DOI:10.3724/cbls.2026032 **CSTR**:32203.14.cbls.2026032
文章编号:1004-0374(2026)02-0236-12

陈大明,研究馆员,长期从事战略研究、产业研究和科技信息分析,主持多项国家科技重大专项、上海市软科学研究项目等,为国家和区域生物科技发展和政策管理提供决策参考,发表论文100余篇,主编著作8部。

陶诚,博士,长期从事战略研究和科技规划编制工作,先后参与组织“创新2050中国科技发展路线图”“科技发展新态势与面向2020年的战略选择”等战略研究。牵头完成科技部、国家自然科学基金委多项研究项目。

生物数据与知识的双向转化进展与趋势

李 荣[1,2#],葛佳莹[3#],张学博[1],张永娟[1],陈大明[1,4*],陶 诚[5*]

(1 中国科学院上海生命科学信息中心,中国科学院上海营养与健康研究所,上海 200031;2 上海大学文化遗产与信息管理学院,上海 200444;3 上海市生物医药科技产业促进中心,上海 201203;4 中国科学院大学,北京 100049;5 中国科学院武汉文献情报中心,武汉 430071)

摘 要:生命科学研究范式正经历从单向数据挖掘向“数据-模型-知识-数据”闭环协同的深刻变革。人工智能技术的全面渗透,推动生物数据从静态资源向可编程、可设计的智能对象演进,而“学习-设计-构建-测试”循环则构成了这一转型的核心引擎。在数据向知识的转化路径中,符合人工智能就绪标准的生物数据通过机器学习模型实现跨模态融合与深度表征,从海量异构信息中提炼可计算、可演绎的生物学模型,进而转化为可解释、可推理的“知识实体”;在知识向数据的转化路径中,数字孪生、虚拟细胞等计算模型将机制性知识编码为可执行的系统架构,通过仿真模拟主动生成预测性数据并指导实验设计。数据、模型与知识在此框架中构成螺旋上升的循环关系:数据驱动模型学习,模型提炼并深化知识,知识又反哺并生成新数据,进而训练更优模型。这一以人工智能赋能为基础、以系统化闭环为核心的整合范式,正成为生命科学迈向智能化、可预测与可设计时代的重要路径。

关键词:生物数据;智能模型;生物知识;双向赋能;学习-设计-构建-测试
中图分类号:Q819; TP18 **文献标识码**:A

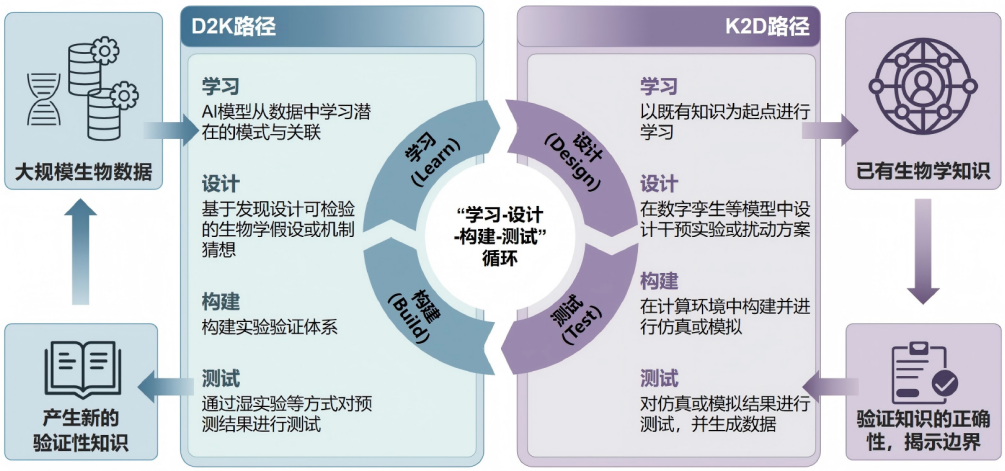
Advances and trends in the bidirectional transformation between biological data and knowledge

LI Rong[1, 2#], GE Jia-Ying[3#], ZHANG Xue-Bo[1], ZHANG Yong-Juan[1], CHEN Da-Ming[1, 4*], TAO Cheng[5*]

收稿日期:2026-01-06; **修回日期**:2026-02-06
基金项目:上海市2025年度“科技创新行动计划”软科学研究项目(25692102800)
#共同第一作者
*通信作者:E-mail:chendaming@sinh.ac.cn(陈大明);taoch@mail.whlib.ac.cn (陶诚)

(1 Shanghai Information Center for Life Sciences, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China; 2 School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, China; 3 Shanghai Center of Biomedicine Development, Shanghai 201203, China; 4 University of Chinese Academy of Sciences, Beijing 100049, China; 5 Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan 430071, China)

Abstract: The life sciences research paradigm is undergoing a profound transformation from unidirectional data analysis towards a synergistic, closed-loop system of ″data-model-knowledge-data″. This evolution is centrally driven by the pervasive integration of artificial intelligence technologies, which are redefining biological data from static repositories into programmable, designable intelligent entities. This paper systematically examines the bidirectional transformation between biological data and knowledge, highlighting the critical roles of AI-ready data, intelligent models, and the ″Learn-Design-Build-Test″ (LDBT) cycle. In the Data-to-Knowledge (D2K) trajectory, the journey begins with ensuring data ″AI-ready″, adhering to FAIR principles, possessing standardized formats, and being semantically aligned with biological knowledge. High-quality, structured data from major databases like PDB, NCBI, and GEO fuel sophisticated models. These models learn patterns to generate statistical or correlative knowledge. The crucial next step, Model-to-Knowledge (M2K), involves translating model outputs into verifiable scientific knowledge, such as mechanistic hypotheses. Enhanced model interpretability and integration into the LDBT cycle are essential for this transformation, moving beyond mere correlations to testable biological insights. Conversely, the Knowledge-to-Data (K2D) trajectory initiates with Knowledge-to-Model (K2M), where established mechanistic, associative, or hypothetical knowledge is encoded into computational model architectures. This is exemplified by digital twins and virtual cell models, which embed biological priors as structural constraints. Subsequently, in Model-to-Data (M2D), these knowledge-informed models including generative AI like diffusion models, cross-omics translators, and single-cell foundation models actively synthesize biologically plausible predictive or synthetic data. This addresses data scarcity and guides experimental design. The LDBT paradigm forms the core operational engine that unifies these bidirectional paths, creating a spiraling iterative relationship. Data drives model learning, models distill knowledge, and knowledge feeds back to generate new data for training superior models. However, challenges remain, including ensuring the reliability and reusability of AI-extracted knowledge, bridging the ″conversion gap″ between computational designs and successful experimental validation, and establishing standardized interfaces between D2K and K2D stages. Looking forward, the bidirectional loop is posited as a fundamental methodological framework for tackling biological complexity and integrating multimodal data. Its systematic engineering, through the continuous optimization of the LDBT cycle within research infrastructure, paves the way for life sciences to advance into an era of predictive and designable intelligence. Future efforts must focus on building a robust AI-ready data foundation, developing next-generation algorithms that deeply integrate data and prior knowledge, and perfecting the dry-wet lab integration for automated scientific discovery.



Key words: biological data; intelligent models; biological knowledge; bidirectional empowerment; Learn-Design-Build-Test (LDBT)

生命科学研究的范式正在经历一场由数据驱动向智能整合的深刻转型。追溯这一转型的历史脉络,可以清晰地看到:从孟德尔遗传定律的发现到人类基因组计划的完成,生命科学研究长期遵循着“假说驱动”的经典范式[1];进入21世纪以来,高通量测序、高分辨率成像、质谱分析等技术的突破性进展,使生物数据的产生速率和积累规模呈指数级增长,推动研究模式全面迈入以“数据密集型科学发现”为特征的第四范式[2]。

然而,范式演进从来不是简单的技术叠加。当前,生命科学领域正面临着一个根本性的科学命题:数据的海量积累与知识的有效产出之间存在显著的张力。一方面,组学数据、影像数据、临床数据等多模态信息的爆发式增长,使得传统的数据分析方法在处理能力、整合效率和解释深度上日益捉襟见肘;另一方面,数据的价值实现不仅取决于其规模,更取决于其能否转化为可计算、可演绎、可验证的生物学知识,并进一步通过知识的反作用来指导更精准、更高效的数据生成与实验设计。这一“数据⇄知识”的闭环协同需求,标志着生命科学研究正从相对单向的分析解读,迈向“数据-模型-知识”三重整合的协同演化新阶段。

在这一转型过程中,人工智能技术的全面渗透起到了关键的催化作用。以AlphaFold[3]、OpenFold[4]为代表的生物模型,通过跨模态数据融合与深度特征表征,实现了从海量异构信息中挖掘深层生物学机制的数据向知识转化(Data-to-Knowledge,D2K);而数字孪生、虚拟细胞等计算模型的兴起,则将机制性知识编码为可执行的系统架构,通过仿真模拟主动生成预测性数据,实现了知识向数据的反向驱动(Knowledge-to-Data,K2D)。模型作为连接数据与知识的核心枢纽,既承载着统计学习的能力,又内嵌着因果推理的机制,从而构成了双向赋能的技术“桥梁”。本文旨在系统梳理近年来生物数据领域在此核心脉络下的关键进展,聚焦于AI就绪(Artificial Intelligence Ready,AI ready)数据基座的建设、D2K与K2D双向转化的进展,并展望其在推动生命科学研究范式进一步向可计算、可预测、可设计方向演进中所面临的挑战与未来趋势。

1 生物数据与知识的双向赋能逻辑

D2K与K2D构成当代生命科学研究的核心闭环。这一双向转化并非自发进行,而是以生物模型为技术中介,以AI就绪数据为基础,遵循特定的认识论逻辑与技术路径。具体来看,AI就绪是指符合可发现、可访问、可互操作、可重用(FAIR)原则[5,6]、数据经规范化处理可供AI工具使用[7]、数据需与生物知识在语义层面对齐[8]。在D2K路径中,机器学习模型从海量数据中“学习”潜在模式,生成统计层面的关联性知识或推断层面的假设性知识,进而通过实验验证形成机制性知识。在K2D路径中,数字孪生等计算模型将已验证的机制性知识“设计”为虚拟实验场景,通过仿真“构建”预测性数据,再经“测试”环节验证其可靠性,从而生成合成数据或指导真实数据采集。

由此,“学习-设计-构建-测试”(Learn-Design-Build-Test,LDBT)范式构成了双向赋能的闭环引擎[9,10]:既驱动自下而上的数据驱动发现,也实现自上而下的知识引导生成。需要指出的是,尽管知识图谱与因果推断技术的发展可能在未来弱化部分专用模型的中介作用,但在当前技术条件下,模型仍是连接数据与知识、实现高效迭代与精准验证不可替代的核心枢纽。这一双向转化机制的建立,标志着生命科学研究正从静态的描述性科学迈向动态的预测性与设计性科学。

1.1 生物数据、模型与知识的三元关系

深入剖析上述转化机制,如图1所示,可以发现生物数据、生物模型与生物知识之间呈现出动态演进、相互反馈的螺旋关系。在D2K转化中,AI工具模型依托强大的模式识别与特征提取能力,从海量、高维、异构的组学或临床数据中挖掘稳定的统计关联,进而提炼可解释的机制性知识或可检验的假设性知识。这类模型的核心功能在于从复杂数据中“解码”生物学规律。

在K2D转化中,数字孪生模型则以已知的生物学机制、通路逻辑和系统动态为基础,在分子、细胞、器官乃至个体层面构建具有生理合理性的虚拟映射[11,12]。通过模拟干预、扰动或演化过程,生成符合特定知识约束的合成数据或预测性表型,从而反哺实验设计、填补数据空白或验证理论假设。这类模型的核心功能在于将抽象知识“编码”为可计算、可操作的数字实体。因此,在当前技术条件下,无论是自下而上的数据驱动发现,还是自上而下的知识引导生成,模型均发挥着不可或缺的中介作用:人工

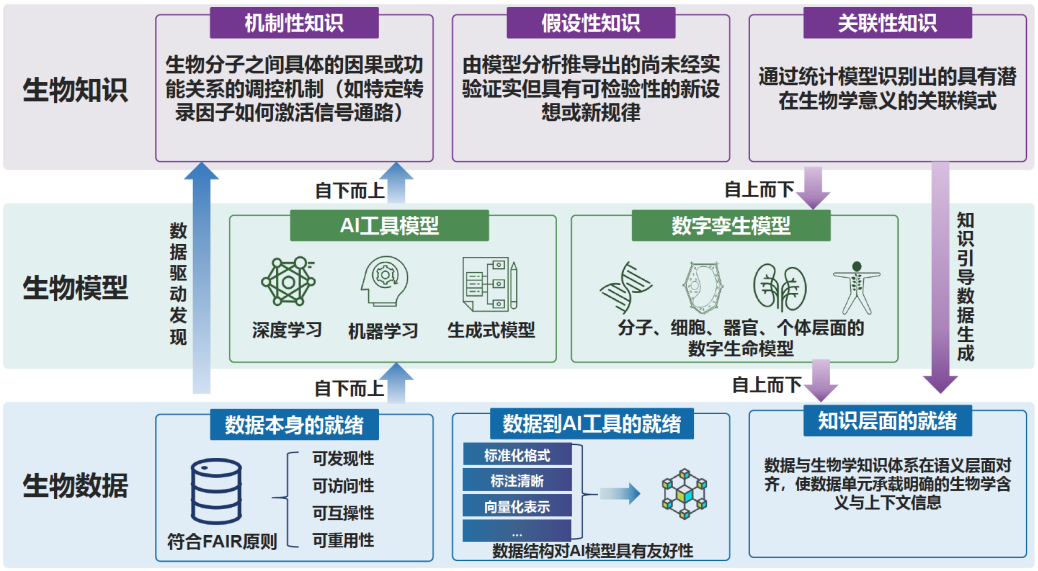


图1　生物数据、生物模型与生物知识的关系图

Figure 1　Relationship diagram of biological data, biological models and biological knowledge

智能工具模型致力于从数据中提炼知识,数字孪生模型致力于用知识生成数据,二者共同构成生物数据与生物知识双向转化的技术桥梁。

然而,这一三元关系并非静态不变。随着数据标准化程度的提高、知识图谱技术的成熟以及因果发现算法的突破,未来的知识发现可能呈现“去模型化”趋势。具体而言,当数据实现高度结构化、语义化并与领域知识深度对齐时,部分知识提取任务可通过数据间的直接关联挖掘、知识图谱推理或因果推断算法完成,无需经过传统意义上的模型训练与参数优化。这种“数据-知识”的直接转化路径,更适用于大规模、多模态、高维度的生物数据场景,能够实现更快速的知识迭代与更精准的因果发现。但必须强调的是,这种趋势并非意味着模型的消亡,而是形成“直接路径”与“模型路径”并存的多元知识发现范式。对于复杂因果机制的解析、多尺度系统的整合、动态过程的模拟等场景,模型仍将是不可或缺的核心工具。因此,数据、模型与知识的关系本质上正从当前的“线性依赖”架构,向未来的“双向互动、模型作为可选工具”的多元架构演进。这一转变的核心驱动力在于数据质量、算法能力与知识表征方式的协同提升,其最终目标是构建更加高效、可解释、可验证的生命科学知识发现体系。

1.2　LDBT循环的理论内涵与实现路径

LDBT范式为生物数据与知识的双向转化提供了系统化、可迭代的操作框架。这一范式借鉴了工程领域的系统开发理念,将其与生命科学的发现逻辑相融合,形成了兼具科学严谨性与技术可操作性的闭环机制。

在D2K路径中,LDBT循环表现为:AI模型首先从大规模生物数据中“学习”潜在的模式与关联,提炼统计规律;基于这些发现,研究者“设计”出可检验的生物学假设或机制猜想;继而“构建”相应的实验验证体系,如利用基因编辑技术构建细胞模型或动物模型;最终通过湿实验对预测结果进行“测试”,产生新的验证性知识。值得注意的是,测试环节产生的数据又作为新的输入反馈至起点,驱动下一轮学习。这一闭环使得知识发现从一次性推断转变为可重复、可优化的迭代过程,显著提升了科学发现的效率与可靠性。

在K2D路径中,LDBT循环遵循不同的逻辑起点与运作方式:研究者以既有的机制性知识为起点进行“学习”,深入理解特定生物学过程,如信号通路的调控逻辑;基于此,在数字孪生模型中“设计”虚拟干预实验或扰动方案;随后在计算环境中“构建”并运行仿真,生成预测性的表型数据或系统行为;最后,通过将模拟结果与真实世界观测进行对比“测试”,验证知识的正确性,揭示其适用范围与边界条件,并据此提出对关键观测数据的新需求。这一循环本质上是将抽象知识转化为可计算、可测试的形式,进而

指导新数据的定向生成或实验方案的优化设计。

尽管D2K与K2D两条路径中的LDBT循环在起点、方向和具体操作上存在差异,但二者共同构成了生命科学研究的双轮驱动机制:前者实现了从经验数据到理论知识的归纳升华,后者实现了从理论知识到预测数据的演绎应用。两个循环的相互嵌套与动态耦合,推动着生命科学研究向可计算、可预测、可设计的智能化范式演进。

2 从生物数据到生物知识的转化

2.1 从数据到模型(D2M)

从数据到模型的转化(Data-to-Model,D2M)构成了D2K路径的逻辑起点,也是LDBT循环得以启动的基础。在这一转化环节中,生物数据质量与结构直接决定了后续知识提取的深度与可靠性。

当前生命科学数据呈现出规模庞大、模态多样、结构复杂的特征。泛基因组数据、临床电子健康记录、医学影像等异构数据源的爆发式增长,对数据的一致性、准确性和可计算性提出了前所未有的要求。数据质量缺陷或格式不兼容将直接导致模型训练的偏差与泛化能力下降[13]。因此,D2M转化的核心前提在于生物数据是否达到“AI就绪”标准。具体而言,AI就绪数据需满足三个维度的要求:其一,数据本身应具备高质量与标准化格式,符合可发现、可访问、可互操作、可重用原则;其二,数据结构应对人工智能模型具有友好性,能够被算法高效解析、调用与处理;其三,数据应与生物学知识体系在语义层面深度对齐,使数据单元承载明确的生物学含义与上下文信息[14,15]。

国际权威数据库的结构化升级实践,为AI就绪数据基座的建设提供了重要参照(图2和表1)。蛋白质结构数据库(PDB)通过采用大分子晶体学信息文件(macromolecular Crystallographic Information Framework,mmCIF)格式[16,17]与高性能应用程序编程接口(API)[18,19],显著提升了结构数据的可编程访问能力,为AlphaFold、OpenFold等深度学习模型提供了高质量的训练数据源。美国国家生物技术信息中心NCBI、欧洲分子生物学实验室EMBL-EBI等基因组学数据库则以标准化的FASTA序列格式(The Fasta Format)[20,21]和变异记录格式(Variant Call Format,VCF)[22]为DeepVariant、DNABERT等基因组分析模型提供了海量、规范的序列输入[23,24]。基因表达综合数据库(GEO)与Reactome等通路数据库,通过系统生物学标记语言(Systems Biology Markup Language,SBML)[25]和SOFT格式[26]等标准化格式,实现了信号通路、基因调控网络等复杂生物过程的语义化表达,支持模型进行从序列到功能的系统性预测。

图2 生物技术与生命健康领域数据库汇总

Figure 2 Summary of databases in biotechnology and life health fields

表1　代表性生物数据库目前状态例举

Table 1　Examples of current status of representative biological databases

序号	数据库名称	数据类型	当前状态	能否公开访问/使用	数据库规模
1	NCBI	综合性数据,包括基因组、文献、医学、序列等	活跃维护	是,绝大部分公开	极高,全球基础性资源,覆盖全部子库
2	GenBank	核酸序列、基因组	活跃维护	是,绝大部分公开	极高,全球最大核酸序列数据库
3	PDB	蛋白质结构	活跃维护	是,公开	极高,所有公开的生物大分子结构
4	UniProt	蛋白质组学	活跃维护	是,公开	极高,蛋白质序列和功能注释
5	Ensembl	基因组、注释	活跃维护	是,公开	极高,主要覆盖脊椎动物
6	dbSNP	遗传变异试剂库	活跃维护	是,公开	极高,全球变异位点数据库
7	ENCODE	表观遗传、转录组、功能基因组	活跃维护	是,公开	很高,数据量庞大
8	TCGA	癌症基因组、临床	活跃维护	部分公开	很高,数据量庞大
9	PubChem	化学分子、药物	活跃维护	是,公开	极高,化学物质库
10	Reactome	生物通路、系统生物学	活跃维护	是,公开	高,全球通路数据库
11	OMIM	遗传疾病、健康	活跃维护	部分受限,需订阅部分内容	高
12	GEO	基因表达、转录组	活跃维护	是,公开	极高,基因表达数据
13	gnomAD	人群基因组变异	活跃维护	是,公开	很高,约30 TB
14	miRBase	miRNA、非编码RNA	活跃维护	是,公开	高
15	SWISS-MODEL	蛋白质同源建模	活跃维护	是,公开	高
16	Allen Brain Atlas	脑图谱、空间转录组	活跃维护	是,公开	高
17	EMBL-EBI	综合性数据	活跃维护	是,公开	极高
18	BioGRID	蛋白质互作、基因互作	活跃维护	是,公开	高
19	HMDB	代谢组学、代谢物	活跃维护	是,公开	高
20	COSMIC	癌症突变、健康	活跃维护	受限访问	很高

这些实践表明,只有当生物数据在格式规范、质量控制和语义标注等层面与目标模型的输入假设、学习机制和任务目标深度对齐时,数据资源才能真正转化为驱动知识发现的生产要素。

随着“AI就绪”数据基座的持续完善,生物信息学领域涌现出一系列具有里程碑意义的深度学习模型(表2)。这些模型的成功实践,深刻揭示了数据质量与模型性能之间的内在关联,为后续研究提供了重要范式参照。

在基因组学领域,变异检测模型DeepVariant与序列表征模型DNABERT的发展,充分说明了高质量标注数据对序列分析模型的决定性作用[27,28]。这类模型对训练数据的要求不仅限于序列信息的完整性,更强调变异注释的精确性与群体遗传背景的代表性。以癌症基因组图谱TCGA与基因组聚合数据库gnomAD为代表的大规模、经严格质控的基因组数据集,通过提供高质量的测序数据与精确的群体频率标注,显著提升了DeepVariant在人类及非模式生物中的变异检测召回率,尤其在低频突变等复杂场景中表现突出[29,30]。这表明,对于序列分析模型而言,AI-ready数据的核心内涵在于:数据不仅要完整,还需具备精准的真值标签与充分的群体代表性。

在蛋白质科学领域,AlphaFold系列模型的持续演进,展示了标准化数据对结构预测精度的基础性支撑,该模型的训练依赖于高度规范化的FASTA序列格式与多序列比对(MSA)数据[31]。截至2024年,AlphaFold数据库已整合超过2.14亿条蛋白质序列[32],其统一的氨基酸编号体系与结构域功能注释,有效降低了模型学习过程中的噪声干扰,直接支撑了原子级精度的三维结构预测。类似地,分子相互作用预测模型,如DimeNet,对配体-受体复合物的结构数据提出了严格要求:不仅需要明确的原子坐标信息,还需精确标注结合界面特征[33]。基于蛋白质结构数据库PDB中经人工校验的高质量复合物数据训练的模型,在药物筛选任务中的预测准确率提升超过30%[34],充分证明了结构数据标准化对图神经网络等复杂模型的关键价值。

在神经科学与医学影像领域,模型对数据AI就绪的要求更为严苛。以nnU-Net为代表的医学影像分割模型,其成功应用依赖于标准化DICOM影像格式与专家级分割标注的协同支撑[35];CheXNet等胸部疾病分类模型的性能提升,则得益于统一拍摄协

表 2 近年来代表性AI模型对应的AI-ready数据类型(举例)

Table 2 Examples of AI-ready data types for representative AI models in recent years

模型类型	代表模型	代表模型对应的AI-ready数据类型(举例)
核酸序列分析模型	DeepVariant、DNABERT	高质量的全基因组测序数据和精确的变异注释
蛋白质结构预测模型	AlphaFold	依赖于高质量的蛋白质序列和同源序列比对数据
分子相互作用预测模型	DimeNet	高质量的分子结构数据,如PDB中的配体-受体复合物
脑电信号处理模型	EEGNet	标准化的EEG时序数据
神经影像分析模型	nnU-Net	标准化的医学影像(DICOM)和精确的分割标签
多模态神经数据融合模型	Transformer	将 fMRI(空间分辨率)与 EEG(时间分辨率)进行对齐的多模态数据
生物声学分析模型	BirdNET	标准化的声学记录和高质量的物种标签
医学影像分析模型	CheXNet	标准化的胸部 X 射线(ChestX-ray14)和精确的疾病标签
临床预测模型	XGBoost、DeepSurv	结构化的临床表格数据和统一的编码标准
手术与行为视频分析模型	SlowFast、3D CNN	标准化的手术视频和帧级别的行为标签
生理信号分析模型	CNN-LSTM	标准化的心电图或肌电图信号

议与国际疾病分类标准编码体系所保障的数据一致性与标签准确性[36]。

基于此,只有当生物数据在格式规范、质量控制、语义对齐和标注精度等维度与目标模型的输入假设、学习机制和任务目标实现深度匹配时,数据资源才能真正达到“AI就绪”标准,从而充分释放AI模型的潜在能力。

2.2 从模型到知识(M2K)

模型训练本身并非科学研究的终极目标,其核心使命在于将算法的预测输出转化为具有明确生物学意义、可被实验验证的科学知识。从模型到知识的转化(Model-to-Knowledge,M2K)构成了D2K路径的关键环节,其实质是实现从计算模型到生物学知识体系的跃迁。

依据认识论层次,M2K所涉及的知识形态可区分为三类:机制性知识,即阐明生物分子间因果或功能关系的调控原理,如特定转录因子激活下游信号通路的分子机制;假设性知识,即由模型分析推导出的、尚未经实验证实但具有可检验性的新规律或新猜想;关联性知识,即通过统计模型识别出的、具有潜在生物学意义的关联模式,如基因突变与疾病表型之间的相关性。

实现M2K的首要前提在于提升模型的可解释性。我们认为,即便预测性能优异的深度学习模型,若无法阐明其决策依据与推理逻辑,其输出往往仅停留在关联性知识层面,难以向机制性知识或假设性知识深化。以单细胞多组学分析为例,Hingerl等[37]构建Scooby模型,以大规模序列学习为基础,并通过细胞特异的解码机制,在单细胞分辨率上生成并解释多模态基因组图谱;Tejada-Lapuerta等[38]则提出Nicheformer模型,对大规模单细胞与空间组学资源进行预训练,使模型学习到能够编码细胞所处的空间邻域的细胞表征,并把这种空间层面的知识用于跨任务、跨数据形态的迁移。在此过程中,LDBT循环发挥关键作用:模型从数据驱动的学习中识别潜在模式后,需经由外部生物学知识或湿实验验证,并将验证结果反馈至模型优化环节,从而形成知识发现的闭环。由此可见,当模型架构能够反映生物学系统的层级组织与功能关联特征,并在迭代循环中持续校正时,其输出才更有可能指向具有生物学意义的机制性解释。

除直接从实验数据中提取知识外,M2K还涵盖从海量非结构化文本中挖掘并结构化知识的重要功能。Wang等[39]基于GeneAgent模型,对1 106个不同来源基因集进行评估,并通过自主与生物数据库交叉验证其输出的可信度,加速了知识发现。以BioBERT[40]、LinkBERT[41]、BioGPT[42]为代表的预训练语言模型,通过编码生物医学文献中基因、疾病、药物等实体间的共现关系与语义关联,构建了隐式的知识网络。因此,这些内隐的关联性知识可被显式化,例如系统性地识别潜在的基因-疾病关联,从而为实验设计提供基于文献证据的研究线索。

基于模型生成假设并通过实验验证以确证知识,是M2K的另一重要实现路径。以RNA m^6A甲基化修饰研究为例,研究者在实现修饰位点高精度预测的基础上,通过解析模型的决策依赖特征,发现某些m^6A修饰变异体倾向于在经典修饰位点下游约50个核苷酸位置富集的规律[43]。这一由模型导出的假设性知识,随后在独立实验数据集中得到验证,从而转化为确证的关联性知识,拓展了人们对RNA表

观遗传调控机制的认知。

3 从生物知识到生物数据的生成

3.1 从知识到模型(K2M)

K2D路径以知识向模型的转化(Knowledge-to-Model,K2M)为逻辑起点。在这一阶段,经过验证的机制性知识、高置信度的关联性知识以及可检验的假设性知识,被系统性地编码并内嵌于人工智能模型的架构设计、训练目标或约束条件之中,从而指导下一代高级模型的构建与优化。在此范式下,计算模型不再仅仅是数据拟合的工具,而是成为承载、验证乃至拓展科学理论的可计算知识载体与模拟平台。数字生命研究正是K2M理念的前沿实践。

机制性知识作为模型的结构性先验与物理约束,在数字孪生构建中发挥基础性作用。生命系统固有的层级组织——从基因、分子、细胞到器官——构成了数字孪生的基本逻辑框架。这种跨尺度的机制性认知,指导着研究者设计能够反映真实生物结构与功能耦合的模型拓扑。例如,基于秀丽隐杆线虫完整神经系统连接图谱构建的数字孪生模型[44],严格遵循已知的突触连接与神经回路功能,将连接权重和模块划分直接映射至生物学事实,从而确保其运动行为具有生理合理性。类似地,心脏数字孪生模型通过内嵌心肌电生理传导方程与力学收缩本构关系,实现对药物诱导心律失常的高保真模拟[45,46]。这些实践表明,将机制性知识编码入模型架构,可显著提升模型在数据稀疏区域的泛化能力与物理解释性[47]。

假设性知识为模型的创新设计提供探索方向,驱动模型结构的拓展与升级。当科学假设提出后,K2M过程需要将抽象假设转化为可计算的模型组件。以人工智能虚拟细胞为例[48],基于Transformer架构的细胞基础模型通过对海量单细胞数据的预训练,将关于基因调控网络的假设转化为可执行的下游任务,如预测细胞对遗传或药物扰动的响应。多智能体仿真平台则能够将关于细胞群体行为的假设转化为高保真的硅基环境模拟[49],从而在虚拟空间中进行大规模参数扫描以优化生物疗法设计。这种从“假设”到“建模”的转化,经由湿实验验证形成闭环,为生命系统的机制探索提供了新途径[50]。

关联性知识常被转化为模型训练的目标函数或正则化约束[51,52]。基因-疾病关联、药物-靶点互作以及跨尺度信号传导等层级间的关联模式,为模型训练提供先验指导。知识图谱技术为关联性知识的模型嵌入提供了系统化路径。通过构建包含药物、靶点和疾病的三元关联知识图谱[53],图神经网络能够优先学习具有生物学意义的路径特征[54],有效减少模型对数据噪声的依赖,提高预测的可信度。此外,分层图神经网络架构可将基因组变异与临床表型之间的多层次关联编码为图结构,实现从基因型到表型的端到端预测,在复杂疾病风险预测中展现出优越性能。

3.2 从模型到数据(M2D)

知识引导的生成式AI技术正成为突破数据稀缺瓶颈、增强机制探索能力的重要途径。传统多组学研究常因样本量有限、数据模态缺失或隐私保护限制而难以开展有效的统计分析[55]。将生物学先验知识深度融入生成模型,能够合成既符合统计分布又具备生物学合理性的多模态数据,为深入理解复杂生命系统提供新的方法论支撑。

当前可生成生物学数据的模型大体可分为四类。一是以生成对抗网络(GAN)、变分自编码器(VAE)和扩散模型为代表的生成式深度学习模型,用于蛋白质结构生成、细胞数据生成、图像去噪等领域,其中扩散模型作为先进的生成式AI框架,通过从噪声到数据的逆向扩散过程,能够克服GAN常见的模式崩溃问题,在条件控制下生成高保真生物数据,正推动多组学数据增强向动态和连续方向发展[56-58]。二是跨组学翻译型模型,利用不同组学之间的统计与语义对应关系,通过模态间映射生成数据。例如,Wang等[59]提出的NicheTrans模型利用模态间映射,实现从基因表达翻译到蛋白质表达、从空间转录组翻译到形态特征,缓解了高成本组学数据的获取瓶颈。三是序列驱动模型,不仅可以从DNA序列预测并生成下游分子表型,还可以通过多模态集成,结合DNA甲基化、RNA二级结构等信息提升数据预测的精度,以AlphaGenome、AlphaFold为代表的模型已经能够从长达1 Mb的DNA序列中预测包括基因表达、染色质可及性在内的数千种功能组学特征[60]。四是单细胞与多组学基础模型,通过学习高维的潜空间,能够捕捉细胞的离散类型、细胞的连续动态变化,进而用于细胞注释、轨迹推断、扰动预测等任务,并通过微调适配不同的实验技术,例如,scGPT[61]、Geneformer[62]、scBERT[63]等通过大

规模的自监督学习,学习通用的细胞状态表征,用于下游任务的微调。

4 总结与展望

生命科学研究范式正经历从单向分析向闭环协同的深刻转型。回溯传统研究模式,其长期遵循“数据积累→信息提取→知识发现”的线性路径:通过高通量实验获取数据,再借助生物信息学工具进行解读。然而,这一范式面临两大结构性瓶颈:其一,对海量、高维、异构数据的深度挖掘与机制性解释能力不足,难以从复杂关联中提炼稳健、可演绎的因果规律;其二,从已有知识反向指导数据生成与实验设计的能力薄弱,导致知识发现与数据积累之间形成断裂。随着人工智能虚拟细胞[64]、数字孪生大脑[65]等前沿概念的初步实践,学界正致力于将生命系统转化为可计算、可模拟的数字实体,构建“数据⇄知识”的双向转化闭环。然而,这一闭环的高效运行仍面临多重挑战。

在数据向知识(D2K)的转化路径上,核心难点在于知识的可靠性与可复用性。当前人工智能模型虽能从单细胞多组学等复杂数据中识别细胞亚群或推断发育轨迹,但这些结果往往呈现为算法依赖性的统计模式,而非明确的因果机制[66,67]。不同分析方法对同一数据集可能得出差异显著的结论,使得这些“知识”难以作为稳固的基础向下游传递,制约其指导后续实验设计的可信度。

在知识向数据(K2D)的转化路径上,突出瓶颈体现为理论设计与实验验证之间的“转化鸿沟”。以人工智能辅助蛋白质设计为例,尽管计算模型能生成大量理论上合理的候选结构,但其中能在湿实验中成功表达并展现预期功能的仍然有限[68,69]。这一差距揭示了当前模型内化的知识与真实生物系统复杂性之间的本质差异,导致设计环节产出在验证阶段出现大量失败,使LDBT循环难以进行多轮有效迭代。

此外,D2K与K2D两个阶段之间缺乏统一、可机器操作的标准接口。D2K阶段产出的是神经网络的特征权重或复杂关联网络,而K2D阶段需要的却是明确的、可被实验设备执行的设计规则与参数化约束。这种不匹配性使得双向转化在关键节点仍需依赖研究者的主观解读,阻碍全自动化智能闭环的形成。同时,闭环的有效运行对数据基础设施与治理水平提出了更高要求:数据的可追溯性、实验上下文的完整性以及负结果的记录,都是保障模型训练质量与体系稳健性的基础条件。

面向未来,生物数据与知识的双向循环将成为应对生命系统复杂性、驾驭数据多样性的必然选择,其实现路径在于LDBT循环的系统性工程化。双向循环是揭示生命复杂性的方法论基础:生物系统并非静态的机械组合,而是多层次、多尺度动态耦合的复杂适应系统,具有高度的自适应、自组织、自演化特征。单向的“数据→知识”分析仅能捕捉特定条件下的静态片段或统计关联,难以揭示驱动动态过程的底层机制;而单向的“知识→数据”应用若缺乏真实数据的持续反馈,其设计易脱离生物实际。LDBT循环通过“学习”提取潜在规律,通过"设计"生成可验证假设,通过"构建"将设计转化为实体,再通过"测试"生成验证数据并反馈优化,形成“分析-生成-验证-优化”的持续闭环。这一机制使人工智能模型能在D2K与K2D的交替迭代中持续学习,通过主动干预生成的数据探索因果关系,趋近对生命复杂性的机制性理解。

双向循环是整合多模态数据的核心框架。当前生命科学研究产生的数据覆盖分子、细胞、个体到生态多个层级,来源多样、格式各异、尺度悬殊。双向循环提供了系统性的整合路径:在“学习”阶段借助知识图谱与跨模态表示学习实现语义融合;在“设计-构建-测试”阶段利用融合知识指导高价值数据的靶向生成。每一次LDBT迭代产生的闭环数据不仅验证假设,更持续反哺模型,增强其处理异构数据的鲁棒性,推动系统向全面AI就绪进化。

双向循环是催生颠覆性创新的范式引擎。当LDBT循环深度集成于科研基础设施,科学发现将从“假说驱动”或“数据驱动”迈向“智能体驱动”的自主演化模式。自动化科学智能体可在闭环中自主设定目标、生成假说、设计实验、执行测试并迭代学习,通过干湿闭环融合加速药物研发、精准医疗等领域的创新突破。

因此,未来的核心任务在于精心构建并持续优化这一智能循环的各个环节:建设高质量、标准化、治理可信的AI就绪数据基座;发展深度融合多模态数据与先验知识的下一代算法框架;完善自动化科学智能体与实验平台的干湿闭环集成;建立适配的伦理规范与协同创新生态体系。

参考文献

[1] 陈竺. 生命科学的发展趋势及我院的战略思考. 中国科学院院刊, 2003, 18: 170–5.
Chen Z. Trends in life sciences and strategic considerations for our institute. Chin Acad Sci J, 2003, 18: 170–5.

[2] Abriata LA. The Nobel Prize in Chemistry: past, present, and future of AI in biology. Commun Biol, 2024, 7: 1409.

[3] Nussinov R, Zhang M, Liu Y, et al. AlphaFold, artificial intelligence (AI), and allostery. J Phys Chem B, 2022, 126: 6372–83.

[4] Ahdritz G, Bouatta N, Floristean C, et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Nat Methods, 2024, 21: 1514–24.

[5] Welter D, Juty N, Rocca-Serra P, et al. FAIR in action-a flexible framework to guide FAIRification. Sci Data, 2023, 10: 291.

[6] David R, Rybina A, Burel JM, et al. “Be sustainable”: EOSC-Life recommendations for implementation of FAIR principles in life science data handling. EMBO J, 2023, 42: e115008.

[7] Sujon KM, Hassan RB, Towshi ZT, et al. When to use standardization and normalization: empirical evidence from machine learning models and XAI. IEEE Access, 2024, 12: 135300–14.

[8] Hao Z, Mayer W, Xia J, et al. Ontology alignment with semantic and structural embeddings. J Web Semantics, 2023, 78: 100798.

[9] Matzko R, Konur S. Technologies for design-build-test-learn automation and computational modelling across the synthetic biology workflow: a review. Netw Model Anal Health Inform Bioinform, 2024, 13: 22.

[10] Liao X, Ma H, Tang YJ. Artificial intelligence: a solution to involution of design-build-test-learn cycle. Curr Opin Biotechnol, 2022, 75: 102712.

[11] Tao F, Xiao B, Qi Q, et al. Digital twin modeling. J Manuf Syst, 2022, 64: 372–89.

[12] Alsalloum GA, Al Sawaftah NM, Percival KM, et al. Digital twins of biological systems: a narrative review. IEEE Open J Eng Med Biol, 2024, 5: 670–7.

[13] Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. Nature, 2023, 617: 312–24.

[14] Caufield H, Ghosh S, Kong SW, et al. Standards in the preparation of biomedical research metadata: a bridge2AI perspective. arXiv, 2025, https://doi. org/10. 48550/arXiv. 2509.10432.

[15] Caufield JH, Putman T, Schaper K, et al. KG-Hub—building and exchanging biological knowledge graphs. Bioinformatics, 2023, 39: btad418.

[16] Westbrook JD, Fitzgerald PM. The PDB format, mmCIF, and other data formats. Methods Biochem Anal, 2003, 44: 161–79.

[17] Zhang C. BeEM: fast and faithful conversion of mmCIF format structure files to PDB format. BMC Bioinformatics, 2023, 24: 260.

[18] Bittrich S, Bhikadiya C, Bi C, et al. RCSB protein data bank: efficient searching and simultaneous access to one million computed structure models alongside the PDB structures enabled by architectural advances. J Mol Biol, 2023, 435: 167994.

[19] Nair S, Váradi M, Nadzirin N, et al. PDBe aggregated API: programmatic access to an integrative knowledge graph of molecular structure data. Bioinformatics, 2021, 37: 3950–2.

[20] Goldfarb T, Kodali VK, Pujar S, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res, 2025, 53: D243–57.

[21] Dyer SC, Austine-Orimoloye O, Azov AG, et al. Ensembl 2025. Nucleic Acids Res, 2025, 53: D948–57.

[22] Hunt S E, Moore B, Amode R M, et al. Annotating and prioritizing genomic variants using the Ensembl Variant Effect Predictor—A tutorial. Hum Mutat, 2022, 43: 986–97.

[23] Abdelwahab O, Torkamaneh D. Artificial intelligence in variant calling: a review. Front Bioinform, 2025, 5: 1574359.

[24] 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: 2112–20.

[25] Paparozzi V, Nardini C. tidysbml: R/Bioconductor package for SBML extraction into dataframes. Bioinform Adv, 2024, 4: vbae148.

[26] Sinno RM, Baldock G, Gleason K, et al. The regulatory dialectic and innovation in service-based money laundering. J Financ Crime, 2025, 32: 245–54.

[27] Gurianova A, Pestruilova A, Beliaeva A, et al. Rethinking DeepVariant: efficient neural architectures for intelligent variant calling. Int J Mol Sci, 2026, 27: 513.

[28] Zhou Z, Wu W, Ho H, et al. DNABERT-S: learning species-aware DNA embedding with genome foundation models. arXiv, 2024, https://doi.org/10.48550/arXiv.2402.08777.

[29] Xu AG, Xu Y, Xing Y, et al. Towards a universal foundation model for biobank-scale human genome variation. bioRxiv, 2025, https://doi.org/10.1101/2025.01.29.635579.

[30] He R, Sarwal V, Qiu X, et al. Generative AI models in time-varying biomedical data: scoping review. J Med Internet Res, 2025, 27: e59792.

[31] Bertoni D, Tsenkov M, Magana P, et al. AlphaFold Protein Structure Database 2025: a redesigned interface and updated structural coverage. Nucleic Acids Res, 2026, 54: D358–62.

[32] Varadi M, Bertoni D, Magana P, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res, 2024, 52: D368–75.

[33] Wang Z, Wang X, Wang D, et al. DIME-Net: a dual-illumination adaptive enhancement network based on Retinex and mixture-of-experts[C]//Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 2025: 8184–93.

[34] Xu A, Tang K, Xu Y, et al. Towards a universal foundation model for biobank-scale human genome variation. bioRxiv, 2025, https://doi.org/10.1101/2025.01.29.635579.

[35] Li X, Xu ZQJ, Ren Y, et al. Towards reliable pediatric brain tumor segmentation: task-specific nnU-net enhancements. arXiv, 2025, https://doi.org/10.48550/arXiv.2511.00449.

[36] Strick D J, Garcia C, Huang A, et al. Reproducing and improving chexnet: deep learning for chest X-ray disease classification. arXiv, 2025, https://doi.org/10.48550/ arXiv.2505.06646.

[37] Hingerl JC, Martens LD, Karollus A, et al. scooby: modeling multimodal genomic profiles from DNA sequence at single-cell resolution. Nat Methods, 2025, 22: 2275–85.

[38] Tejada-Lapuerta A, Schaar AC, Gutgesell R, et al. Nicheformer: a foundation model for single-cell and spatial omics. Nat Methods, 2025, 22: 2525–38.

[39] Wang Z, Jin Q, Wei CH, et al. GeneAgent: self-verification language agent for gene-set analysis using domain databases. Nat Methods, 2025, 22: 1677–85.

[40] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020, 36: 1234–40.

[41] Yasunaga M, Leskovec J, Liang P. LinkBERT: pretraining language models with document links[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 2022: 8003–16.

[42] Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform, 2022, 23: bbac409.

[43] Hwang H, Jeon H, Yeo N, et al. Big data and deep learning for RNA biology. Exp Mol Med, 2024, 56: 1293–321.

[44] Zhao M, Wang N, Jiang X, et al. An integrative data-driven model simulating *C. elegans* brain, body and environment interactions. Nat Comput Sci, 2024, 4: 978–90.

[45] Armeni P, Polat I, De Rossi LM, et al. Digital twins in healthcare: is it the beginning of a new era of evidence-based medicine? A critical review. J Pers Med, 2022, 12: 1255.

[46] De Benedictis A, Mazzocca Z, Somma A, et al. Digital twins in healthcare: an architectural proposal and its application in a social distancing case study. IEEE J Biomed Health Inform, 2023, 27: 5143–54.

[47] Sun T, He X, Song X, et al. The digital twin in medicine: a key to the future of healthcare? Front Med (Lausanne), 2022, 9: 907066.

[48] Ladune T, Philippe P. AIVC: artificial intelligence based video codec[C]//2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022: 316–20.

[49] Toscano E, Cimmino E, Boccia A, et al. Cell populations simulated *in silico* within SimulCell accurately reproduce the behaviour of experimental cell cultures. NPJ Syst Biol Appl, 2025, 11: 48.

[50] Lu W, Du X, Wang J, et al. Simulation and assimilation of the digital human brain. Nat Comput Sci, 2024, 4: 890–8.

[51] Long M, Wang J, Ding G, et al. Adaptation regularization: a general framework for transfer learning. IEEE Trans Knowl Data Eng, 2013, 26: 1076–89.

[52] Sapoval N, Aghazadeh A, Nute MG, et al. Current progress and open challenges for applied deep learning across the biosciences. Nat Commun, 2022, 13: 1728.

[53] Oğuztüzün Ç, Gao Z, Li H, et al. KGiA: drug repurposing through disease-aware knowledge graph augmentation. J Biomed Inform, 2025, 168: 104857.

[54] Li H, Shan Z, Fu H, et al. PRSNet-2: end-to-end genotype-to-phenotype prediction via hierarchical graph neural networks. bioRxiv, 2025, https://doi.org/10.1101/2025.11.22.689899.

[55] Baranger DAA, Hatoum AS, Polimanti R, et al. Multi-omics cannot replace sample size in genome-wide association studies. Genes Brain Behav, 2023, 22: e12846.

[56] Guo Z, Liu J, Wang Y, et al. Diffusion models in bioinformatics and computational biology. Nat Rev Bioeng, 2024, 2: 136–54.

[57] Cao H, Tan C, Gao Z, et al. A survey on generative

diffusion models. IEEE Trans Knowl Data Eng, 2024, 36: 2814–30.

[58] Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. arXiv, 2019, https://doi.org/10.48550/arXiv.1912.01703.

[59] Wang Z, Lin S, Zou Q, et al. NicheTrans: spatial-aware cross-omics translation. bioRxiv, 2024, https://doi.org/10.1101/2024.12.05.626986.

[60] Arun A. DNA foundation models and their applications [EB/OL]. (2025-09-05)[2026-02-06]. https://www.aditharun.com/p/dna-foundation-models.

[61] Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods, 2024, 21: 1470–80.

[62] Liu X, Li G, Liao Y, et al. A Geneformer-based model for identifying early gastric cancer therapeutic targets [C]//2025 2nd International Conference on Algorithms, Software Engineering and Network Security (ASENS), Guangzhou, China, 2025: 139–43.

[63] Yang F, Wang W, Wang F, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell, 2022, 4: 852–66.

[64] Yang T, Wang YY, Ma F, et al. Build the virtual cell with artificial intelligence: a perspective for cancer research. Mil Med Res, 2025, 12: 4.

[65] Xiong H, Chu C, Fan L, et al. The digital twin brain: a bridge between biological and artificial intelligence. Intell Comput, 2023, 2: 0055.

[66] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell, 2019, 1: 206–15.

[67] Kim H, Choi E, Shim Y, et al. PriorCCI: interpretable deep learning framework for identifying key ligand-receptor interactions between specific cell types from single-cell transcriptomes. Int J Mol Sci, 2025, 26: 7110.

[68] Kang C, Song H, Xu W. Structural and computational biology: compete or complement? J Med Chem, 2025, 68: 24721–3.

[69] Fu C, Chen Q. The future of pharmaceuticals: artificial intelligence in drug discovery and development. J Pharm Anal, 2025, 15: 101248.