AI+ 生物大数据推动生命科学研究范式变革

◉本报记者 赵宇彤

11 月下旬，初冬的北京，香山红叶正浓。在第 798 次香山科学会议的会场内，一场关于生物大数据与人工智能如何颠覆生命科学研究范式的讨论正在激烈进行。近五十位来自生命健康、人工智能领域的顶尖专家学者及科技管理部门负责人齐聚一堂，共同把脉这一关乎未来科技竞争格局的战略领域。

“大数据、AI（人工智能）技术正孕育着深刻变革，生命科学领域也应作出调整。”中国科学院院士、中国科学院生物物理研究所研究员陈润生在会议上强调，“生物大数据与 AI 的深度融合，将系统性重塑整个生命科学研究体系。”

尽管前景广阔，现实却提出了严峻的挑战：数据“孤岛化”现象普遍、AI 模型可解释性不足、从技术到转化的断层明显。这些瓶颈正制约着我国在该领域的创新步伐。本次以“生物大数据驱动的智能科学变革”为主题的会议，旨在凝聚共识，破局攻坚。

数据的双重困境

伴随人类基因组计划的完成，组学技术迎来爆发式发展。基因组学、转录组学、蛋白质组学、代谢组学等多分支领域的技术迭代，催生了海量生物数据，为生物大数据生态的形成奠定了基石。

“当前生物数据的复杂性已远超传统基因序列范畴。”陈润生指出，数据维度覆盖基因表达谱、蛋白质互作网络、代谢物动态变化、表观遗传修饰等多层次信息，构成一套全景式反映生命活动的复杂体系。

数据规模正经历指数级跃升。随着高通量测序技术的普及，单台设备日产出量已达数十 GB 至 TB 级别，全球科研与医疗机构持续汇交的数据总量早已突破 PB 量级，对存储、传输与计算能力提出空前挑战。

然而，数据爆炸的背后，“数据孤岛”问题日益凸显。

“生物数据是关乎科学突破、生命健康与产业竞争力等的核心战略资源。”国家生物信息中心主任杨运桂研究员强调，全球生物数据长期维持着美国国家生物技术信息中心（NCBI）、欧洲生物信息研究所（EBI）和日本 DNA 数据库（DDBJ）“三足鼎立”的格局。

尽管我国通过集中与分布式相结合的网络架构积极推动数据共享，仍面临数据库国际影响力有限、数据共享机制不健全、高质量数据集匮乏、数据存储机构可持续发展机制不健全等现实瓶颈。

面对这一局面，我国正加快推进生物数据资源体系化建设。杨运桂表示，以国家生物信息中心为核心的数据体系正在形成，整体架构加速从分散的“数据孤岛”向集约化、标准化的“数据基座”演进，推动数据完成从资源到资产、再到产品的价值跃迁，全面释放其科学价值和应用潜力。

机遇与挑战并存

近年来，以大语言模型为代表的人工智能技术，为破解生物大数据难题开辟了全新路径。

“与传统生物信息学方法相比，AI 技术具有显著优势。”陈润生深入阐释，AI 不仅能自主从海量数据中学习规律，无需依赖预设的先验知识即可挖掘深层关联，“更重要的是，它具备知识创造能力——基于已学规律生成新知识，并通过智能体实现自我迭代与持续进化”。

在实践层面，AI 技术的潜力正在多个领域显现。中国科学院院士曹晓风从农业与健康角度指出：“今天我们吃得越来越丰盛，却未必吃得更健康。”她提出，通过构建 AI 驱动的农业与土地数据采集机制，共建高质量农业数据库，将“种养循环”的生态理念与“大健康”的民生关切融入技术方案，可系统化保障从农田到餐桌的食物安全与公众健康。

“生物大数据与 AI 的深度融合正在推动生物医学研究范式的根本转变。”中国科学院院士、北京昌平实验室主任谢晓亮强调，“高质量数据是‘AI for 生物医学’的核心基石。”他透露，实验室近期开发的 FOODIE 底层技术实现了转录因子结合位点的精准测量，其升级版 ivtFOODIE 更进一步，通过机器学习与大模型预训练，可直接依据蛋白与 DNA 序列预测结合常数。

与此同时，一系列国家主导的大科学计划正稳步推进。中国科学院院士贺福初介绍的人体蛋白质组导航（π-HuB）计划，以 DIKW（数据 - 信息 - 知识 - 智慧）为路径，依托广州“慧眼”大科学设施，致力于构建全球最大的蛋白质组动态图谱，实现从“描述生命”到“预测生命”的跨越。

中国科学院院士金力则提出了开放人体生物特征通用数据模型体系的构想。该体系以构建个体生命状态的数字孪生为目标，推动多源数据的标准化交互，加速形成高质量、AI 友好的数据集。贺福初补充道：“最终将形成能够动态演化、涌现群体智慧的‘智能共生’网络，完成从描述、预测到决策的完整闭环。”

然而，机遇与挑战并存。国家生物信息中心研究员韩大力指出，当前生命科学领域的基础大模型，其训练语料仍局限于序列信息或单细胞转录组数据等单一模态，高价值的跨维度组学数据尚未实现系统化整合与深度应用。“如何让 AI 模型真正理解和融合多模态数据，是当前面临的核心技术难题，也是未来实现突破的关键所在。”

变革在路上

“生物大数据正在驱动一场深刻的智能科学变革，一个由‘AI 智能体设计实验、自动化实验室执行、数据结果闭环反馈’构成的全新科研范式正加速形成。”北京中关村学院党委书记、院长刘铁岩表示，这场变革的核心在于构建一个融合跨模态、跨学科的数据、物理规律和科学知识的“统一科学基础模型”，同时研发自主可控的软硬件协同设计基座，以充分释放统一模型与国产硬件的性能潜力。

然而，这场变革仍面临多重战略瓶颈。

在科学范式层面，AI 模型的优化目标与真实生物学问题之间存在显著的“对齐鸿沟”——AI 模型往往基于单一指标优化，而真实世界需要多目标、多约束的复杂平衡。在基础设施方面，我国在高端生物信息软件和高精度生物模拟计算硬件上仍依赖国外技术。在数据资源层面，缺乏国家级统一战略部署，导致数据质量参差不齐，难以支撑系统性突破。在人才培养方面，现有评价体系与科研组织模式亟待优化，以适应跨学科创新需求。

面对这些挑战，多位专家提出了具体路径。中国科学院遗传与发育生物学研究所研究员王秀杰强调，应加快发展生命科学多模态基础大模型。“我们正处在从‘序列’走向‘细胞’的关键爬坡期，需要精准定位 AI 可解决的科研问题，创新生物机制驱动的 AI 算法，建设自主可控的生物智能算法体系。”

中国科学院院士鄂维南指出，“科研范式的变革离不开底层基础设施的支撑。”他建议构建智能化科研平台与门户基础设施，发展面向科学推理的专业大模型与智能体，建设自动化实验操作系统，完善数据与工具基础设施。

陈润生则着眼于应用落地，指出“开发适配的未来大数据成为关键需求”。他建议推进高质量数据集建设，建立标准化的数据采集与存储规范，提升 AI 模型的泛化能力与应用可靠性，并考虑发起国家主导的生物数据基建大科学计划。

杨运桂进一步建议强化顶层设计，设立国家级生物数据管理委员会，建立统一的数据汇交与共享平台，完善国家生物数据治理体系。同时，依托国家重大需求和大科学设施，建设国家生物信息学基地，培养跨学科复合型领军人才。

在推进技术发展的同时，陈润生特别强调要加快构建完善的 AI 约束体系。“当前过度强调 AI 技术的能力赋予，却忽视了对应的约束技术体系发展。”这不仅需要建立法律法规与伦理准则，明确应用边界与责任，还要研发可解释性分析技术，确保 AI 决策透明可追溯，开发安全防护技术，防范技术滥用与系统风险。

“通过‘发展’与‘约束’的协同推进，我们才能实现 AI 与生物医药领域的深度融合，为人类健康事业提供更有力的支撑。”陈润生总结道。