DOI: 10.13376/j.cbls/2025148

文章编号：1004-0374(2025)12-1493-12

杨帆，山东大学教授、博士生导师。山东省泰山学者青年专家，国家健康医疗大数据研究院副院长，科技部火炬人才项目会评专家，国家数据局“数据要素”评审专家，工业和信息化部生物医药产业人才岗位评审专家。兼任 *Biomedical Informatics* 期刊副主编、*Statistics Innovation* 期刊副主编、中国计算机学会数字医学分会常委、山东省预防医学会健康大数据与数字智能公共卫生分会主委。获科技部关于“抗击新冠科技攻关”书面嘉奖、教育部工程技术奖二等奖。主要研究方向：基于因果推断驱动的可信深度学习、肿瘤信息学。先后以负责人主持国家重点研发计划课题 / 子课题等国家 / 省部级项目 10 项，发表 SCI/EI 论文 30 余篇。授权 / 受理国家发明专利 25 项、行业标准 / 团体标准 2 项。

基于生物学机制驱动的人工智能与基础模型

杨　帆

(山东大学公共卫生学院，济南 250012)

摘　要：本综述审视了系统生物学中机理建模与先进机器学习的融合，我们将其定义为“知识注入式学习”(knowledge-infused learning) 范式。我们剖析了四种将生物学知识整合到计算模型中的主要模式：(1) 使用神经普通微分方程 (neural ordinary differential equations) 和物理信息神经网络 (physics-informed neural networks) 编码连续动态过程；(2) 利用基于生物通路的图神经网络 (graph neural networks) 表示结构关系；(3) 通过因果发现算法从观测数据中推断有向相互作用；以及 (4) 通过大规模、自监督的基础模型 (foundation models) 学习“生物学语言”。对于每种模式，我们都分析了其基础方法论，重点介绍了里程碑式的应用，并对其基本假设和实践局限性进行了严格的批判。我们认为，尽管每种方法都为特定挑战 (如处理不规则时间序列数据、整合多组学数据集或生成新颖假设) 提供了强大的解决方案，但它们的真正潜力并非孤立存在，而是作为统一的神经 - 符号框架的组成部分时才能得以释放。本综述最后综合了这些主题，并为构建整合动态、结构、因果和大规模学习表征的混合模型指明了方向，旨在超越单纯的预测，实现真正的机理洞察。

关键词：生物学机制驱动；人工智能；基础模型；生物医学大模型；多模态学习；因果推理；生成式模型；计算生物学

中图分类号：Q-31；TP18　　**文献标志码：**A

Biology-driven artificial intelligence and foundation models

YANG Fan

(School of Public Health, Shandong University, Jinan 250012, China)

Abstract: This review examines the integration of mechanistic modeling and advanced machine learning in systems biology, which we define as the "knowledge-infused learning" paradigm. We dissect four primary modes of

收稿日期：2025-10-23；**修回日期：**2025-12-12

基金项目：国家自然科学基金面上项目(82273736)

通信作者：E-mail: fanyang@sdu.edu.cn

integrating biological knowledge into computational models: (1) encoding continuous dynamic processes using neural ordinary differential equations and physics-informed neural networks; (2) representing structural relationships with graph neural networks based on biological pathways; (3) inferring directed interactions from observational data via causal discovery algorithms; and (4) learning the "biological language" through large-scale, self-supervised foundation models. For each mode, we analyze its underlying methodology, highlight landmark applications, and provide a critical assessment of its fundamental assumptions and practical limitations. We argue that while each approach offers powerful solutions to specific challenges (such as handling irregular time-series data, integrating multi-omics datasets, or generating novel hypotheses), their true potential lies not in isolation but as components of a unified neuro-symbolic framework. This review concludes by synthesizing these themes and charting a course for building hybrid models that integrate dynamic, structural, causal, and large-scale learning representations, aiming to move beyond mere prediction and achieve genuine mechanistic insight.

Key words: biology-driven mechanisms; artificial intelligence; foundation models; biomedical large models; multimodal learning; causal inference; generative models; computational biology

1 引言：现代生物学中数据与机制的交汇

1.1 现代系统生物学的两大支柱

现代生物学的核心张力与机遇在于，一方面是高通量多组学技术带来的前所未有的“大数据”爆炸式增长[1]，另一方面是长期以来基于假设驱动的机理建模传统[2]。这两种范式通常被独立研究，但如今正逐渐融合，开创了科学机器学习 (scientific machine learning, SciML) 的新前沿[3]。机器学习的优势在于从高维数据中进行统计模式识别，而这恰恰是机理建模的弱点 (难以扩展、参数化困难)；反之亦然，这使得它们的整合成为一个极具前景的方向[3]。

1.2 定义“知识注入式学习”(KiL)

我们正式引入本综述的核心主题：知识注入式学习 (knowledge-infused learning, KiL)。它被定义为一系列方法论，旨在将人类产生的先验知识 —— 无论是物理定律、生物网络拓扑、因果假设还是程序性指南 —— 明确地整合到机器学习模型的架构或训练过程中[4]。这种方法超越了纯粹数据驱动的“黑箱”模型，旨在创建更稳健、数据效率更高且更具可解释性的系统[5]。我们将简要介绍不同层次的知识注入 —— 浅层、半深层和深层 —— 为后续的讨论构建框架[5]。

这种融合并非仅仅是技术上的改进，更是一种哲学上的和解。它体现了科学领域对纯粹统计性“黑箱”人工智能的普遍反思，转而青睐于结合了深度学习的感知能力与符号知识的推理能力的混合模型。科学机器学习[3]、神经 - 符号人工智能[4] 和知识注入式学习[5] 等看似不同的领域，实际上都反映了这一根本性的转变。这一趋势解决了生物学研究中长期存在的、介于数据不可知论的假设驱动方法与假设自由的“发现科学”方法之间的紧张关系[6]。KiL 提供了一个框架，使得大规模数据分析能够被现有的生物学知识明确引导和约束，从而形成一个良性循环，而非二元对立。

1.3 知识注入的四类分类法

我们提出，系统生物学中的 KiL 主要通过四种不同但相互关联的方式体现，每种方式都针对生物系统的一个基本方面：(1) 动态知识：建模系统如何随时间变化；(2) 结构知识：表示组件之间的关系和相互作用；(3) 因果知识：推断有向的、因果效应机制；(4) 数据驱动的先验：从海量的、无标签的数据语料库中学习基础生物学原理。这个分类法将作为本文主体部分的路线图 (表 1)。

表1 系统生物学中知识注入方法的比较分类

范式	核心方法论	知识来源	主要应用	关键挑战/局限性
动态知识	神经ODE/PINN	微分方程	建模不规则时间序列	刚性问题/模型错误设定
结构知识	图神经网络(GNN)	通路数据库(KEGG、STRING)	多组学整合	图构建偏见
因果知识	基于约束/评分的算法	条件独立性约束	假设生成	不可检验的假设
数据驱动先验	Transformer/自监督学习	大规模无标签序列/组学语料库	零样本预测	可解释性/幻觉

2　建模连续生物动态：从离散步骤到连续流

2.1　动态生物数据的挑战

生物过程，如疾病进展或代谢调控，本质上是随时间演变的动态系统[2]。对这些过程进行建模面临着独特的挑战，尤其是当处理真实世界的临床时间序列数据时。这类数据通常是稀疏的、不规则采样的，并且含有噪声[7]。传统的序列模型，如循环神经网络 (RNN)，其设计假设是数据在离散且规则的时间步长上采样，因此在处理不规则的临床数据时会遇到很大困难[7]。

2.2　神经普通微分方程：学习动态过程

神经普通微分方程 (NODEs) 为连续时间建模提供了一个根本性的范式转变[7]。其核心思想是用一个神经网络来参数化一个隐藏状态的导数，从而学习系统动态的连续时间模型[7]。模型直接从数据中学习向量场，其中 $z(t)$ 是时间 t 的隐藏状态，而 $f(\cdot)$ 是由参数 θ 定义的神经网络[7]。

这一框架具有显著优势。首先，它天生能够处理不规则采样的数据，因为动态过程是连续定义的，数据点可以在任意时间点用于更新模型[7]。其次，通过伴随灵敏度方法 (adjoint sensitivity method)，NODEs 可以在训练期间实现恒定的内存成本，这解决了深度模型的一个主要瓶颈[8]。此外，模型可以在数值精度和计算速度之间进行权衡[8]。这些特性使其在时间序列重建和外推等任务中表现出色[8]。

为了更好地处理复杂数据，研究人员开发了多种架构变体。例如，潜在常微分方程 (latent ODEs、LODE-VAE) 采用编码器 - 解码器结构，在低维潜在空间中学习动态过程，这有助于从高维噪声数据中提取有意义的动态特征[8]。另一类模型，如 LT-NODE，通过将积分的终止时间视为一个潜在变量来捕捉模型的不确定性，这对于医疗等高风险应用至关重要[9, 10]。

NODEs 虽能通过学习高维向量场刻画系统连续演化，理论逼近能力强，但在真实生物数据场景中存在局限：一是多组学、临床纵向数据样本有限、观测稀疏且噪声大，易导致参数化向量场过拟合，产生不合理动态，且系统可识别性不足，“学到的动力学”不唯一，削弱解释力；二是生物系统多具多时间尺度和刚性特征，ODE 求解需极小步长显式求解器或高开销隐式自适应求解器，训练与内存成本高，伴随法反向积分还可能面临数值不稳定、梯度爆炸或消失问题，限制其可扩展性与实用性。

2.3　物理与生物信息神经网络(PINNs/BINNs)：约束动态过程

与 NODEs 从数据中学习全部动态不同，物理信息神经网络 (PINNs) 或生物信息神经网络 (BINNs) 采取了一种注入先验知识的方法[11]。这类模型将已知的物理或生物学定律 —— 通常以偏微分方程 (PDEs) 或普通微分方程 (ODEs) 的形式 —— 直接嵌入到神经网络的损失函数中，作为一个“物理损失”项[11]。这种方法的主要优点在于提高了数据效率，尤其是在生物学研究中常见的小数据场景下[12]。通过物理约束，模型不再仅仅依赖于稀疏的数据点，而是被引导去学习符合已知科学原理的解。此外，由于模型被约束遵循已知的机理，其可解释性也得到了增强[13]。这不仅使得模型能够进行预测，还能够用于发现或估计已知方程中的未知参数[11]。

一个典型的应用案例是血糖 - 胰岛素动态调节。生物学家已经建立了描述这一过程的“最小模型”，它由一组常微分方程构成[14]。一个 BINN 可以利用这个已知的模型结构，从稀疏的患者血糖和胰岛素测量数据中学习出个体化的模型参数，从而实现对患者代谢状态的个性化建模。

PINNs/BINNs 的局限性主要包括：依赖先验机理模型的结构正确性，基础方程、边界条件或源项设定偏差会导致观测外预测出现系统性偏差；需精细平衡多类损失项权重，生物场景中高噪声、有限样本的特点会加剧权重选择敏感性，易引发训练收敛问题；面对高维、多尺度生物 PDE 时，依赖自动微分导致计算成本高昂，且对初始和边界条件极为敏感；即便 BINNs 可通过子网络学习未知机理项，仍难以规避控制方程结构与真实生物过程存在根本性偏差带来的预测能力限制。

2.4　灵活性与保真度以及刚性问题的挑战

NODEs 和 PINNs 代表了动态系统知识注入的两种对立哲学。NODEs 提供了最大的灵活性，理论上可以学习任何动态系统，但这也带来了过拟合和学习到生物学上不合理动态的风险。它执行的是一种后验知识发现 (学习到的动态本身就是知识)。相比之下，PINNs/BINNs 强制模型遵循生物学现实，但其性能受限于先验知识的准确性；如果预设的“物理”模型是错误的，那么网络的预测将从根本上受到限制。它应用的是先验知识约束 (已知的方程指导学习)。

在实际应用中，一个常被忽视但至关重要的挑战是动力系统的“刚性”问题[15]。当一个系统中存

在发生在截然不同时间尺度上的过程时 (例如，快速的酶促反应与缓慢的基因表达变化)，该系统就被认为是刚性的 [16]。刚性为 NODEs 中使用的常微分方程求解器带来了巨大的数值挑战，通常需要计算成本高昂的隐式求解器，并可能导致不稳 [7]。这一实际约束是阻碍原生 NODEs 在系统生物学中广泛应用的主要障碍之一。为缓解这一问题，实践中通常结合两类策略：在数值层面，优先选择适合刚性系统的隐式或自适应步长求解器，配合误差控制和检查点技术，以在稳定性与计算成本之间取得平衡；在建模层面，则通过对快慢过程进行显式分解，仅用神经网络去拟合相对平滑的慢过程或校正项，将刚性最强的部分保留在结构明确的机理子模型中，从而减轻学习到的向量场的难度，这在药效建模和复杂信号通路建模中已被证明可以显著改善训练稳定性和求解效率。

这两种方法的局限性共同指向了一个更有前景的未来方向：混合“灰箱”模型。通用微分方程 (universal differential equations, UDEs) 框架为此提供了一个范例 [17]。在这种框架下，神经网络只学习系统的未知部分，以补充一个已知的机理骨干。灰箱 UDE 融合机理模型与神经网络优势：在葡萄糖 - 胰岛素模型中，ODE 描述基础动力学，神经网络学习未建模异质性等校正项，兼顾血糖拟合预测精度与参数生理可解释性；在细胞迁移和组织修复模型中，PDE 保留扩散与基本反应结构，神经网络替代难建模的增殖率等函数，凭稀疏数据反推形式，更精准重现细胞密度变化并提升预测力。其刚性集中于机理子系统，可通过成熟求解器处理，神经网络校正部分设计温和，缓解训练刚性问题。这种方式统一了 NODEs 的灵活性与 PINNs/BINNs 的保真度，为知识不完全的复杂生物系统提供稳健的连续动态建模方案。

2.5 生物时间序列任务中NODEs与PINNs泛化能力的实证

在已有工作中，NODEs 和 PINNs 在多个具体生物系统中已经给出了较为扎实的“时间序列重建与短期外推”实证结果，但这些证据主要仍局限于系统内部的泛化。例如，在单细胞水平，scNODE[18]、RNAForecaster[19] 和 DeepVelo[20] 等方法在发育过程与细胞周期数据上，能够从部分时间点的观测出发重建未测时间点的基因表达分布和细胞轨迹，并在 Wasserstein 距离、相关系数等指标上优于传统方法；在基因调控网络和药代动力学建模中，Biologically informed NeuralODEs[21] 以及基于 Neural ODE[22] 的 PK 模型，可以在不同实验条件或给药方案之间实现对未观测时间窗和新剂量方案的浓度 – 时间曲线预测，表现出相对于 LSTM、NLME 等基线更好的外推能力；在生态动力学中，Neural ODE 及其变体能够从种群密度时间序列中恢复底层动力系统，并在未来时间段预测上达到或优于经典微分方程与时间序列模型的精度。对于 PINNs，一方面，在生理信号建模中，基于 PINNs 的无袖带血压估计在跨时段、跨诱导条件和跨个体的测试中，相比纯数据驱动的深度模型保持了更低的误差和更稳定的泛化表现 [23]；另一方面，在肿瘤生长、基因表达和流行病动力学等 ODE 场景，以及基于 PI-SDE 的单细胞动力学建模中，PINNs 物理约束 SDE 能在不同初值和参数下稳定重建轨迹并改善对未观测时间点和扰动条件的预测 [17]。然而，这些研究大多是在针对每一种具体生物系统单独训练一个模型的前提下考察泛化能力，目前仍缺乏统一基准下、系统性比较 NODEs/PINNs 在跨不同类型生物系统之间的迁移与泛化性能的大规模实证评估，因此我们在文中明确将上述结论限定在“各自任务和数据集范围内”的经验性证据，而不是将其简单推广为对所有生物系统的普适性结论。

在模型可解释性方面，已有多项具体工作给出了 PINNs/BINNs 在真实生物任务中的实证证据：在细胞迁移中，BINNs 能从划痕实验时间序列中自动学习出与密度依赖扩散和增殖相符的反应 – 扩散项，用以构建可解释 PDE/ABM 模型 [24]；在肿瘤生长和胶质瘤反应 – 扩散建模中，PINNs 由体积或影像时间序列估计的增殖率、扩散系数等参数既能重现实测曲线，其数值范围也与文献报道相符，可直接解释肿瘤扩散速度和侵袭性差异 [25]；在多组学预测中，biology-informed 网络通过编码通路结构，在提升性状预测精度的同时，显式突出与目标性状显著相关的通路和基因簇，为后续功能验证提供了具体候选 [26]。这些结果表明，物理 / 生物约束确实在多种实际场景中转化为可检验、可利用的解释性，但系统比较不同先验设定对解释稳定性及实验可指导性的工作仍相对不足。

3 编码生物结构：用于多组学整合的图神经网络

3.1 以网络为中心的生物学视角

生物学在根本上是一门网络科学 [27]。大量的

公共数据库系统地编码了这一知识 (表 2)，例如，京都基因与基因组百科全书 (KEGG) 记录了代谢通路 [28]，Reactome 数据库详细描述了反应和通路 [29]，而 STRING 数据库则汇集了蛋白质 - 蛋白质相互作用 (PPIs)[30]。这些数据库提供了结构化的生物学知识，是注入到图神经网络 (GNNs) 中的主要信息来源。

3.2　用于生物数据的图神经网络(GNNs)

图神经网络是一类专为处理非欧几里得、图结构化数据而设计的深度学习模型 [31]。其核心操作是“消息传递”，即节点通过迭代地聚合其邻居节点的信息来学习自身的嵌入表示。这些嵌入同时捕获了节点的特征和图的拓扑结构 [31]。

近期工作越来越多地转向异构图，用以同时表示不同层级、不同类型的生物实体及其关系。例如，在同一个图中同时纳入基因、蛋白质、细胞类型、药物或临床表型节点，并通过“基因 - 蛋白质”“蛋白质 - 药物”“细胞 - 表型”等多种边类型刻画其相互作用。此类异构图能够更自然地融合多组学、PPI、药物靶点和临床信息，为多任务联合建模提供结构基础。对应地，异构 GNN 可以对不同关系赋予差异化权重，自动学习哪类跨模态连接对下游任务最为关键，为多模态知识整合提供了比单一同质图更灵活的建模范式。

另一方面，围绕 GNN 可解释性的研究也在快速发展。图解释性模型通常通过节点重要性评分、通路级掩蔽、注意力权重或反事实子图等方式，给出“模型为何做出某个预测”的局部解释。在生物医学场景中，这类方法可用于从训练好的 GNN 中自动提取与疾病标签、药物敏感性或预后风险高度相关的关键基因子网、通路模块或细胞间相互作用模式，为后续的机制研究和实验验证提供候选假设。将异构图建模与图解释性技术结合，有望在提高预测性能的同时，显式给出跨模态、跨层级的可视化因子和可检验假设，更好地服务于生物学发现。

文献中讨论了几种关键的 GNN 架构 [32]：(1) 图卷积网络 (GCNs)：通过聚合邻域信息来更新节点表示，对于捕捉局部模式非常有效 [33]。(2) 图注意力网络 (GATs)：引入注意力机制，使模型能够学习不同邻居节点的重要性，从而在异构图上表现更佳 [32]。(3) 图 Transformer 网络 (GTNs)：将 Transformer 架构应用于图数据，能够捕捉图中节点之间的长程依赖关系 [32]。

3.3　应用：用于疾病分型的多组学整合

整合多组学数据 (如转录组、蛋白质组、甲基化组) 以获得对癌症或 2 型糖尿病等复杂疾病的整体视图是一个巨大的挑战 [1]。简单地将不同组学的特征拼接在一起，往往无法捕捉到组学间复杂的相互关系 [33]。

GNN 为此任务提供了自然的解决方案。在模型中，节点可以代表基因或蛋白质，而边则可以代表来自数据库的已知相互作用 (知识驱动) 或从数据中学到的关系 (数据驱动)[32]。GNN 随后学习强大的、整合的节点表示，用于下游任务，如患者分类或生物标志物发现 [34]。研究一致表明，基于 GNN 的多组学模型性能优于单组学方法 [32]。

3.4　图构建的首要性

在应用 GNN 时，最重要也往往是最具任意性的一步是定义图的结构。这一选择对模型性能有深远影响，并引入了显著的归纳偏见。主要有两种方法：(1) 知识驱动的图：直接使用 PPI 网络或 KEGG 通路作为图的邻接矩阵。这种方法直接注入了生物学知识 [32]。(2) 数据驱动的图：基于样本间的相似性 (如皮尔逊相关系数) 构建图 [32]。

我们建议从三个方面权衡图构建策略：(1) 任务目标：若侧重预测性能且对可解释性要求较低，可偏向数据驱动或自动图结构学习；若强调机制解释与通路假设，则更宜采用知识驱动或知识 – 数据混合图。(2) 数据特点：样本量小、噪声大的队列宜以 STRING、KEGG 等知识图为骨架，并用数据

表2　用于知识注入建模的公共数据库和队列研究

类别	资源	数据类型	主要用例	访问方式
生物知识图谱	KEGG、STRING、Reactome	代谢通路、蛋白质-蛋白质相互作用(PPIs)、反应网络	定义GNN的图结构，通路富集分析	API、下载格式(BioPAX、SBML等)[45]
大规模观测队列	UKBiobank、AllofUs、iPoP	纵向多组学(基因组、蛋白质组、代谢组)、电子健康记录(HER)、可穿戴设备、调查问卷	训练动态模型(NODEs)、因果发现、疾病风险预测(如T2D)	分级访问模型、数据标准化(如OMOP-CDM)[46]

进行适度增强或加权；样本量大、表型高度异质时，从数据学习图结构更有利于发现跨通路新互作。(3) 研究阶段：早期探索可更多依赖数据驱动图以拓展新模式空间，后期验证则转向较保守的知识图或混合图以进行精炼建模和机制巩固。

现有研究的证据是矛盾的：一些研究发现基于相关性的图在识别癌症特征方面更优[32]，而另一些研究则通过通路知识取得了成功[27]。这表明 GNN 在多组学中的成功，与其说是神经网络架构的复杂性 (GCNvs.GAT)，不如说是注入的图结构的质量和相关性。知识注入主要发生在数据表示阶段，而非消息传递过程。GNN 本质上是一个强大的特征提取器，它学习利用研究者提供的关系归纳偏见。然而，使用预定义的知识图 (如 KEGG) 进行发现存在一个根本性的矛盾。如果我们构建一个受限于已知通路的模型，它在重新发现这些通路内的信号方面会表现得异常出色。但是，它在结构上被蒙蔽了，无法发现数据库中不存在的全新通路或交互。这造成了一种“路灯效应”，即我们只在已经有光的地方寻找钥匙。数据驱动的图 (如样本相关性) 摆脱了这种束缚，已有多项实证研究直接采用数据驱动的样本相似图来进行多组学整合，并给出了量化证据：MOGONET 通过各组学内部的样本相似度构建患者图，在多种癌症队列上稳定优于多种非图基线方法，并能识别具有生物学意义的标志物[35]；DeepMoIC 基于相似网络融合构建患者相似网络，在多套 TCGA 癌症亚型数据上，深层 GCN 一致优于现有多组学整合模型[36]；MOGLAM 等方法进一步从多组学特征中自适应学习患者关系图，在三个癌症数据集上整体优于一系列主流算法[37]；而系统比较研究显示，用相关矩阵构建的相关性图在癌症类型的分类任务中，普遍优于基于 PPI 的知识图结构，为“数据驱动图在多组学整合中是有效且具有竞争力的选择”提供了直接的实证支持，但其缺点是失去了生物学的可解释性。与其试图在综述中给出哪种构图方式更优的简单结论，本工作更倾向于提出一个可操作的实验规划思路：在统一的基准框架下固定 GNN 主干与训练策略，仅改变图构建方式，系统比较纯知识图、纯相似度图，以及知识图与相似度图的混合或可学习图 (例如在知识图上增加数据驱动边、或用可学习权重标定边强度)。在单细胞谱系推断、多组学疾病分型、药物响应预测等代表性任务中，通过交叉验证和消融实验同时考察预测性能、结果稳定性以及关键通路和枢纽基因的一致性，从而在任务、数据类型和构图策略的空间中逐步勾勒出哪类图在何种情境下更为有效。

未来的方向聚焦于利用数据驱动的边来对知识图谱进行优化或增强，从而结合两者的优点。在实现层面，知识图与数据驱动图的结合已经有若干可借鉴的范式。其一是在 KEGG 通路或 PPI 等知识图上，用多组学数据学习可调的边权或新增少量数据驱动边，对原始知识图进行重标定，例如 GNNRAI 以通路为骨架，通过多组学特征学习边权并对潜在交互进行补充，在多种 TCGA 癌症队列中优于单纯知识图或单纯数据图模型[38]。其二是采用多通道或多图结构，将通路图与样本相似度图或统计关联图并行输入 GNN，由注意力分配不同图的权重，如 Multilevel-GNN 将基因调控网络和通路图与多组学表达特征分层结合用于肿瘤生存风险预测[39]。这些工作表明，“以知识图为结构先验，再由数据驱动进行边权调整或引入补充边”，以及“知识图与数据驱动图的多通道联合建模”，是目前较为可行且具有实证支持的融合路径。

4 从观测数据中探寻因果关系

4.1 生物学中的“为什么”问题

标准机器学习主要解决预测和关联任务，但要理解疾病的病因和治疗效果，我们需要回答“为什么”的问题[40]。因果推断的目标是从数据中推断出因果效应的有向图，这对于识别治疗靶点和设计有效干预至关重要[41]。

4.2 因果发现的范式

从观测数据中学习有向无环图 (DAG) 的算法主要分为两大家族[42]：(1) 基于约束的方法：以 PC 算法为代表，这类方法利用条件独立性检验从一个全连接图中剪除边[40]。由于统计上的不可区分性，它们通常输出一个马尔可夫等价类 (即一组共享相同条件独立性关系的可能 DAGs)[43]。(2) 基于评分的方法：这类方法定义一个评分函数 (如似然度)，然后在所有可能的 DAG 空间中搜索得分最高的图。一个里程碑式的工作是 NOTEARS 算法，它巧妙地将 DAG 的组合性无环约束转化为一个连续可微的约束，从而允许使用基于梯度的优化方法进行求解[44]。

4.3 因果发现的脆弱假设

这些算法的强大能力建立在一系列脆弱且在生物学背景下往往不切实际的假设之上。违反这些假设将导致严重的后果 (表 3)[43]。(1) 因果充足性 (无未测量混杂)：该假设要求所有变量对的共同原因

都已被观测。在生物学中，这几乎永远无法满足，因为潜在变量 (如环境暴露、未知的遗传因素、表观遗传修饰) 普遍存在 [43]。(2) 无环性：该假设排除了反馈回路的存在。然而，许多生物系统，如具有反馈控制的基因调控网络，从根本上违反了这一假设 [42]。(3) 线性与噪声分布 (针对 NOTEARS)：基础的 NOTEARS 算法假设变量间存在线性关系，并且噪声通常为高斯分布，这是对复杂、非线性生物相互作用的过度简化 [47]。(4) 忠实性：该假设认为数据中所有的条件独立性都源于因果结构，但这可能因参数抵消而被违反。(5) 高维挑战：随着变量数量 (基因、蛋白质) 的增加，搜索空间变得难以处理，尽管新方法正试图解决这一问题 [43]。

为缓解因果建模假设的脆弱性，可从三方面着手：针对未测量混杂因素，既在观测数据中纳入更多已知协变量、用工具变量显式建模部分未测量混杂，也结合 CRISPR 基因扰动、药物处理等干预数据对关键因果边做外部校准；针对非线性关系，优先采用非线性因果发现框架，并通过模拟实验与反事实检验，评估不同函数族下因果方向的一致性；针对高维挑战，借助通路、共表达模块等生物先验聚类或筛选变量，先在模块层面学习稀疏因果结构再回溯至基因蛋白层，同时通过正则化、稳定选择与重采样评估边的置信度，减少高维噪声和偶然相关导致的虚假因果关系。

尽管我们在文中强调了因果发现算法在生物学场景中的诸多理想化假设，近期的系统基准也提示在特定条件下它们仍具有一定鲁棒性。例如，CausalCell 在多种单细胞数据集中系统比较 11 种因果发现算法，发现基于约束的 PC 算法结合条件独立性检验在不同细胞类型和子采样比例下性能较为稳定 [48]；CICT 在多套模拟与真实 scRNA-seq 基准上，相比十余种 GRN 推断方法在 rpAUPR 等指标上保持领先，对训练样本量和相关性度量选择相对不敏感 [49]；CausalBench 则利用大规模 CRISPR 单细胞扰动数据评估 20 余种因果结构学习方法，结果显示 SparseRC、Mean Difference 等方法在不同细胞系和干预覆盖度下表现相对稳健，而多数经典因果发现算法在真实扰动数据上的性能明显弱于其在合成数据上的表现 [50]。这些实证结果一方面为因果发现算法在不同生物系统中的鲁棒性提供了量化证据，另一方面也从反面印证了我们在解释其输出网络时保持审慎态度的必要性。

在电子健康记录数据中，因果推断方法已经在多个具体医疗场景中给出了定量证据。Doutreligne 等 [51] 以 MIMIC-ICU 患者为对象，围绕“白蛋白联合晶体液和单纯晶体液”液体复苏方案，按目标试验框架构建队列，并采用随机森林和 AIPW 的双重稳健估计，在充分控制混杂后得到的 28 天死亡率效应与随机对照试验结果高度一致，同时揭示高龄、休克患者对白蛋白更获益的异质性，为 EHR 中复制 RCT 结论提供了实例化范式。Chen 等 [52] 利用 736 例心力衰竭住院患者 EHR，提出多任务深度表征 KNN 的 ITE 估计模型，在真实数据上心衰治疗获益分类的准确率和 F1 分别达到 0.703 和 0.796，并给出与既有临床知识相符的用药获益模式，说明基于 EHR 的个体化治疗效应估计在实际决策中具有可用信号 [52]。但是在 EHR 数据中应用因果推断时，上述挑战变得更加严峻。EHR 数据本身存在选择性偏倚、信息性缺失和差异性监测等固有问题，进一步混淆了对真实因果效应的探究 [53]。

将这些算法天真地应用于观测生物数据充满了风险，可能会产生虚假的因果声明 [54]。那种认为算法是能将观测数据直接转化为因果黄金的“点金石”

表3 因果发现算法的假设与脆弱性

假设	PC算法	NOTEARS算法	生物学现实与违反后果
因果充足性	假设无潜在混杂；FCI是允许混杂的扩展	假设无潜在混杂	生物系统充满未测量的混杂因素。违反会导致伪边和错误的方向性。
无环性	输出无环图(或等价类)	通过连续约束强制无环	基因调控网络常包含反馈回路。违反会导致算法失败或错误的无环近似。
线性	不要求线性(取决于CI检验)	基础形式假设线性；存在非线性扩展但增加复杂性	生物相互作用高度非线性。违反会导致无法发现非线性关系。
噪声分布	不要求特定分布(取决于CI检验)	基础形式假设高斯噪声	生物噪声分布复杂。违反可能影响评分函数的准确性。
忠实性	假设忠实性	假设忠实性	参数抵消可能导致违反。违反会遗漏真实的因果边。

的观点，是一种危险的反模式 [54]。

相反，我们应该重新定义这些方法的效用。它们的真正价值不在于“发现”唯一的真实因果图，而在于作为强大的假设生成引擎。它们可以从数千种潜在关系中筛选出少数几个看似合理的因果联系，然后可以优先进行实验验证 [6]。这将计算方法整合到了假设 - 实验的经典科学方法中。因果发现领域本身似乎正在分化。一个分支致力于开发日益复杂、数学上更精密的算法来处理假设违规 (例如动态、非线性、潜在变量模型)[55]；另一个更务实的分支则专注于整合领域知识和实验 (干预) 数据来约束问题，使其更易于处理 [56]。这反映了纯粹的数学雄心与实际科学效用之间的张力。

在实际生物系统中，因果发现算法已经被用作假设生成引擎，并通过后续实验得到验证。CellBox 从黑色素瘤细胞的药物扰动数据出发，直接学习一个可解释的微分方程因果网络，在仅以 89 个扰动条件训练的前提下，对约 11 万个未见单药和双药组合的分子与表型响应预测与实验测量的相关系数达到 0.93，并据此提出 MEKi+c-Myc、RAFi+c-Myc 等新组合，其抑制细胞增殖的效果在后续实验中得到验证 [57]。BETS 则从糖皮质激素刺激下的转录组时间序列推断出包含 31 945 条有向边的因果网络，并利用 10 个转录因子过表达实验和 GTEx 肺组织 trans-eQTL 显著富集对网络边进行外部验证，最终得到 340 对候选远端调控位点，数量较原始 GTEx 研究提高约 170 倍，作为可优先跟进的功能验证假设集 [57]。在单细胞扰动尺度上，PerturbDB 整合 66 套 Perturb-seq 数据和化学扰动转录组，基于基因调控网络推断出 552 种潜在抑制剂靶点配对，并在后续功能实验中得到证实，说明通过因果调控网络进行靶点 – 化合物假设生成在药物发现场景中具有实际可用性 [58]。

然而，生物学中的观测性因果发现的整个事业可能从根本上受到可识别性问题的限制。如果没有能力进行干预 (或观察自然发生的干预)，许多因果结构在统计上是无法区分的 (马尔可夫等价类问题)。因此，生物学中计算因果推断的未来与高通量实验扰动技术 (如 CRISPR 筛选) 的未来密不可分。未来最有影响力的“因果算法”可能不是那些处理静态观测数据的算法，而是那些为优化实验设计而生的算法 —— 能够智能地选择下一个、信息量最大的干预措施，以最有效的方式解析出真实的因果图 [56]。

5 新前沿：基础模型与规模的力量

5.1 基础模型范式

基础模型是一类大规模模型，通常基于 Transformer 架构，使用自监督目标在海量的、广泛的、无标签的数据上进行预训练 [59]。其核心思想是“预训练 - 微调”范式：模型在预训练阶段学习一个领域的通用表示 (例如“生物学的语言”)，然后只需少量额外的训练数据即可适应众多特定的下游任务 [60]。这种方法代表了从知识注入到知识归纳的范式转变。之前的 PINNs 和 GNNs 方法需要人类将明确的知识注入模型，而基础模型则旨在从数据的巨大规模中隐式地归纳出领域的基本原理和“知识”。它们是学习规则，而不是被告知规则。

5.2 案例研究1：GeneFormer与转录组语言

GeneFormer 是一个基于 Transformer 的基础模型，它在数千万个单细胞转录组上进行了预训练 [61]。它通过掩码语言建模 (masked language modeling) 的目标来学习基因调控网络的动态，将细胞视为“句子”，基因视为“单词”[61]。GeneFormer 最强大的应用之一是零样本学习 (zero-shot learning) 和计算机模拟 (*in silico*) 扰动。该模型可以在从未见过任何扰动数据的情况下，预测基因敲除的效应。这使得研究人员能够快速筛选致病基因和潜在的治疗靶点，其中一些预测已经得到了实验验证 [61]。

5.3 案例研究2：ESMFold与蛋白质结构语言

ESM (evolutionary scale modeling) 系列是蛋白质语言模型领域的代表 [62]。我们重点关注 ESMFold，它利用从一个在数亿条蛋白质序列上训练的巨大语言模型中提取的嵌入，直接从单个氨基酸序列预测蛋白质的三维结构 [63]。与 AlphaFold[64] 相比，ESMFold 具有独特的优势。AlphaFold 严重依赖多重序列比对 (MSAs) 来识别共进化模式，而 ESMFold 在语言模型内部隐式地学习了这些模式。这使得 ESMFold 的推理速度比 AlphaFold 快几个数量级，并且对于没有已知同源物的蛋白质 (孤儿蛋白)—— 这是蛋白质宇宙中一个巨大且未被探索的部分 —— 尤其具有优势 [65]。然而，对于具有丰富进化信息 (即大量同源序列) 的蛋白质，其准确性可能低于 AlphaFold[63]。

5.4 模型参数规模的希望与风险

基础模型代表了“黑箱”建模范式的顶峰，其巨大的复杂性带来了严峻的挑战。

(1)“黑箱”问题：模型的复杂性使其难以解释，引发了对其可靠性和预测的科学有效性的担忧 [66]。

我们如何能信任一个我们不理解的模型提出的治疗靶点？

(2)“幻觉”与事实不一致：与其自然语言处理的同类模型一样，生物学基础模型也可能产生“幻觉”——生成看似合理但实际上错误的输出 (例如一个能量上不可能存在的蛋白质结构)。如果未经仔细验证，这将对科学诚信构成重大威胁 [67]。确保事实基础是当前的一个主要研究领域 [68]。

(3) 数据偏见与泛化：这些模型是在现有的公共数据上训练的，这些数据本身就存在偏见 (例如对研究充分的蛋白质或疾病的过度代表)。模型可能会学习并放大这些偏见，从而限制其向生物学研究不足领域的泛化能力 [69]。

(4) 基准测试与更简单的模型：在不同生物系统上，预训练 – 微调范式已经通过系统基准得到了量化验证。从单细胞转录组出发，Geneformer 在约 3 000 万个单细胞转录组上预训练后，仅用有限任务特异数据微调，便能在多种与染色质和网络动力学相关的下游任务中稳定提升预测准确率，并在心肌病小样本患者数据上成功优先排序出候选治疗靶点，体现了在跨组织、跨任务场景下的迁移能力 [70]。但是一个关键问题是，基础模型的巨大规模和计算成本是否总是必要的。最近的批判性评论表明，在某些情况下，更简单、更传统的机器学习模型可以胜过或匹敌大型基础模型的性能，尤其是在微调数据充足时 [71]。在基因组序列建模方面，HyenaDNA 利用长序列建模架构，在 Nucleotide Transformer 的 18 个下游任务中，以远少于 Transformer 的参数量和预训练数据，在 12 个任务上取得新的 SOTA，并在 GenomicBenchmarks 上平均提升约 10 个百分点的分类准确率，表明在合理的架构与归纳偏置下，中等规模基础模型在跨数据集和任务的泛化上可以匹敌甚至超过更大的 Transformer 模型 [72]。这引发了对“预训练 - 微调”范式的根本性质疑，并强调了进行严格、公平的基准测试的必要性 [59]。

为提升生物医学基础模型的可靠性，需引入系统性控制措施：验证层面，用独立外部队列数据评估关键任务性能，通过敏感性分析检验模型输出对数据和先验的稳定性；不确定性控制层面，结合置信度估计、集成模型或贝叶斯方法，对高不确定性样本拒答，生成式任务中明确区分“证据型回答”与“推断性猜测”，约束关键结论仅基于可追溯文献或知识库；可解释性层面，借助注意力权重等方法将模型决策映射到基因、通路等变量，对照既有生物知识系统比对，统计解释在外部队列中的复现性，以“解释一致性”和“可验证性”约束模型。将三者纳入闭环人机协同流程，可在保留模型表达能力的同时，降低其不确定性与错误对下游结论的影响。

为减轻数据偏见与泛化问题，建模流程需设计显式检测与纠正环节：数据准备阶段，对性别、年龄等关键变量分层统计并可视化审计，识别批次效应；训练过程中，通过样本重采样、领域对抗学习等方式剥离技术变异，弱化其在表征中的可分性；优化目标中加入子群鲁棒性约束，惩罚不同亚群的性能差距；评估阶段系统报告各队列的性能，经外部独立数据集验证，将跨队列性能一致性作为模型落地的前置条件。

GeneFormer 和 ESMFold 等模型的成功表明，生物系统中可能存在一种可从大规模数据中学习的“通用语法”。然而，它们的局限性 (幻觉、偏见) 也揭示了这种学习到的语法是不完整且有缺陷的。该领域最关键的挑战不仅仅是扩大模型规模，而是开发用于验证、解释和纠正这些模型所归纳出的知识的方法。这预示着一个“AI 辅助科学”的未来，其中基础模型扮演着才华横溢但并非完美无缺的合作者角色，提出新颖的假设 (如计算机模拟扰动)，然后必须由人类科学家进行严格的测试和验证，从而在计算和实验之间建立一个新的反馈循环。

6　综合与未来方向：迈向统一框架

6.1　综合四种模式

本节将把前述讨论的线索编织在一起。我们认为，动态 (NODEs/PINNs)、结构 (GNNs)、因果 (发现) 和大规模先验 (基础模型) 并非相互竞争，而是互为补充。一个真正智能的生物发现系统需要将这四者融为一体。例如，未来的模型可能会使用基础模型提出一个基因调控网络的草图，然后将其形式化为一个 GNN 结构的神经 ODE 来模拟动态扰动响应，并使用因果推断原则来精炼边的方向。

为落地动态、结构、因果及大规模先验的统一框架，可分三阶段循序推进：一是对接多尺度先验并模块化建模，以通路图等结构先验为底座，叠加动态信息，实现结构关系与时间依赖的联合编码；二是在统一图表示上引入可学习因果结构层，通过可学习边权等方法自动调整先验图边的方向与强度，结合已知扰动数据做因果一致性校准；三是构建可扩展预训练 – 微调流程，先基于大规模生物先

验图和无监督任务预训练，再轻量微调，叠加不确定性估计、子群鲁棒性评估模块，提升跨任务、跨系统泛化能力。

融合层面通过模块化设计兼容四者：在结构先验图上叠加动态特征，结合可学习因果边权或可干预 GNN，让模型同时利用结构约束、动态演化与因果信息。验证层面采用渐进流程：先在已知通路小图验证推理一致性，再用扰动实验数据检验因果层对实验趋势的重现性，最后依托预训练 - 微调范式在更大规模、更多任务上评估泛化能力。该分阶段路径明确了多尺度图构建、因果一致性校准、跨数据集验证等关键技术点，也指出先验冲突处理、噪声传播、尺度不一致等潜在挑战，为统一框架落地提供可操作技术路线。

6.2 神经-符号与多模态AI的兴起

我们将这一愿景与更广泛的神经 - 符号 AI[4] 和多模态深度学习 [73] 领域明确联系起来，目标是创建能够对离散的符号知识 (如通路图) 进行推理，同时从连续的高维数据 (如单细胞表达谱或医学影像) 中学习的模型。我们将强调这种整合面临的挑战，包括数据异质性、计算复杂性和对齐问题 [73]。

针对神经 - 符号 AI 与多模态深度学习应用中的技术挑战，建模流程需针对性应对：数据异质性方面，利用模态特异编码器、域对齐损失、可学习模态权重，缓解基因组、影像等多模态的分布差异，实现统一潜在空间的比对；计算复杂性方面，通过稀疏注意力、子图采样、分层图结构或参数高效微调，降低大规模知识图与高维模态带来的计算负担；符号 - 神经对齐方面，借助可学习约束、知识图嵌入、可解释注意力，兼顾符号规则遵循与神经网络表示能力。这类设计已提升模型稳定性与跨模态一致性，但仍面临符号规则噪声、模态缺失、跨尺度对齐难等挑战，未来需在可扩展表示对齐、逻辑一致性约束、跨模态鲁棒性评估上进一步完善方法论。

6.3 从预测到决策：终极目标

我们将重申，这项研究的最终目标是从描述性或预测性模型转向指导性模型 —— 即能够为达到期望结果而推荐最佳干预措施的模型 (例如，个性化治疗方案)[74]。例如，基于目标试验框架的 EHR 决策模型在液体复苏及抗凝策略中取得与 RCT 接近的一致效应；因果强化学习在脓毒症与心衰管理中展示出比临床常规更优的策略价值，并在独立人群中验证了临床获益。这些案例说明，在严格的因果与验证框架下，决策型模型具备可在临床场景中测试和验证的现实途径。但是这需要对系统有深刻的、因果的、机理性的理解，而这是任何单一的知识注入学习模式都无法单独提供的。整个知识注入学习领域正在创造一种新型的科学产物：生物系统的“计算机模拟孪生”(in silico twin) 或“数字孪生”(digital twin)[75]。这不仅仅是一个预测模型，而是一个可模拟、可交互的生物过程表示。例如，在一个患者队列的多组学数据上训练的 GNN 结构的神经 ODE，实际上就是该疾病过程的数字孪生。最终目标是将这些孪生用于个性化医疗：在给真实患者用药之前，在计算机中模拟数千种潜在的药物干预，为特定患者的数字孪生找到最佳方案。这将整个事业从数据分析重塑为虚拟生物世界的构建。

6.4 结论

构建这些强大的新模型不仅仅是一个计算挑战。它需要机器学习专家 (他们构建工具) 与生物学领域专家 (他们拥有关键知识来指导模型构建、验证输出并以科学上有意义的方式解释结果) 之间深入、持续的合作。为提升跨学科团队在复杂生物医学问题上的协同效率，可建立多阶段协作机制：项目初期由生物医学专家提出可检验的生物学问题空间，机器学习专家将其转化为可操作建模任务，避免技术驱动或题目漂移；通过嵌入式协作让算法研究者参与实验室讨论、生物学家参与模型评审，持续对齐数据特征、模型假设与实验设计；项目执行阶段搭建共享文档、版本控制、标准化数据模式及自动化分析流水线，降低跨领域沟通成本；结果解释与验证阶段遵循“模型 - 假设 - 实验”闭环，模型生成的假设经生物学家筛选可行性并设计验证实验，实验结果反哺模型迭代。未来不仅仅是 AI 为生物学服务，而是 AI 与生物学同行。

[参 考 文 献]

[1] Subramanian I, Verma S, Kumar S, et al. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights, 2020, 14: 1177932219899051

[2] Rthee S, Nilam. ODE models for the management of diabetes: a review. Int J Diabetes Dev Cries, 2017, 37: 4-15

[3] Noordijk B, Garcia Gomez ML, Ten Tusscher KHWJ, et al. The rise of scientific machine learning: a perspective on combining mechanistic modelling with machine learning for systems biology. Front Syst Biol, 2024, 4: 1407994

[4] Bhuyan BP, Ramdane-Cherif A, Tomar R, et al. Neuro-symbolic artificial intelligence: a survey. Neural Comput Appl, 2024, 36: 12809-44

[5] Gaur M. Knowledge-Infused Learning[D]. Columbia, USA: University of South Carolina, 2022
[6] Davidson EH. Genomics, "discovery science, "systems biology, and causal explanation: what really works? Perspect Biol Med, 2015, 58: 165-81
[7] Oh YK, Kam S, Lee J, et al. Comprehensive review of neural differential equations for time series analysis. arXiv, 2025, doi: 10.48550/arXiv. 2502.09885
[8] Androsov DV. Neural ordinary differential equations for time series reconstruction. Radio Electron Comput, 2023: 69
[9] Anumasa S, Srijith PK. Latent time neural ordinary differential equations. Proceedings of the AAAI conference on artificial intelligence, 2022, 36: 6010-8
[10] Wendland P, Birkenbihl C, Gomez-Freixa M, et al. Generation of realistic synthetic data using multimodal neural ordinary differential equations. NPJ Digit Med, 2022, 5: 122
[11] Lagergren JH, Nardini JT, Baker RE, et al. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. PLoS Comput Biol, 2020, 16: e1008462
[12] Shukla K, Xu M, Trask N, et al. Scalable algorithms for physics-informed neural and graph networks. Data-Centric Eng, 2022, 3: e24
[13] Li Z. A review of physics-informed neural networks. Appl Comput Eng, 2025, 133: 165-73
[14] Kumnungkit K, Likasiri C, Pongvuthithum R, et al. Universal minimal model for glucose-insulin relationship with the influence of food dynamic. Comput Mathemat Methods Med, 2022, 2022: 8990767
[15] Caldana M, Hesthaven JS. Neural ordinary differential equations for model order reduction of stiff systems. Int J Numer Meth Eng, 2025, 126: e70060
[16] van Lent P, Bunkova O, Magyar B, et al. Jaxkineticmodel: neural ordinary differential equations inspired parameterization of kinetic models. PLoS Comput Biol, 2025, 21: e1012733
[17] Arroyo-Esquivel J, Klausmeier CA, Litchman E. Using neural ordinary differential equations to predict complex ecological dynamics from population density data. J Roy Soc Interface, 2024, 21: 20230604
[18] Zhang J, Larschan E, Bigness J, et al. scNODE: generative model for temporal single cell transcriptomic data prediction. Bioinformatics, 2024, 40: ii146-54
[19] Erbe R, Stein-O'Brien G, Fertig EJ. Transcriptomic forecasting with neural ordinary differential equations. Patterns, 2023: 100793
[20] Chen Z, King WC, Hwang A, et al. DeepVelo: single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. Sci Adv, 2022, 8: eabq3745
[21] Hossain I, Fanfani V, Fischer J, et al. Biologically informed NeuralODEs for genome-wide regulatory dynamics. Genome Biol, 2024, 25:127
[22] Lu J, Deng K, Zhang X, et al. Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens. Iscience, 2021, 24: 102804
[23] Bräm DS, Nahum U, Schropp J, et al. Low-dimensional neural ODEs and their application in pharmacokinetics. J Pharmacokinet Pharmacodyn, 2024, 51:123-40
[24] Nardini JT. Forecasting and predicting stochastic agent-based model data with biologically-informed neural networks. Bull Math Biol, 2024, 86: 130
[25] Rodrigues JA. Using physics-informed neural networks (PINNs) for tumor cell growth modeling. Mathematics, 2024, 12: 1195
[26] Kontolati K, Gladstone RJ, Davis I, et al. Biology-informed neural networks learn nonlinear representations from omics data to improve genomic prediction and interpretability. arXiv, 2025, doi: 10.48550/arXiv: 2510.14970
[27] Selby D, Jakhmola R, Sprang M, et al. Visible neural networks for multi-omics integration: a critical review. Front Artif Intell, 2025, 8: 1595291
[28] KEGG: Kyoto Encyclopedia of Genes and Genomes, accessed October 13, 2025
[29] Details Panel - Reactome Pathway Database, accessed October 13, 2025
[30] API- STRING Help, accessed October 13, 2025
[31] Ahmad W, Tayara H, Chong KT. Attention-based graph neural network for molecular solubility prediction. ACS Omega, 2023, 8: 3236-44
[32] Alharbi F, Vakanski A, Zhang B, et al. Comparative analysis of multi-omics integration using graph neural networks for cancer classification. IEEE Access, 2025, 13: 37724-36
[33] Li B, Nabavi S. A multimodal graph neural network framework for cancer molecular subtype classification. BMC Bioinform, 2024, 25: 27
[34] Tanvir RB, Islam MM, Sobhan M, et al. Mogat: a multi-omics integration framework using graph attention networks for cancer subtype prediction. Int J Mol Sci, 2024, 25: 2788
[35] Wang T, Shao W, Huang Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun, 2021, 12: 3445
[36] Wu J, Chen Z, Xiao S, et al. DeepMoIC: multi-omics data integration via deep graph convolutional networks for cancer subtype classification. BMC Genomics, 2024, 25: 1209
[37] Ouyang D, Liang Y, Li L, et al. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. Comput Biol Med, 2023, 164: 107303
[38] Tripathy RK, Frohock Z, Wang H, et al. Effective integration of multi-omics with prior knowledge to identify biomarkers via explainable graph neural networks. NPJ Syst Biol Appl, 2025, 11: 43
[39] Yan H, Weng D, Li D, et al. Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. Brief Bioinform, 2024, 25: bbae184

[40] Huber M. An introduction to causal discovery. Swiss J Economics Statistics, 2024, 160: 14
[41] Raghu VK, Poon A, Benos PV. Evaluation of causal structure learning methods on mixed data types. PMLR, 2018, 92: 48-65
[42] Geffner T, Antoran J, Foster A, et al. Deep end-to-end causal inference. arXiv, 2022, doi: 10.48550/arXiv: 2202.02195
[43] Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet, 2019, 10: 524
[44] Lopez R, Hütter JC, Pritchard J, et al. Large-scale differentiable causal discovery of factor graphs. Adv Neural Inform Process Syst, 2022, 35:19290-303
[45] API- STRING Help, accessedOctober13, 2025
[46] Rohrer C, Gräf JF, Pielies Avelli M, et al. Unsupervised learning of multi-omics data enables disease risk prediction in the UK Biobank. bioRxiv, 2025, doi: 10.1101/2025. 10.02. 679853
[47] Causal Discovery — Causal Decision Making, accessed October 13, 2025
[48] Wen Y, Huang J, Guo S, et al. Applying causal discovery to single-cell analyses using CausalCell. Elife, 2023, 12: e81464
[49] Chevalley M, Roohani YH, Mehrjou A, et al. A large-scale benchmark for network inference from single-cell perturbation data. Commun Biol, 2025, 8: 412
[50] Yuan B, Shen C, Luna A, et al. CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. Cell Syst, 2021, 12: 128-140.e4.
[51] Doutreligne M, Struja T, Abecassis J, et al. Step-by-step causal analysis of EHRs to ground decision-making. PLoS Digital Health, 2025, 4: e0000721
[52] Chen P, Dong W, Lu X, et al. Deep representation learning for individualized treatment effect estimation using electronic health records. J Biomed Inform, 2019, 100: 103303
[53] Causal Inference in Electronic Health Record Analysis, accessed October 13, 2025
[54] PC algorithm for causal discovery from observational data without latent confounders — dodiscover v0.0.0 - PyWhy, accessed October 13, 2025
[55] Wang J, Song R. Dynamic causal structure discovery and causal effect estimation. arXiv, 2025, doi: 10.48550/ arXiv:2501.06534
[56] Shah A, Ramanathan A, Hayot-Sasson V, et al. Causal Discovery and Optimal Experimental Design for Genome-Scale Biological Network Recovery[M]//Proceedings of the Platform for Advanced Scientific Computing Conference, New York: Association for Computing Machinery, 2023: 1-11
[57] Lu J, Dumitrascu B, McDowell IC, et al. Causal network inference from gene transcriptional time-series response to glucocorticoids. PLoS Comput Biol, 2021, 17: e1008223
[58] Yang B, Zhang M, Shi Y, et al. PerturbDB for unraveling gene functions and regulatory networks. Nucleic Acids Res, 2025, 53: D1120-31
[59] Wong DR, Hill AS, Moccia R. Simple controls exceed best deep learning algorithms and reveal foundation model effectiveness for predicting genetic perturbations. Bioinformatics, 2025, 41: btaf317
[60] Foundation Models for Biomedical Research - Watershed Bio, accessed October 13
[61] Ito K, Hirakawa T, Shigenobu S, et al. Mouse-Geneformer: a deep learning model for mouse single-cell transcriptome and its cross-species utility. PLoS Genet, 2025, 21: e1011420
[62] Mapping the Landscape of AI-Enabled Biological Design - The Age of AI in the Life Sciences, accessed October 13, 2025
[63] Garcia M, Dixit SM, Rocklin GJ. Evaluating zero-shot prediction of protein design success by AlphaFold, ESMFold, and ProteinMPNN. bioRxiv, 2025, doi: 10.1101/2025.07. 29.667290
[64] AlphaFold - Google DeepMind, accessed October 13, 2025
[65] Lin Z, Akin H, Rao R, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. BioRxiv, 2022, doi: 10.1101/ 2022.07.20.500902
[66] Beyond the lab: How AI is redefining drug discovery - Abcam, accessed October 13, 2025
[67] Binz M, Alaniz S, Roskies A, et al. How should the advancement of large language models affect the practice of science? Proc Natl Acad Sci USA, 2025, 122: e2401227121
[68] Xu Z, Zhao Y, Patwardhan M, et al. Can LLMs identify critical limitations within scientific research? a systematic evaluation on AI research papers. arXiv, 2025, doi: 10.48550/arXiv: 2507.02694
[69] The Risks of Using Large Language Models in Scientific Research, accessed October 13, 2025
[70] Nguyen E, Poli M, Faizi M, et al. Hyenadna: long-range genomic sequence modeling at single nucleotide resolution. Adv Neural Inform Process Syst, 2023, 36: 43177-201
[71] Wu J, Ye Q, Wang Y, et al. Biology-driven insights into the power of single-cell foundation models. Genome Biol, 2025, 26: 1-39
[72] Theodoris CV, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. Nature, 2023, 618: 616-24
[73] Waqas A, Tripathi A, Ramachandran RP, et al. Multimodal data integration for oncology in the era of deep neural networks: a review. Front Artif Intell, 2024, 7: 1408843
[74] Dai J, Xu H, Chen T, et al. Artificial intelligence for medicine 2025: navigating the endless frontier. The Innovation Medicine, 2025, 3: 100120-1-100120-15
[75] Antonelli L, Guarracino M R, Maddalena L, et al. Integrating imaging and omics data: a review. Biomed Signal Process Control, 2019, 52: 264-80