类方法在低分辨率样本、结构复杂的组织中往往效果一般. 近年来, 深度学习方法开始被广泛用于空间转录组信息的解析, 例如借助卷积神经网络提取组织形态特征, 再结合图神经网络或注意力机制建模细胞间的空间邻域关系, 实现表达信息与空间结构的联合学习. 这类方法可以同时整合组织形态、局部空间上下文与基因表达三类信息, 进而实现更精细的细胞边界识别与细胞空间定位. 例如在 HuBMAP[105]等包含精细组织结构标注的数据集上训练后得到的模型能够在空间转录组数据中获得更可靠的细胞定位与组织形态解析结果.

在空间域识别任务中, 传统聚类方法大多会根据基因表达的相似性对空间位置分组, 却往往容易忽略细胞或测序点之间存在的空间邻近关系, 这就导致其最终难以真实反映组织的空间结构域与微环境分布. 基于图神经网络的人工智能模型, 恰好能弥补这一不足, 通过将基因表达模式与空间拓扑结构结合, 模型能实现对组织内功能区域及微环境特征的精准识别. SpaGCN[111]和 STAGATE[112]这类代表性方法, 先构建空间邻接图, 再利用图卷积或图注意力机制聚合邻域信息, 在多个空间转录组数据集的测试中, 均展现出了十分出色的空间区域识别表现.

然而, 在单细胞与空间组学相关研究以及跨物种功能预测、罕见疾病相关分析等场景中, 大量存在样本量有限、标注数据不足等问题. 为了解决这类问题, 研究人员提出迁移学习与小样本学习的思路. 迁移学习通过先在海量无标注数据上完成预训练, 让模型捕捉到具有普适性的生物序列特征, 再把这些已经学到的特征迁移到具体下游任务中做进一步微调. 已有研究在 2.5 亿余条蛋白质序列上开展无监督预训练, 随着训练数据规模的逐步扩大, 模型逐渐掌握了与蛋白质结构、功能紧密相关的关键信息, 后续应用到多种预测任务时, 仅需简单微调就能取得较为理想的效果[113]. 这种大规模预训练的思路为跨物种功能预测这类标注数据匮乏的研究提供了有效的解决方案.

小样本学习主要关注少量标注样本情况下的可靠分类与预测问题. 原型网络(Prototypical Networks)方法就是其中一类典型方法, 它先构建起一个特征嵌入空间, 让同类样本在这个空间里聚集, 同时把少量支持样本的均值当作类别原型, 可以实现对新类别的快速识别. 有了这一基础框架, 模型即使遇到训练阶段完全没接触过的类别, 也能通过这种思路顺利完成分类. 值得一提的是, 该方法还能进一步延伸到零样本学习场景, 利用类别元数据构建原型表示, 模型即便没有直接的训练样本, 也能实现有效的预测[114]. 这些方法为生物医学领域普遍存在的小样本问题以及跨物种功能预测、罕见疾病相关研究等标注数据稀缺等情况提供了切实可行的解决方案.

人工智能方法正推动单细胞与空间组学研究从传统的基因间的表达差异分析向细胞状态与组织空间结构向细胞状态与组织空间结构的更深层解析发展. 大规模单细胞测序技术的成熟加上空间组学数据库的持续积累, 为相关 AI 模型提供了关键的数据支撑, 既保障了模型训练与评估工作的顺利开展, 也为跨数据集、跨物种研究的推进创造了有利条件. 这些数据库也因此逐渐成为 AI 驱动生物信息学研究中不可或缺的核心基础平台. 结合这样的研究背景, 下一节将详细梳理当前 AI 驱动生物医学研究中常用的数据库资源, 以及它们在模型训练与实际应用中所发挥的具体支撑价值.

## 3 AI 驱动的生物医学数据库研究

生物医学数据库是人工智能应用于生物信息学研究中不可缺少的数据支撑. 随着高通量测序技术的不断发展与多组学相关研究的逐步深入, 海量研究数据借助各类公开数据库实现共享, 为生物信息学研究提供重要支撑[115, 116]. 这些资源不仅可以用于人工智能模型的预训练工作, 还能为模型评估、多模态整合以及知识约束等多个关键环节提供依据. 此外, 一部分平台搭建了标准化的基准数据集, 使不同算法间能够开展客观可靠的性能对比.

在实际开展的研究工作中, 许多人工智能模型的训练与验证流程都在很大程度上依赖于这些规模庞大的生物医学数据库. Protein Data Bank(PDB)[117]收录了海量经由实验验证的蛋白质结构数据, 这些数据为 AlphaFold[55, 118]及 RoseTTAFold[11]等经典的结构预测模型提供了关键数据支撑, 帮助模型学习氨基酸序列和三维构象之间复杂的映射关系. 通用蛋白质知识库(UniProt Knowledgebase, UniProtKB)[119]

中存储的海量蛋白质序列也为 ESM-2[46]这类蛋白质语言模型的自监督预训练提供了核心数据，让模型可以依托序列上下文的内在关联，学习蛋白质的结构与功能特征.

除了蛋白质相关研究领域，癌症基因组图谱(The Cancer Genome Atlas, TCGA)[120]等癌症多组学数据库在疾病预测、肿瘤亚型分类模型的搭建工作中也有着十分广泛的应用. 已有相关研究构建过一种深度学习多组学整合预后模型，通过对 TCGA 中的 RNA-seq、miRNA 表达以及 DNA 甲基化数据进行综合分析，在肝细胞癌患者的预后评估和生存亚型识别任务中均取得了较为显著的进展[121]. 在这类实际应用中，多组学数据通常需要先进行统一预处理与特征整合，再输入模型，达到对复杂疾病表型联合建模与预测的目的. 这些典型案例证明，生物医学数据库的价值远不止简单的数据平台，而是在 AI 模型的训练与验证全流程中作为核心知识载体而存在. 这些数据库不仅保障了大规模数据的供给，还通过不同的数据类型与标注方式影响模型的实际应用，是连接数据与算法的重要媒介.

总体来看，现有生物医学数据库大体上可以根据其对应的数据类型和研究应用分为多个类别，包括基因组学、转录组学、表观遗传学与非编码 RNA、蛋白质组学与结构生物学、药物与精准医学、临床与多模态数据、单细胞组学、空间组学以及知识图谱与通路等. 这些数据库从不同层面为 AI 模型在生物信息学领域及跨领域任务中提供了重要的数据基础. 相关数据库具体数据类型及其典型应用场景汇总详见表 6.

**表 6 常用数据库汇总**

**Table 6 Summary of Common Databases**

| 分类 | 数据库名称 | 数据类型 | AI 应用场景 | 最近更新时间 | 官方网址 |
|---|---|---|---|---|---|
| 基因组学(Genomics) | TCGA[120] | 33 种癌症类型，约 11k+样本(整合基因组、转录组和临床数据) | 癌症亚型分类、预后预测、泛癌分析 | 持续更新 | https://www.cancer.gov/ccg/research/genome-sequencing/tcga |
| | ICGC[122] | 全球多癌种基因组数据(约 225k 样本、70+癌种类型) | 遗传变异分析、跨人群癌症遗传学习 | 持续更新 | https://dcc.icgc.org/ |
| | gnomAD[123] | 125k WES+15.7k WGS(多族裔人群) | 变异致病性评估、群体遗传分析 | v4.1 (2024) | https://gnomad.broadinstitute.org/ |
| | IGSR[124] | 1000 Genomes SV 数据(长读长测序) | 群体遗传学研究、跨族群变异识别 | 持续更新 | https://www.internationalgenome.org |
| | dbSNP[125] | 11.4M+变异，包括 5 种模式生物：人类、小鼠(*Mus musculus*)、大鼠(*Rattus norvegicus*)、斑马鱼(*Danio rerio*)、果蝇(*Drosophila melanogaster*) | 变异致病性初步筛选、基础基因组研究 | 持续更新 | https://www.ncbi.nlm.nih.gov/snp |
| | Ensembl[126] | 多物种基因组注释(200+物种) | 基因注释、比较基因组学、跨物种学习 | Release 114 (2025) | https://www.ensembl.org/ |
| | COSMIC[127] | 癌症体细胞突变数据库 | 癌症驱动突变识别 | 持续更新 | https://cancer.sanger.ac.uk/cosmic |
| | OMIM[128] | 人类基因与遗传疾病数据库 | 遗传病基因研究 | 持续更新 | https://www.omim.org |
| | dbVar[129] | 结构变异数据库 | 结构变异分析 | 持续更新 | https://www.ncbi.nlm.nih.gov/dbvar |
| 转录组学(Transcriptomics) | GEO[130] | 高通量表达/杂交数据(多组织、多疾病) | 表达预测、基因功能注释、转录组分析 | 持续更新 | https://www.ncbi.nlm.nih.gov/geo/ |
| | CCLE[131] | 肿瘤细胞系基因组数据(表达、拷贝数等) | 药物筛选、基因型-表型关联分析 | 持续更新 | https://sites.broadinstitute.org/ccle |
| | GTEx[132] | 15k+ RNA-seq(49 种人类组织、健康个体) | 组织特异性表达分析、eQTL 分析 | v8 (latest public release) | https://www.gtexportal.org |
| | ArrayExpress[133] | 12k+ 杂交数据(35+物种) | 转录组整合分析、基因功能预测 | 持续更新 | https://www.ebi.ac.uk/biostudies/arrayexpress |
| | Expression Atlas[134] | 基因表达数据库 | 基因表达预测 | 持续更新 | https://www.ebi.ac.uk/gxa |
| 表观组学与非编码 RNA(Epigenomics & | ENCODE[135] | DNA 可及性、染色质修饰(1.6k+ 细胞系/组织) | 调控元件预测、RNA 调控网络、转录因子结合分析 | 持续更新 | https://www.encodeproject.org/ |
| | Roadmap Epigenomics[136] | 组蛋白修饰、DNA 甲基化、RNA 表达(127 种细胞类型，约 | 跨组织表观组分析、调控网络构建 | 2015 | https://www.ncbi.nlm.nih.gov/geo/roadmap/epigeno |



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

| | | | | | |
|---|---|---|---|---|---|
| Non-coding RNA) | ] | 10k 样本) | | | mics/ |
| | MethBank[137] | DNA 甲基化数据库 | 表观遗传调控研究 | 持续更新 | https://ngdc.cncb.ac.cn/methbank |
| | miRBase[138] | miRNA 测序与功能注释(多物种) | miRNA 功能预测、调控机制研究 | v22.1 (2018) | https://www.mirbase.org/ |
| | LncBook[139] | lncRNA 测序与功能注释(以人类为主) | lncRNA 调控网络、功能研究 | v2.0 (2021) | https://ngdc.cncb.ac.cn/lncbook/home |
| | LncRNADisease[140] | lncRNA 疾病数据库 | 疾病机制研究 | 持续更新 | http://www.rnanut.net/lncrnadisease/ |
| 蛋白质组学与结构生物学 (Proteomics& Structural Biology) | Human Protein Atlas[141] | 多组学 + 空间蛋白质组(人类组织/细胞) | 蛋白表达/功能分析、空间分布研究 | v23 (2023) | https://www.proteinatlas.org/ |
| | STRING[142] | 蛋白质-蛋白质相互作用(5k+物种，百万级交互) | 蛋白相互作用分析、功能关联研究 | v12 (2023) | https://string-db.org/ |
| | UniProtKB[119] | 190M+ 蛋白序列(多物种，详细注释) | 蛋白结构预测、功能分类、分子相互作用 | Release 2026_01 | https://www.uniprot.org/uniprotkb |
| | PDB[117] | 3D 生物大分子结构(200k+ 条目) | 蛋白结构预测/验证、分子相互作用研究 | 每周更新 | https://www.wwpdb.org/ |
| | Pfam[143] | 19k+ 蛋白家族(保守结构域) | 蛋白家族分析、结构域预测、进化分析 | v35 (2022) | https://pfam.xfam.org/ |
| | PRIDE[144] | 蛋白质组质谱数据 | 蛋白质组分析 | 持续更新 | https://www.ebi.ac.uk/pride |
| | PeptideAtlas[145] | 肽段数据库 | 蛋白鉴定 | 持续更新 | https://www.peptideatlas.org |
| | AlphaFold DB[146] | AI 预测蛋白结构数据库 | 蛋白结构预测 | 持续更新 | https://alphafold.ebi.ac.uk |
| | SCOP[147] | 蛋白结构分类数据库 | 蛋白结构分类 | 持续更新 | https://scop.mrc-lmb.cam.ac.uk |
| | BioGRID[148] | 蛋白互作数据库 | 分子互作网络分析 | 持续更新 | https://thebiogrid.org |
| | IntAct[149] | 分子互作数据库 | 网络生物学研究 | 持续更新 | https://www.ebi.ac.uk/intact |
| 药物与精准医学(Drug & Precision Medicine) | PharmGKB[150] | 基因-药物-表型关联、药物基因组学数据 | 个性化药物设计、药物反应预测 | 持续更新 | https://www.pharmgkb.org/ |
| | DrugBank[151] | 10k+药物/靶点(小分子、生物制剂等) | 药物筛选、药物再利用(重定位) | v5.1.11 (2024) | https://www.drugbank.com/ |
| | ChEMBL[152] | 2M+ 生物活性(1.8M 化合物) | 化合物活性预测、药物靶点注释 | v34 (2024) | https://www.ebi.ac.uk/chembl/ |
| | BindingDB[153] | 1.1M+ 蛋白质-小分子结合亲和力数据 | 药物-靶点亲和力预测、虚拟筛选 | 持续更新 | https://www.bindingdb.org/rwd/bind/index.jsp |
| | Open Targets | 药物靶点数据库 | 靶点发现 | 持续更新 | https://www.opentargets.org |
| 临床与多模态数据 (Clinical & Multi-modal Data) | ClinVar[154] | 临床显著性变异(致病性、良性、不确定) | 变异致病性评估、疾病相关性分析 | 每月更新 | https://www.ncbi.nlm.nih.gov/clinvar/ |
| | DisGeNET[155] | 基因-疾病关联(人工整理及 GWAS 数据) | 疾病相关基因分析、变异致病性研究 | v7.0 (2020) | https://www.disgenet.org/ |
| | MIMIC-IV[156] | 100k+ICU 患者电子病历(生命体征、实验室结果等) | 临床风险预测、疾病类型分类 | v2.2 (2023) | https://physionet.org/content/mimiciv/ |
| | TCIA[157] | 20k+癌症影像数据集(CT、MRI 等) | 癌症影像分析、分割与检测 | 持续更新 | https://tcia.at/home |
| | OMOP CDM[158] | 临床数据模型 | 临床预测模型 | 持续更新 | https://www.ohdsi.org |
| 单细胞组学 (Single-cell ) | Human Cell Atlas[103] | 33M+单细胞转录组(多组织/器官) | 细胞类型识别、细胞异质性分析 | 持续更新 | https://www.humancellatlas.org/ |
| | Tabula Sapiens[104] | 单细胞转录组(28 个器官/组织的 1.1M+人类细胞详细注释) | 细胞类型分类、组织特异性功能分析 | 2022 | https://tabula-sapiens.sf.czbiohub.org/ |
| | CellMarker[159] | 细胞标记基因数据库 | 细胞注释 | 持续更新 | http://bio-bigdata.hrbmu.edu.cn/CellMarker |
| 空间组学 (Spatial Omics) | HuBMAP[105] | 空间解析单细胞数据(约 4.5k 解剖结构) | 细胞特异性功能分析、组织结构研究 | 持续更新 | https://hubmapconsortium.org/hubmap-data/ |
| | SpatialDB[160] | 空间分辨单细胞数据(多组织/疾病) | 空间转录组分析、细胞定位研究 | 2021 | https://spatialomics.org/SpatialDB/ |
| 知识图谱与 | Reactome[161] | 人工整理的生物通路(信号转 | 信号通路分析、通路预测、信 | v88 (2024) | https://reactome.org/ |

| | | | | | |
|---|---|---|---|---|---|
| 通路 (Knowledge Graphs & Pathways) | KEGG[162] | 导、代谢等) 1.8k+通路、分子相互作用 | 号传导研究 通路分析、药物重定位 | 持续更新 | https://www.genome.jp/kegg/ |
| | Hetionet[163] | 基因-疾病-药物网络(500k+节点) | 三元关联分析、跨网络学习 | 2017 | https://het.io/ |
| | WikiPathways[164] | 社区维护通路数据库 | 通路挖掘 | 持续更新 | https://www.wikipathways.org |

注：数据库更新时间信息来源于各数据库官方网站. 部分数据库采用持续更新机制，而部分数据库采用周期性版本发布.

缩写说明： ICGC, 国际癌症基因组联盟(International Cancer Genome Consortium); gnomAD, 人群基因组变异数据库(Genome Aggregation Database); IGSR, 国际基因组样本资源库(International Genome Sample Resource); dbSNP, 单核苷酸多态性数据库(database of Single Nucleotide Polymorphisms); COSMIC, 癌症体细胞突变目录(Catalogue of Somatic Mutations in Cancer); OMIM, 人类孟德尔遗传数据库(Online Mendelian Inheritance in Man); dbVar, 结构变异数据库(database of Genomic Structural Variation); GEO, 基因表达综合数据库(Gene Expression Omnibus); CCLE, 癌症细胞系百科全书(Cancer Cell Line Encyclopedia); GTEx,, 基因型-组织表达项目(Genotype-Tissue Expression); WES, 全外显子组测序(Whole Exome Sequencing); WGS, 全基因组测序(Whole Genome Sequencing); eQTL, 表达数量性状位点(expression Quantitative Trait Locus); Pfam, 蛋白质结构域家族数据库(Protein family); PRIDE, 蛋白质组学鉴定数据库(Proteomics Identifications Database); SCOP, 蛋白质结构分类数据库(Structural Classification of Proteins); BioGRID, 生物通用相互作用数据集(Biological General Repository for Interaction Datasets); GWAS, 全基因组关联分析(Genome-Wide Association Study); ICU, 重症监护病房(Intensive Care Unit); CT, 电子计算机断层扫描(Computed Tomography); MRI, 磁共振成像(Magnetic Resonance Imaging); OMOP CDM, 观测医疗结果伙伴通用数据模型(Observational Medical Outcomes Partnership Common Data Model); KEGG, 京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes).

依托这些数据库，AI 在生物医学研究中实现了从上游的数据整合与知识建模（如多源异质数据的清洗、标准化、融合与知识图谱构建），到中游的信息挖掘与机制解析（如特征提取、疾病预测、多组学数据整合与生物学机制发现），再到下游的临床转化研究与药物开发（如靶点筛选、分子设计、疾病诊断和个性化治疗），从而为基础研究、药物开发、疾病诊断和精准医疗等领域的广泛应用奠定了坚实基础.

必须要注意的是，虽然大规模生物医学数据库极大地促进了 AI 在生物信息领域的应用，但由于不同实验流程会引入明显的技术噪声和批次效应，以及单细胞数据中普遍存在的表达矩阵稀疏与随机失活现象，数据质量的不稳定在 AI 的实际应用过程中依旧是重要挑战. 研究人员尝试通过 AI 方法改善数据质量问题. 例如，深度自编码模型 DCA[165]通过对基因表达分布特征的学习，实现了有效去噪与表达值的恢复；Harmony[166]等算法则是通过对低维表征的迭代校正消除数据来源的差异，推进跨实验数据的整合；诸如 scVI[167]等深度生成模型则尝试利用概率建模对技术噪声与生物变异进行统一处理. 除此之外，也有研究尝试利用数据增强或域适应(domain adaptation)等策略[107]，来进一步改善模型在不同来源数据间的泛化表现. 针对高噪声与标注不确定性问题，也有不少研究从噪声鲁棒学习的角度入手，通过设计稳健损失函数，或是通过不确定性建模，逐步提升模型对噪声数据的适应能力. 整体而言，这些方法能够在一定程度上提升复杂数据分析的可靠性.

## 4 AI 驱动的跨学科应用研究

医学研究与药物开发领域的重点是如何实现基础科研成果向临床实际应用转化，人工智能工具不仅可以给出解决方案，也能为分子机制的深度解析和个体化诊疗干预提供技术支持. 单细胞层面的表观遗传研究工具 DeepCpG[89]和 EpiScanpy[168]可以精准解析细胞类型特异性表观调控机制；肿瘤相关的 PathAI[169]模型能够完成肿瘤免疫微环境内细胞特征和区域特征的量化分析，进而实现肿瘤样本的精准分型判断. 借助深度学习模型的优化算法，DeepSurv[170]可以完成肿瘤患者个体的患病风险评估，同时给出更贴合患者情况的个性化治疗方案建议. 人工智能在新药研发和疫苗开发环节也展现出了巨大的应用潜力，IgLM[66]通过序列填充的技术思路优化抗体设计流程，Thera-SAbDab[171]则整合现有治疗性抗体的序列与结构相关数据，为抗体研发提供数据层面的支

撑，有效提升免疫分子的优化效率.

人工智能技术的融入为合成生物学领域的基因编辑技术的优化升级提供了新的思路. 传统CRISPR/Cas9 基因编辑技术会受到离靶效应的困扰，而借助人工智能模型的预测能力可以提前预判离靶效应的发生概率，从源头提升基因编辑操作的整体精度与安全性. CRISPR-SGRU[27]模型结合双向门控循环单元(Bidirectional Gated Recurrent Unit, BiGRU)搭建深度学习框架，能够精准预测包含错配、插入以及缺失在内的多种基因编辑最终结果，可以应用于更复杂的编辑场景. DeepCRISPR[36]模型主要关注导向 RNA 的设计环节，在设计阶段同步完成靶向活性和潜在离靶效应的双重预测，这使得基因编辑操作变得更可靠.

如何筛选优良性状是农业育种的核心问题，人工智能技术被广泛应用于高通量表型分析工作，同时也为精准育种的决策制定提供了大量的数据支持. 作物表型监测方面，PlantCV[172]能够高效提取植物图像中的各类关键特征，支撑高通量作物表型的常态化监测工作并节约大量人工成本；多组学整合工具 OmicsPipe[173]可实现 RNA-seq、ChIP-seq 等多组学数据的自动化分析处理，可精准找到基因、环境与表型三者之间的内在关联，为育种研究提供数据支持；作物生长模拟系统 DSSAT[174]能够整合土壤条件、气候因素、田间管理模式等各类环境变量，精准预测农作物的生长态势和最终产量，为优良品种筛选和个性化种植策略的制定提供参考.

微生物学研究与环境监测工作中，人工智能技术通过高通量测序数据的深度分析来解析微生物群落的整体结构. 目前应用广泛的开源微生物组多组学数据科学平台 QIIME 2[175]可完成序列质量控制、ASV/OTU 构建以及群落多样性分析等一系列基础操作，还能通过配套插件实现机器学习算法调用和多组学数据整合分析，适配多样化的研究需求，配合 PICRUSt2[176]这类人工智能工具，还可以依托 16S、18S 核糖体 RNA 等标记基因，进一步推断微生物群落的潜在功能，刚好弥补了传统测序技术只能获取群落组成、无法直接解读功能信息的短板. 在环境监测场景中，人工智能技术还能针对空气、水体等各类环境样本，实时分析微生物污染情况，预判对应的生态风险，及时发出预警信号，整体监测效率和结果准确性都得到了明显提升[177]. 从整体应用效果来看，人工智能推动了微生物学与环境科学相关研究向智能化、自动化方向转型，也为生态环境保护和公共健康保障提供了实用的技术支撑.

为了更直观地梳理不同应用领域和人工智能相关技术、科研工具之间的对应关系，我们构建了多领域科研工具/方法应用分类图，如图 3.



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

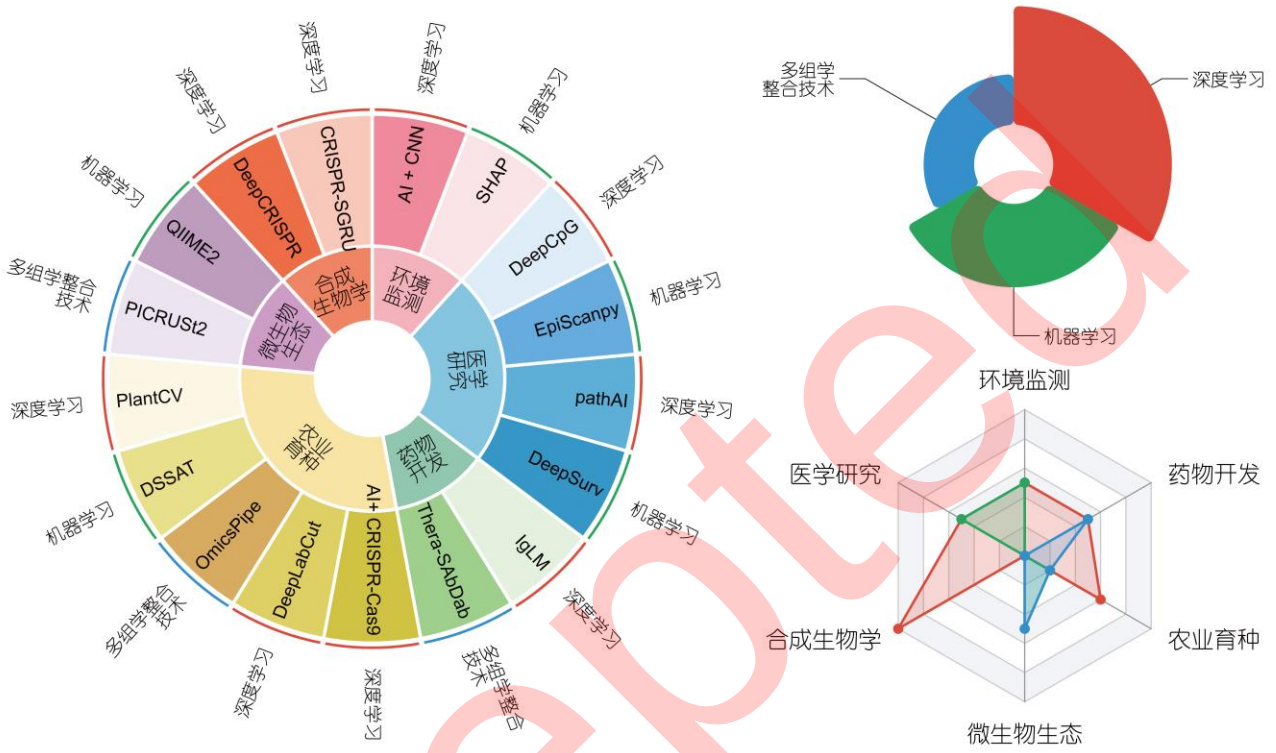


**图 3　人工智能工具与方法在多学科科研领域的分类与应用结构可视化.**

该图整合展示了人工智能核心技术（深度学习、机器学习、多组学整合技术）在六大跨学科领域中的分布特征及应用差异. 左侧旭日图呈现了各领域代表性 AI 工具的层级结构及其与不同技术路径之间的关联关系；右上角饼图展示了三类核心 AI 技术的整体应用占比；右下角雷达图对比了不同技术在环境监测、医学研究、合成生物学、微生物生态学、农业育种及药物开发等领域中的应用分布差异.

**Figure 3 Visualization of the classification and application structure of AI tools and methods across interdisciplinary research domains.**

This figure comprehensively shows the distribution characteristics and application differences of core artificial intelligence technologies, including deep learning, machine learning, and multi-omics integration, in six major interdisciplinary fields. The sunburst chart on the left presents the hierarchical structure of representative AI tools in each field and their relationships with different technical paradigms. The pie chart in the upper right shows the overall application proportion of the three core AI technologies. The radar chart in the lower right compares the application distribution differences of different technologies in environmental monitoring, medical research, synthetic biology, microbial ecology, agricultural breeding, and drug development.

## 5 人工智能在生物信息学中的挑战与发展方向

### 5.1 AI 模型的失败案例与局限性分析

AI 方法在提升生物信息学任务效率方面表现出色，但不可否认的是，模型的复杂度提升并不意味着输出的结果一定更加可靠. 这种不可靠性主要来源于生物系统的动态性、训练数据覆盖范围以及目标函数偏好等因素. 因此我们必须清楚指出 AI 工具存在的局限性.

从基因组层面来看，基因组基础模型在面对跨物种或是弱监督、无监督等特定场景时，仍可能出现表征不足或由分布偏移所带来的泛化问题. DNABERT-S 在讨论中指出，模型在下游标注匮乏或数据分布发生变化时难以有效区分不同物种来源的序列，因此研究者们尝试引入物种感知的对比学习策略，以此来优化嵌入空间的聚类效果与区分能力[178]. 在将模型应用于跨物种应用或训练语料覆盖不足的



https://www.sciengine.com/doi/10.1360/SSV-2025-0378

基因组区域时，有极大可能伴随着隐性偏差，必须进一步通过外部验证来界定适用的边界

在 RNA 三维结构预测领域，尽管 RhoFold 模型在相关任务上表现出优异性能，但该模型在处理 RNA 连接区(junction)时更容易在螺旋相对取向与堆叠方式方面出现偏差，这种偏差往往与连接区本身的动态性及复杂的多构象特征紧密相关. 这类模型的预测效果往往与 MSA 质量、结构本身的复杂度与序列长度息息相关，当模型在面对假结、多螺旋交汇拓扑等复杂结构以及比较长的 RNA 序列时，性能可能会出现明显的退化[10]. 因此在处理 RNA 任务时，尤其是当体系受到离子、配体或蛋白的调控时，模型输出的结果不能直接视作真实构象，只能被视作假设，等待进一步的实验验证提供可靠的生物学解释.

在蛋白质结构预测方面，AlphaFold 系列模型在面对本质无序蛋白/无序区域(IDPs/IDRs)及多构象体系时存在着明显局限. AlphaFold 对人类蛋白组的大规模预测结果中存在大量与无序区域高度重合的低置信度（例如 pLDDT 较低）的片段，这是因为生理状态下该区域接近于以构象集合的状态存在，而不是稳定保持在某种单一的稳定构型状态，如果直接把 AlphaFold 输出结果中的低置信度的静态构型用于后续的推断，自然会引入一定的偏差[118]. 此外，对于一些存在明显的构象切换、域间相对取向变化或条件依赖型折叠结合过程的蛋白体系，仅凭 AlphaFold 提供的单一结构快照，并不能满足真实构象分布的全覆盖需求. 在这种情况下，我们仍然需要结合湿实验结果进行进一步补充佐证.

如果说结构预测的难点主要在于对动态体系的处理能力及泛化能力的不足，那么在使用生成式 AI 进行生物序列与分子设计时的难点则更多集中在由于缺乏生物学可行性而导致的幻觉. 在蛋白序列生成与结构设计中，研究者可以利用模型可以基于先验知识从零设计出大量新序列与结构候选，但在后续的实验验证中往往仅有部分分子能实现单分散与稳定折叠. 除此之外，这类设计中还可能由于寡聚体或复合物结构先验的干扰而出现寡聚化或聚集倾向[179]. 在小分子生成与药物设计领域，深度生成模型提供的结果虚拟评分较高，实际存在化学结构不稳定、违反基本化学规则导致合成难度极高等问题. 未来可以加强相关约束并引入合成可行性评估与规则过滤等手段来降低不可合成分子的比例[80].

针对这些失败模式，目前的改进方向并不是继续扩大模型规模，而应该是把物理化学约束与可验证的后处理流程结合到完整的工作流程中. 具体来说，在生成或优化阶段就需要引入能量函数、结构稳定性评分、化学价态及反应规则、合成可行性评分等一系列约束条件，这样才能有效减少不合理构象与难以合成的候选分子产生. 而在推断完成后，还需建立多级筛选与一致性验证机制，比如结构预测结果与能量评估的交叉校验、合成路径评估等步骤等. 通过这种方式来将实验验证加入到完整闭环中，让结果的生物学合理性与可转化性得到切实提升[80, 179]. 基因编辑这类对安全性要求极高的应用场景，数据分布偏移引发的泛化能力下降问题更是格外值得重视. 基于 TrueOT 基准的相关研究发现部分 CRISPR 脱靶预测深度模型在代理数据上表现尚可，但在经过实验验证的脱靶数据集上则会出现明显的性能下降. 这一现象证实了外部基准验证与分布外评估的重要意义，同时表明通过更稳健的表示学习方法或许能使模型的跨数据集泛化能力得到改善[180].

此外，公共生物医学数据库自身的数据来源偏倚，也在一定程度上影响了人工智能模型的可靠性，例如 TCGA、gnomAD 及大规模 GWAS 数据库等基因组学和临床队列数据库在样本构成上明显偏向欧洲裔人群[181, 182]. 已有研究证实，超 80%的 GWAS 数据来源于欧洲祖源人群，然而欧洲裔人口仅占全球比例的小部分，针对非欧洲裔群体、少数族裔以及罕见病患者的数据则长期处于不足状态，这使得人工智能模型的训练基础存在系统性偏差[183, 184]. 实际上，不同人群间的等位基因频率、疾病易感位点、临床表型以及情况差异显著，模型如果过度依赖单一祖源背景数据，往往只能捕捉到特定人群中的统计关联，得出的生物学规律不具有普适性. 由于泛化能力不足，当这类模型被直接应用于训练数据覆盖不足的人群或罕见病亚群时，往往会面临风险评估失准、诊断性能下降以及错误分类等问题[185]. 除此之外，这类底层数据偏差在临床转化环节中还可能产生“算法偏见”，例如部分临床预测模型在预测女性、少数族裔及弱势群体时，准确性远远不如优势人群[181]. 实际部署这些模型后不仅无法缓解现有的健康预测难题，还可能会加剧医疗资源分配的不公平. 由于群体代表性在

训练集中的缺失，人工智能模型在实际应用中暴露出泛化能力薄弱、误诊率攀升等现象，甚至会演变为直接的临床安全风险[182].

因此，AI 工具更适合被视作高效提出可检验假设的手段，而无法得出能够直接替代实验事实的结论. 了解失败案例与局限性，可以帮助研究人员在结构预测、序列表征与分子生成的不同任务中明确预期结果，并进一步完善在物理化学约束、评测标准与实验闭环验证方面的研究. 基于这些内容，如何理解模型在给出预测结果时所依赖的关键特征与决策依据，也逐渐成为人工智能在生物信息学应用中的一个重要研究问题.

## 5.2 可解释人工智能方法及其在生物信息学中的应用

深度学习在基因组学、转录组学以及蛋白质结构预测等多项任务中得到了越来越广泛的应用，模型预测结果的可解释性问题，也逐渐成为关注重点. 相较于传统的统计模型，深度神经网络本身带有大量的参数，同时还具备复杂的非线性结构，整体的预测过程很难被直接解读. 针对这一情况，找准模型做出预测时所依赖的关键特征，也就成为当下的研究重点. 为了解决这一难点，目前已经有一系列可解释人工智能(Explainable Artificial Intelligence，XAI)方法被相继提出，这些方法能够从多个不同角度对复杂模型的内在预测依据展开解析.

目前应用较为普遍的解释策略是特征归因(feature attribution)方法. 以积分梯度法(Integrated Gradients, IG)方法为例，这种方法依靠输入与基线之间的路径累积梯度，从而实现不同特征对预测结果影响程度的量化[186]. 和这一思路相近的还有可解释性 AI 方法(SHapley Additive exPlanations, SHAP)，它依托博弈论里 Shapley 值的核心思想，把模型输出转化成各类输入特征贡献的加性组合形式，进而为不同特征给出具备一致性、可直接相互比较的重要性评估结果[187]. 这些方法不需要改动原有模型结构就可以在模型预测完成后开展事后解释，因此在生物序列建模的相关任务中得到了十分广泛的应用.

实际开展研究工作时，借助各类可解释方法，不仅能理清模型内部的决策逻辑，还能帮助研究者精准捕捉到潜在的生物学关键信号. 以深度学习驱动的基因调控预测任务为例，分析模型针对不同序列位置给出的归因结果，就能定位到和转录因子结合、染色质状态相关的核心序列区域，从而为后续调控元件的功能解析提供对应的参考依据. 代表性模型如 DeepSEA，通过对输入序列开展系统性的计算机突变分析(in silico mutation analysis)，逐一评估单个碱基改变，对转录因子结合、染色质状态预测结果产生的具体影响，从而挖掘出潜在的调控序列元素[20]. 针对 RNA 结合蛋白结合位点预测的 DeepRiPe 模型，对模型输出开展对应的归因分析，也能揭示模型所捕获的关键序列基序和对应的上下文依赖关系，进而为 RNA 转录后调控机制的相关研究提供可参考的思路[29].

需要注意的是，包括注意力权重在内的部分模型内部运行机制即使能提供一部分可供可视化的参考信息，但这类结果并不能直接当作核心依据. 例如在依托 Transformer 完成的序列建模任务里，对注意力权重做可视化处理，能够直观看到模型预测阶段重点关注的序列区域，也能辅助定位一些潜在的功能位点和结构特征. 但这些可视化结果往往无法直接体现输入特征对最终预测结果的真实贡献程度，实际研究中大多需要结合梯度归因、Shapley 值这类方法，开展多方面的综合分析才能得出可靠结论[186]. 可解释 AI 相关方法的应用，一方面能够有效提升模型预测结果的整体可信度，另一方面也为复杂组学数据的挖掘提供了新的分析思路，方便研究者从中挖掘潜在的生物学规律. 将模型解释得到的结果与已有的生物学知识相互比对，还能从深度学习模型中提取出具备实际生物学意义的相关模式，帮助研究者进一步深化对复杂生物调控机制的理解.

## 5.3 发展趋势与未来研究方向

通过前文对模型局限性与可解释性问题的分析，不难发现，目前人工智能在生物信息学领域的应用重点已经从单纯追求性能指标的提升，逐渐过渡到探究如何在可靠性可解释性以及临床或实验转化率之间达到平衡. 人工智能模型的突出优势，是其能够从大规模数据中快速提出候选结构、功能位点、调控关系或药物分子这类可检验的研究假设，但这些结果必须经过湿实验的验证，才能落地转化为具备可信度的生物学知识[179]. 经过湿实验获得的数据不应该被视作

这一流程的终点，还可以反向融入模型的优化过程，达到修正训练偏差、更新相关参数并进一步完善目标函数的作用. 这种模式在蛋白质设计、RNA 结构解析、CRISPR 脱靶评估和药物筛选等任务中已经有了初步的探索[79]. 在这样的研究背景下，未来我们应该将人工智能预测、实验验证与模型更新几个环节真正结合起来，让人工智能在最大限度上发挥优势.

要使人工智能在真实生物医学场景中发挥作用，还需要从标准化评测体系，使用隐私保护技术确保模型不会泄露敏感信息，推动资源普惠化并降低算力门槛等方向入手. 为了有效缓解当前机器学习领域广泛关注的可重复性问题，在开展模型评估时，我们有必要加强外部独立测试与分布外评估，同时完善统一基准的建设工作. 研究人员还应该规范数据版本、预处理流程、代码与参数配置，为后续工作提供可对比复现的基础条件[188].

在将人工智能模型应用于临床领域的过程中，模型的稳健性受生物医学数据本身存在的高异质性、分布偏移问题，罕见样本数量不足等因素影响，域适应(Domain Adaptation, DA)、少样本学习(Few-shot Learning, FSL)与联邦学习(Federated Learning, FL)等前沿方法具备良好的应用前景. 域适应技术能够为跨中心、跨族裔的“分布偏移”问题的解决提供有效思路. 通过对抗学习或特征空间对齐等方式，模型在不依赖特定设备、也不局限于特定人群背景的情况下，依然可以学习到稳健的特征，从而达到有效缩小源域（如高通量公共数据库）与目标域（如临床特定亚群）之间数据分布差异的目的. 通过这种方式，在应对多样化的实际部署场景时，模型依然能够保持稳定、一致的预测性能[189, 190]. 针对罕见病或罕见突变位点带来的小样本难题，少样本学习及其核心的元学习框架具有很高的实用价值. 借助在通用任务中学习到的先验知识，模型可以形成更高效的学习能力，即使只依靠少量临床标注数据，也能通过参数迁移有效提取深层生物学特征，这对推动医疗资源普惠、拓展精准医疗的覆盖范围具有重要意义[191, 192]. 而联邦学习则是能够从整体架构层面缓解数据不平衡问题. 通过使用联邦学习方法，不同医疗机构可以在严格保护数据隐私的前提下，仅通过共享模型参数而非原始数据的方式进行联合建模. 联邦学习不仅能够打破长期存在的数据孤岛问题，还能让模型学到更全面、更多样化的数据特征，从根本上降低因数据来源单一造成的系统性偏差[193, 194]，从而使人工智能系统更加适配临床转化在安全性与普惠性上的需求，真正成为可落地、可信的临床辅助决策工具[195].

当前大模型在实际应用中仍面临训练成本偏高、部署条件受限等问题. 通过轻量化结构设计、高效微调策略与开源工具链的搭建，能够有效降低使用门槛，让更多研究团队可以便捷地使用相关 AI 方法，进一步提升这类技术的适用范围与实际应用价值[196]. 此外，“通用大模型”与“垂直领域模型”之间的博弈也是需要重点关注的方向. 以 DNABERT-2、AlphaFold3 等为代表的大规模预训练模型的确在跨任务迁移表现上展现出了极强的适应性，并能够在生信任务的复用中展现出统一建模的优势. 但一些专门针对特定任务或数据分布设计的轻量化方案在计算成本和可解释性方面也有着明显优势. 因此，未来的技术发展方向可能并非单一依赖通用大模型，而是进一步尝试结合基础模型的通用能力与下游任务的适配策略，在通用和专用之间找到平衡点，使人工智能的实际可及性与其推广价值得到进一步提升，从而更好地满足生物信息学研究需求.

总体而言，人工智能在生物信息领域绝不能完全取代传统实验手段，而是能够更快提出各种科学假设，提高对复杂数据的处理与解读效率并推动多尺度信息的整合分析. 人工智能时代，这些工具在生物信息学领域想要实现长期稳定的发展，需要在模型性能优化、机制可解释性构建、实验闭环搭建、隐私安全管理以及评测体系建设等方面同步推进，只有通过这种方式，人工智能才能更加稳步地成为生命科学研究与临床应用中的核心支撑力量.

## 参　考　文　献

1 Baião A R, Cai Z, Poulos R C, et al. A technical review of multi-omics data integration methods: From classical statistical to deep generative approaches. Briefings Bioinf, 2025, 26: bbaf355
2 Guo Z, Liu J, Wang Y, et al. Diffusion models in bioinformatics: A new wave of deep learning revolution in action. arXiv preprint arXiv:2302.10907, 2023. Unpublished
3 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583-589
4 Brixi G, Durrant M G, Ku J, et al. Genome modelling and design across all domains of life with Evo 2. Nature, 2026, 1-13: in press
5 Altschul S F, Madden T L, Schäffer A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res, 1997, 25: 3389-3402
6 Eddy S R. Accelerated profile HMM searches. PLoS Comput Biol, 2011, 7: e1002195
7 Hollingsworth S A, Dror R O. Molecular dynamics simulation for all. Neuron, 2018, 99: 1129-1143
8 Ji Y, Zhou Z, Liu H, et al. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in

genome. Bioinformatics, 2021, 37: 2112-2120
9 Li Z, Zhang Y, Peng B, et al. A novel interpretable deep learning-based computational framework designed synthetic enhancers with broad cross-species activity. Nucleic Acids Res, 2024, 52: 13447-13468
10 Shen T, Hu Z, Sun S, et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. Nat Methods, 2024, 21: 2287-2298
11 Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, 373: 871-876
12 Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, et al. Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. Nat Methods, 2025, 22: 287-297
13 Deng L, Wu H, Liu X, et al. DeepD2V: A novel deep learning-based framework for predicting transcription factor binding sites from combined DNA sequence. Int J Mol Sci, 2021, 22: 5521
14 Li C, Zheng X, Chen C, et al. Tiny Data Is Sufficient: A Generalizable CNN Architecture for Temporal Domain Long Sequence Identification. IEEE Trans Neural Networks Learn Syst, 2025, 36: 12977-12990
15 Zhou Z, Ji Y, Li W, et al. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. ICLR, 2024: 1–21
16 Wang Y, Liu T, Xu D, et al. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. Sci Rep, 2016, 6: 19598
17 Teragawa S, Wang L, Liu Y. DeepPGD: A deep learning model for DNA methylation prediction using temporal convolution, BiLSTM, and attention mechanism. Int J Mol Sci, 2024, 25: 8146
18 Yan Y, Chai X, Liu J, et al. DeepMethyGene: A deep-learning model to predict gene expression using DNA methylations. BMC Bioinf, 2025, 26: 99
19 Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics, 2015, 31: 761-763
20 Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods, 2015, 12: 931-934
21 Hassanzadeh H R, Wang M D. DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins. In: eds. IEEE International Conference on Bioinformatics and Biomedicine. 2016. 2016. 178-183
22 Ng P. dna2vec: Consistent vector representations of variable-length k-mers. arXiv preprint arXiv:1701.06279, 2017. Unpublished
23 Jiang W J, Hu C, Lai F, et al. Assessing base-resolution DNA mechanics on the genome scale. Nucleic Acids Res, 2023, 51: 9552-9566
24 Nguyen E, Poli M, Faizi M, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. Advances in neural information processing systems, 2023, 36: 43177-43201
25 Li Z, Ni Y, Beardall W A, et al. Discdiff: Latent diffusion model for dna sequence generation. arXiv preprint arXiv:2402.06079, 2024. Unpublished
26 Zhao Q, Zhang C, Zhang W. DnaGrinder: a lightweight and high-capacity genomic foundation model. arXiv preprint arXiv:2409.15697, 2024. Unpublished
27 Zhang G, Luo Y, Xie H, et al. Crispr-SGRU: Prediction of CRISPR/Cas9 Off-Target Activities with Mismatches and Indels Using Stacked BiGRU. Int J Mol Sci, 2024, 25: 10945
28 Chen X, Li Y, Umarov R, et al. RNA secondary structure prediction by learning unrolled algorithms. arXiv preprint arXiv:2002.05810, 2020. Unpublished
29 Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Research, 2020, 30: 214-226
30 Arraiano C M. Regulatory noncoding RNAs: Functions and applications in health and disease. FEBS J, 2021, 288: 6308-6309
31 Yang C, Yang L, Zhou M, et al. LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning. Bioinformatics, 2018, 34: 3825-3834
32 Hendra C, Pratanwanich P N, Wan Y K, et al. Detection of m6A from direct RNA sequencing using a multiple instance learning framework. Nat Methods, 2022, 19: 1590-1598
33 Tu G, Wang X, Xia R, et al. m6A-TCPred: A web server to predict tissue-conserved human m(6)A sites using machine learning approach. BMC Bioinf, 2024, 25: 127
34 Liu C, Liang H, Wan A H, et al. Decoding the m(6)A epitranscriptomic landscape for biotechnological applications using a direct RNA sequencing approach. Nat Commun, 2025, 16: 798
35 Biswas A, Gagnon J N, Brouns S J, et al. CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. RNA Biol, 2013, 10: 817-827
36 Chuai G, Ma H, Yan J, et al. DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. Genome Biol, 2018, 19: 80
37 Kalicki C H, Haritaoglu E D. RNAbert: RNA family classification and secondary structure prediction with BERT pretrained on RNA sequences. 2020.
38 Chen J, Hu Z, Sun S, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. arXiv preprint arXiv:2204.00300, 2022. Unpublished
39 Zeng C, Jian Y, Vosoughi S, et al. Evaluating native-like structures of RNA-protein complexes through the deep learning method. Nat Commun, 2023, 14: 1060
40 Fu Z C, Gao B Q, Nan F, et al. DEMINING: A deep learning model embedded framework to distinguish RNA editing from DNA mutations in RNA sequencing data. Genome Biol, 2024, 25: 258
41 Franke J K H, Runge F, Koksal R, et al. RNAformer: Axial-Attention for Homology-Aware RNA Secondary Structure Prediction. ICLR, 2025
42 Kagaya Y, Zhang Z, Ibtehaz N, et al. NuFold: End-to-end approach for RNA tertiary structure prediction with flexible nucleobase center representation. Nat Commun, 2025, 16: 881
43 Riley A T, Robson J M, Ulanova A, et al. Generative and predictive neural networks for the design of functional RNA molecules. Nat Commun, 2025, 16: 4155
44 Camacho C, Coulouris G, Avagyan V, et al. BLAST+: Architecture and applications. BMC Bioinf, 2009, 10: 421
45 Weissenow K, Rost B. Are protein language models the new universal key? Curr Opin Struct Biol, 2025, 91: 102997
46 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
47 Elnaggar A, Heinzinger M, Dallago C, et al. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 7112-7127
48 Behzadi P, Gajdacs M. Worldwide Protein Data Bank (wwPDB): A virtual treasure for research in biotechnology. Eur J Microbiol Immunol, 2022, 11: 77-86
49 Kermani A A, Aggarwal S, Ghanbarpour A. Advances in X-ray crystallography methods to study structural dynamics of macromolecules. In: ed. Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics. Place-Published: 2023. 309-355
50 Salomon-Ferrer R, Case D A, Walker R C. An overview of the Amber biomolecular simulation package. WIREs Comput Mol Sci, 2012, 3: 198-210
51 Muhammed M T, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. Chemical Biology & Drug Design, 2019, 93: 12-20
52 Lemer C M, Rooman M J, Wodak S J. Protein structure prediction by threading methods: Evaluation of current techniques. Proteins: Structure, Function, and Genetics, 1995, 23: 337-355
53 Simons K T, Kooperberg C, Huang E, et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol, 1997, 268: 209-225
54 Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature, 2020, 577: 706-710
55 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630: 493-500
56 Yin Q, Wu M, Liu Q, et al. DeepHistone: A deep learning approach to predicting histone modifications. BMC Genomics, 2019, 20: 193
57 Brandes N, Ofer D, Peleg Y, et al. ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics, 2022, 38: 2102-2110
58 Nijkamp E, Ruffolo J A, Weinstein E N, et al. ProGen2: Exploring the boundaries of protein language models. Cell Syst, 2023, 14: 968-978 e963
59 Su J, Han C, Zhou Y, et al. SaProt: Protein language modeling with structure-aware vocabulary. ICLR, 2024
60 Heinzinger M, Weissenow K, Sanchez J G, et al. Bilingual language model for protein sequence and structure. NAR Genomics Bioinf, 2024, 6: lqae150
61 Odum M T, Teufel F, Thumuluri V, et al. DeepLoc 2.1: Multi-label membrane protein type prediction using protein language models. Nucleic Acids Res, 2024, 52: W215-W220
62 Wang X, Zheng Z, Ye F, et al. DPLM-2: A Multimodal Diffusion Protein Language Model. ICLR, 2025
63 Chen B, Cheng X, Li P, et al. xTrimoPGLM: Unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. Nat Methods, 2025, 22: 1028-1039
64 Kim H, Kim E, Lee I, et al. Artificial Intelligence in Drug Discovery: A Comprehensive Review of Data-driven and Machine Learning Approaches. Biotechnol Bioprocess Eng, 2020, 25: 895-930
65 Corso G, Stärk H, Jing B, et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. OpenReview, 2023:
66 Shuai R W, Ruffolo J A, Gray J J. IgLM: Infilling language modeling for antibody sequence design. Cell Syst, 2023, 14: 979-989 e974
67 Moingeon P, Kuenemann M, Guedj M. Artificial intelligence-enhanced drug design and development: Toward a computational precision medicine. Drug Discovery Today, 2022, 27: 215-222
68 Özçelik R, van Tilborg D, Jiménez-Luna J, et al. Structure-Based Drug Discovery with Deep Learning. ChemBioChem, 2023, 24: e202200776
69 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, et al., eds. Advances in Neural Information

Processing Systems. Long Beach, California, USA. 2017. Red Hook, NY, USA: Curran Associates, Inc., 2017. 6000 - 6010
70 Gilmer J, Schoenholz S S, Riley P F, et al. Neural Message Passing for Quantum Chemistry. In: Doina P, Yee Whye T, eds. Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research. 2017. PMLR, 2017. 1263--1272
71 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016. Unpublished
72 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020, 33: 6840-6851
73 Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 2021, 3: 1-23
74 Luo R, Sun L, Xia Y, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings Bioinf, 2022, 23: bbac409
75 Taylor R, Kardas M, Cucurull G, et al. Galactica: A large language model for science. arXiv preprint arXiv:2211.09085, 2022. Unpublished
76 Bolton E, Venigalla A, Yasunaga M, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. arXiv preprint arXiv:2403.18421, 2024. Unpublished
77 Olivecrona M, Blaschke T, Engkvist O, et al. Molecular de-novo design through deep reinforcement learning. J Cheminf, 2017, 9: 48
78 Zhou Z, Kearnes S, Li L, et al. Optimization of molecules via deep reinforcement learning. Sci Rep, 2019, 9: 10752
79 Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discovery Today, 2021, 26: 511-524
80 Zeng X, Wang F, Luo Y, et al. Deep generative molecular design reshapes drug discovery. Cell Rep Med, 2022, 3: 100794
81 Barabasi A L, Oltvai Z N. Network biology: Understanding the cell's functional organization. Nat Rev Genet, 2004, 5: 101-113
82 Yuan Q, Duren Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. Nat Biotechnol, 2025, 43: 247-257
83 Marbach D, Prill R J, Schaffter T, et al. Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107: 6286-6291
84 Wang J, Chen Y, Zou Q. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. PLos Genet, 2023, 19: e1010942
85 Chang Z, Xu Y, Dong X, et al. Single-cell and spatial multiomic inference of gene regulatory networks using SCRIPro. Bioinformatics, 2024, 40: btae466
86 Tian X, Patel Y, Wang Y. TRENDY: Gene regulatory network inference enhanced by transformer. Bioinformatics, 2025, 41: btaf314
87 Zhao M, He W, Tang J, et al. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. Briefings Bioinf, 2022, 23: bbab568
88 Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat Genet, 2003, 33 Suppl: 245-254
89 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
90 de Lima Camillo L P, Sehgal R, Armstrong J, et al. CpGPT: A foundation model for DNA methylation. bioRxiv preprint, 2024. doi: Unpublished
91 Lundberg S M, Tu W B, Raught B, et al. ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data. Genome Biol, 2016, 17: 82
92 Armingol E, Officer A, Harismendy O, et al. Deciphering cell–cell interactions and communication from gene expression. Nat Rev Genet, 2021, 22: 71-88
93 Armingol E, Baghdassarian H M, Lewis N E. The diversification of methods for studying cell–cell interactions and communication. Nat Rev Genet, 2024, 25: 381-400
94 Raredon M S B, Yang J, Kothapalli N, et al. Comprehensive visualization of cell–cell interactions in single-cell and spatial transcriptomics with NICHES. Bioinformatics, 2023, 39: btac775
95 Armingol E, Baghdassarian H M, Martino C, et al. Context-aware deconvolution of cell-cell communication with Tensor-cell2cell. Nat Commun, 2022, 13: 3665
96 Cang Z, Zhao Y, Almet A A, et al. Screening cell-cell communication in spatial transcriptomics via collective optimal transport. Nat Methods, 2023, 20: 218-228
97 Dibaeinia P, Sinha S. CIMLA: Interpretable AI for inference of differential causal networks. arXiv preprint arXiv:2304.12523, 2023. Unpublished
98 Almet A A, Tsai Y C, Watanabe M, et al. Inferring pattern-driving intercellular flows from single-cell and spatial transcriptomics. Nat Methods, 2024, 21: 1806-1817
99 Megas S, Chen D G, Polanski K, et al. Estimation of single-cell and tissue perturbation effect in spatial transcriptomics via Spatial Causal Disentanglement. ICLR, 2025
100 Li L, Xia R, Chen W, et al. Single-cell causal network inferred by cross-mapping entropy. Briefings Bioinf, 2023, 24: bbad281
101 Luecken M D, Theis F J. Current best practices in single-cell RNA-seq analysis: A tutorial. Mol Syst Biol, 2019, 15: MSB188746
102 Rao A, Barkley D, França G S, et al. Exploring tissue architecture using spatial transcriptomics. Nature, 2021, 596: 211-220
103 Rozenblatt-Rosen O, Stubbington M J, Regev A, et al. The Human Cell Atlas: From vision to reality. Nature, 2017, 550: 451-453
104 Tabula Sapiens C, Jones R C, Karkanias J, et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science, 2022, 376: eabl4896
105 Börner K, Blood P D, Silverstein J C, et al. Human biomolecular atlas program (hubmap): 3D human reference atlas construction and usage. Nat Methods, 2025, 1-16
106 Hao M, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods, 2024, 21: 1481-1491
107 Xu C, Lopez R, Mehlman E, et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol, 2021, 17: MSB20209620
108 Domínguez Conde C, Xu C, Jarvis L B, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science, 2022, 376: eabl5197
109 Saelens W, Cannoodt R, Todorov H, et al. A comparison of single-cell trajectory inference methods. Nat Biotechnol, 2019, 37: 547-554
110 Bergen V, Lange M, Peidli S, et al. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol, 2020, 38: 1408-1414
111 Hu J, Li X, Coleman K, et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods, 2021, 18: 1342-1351
112 Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nat Commun, 2022, 13: 1739
113 Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118: e2016239118—e2016239129
114 Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Guyon I, von Luxburg U, Bengio S, et al., eds. Advances in Neural Information Processing Systems. Long Beach, California, USA. 2017. Red Hook, NY, USA: Curran Associates, Inc., 2017. 4080—4090
115 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
116 Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol, 2017, 18: 83
117 Burley S K, Berman H M, Kleywegt G J, et al. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. Methods Mol Biol, 2017, 1607: 627-641
118 Ruff K M, Pappu R V. AlphaFold and Implications for Intrinsically Disordered Proteins. J Mol Biol, 2021, 433: 167208
119 UniProt C. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res, 2021, 49: D480-D489
120 Cancer Genome Atlas Research N, Weinstein J N, Collisson E A, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet, 2013, 45: 1113-1120
121 Chaudhary K, Poirion O B, Lu L, et al. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical cancer research, 2018, 24: 1248-1259
122 International Cancer Genome C, Hudson T J, Anderson W, et al. International network of cancer genome projects. Nature, 2010, 464: 993-998
123 Karczewski K J, Francioli L C, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 2020, 581: 434-443
124 Clarke L, Fairley S, Zheng-Bradley X, et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res, 2017, 45: D854-D859
125 Sherry S T, Ward M-H, Kholodov M, et al. dbSNP: The NCBI database of genetic variation. Nucleic Acids Res, 2001, 29: 308-311
126 Flicek P, Amode M R, Barrell D, et al. Ensembl 2014. Nucleic Acids Res, 2014, 42: D749-755
127 Forbes S, Clements J, Dawson E, et al. COSMIC 2005. Br J Cancer, 2006, 94: 318-322
128 Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.

Nucleic Acids Res, 2005, 33: D514-D517
129 Lappalainen I, Lopez J, Skipper L, et al. DbVar and DGVa: Public archives for genomic structural variation. Nucleic Acids Res, 2012, 41: D936-D941
130 Edgar R, Domrachev M, Lash A E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 2002, 30: 207-210
131 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
132 Consortium G T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 2020, 369: 1318-1330
133 Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res, 2003, 31: 68-71
134 Papatheodorou I, Fonseca N A, Keays M, et al. Expression Atlas: Gene and protein expression across multiple studies and organisms. Nucleic Acids Res, 2018, 46: D246-D251
135 Consortium E P, Moore J E, Purcaro M J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 2020, 583: 699-710
136 Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. Nature, 2015, 518: 317-330
137 Zhang M, Zong W, Zou D, et al. MethBank 4.0: An updated database of DNA methylation across a variety of species. Nucleic Acids Res, 2023, 51: D208-D216
138 Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: From microRNA sequences to function. Nucleic Acids Res, 2019, 47: D155-D162
139 Li Z, Liu L, Feng C, et al. LncBook 2.0: Integrating human long non-coding RNAs with multi-omics annotations. Nucleic Acids Res, 2023, 51: D186-D191
140 Chen G, Wang Z, Wang D, et al. LncRNADisease: A database for long-non-coding RNA-associated diseases. Nucleic Acids Res, 2012, 41: D983-D986
141 Uhlen M, Fagerberg L, Hallstrom B M, et al. Proteomics. Tissue-based map of the human proteome. Science, 2015, 347: 1260419
142 Szklarczyk D, Gable A L, Nastou K C, et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res, 2021, 49: D605-D612
143 Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. Nucleic Acids Res, 2021, 49: D412-D419
144 Perez-Riverol Y, Bandla C, Kundu D J, et al. The PRIDE database at 20 years: 2025 update. Nucleic Acids Res, 2025, 53: D543-D553
145 Deutsch E W, Lam H, Aebersold R. PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. The EMBO Reports, 2008, 9: 429-434
146 Varadi M, Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences. Proteomics, 2023, 23: 2200128
147 Murzin A G, Brenner S E, Hubbard T, et al. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol, 1995, 247: 536-540
148 Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci, 2021, 30: 187-200
149 Hermjakob H, Montecchi-Palazzi L, Lewington C, et al. IntAct: An open source molecular interaction database. Nucleic Acids Res, 2004, 32: D452-D455
150 Barbarino J M, Whirl-Carrillo M, Altman R B, et al. PharmGKB: A worldwide resource for pharmacogenomic information. Wiley Interdiscip Rev Syst Biol Med, 2018, 10: e1417
151 Wishart D S, Feunang Y D, Guo A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res, 2018, 46: D1074-D1082
152 Mendez D, Gaulton A, Bento A P, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res, 2019, 47: D930-D940
153 Gilson M K, Liu T, Baitaluk M, et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res, 2016, 44: D1045-1053
154 Landrum M J, Lee J M, Riley G R, et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res, 2014, 42: D980-985
155 Pinero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res, 2017, 45: D833-D839
156 Johnson A E W, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data, 2023, 10: 8
157 Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. J Digital Imaging, 2013, 26: 1045-1057
158 Ahmadi N, Peng Y, Wolfien M, et al. OMOP CDM can facilitate data-driven studies for cancer prediction: A systematic review. Int J Mol Sci, 2022, 23: 11834
159 Zhang X, Lan Y, Xu J, et al. CellMarker: A manually curated resource of cell markers in human and mouse. Nucleic Acids Res, 2019, 47: D721-D728
160 Fan Z, Chen R, Chen X. SpatialDB: A database for spatially resolved transcriptomes. Nucleic Acids Res, 2020, 48: D233-D237
161 Milacic M, Beavers D, Conley P, et al. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res, 2024, 52: D672-D678
162 Kanehisa M, Furumichi M, Sato Y, et al. KEGG: Biological systems database as a model of the real world. Nucleic Acids Res, 2025, 53: D672-D677
163 Himmelstein D S, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. elife, 2017, 6: e26726
164 Agrawal A, Balcı H, Hanspers K, et al. WikiPathways 2024: Next generation pathway database. Nucleic Acids Res, 2024, 52: D679-D689
165 Eraslan G, Simon L M, Mircea M, et al. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun, 2019, 10: 390
166 Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods, 2019, 16: 1289-1296
167 Lopez R, Regier J, Cole M B, et al. Deep generative modeling for single-cell transcriptomics. Nat Methods, 2018, 15: 1053-1058
168 Danese A, Richter M L, Chaichoompu K, et al. EpiScanpy: Integrated single-cell epigenomic analysis. Nat Commun, 2021, 12: 5228
169 Diao J A, Wang J K, Chui W F, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nat Commun, 2021, 12: 1613
170 Katzman J L, Shaham U, Cloninger A, et al. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol, 2018, 18: 24
171 Raybould M I J, Marks C, Lewis A P, et al. Thera-SAbDab: The Therapeutic Structural Antibody Database. Nucleic Acids Res, 2020, 48: D383-D388
172 Gehan M A, Fahlgren N, Abbasi A, et al. PlantCV v2: Image analysis software for high-throughput plant phenotyping. PeerJ, 2017, 5: e4088
173 Fisch K M, Meissner T, Gioia L, et al. Omics Pipe: A community-based framework for reproducible multi-omics data analysis. Bioinformatics, 2015, 31: 1724-1728
174 Jones J W, Hoogenboom G, Porter C H, et al. The DSSAT cropping system model. Eur J Agron, 2003, 18: 235-265
175 Bolyen E, Rideout J R, Dillon M R, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol, 2019, 37: 852-857
176 Douglas G M, Maffei V J, Zaneveld J R, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol, 2020, 38: 685-688
177 Olawade D B, Wada O Z, Ige A O, et al. Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions. Hygiene and Environmental Health Advances, 2024, 12: 100114
178 Zhou Z, Wu W, Ho H, et al. DNABERT-S: Pioneering species differentiation with species-aware DNA embeddings. Bioinformatics, 2025, 41: i255-i264
179 Anishchenko I, Pellock S J, Chidyausiku T M, et al. De novo protein design by deep network hallucination. Nature, 2021, 600: 547-552
180 Bhargava N, Goswami A. Decoupled Representation Learning Improves Generalization in CRISPR Off-Target Prediction. bioRxiv preprint, 2026. doi: 10.64898/2026.01.04.697551. Unpublished
181 Chen I Y, Pierson E, Rose S, et al. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci, 2021, 4: 123-144
182 Norori N, Hu Q, Aellen F M, et al. Addressing bias in big data and AI for health care: A call for open science. Patterns, 2021, 2: 100337
183 Gao Y, Sharma T, Cui Y. Addressing the challenge of biomedical data inequality: An artificial intelligence perspective. Annu Rev Biomed Data Sci, 2023, 6: 153-171
184 Sirugo G, Williams S M, Tishkoff S A. The missing diversity in human genetic studies. Cell, 2019, 177: 26-31
185 Yuan J, Hu Z, Mahal B A, et al. Integrated analysis of genetic ancestry and genomic alterations across cancers. Cancer cell, 2018, 34: 549-560. e549
186 Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. PMLR, 2017: 3319-3328
187 Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, et al., eds. Advances in Neural Information Processing Systems. Long Beach, California, USA. 2017. Red Hook, NY, USA: Curran Associates, Inc., 2017. 4768 - 4777
188 Pineau J, Vincent-Lamarre P, Sinha K, et al. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). J Mach Learn Res, 2021, 22: 1-20
189 Wang M, Deng W. Deep visual domain adaptation: A survey. Neurocomputing, 2018, 312: 135-153
190 Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of

neural networks. J Mach Learn Res, 2016, 17: 1-35
191 Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 2020, 53: 1-34
192 Hospedales T, Antoniou A, Micaelli P, et al. Meta-learning in neural networks: A survey. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 5149-5169
193 Guan H, Yap P-T, Bozoki A, et al. Federated learning for medical image analysis: A survey. Pattern Recognit, 2024, 151: 110424
194 Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. npj Digital Med, 2020, 3: 119
195 Li T, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions. IEEE Signal Process Mag, 2020, 37: 50-60
196 Kaissis G A, Makowski M R, Rückert D, et al. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell, 2020, 2: 305-311