ChatGPT

AI技术在多组学整合与基因型-环境-表型预测中的应用综述

近年来，人工智能（AI）驱动的多组学数据整合和多尺度基因型-环境-表型关系预测模型成为生物医学研究的热点 [1] [2]。该领域旨在利用深度学习、图算法和迁移学习等技术，将不同层次的组学数据（基因组、转录组、蛋白质组、代谢组等）与环境因素相结合，以预测复杂表型结果。本文针对Wu等人（2025）的综述论文《AI-driven multi-omics integration for multi-scale predictive modeling of causal genotype-environment-phenotype relationships》进行了总结，重点介绍其中应用的关键AI/机器学习技术，包括**深度学习方法**（如自动编码器、Transformer、自监督对比学习等）、**图神经网络（GNN）**、**知识图谱**、**跨物种迁移学习**、**预训练基础模型**，以及这些技术如何应对**数据稀缺**、**分布外泛化（OOD）**、**高维异构数据融合**等挑战。我们将突出这些技术的作用、代表性模型，以及它们如何组合应用于实际生物医学场景（如药物发现、疾病表型预测等）。

无监督与自监督深度学习方法

多组学数据通常维度高、噪声大且标注匮乏。无监督和自监督深度学习方法能够从大量未标注的多组学数据中学习有用模式和低维表示，在**消除批次效应**、**降维去噪**以及**跨模态对齐**等方面展现出强大能力 [3] [4]。这类方法不依赖稀缺的表型标签，通过学习数据本身的结构来提取特征，有效缓解了因**标注数据稀少**导致的建模困难 [3]。以下介绍几类重要的无监督深度学习技术：

自动编码器方法

自动编码器（Autoencoder）及其变体（尤其是**变分自动编码器**，VAE）在单细胞多组学数据分析中发挥着核心作用 [5]。它通过编码器-解码器结构将高维输入压缩到低维**潜在空间**再重构输入，实现对关键特征的提取。典型代表如**scVI**模型，使用VAE对单细胞RNA测序（scRNA-seq）表达矩阵建模，并针对批次信息和细胞测序深度等因素进行校正 [4]。scVI成功地**去除了批次效应**，生成了归一化的基因表达潜在表示，可用于差异表达分析和缺失值推断 [4]。在此基础上，**scANVI**进一步结合半监督学习，利用部分细胞类型标签来指导VAE训练，从而实现**细胞类型标签的迁移学习**和不确定度评估，对于层次化的细胞类型注释尤为有用 [6]。不过，这两种模型目前仅限于**单一模态**（RNA测序）的数据整合 [6]。

为整合多模态单细胞数据，后续出现了结合多组学的变分自动编码器模型。例如**TotalVI**利用同一细胞上同时测量的转录组和细胞表面蛋白（CITE-seq技术）数据训练联合VAE模型，能够学习RNA和蛋白质的**联合概率表示**，并校正两种模态各自的技术噪声 [7]。TotalVI在RNA部分与scVI类似，同时针对蛋白质数据显式建模了背景噪声等技术因素，从而提供降噪的多模态视图 [7]。其局限是需要**配对的RNA-蛋白测量**，且缺乏不同实验间域对齐机制 [8]。**Cobolt**和**MultiVI**则是对多模态VAE的改进：Cobolt采用了**对称多模态VAE**架构，并用**专家乘积模型（PoE）**融合多个模态的后验分布，在有配对RNA-seq和ATAC-seq数据的前提下指导未配对数据的整合 [9]。通过这种方式，Cobolt能学到单细胞转录组与染色质开放区(ATAC)的联合表示，用于下游分析 [9]。但Cobolt假设不同模态服从简单的多项分布，可能忽视模态间的生物差异 [10]。相较之下，MultiVI为不同模态引入了**各自的概率分布和噪声模型**（如对基因表达使用负二项分布，对ATAC使用伯努利分布），并通过惩罚项优化两模态的融合，从而改善潜在表示的整合效果 [11]。此外，MultiVI还能纳入细胞表面蛋白信息，使其对细胞特性的刻画更加丰富 [12]。尽管Cobolt和MultiVI有效缓解了未配对数据的整合问题，但**极端稀疏性和高噪声**仍给VAE整合带来挑战 [13]。为增强鲁棒性，**scMVP**提出**非对称多视图VAE**框架，引入**聚类一致性约束**，并结合**多头自注意力机制**和**cycle-GAN循环一致性模块**，显著提升了RNA-seq与ATAC-seq跨模态映射的稳定性 [14]。不过，scMVP仍需要单细胞同时具有多模态测量的数据来训练 [15]。另一项工作**GLUE**则在VAE中融合**先验生物网络知识**：它利用模态特异的图结构VAE，将染色质区域与基因表达的已知调控关系纳入模型，以改进特征转换过

程 [16]。GLUE不仅能学习局部邻近信息，还能捕捉全局跨基因组的关联，并通过可扩展的**对抗训练**实现了基因表达-开放染色质-DNA甲基化**三组学数据**的联合对齐 [16]。

除了用于表征学习，生成式自动编码器还被用于**预测任务**。例如，**BioLORD**模型针对**药物处理和基因扰动**下的细胞响应预测，采用了一个分解的自编码器框架：分别编码细胞已知属性（如细胞系、基因型等）和未知影响因素，构建出**解耦的潜在空间**，再由生成模块预测干预后的多组学读数 [17]。这种设计旨在**因子分解**细胞响应中的已知和未知成分，提高对未知因素（如未观测通路活动）的泛化能力 [17]。BioLORD在未见过的新药物或基因组合上表现出良好的预测性能，但作者指出仍需进一步挖掘“未知属性”潜在表示，以提升模型泛化性 [18]。又如**ChemCPA**，该模型融合了**化合物结构信息**，并利用对抗训练策略将来自**大规模体细胞转录组**的数据知识迁移到**单细胞数据**上 [19]。具体来说，ChemCPA通过编码器-解码器架构对不同属性（如药物剂量、细胞类型）进行解耦表示，并通过加入对抗损失来使模型生成的单细胞表型与真实分布接近 [19]。结果表明，该模型在**未见过的新化合物**上能够预测其引起的基因表达变化 [19]。然而，目前ChemCPA主要验证了对新药物的泛化能力，尚需进一步评估其对**未见过的新细胞类型或细胞系**的适用性 [20]。

总体而言，基于自动编码器的无监督表示学习为多组学数据融合提供了强大工具：**有效降维和去噪**，捕获不同组学间的相关结构，从而将**高维异构数据映射到统一潜在空间** [4] [16]。这在一定程度上缓解了因**样本数量有限**而难以训练深度模型的问题 [3]。不过，许多方法仍需要一定**配对的多组学数据**来指导模型学习不同模态之间的对应关系，且面对单细胞组学数据的**高稀疏、高变异**特征时，模型稳健性和生物信号提取仍是挑战 [13]。

Transformer模型与预训练基础模型

Transformer以自注意力机制为核心，在自然语言处理和计算机视觉中引领了**基础模型（Foundation Model）**的发展 [21]。类似地，在生物领域，研究者将基因组/细胞数据类比为“语言”，利用Transformer构建大规模预训练模型，从海量**单细胞组学**数据中学习通用模式 [22]。**scGPT**就是这方面的代表性工作：它采用12层Transformer架构，在超过**1000万单细胞**转录组及多组学数据上预训练，成为第一个单细胞领域的基础模型 [22]。scGPT通过**自注意力**有效学习了基因与基因之间的相互作用关系，并使用“细胞条件token”编码实验批次、测序技术等上下文信息，使模型能够捕获**特定细胞的特征** [23]。经过预训练的scGPT可以支持多种下游任务，并可在有**配对数据**的情况下实现不同组学之间的联合分析 [23]。然而，scGPT的局限在于对**未见过情形的“零样本”预测能力有限**，且需要一定配对的多组学数据来学习跨模态关联 [24]。除了scGPT之外，近期还出现了**Geneformer**、**scFoundation**等类似的大模型，皆旨在构建统一的单细胞转录组表征空间 [22]。

基础模型通过在超大规模数据上自监督预训练，学得对广泛任务通用的**表征能力**。在多组学预测中，它们缓解了**数据稀缺**问题，因为模型已从海量未标数据中获取知识，只需少量标注即可微调到特定任务 [3] [25]。值得注意的是，基础模型概念正在扩展到**跨物种**和多模态领域。例如**GeneCompass**是一种**跨物种的单细胞基础模型**：采用与scGPT类似的Transformer架构，在**超过1.2亿个人类和小鼠单细胞**转录组上联合预训练 [26]。GeneCompass将基因ID、表达值以及先验生物知识一起编码为“基因token”，学会了跨物种的细胞基因表达模式 [27]。该模型在下游可通过监督学习执行多种任务，包括**基因调控网络重构**、**药物作用效应预测**、**基因剂量效应**和**细胞对扰动的响应**等，显示出将不同物种的数据知识整合的巨大潜力 [27]。不过目前GeneCompass仍限于**转录组**数据的整合 [28]。随着更多模态（如蛋白质、代谢物）的基础模型出现，以及结合其它技术桥接模态，Transformer基础模型有望实现对**复杂生物系统**更全面的建模 [29] [30]。

对比学习与多模态对齐

对比学习（Contrastive Learning）是一种自监督技术，通过**拉近匹配样本对的表示、区分非匹配对**，从而学习出判别性特征表示。它在多组学整合中被用于**跨模态数据对齐**和**统一表示空间**构建。当存在不同组学数据的配对观测（例如同一样本的多组学测量）时，可以利用对比损失使模型输出的多模态嵌入在**同一语义空间**中。典型工作如**scCLIP**，受计算机视觉-文本对齐模型CLIP的启发，采用一对**Transformer编码器**分别对单细胞转录组和染色质开放区数据编码，通过对比损失训练模型 [31]。scCLIP成功将scRNA-seq和scATAC-seq数据融合到**统一的嵌入空间**中，而且具备良好的**扩展性**，可整合大规模组织和物种水平的数据 [31]。另一项方法**MatchCLOT**则结合**最优传输（Optimal Transport, OT）**思想，将跨模态对齐转换为一个软匹配问题 [32]。具

体而言，MatchCLOT训练两个模态特定编码器，将多模态测量映射到公共潜在空间，然后利用定制的OT算法在该空间中软匹配不同模态的细胞 [32]。OT匹配过程中引入批次标签等辅助信息以缩小搜索空间、缓解分布偏差 [33]。通过这种方式，MatchCLOT实现了更精细的细胞对应关系识别，有助于推断细胞发育轨迹或多组学变化的一致模式 [32]。

对比学习方法的优势在于**不需要显式的表型标签**即可学习跨模态关联，充分利用未标注数据。同时，它能**校正组学间的系统差异**，使融合后的表示更具可比性，助力下游预测任务 [31] [32]。这对于**高维异构数据融合**和**消除不同数据来源的偏差**非常关键。不过，对比学习通常要求存在**成对的多组学数据**作为训练依据（如同一细胞的两种组学测定值） [34]。在实际应用中，这类完全配对的数据较难大量获取，因此如何利用部分配对甚至无配对的数据进行对比训练仍是研究重点。同时，对比学习模型需要仔细设计**负样本**选择策略和多模态**投影头**结构，以确保模型学到有生物意义的共同模式，而非受到技术噪声或数据集偏差的干扰。

图神经网络在生物网络建模中的应用

生物系统中充满了相互作用关系，例如基因调控网络、蛋白质-蛋白质相互作用网络、细胞-细胞通信网络等。**图神经网络（GNN）**利用图结构数据，可以直接在这些生物网络上进行学习，被广泛应用于多组学和基因型-表型关系的建模 [35]。与将多组学当作平坦特征向量的传统方法相比，图神经网络**显式编码**了生物实体（节点）之间的复杂关系，能够捕获数据中蕴含的**语义联系**和**上下文依赖** [35]。这带来几大好处：首先，GNN天然适合融合异构信息——不同类型的节点和边可以表示不同模态的数据和知识来源，将多源信息无缝集成于统一模型中 [35]。其次，图模型可以根据已有关系**推理出新的联系**，例如通过邻居传播发现潜在的基因功能关联。再次，图表示使模型更具**可解释性**，因为预测结果可以追溯到网络路径或连接的证据 [36]。

一项代表性工作是**GEARS**模型，它将**基因-基因相互作用知识图谱**与单细胞扰动实验数据相结合，利用图神经网络来预测基因扰动的转录反应 [37]。具体而言，GEARS以已知基因交互网络为图结构，对节点（基因）嵌入初始化后，结合单细胞RNA测序的扰动实验数据训练GNN，从而学会基因调控模式 [38]。该模型不仅能预测**单基因**突变对转录组的影响，还能预测**多基因组合扰动**的结果，即使这些组合未在实验中直接测试过 [39]。这表明GNN可以通过在图中传播和组合基因效应，**推断出实验未观察到的情形**。然而，GEARS目前局限于与训练数据相同的细胞类型和实验条件，其对**跨条件泛化**能力有待提高 [40]。此外，由于组合扰动数据可能存在**交互效应的混杂因素**，模型需要精心设计来去除这些混杂，以确保所学代表是真正的因果关系 [41]。除了GEARS外，还有研究利用GNN预测**细胞-细胞通信**对免疫治疗响应的影响。例如Lee等人构建了以**细胞类型为节点、通讯强度为边**的图，从患者肿瘤转录组数据中解卷积得到不同细胞间信号网络，并训练模型预测病人对免疫检查点抑制剂的反应 [42]。这种网络能够识别与疗效相关的关键通信通路，并与单细胞水平的发现相吻合 [42]。尽管如此，该图模型相对简单（如仅利用浅层图结构和线性关系），未来可引入更复杂的GNN变体来揭示更深层的非线性关系 [43]。

总的来说，图神经网络将**生物知识图**融入机器学习，有助于在**数据稀少**的情况下通过先验关系提升预测性能，也为**OOD泛化**提供了一条思路：如果新环境下的实体关系与训练时相似，GNN可依据图结构做出合理推断。而对于完全新颖的节点或边（知识图谱中不存在的部分），GNN本身难以处理，需要结合知识图谱完善与不确定性处理等手段（详见下文知识图谱部分）。尽管如此，GNN在多组学整合中的应用前景非常广阔，将不同层次生物实体关联起来进行建模，对发现潜在机制与因果关系十分有益 [35]。

知识图谱驱动的整合分析

知识图谱（Knowledge Graph, KG）是以图形式表示知识的框架，在生物医学中通常将基因、蛋白质、疾病、药物等实体作为节点，已知相互作用或关系作为边。知识图谱为多组学数据整合提供了一个**结构化先验**：模型可以以知识图谱为背景，将异构数据投射到该网络上，从而将**生物学知识**引入AI模型中 [35]。相比纯数据驱动的方法，利用知识图谱有以下优势：**整合多源信息**（文献、数据库、多种实验等）提高模型的知识面；**引入因果或机制约束**，减少仅凭统计相关性的误判；**挖掘隐含关系**，通过路径联结推理新的知识。

一项有趣的成果是**BioBridge**，它探索了将多个**单模态基础模型**（如分别预训练在蛋白质、化合物、疾病等数据上的模型）通过知识图谱来连接 [44]。在BioBridge中，每种数据模态有各自强大的基础模型，而知识图谱充当不同模态之间的“桥梁” [44]。具体而言，通过一个可训练的小型桥接模块学习不同模态表示间的转换，同时将每个基础模型本身参数固定不变，利用知识图谱中已知对应关系对桥接模块施加监督 [44]。这种方法效率很高，只需训练少量参数，就能实现在**不直接微调大型基础模型**的情况下，让它们的嵌入空间互通 [44]。BioBridge展示了多种任务的可行性，例如**跨模态检索**（在一种模态中查询对应另一模态的实体）、**语义相似性推理**、**蛋白质-蛋白质相互作用预测**，甚至**跨物种的蛋白质-表型匹配**等 [44]。这证明知识图谱可以作为纽带，将不同领域的AI模型联合起来。不过，BioBridge目前在**新分子设计**等生成类任务上尚缺乏定量验证 [45]。

另一个具有前瞻性的思路是近期提出的**OFA（One For All）**框架。OFA尝试用**文本描述**来统一不同领域的图数据，将每个节点的属性和关系用一段文字描述表示，再利用强大的**语言模型**将这些描述编码到同一向量空间 [46]。同时，它在图上增加特殊结构作为“提示”（prompt），让模型不经微调就能对不同任务进行推理 [47]。通过这种**图提示学习**，OFA希望构建一个通用模型，能同时处理文献引文网络、分子结构、知识库等各种领域的图任务 [46]。然而，初步结果显示，该方法在各具体任务上的性能仍低于专门训练的模型 [48]。尽管如此，OFA体现了利用**大模型和知识图谱融合**以实现“一模多能”的探索方向。

值得注意的是，在知识图谱应用中必须应对**不完整和噪声知识**的问题 [49] [50]。现实中的生物医学知识库往往**残缺不全**：大量基因、蛋白尚无注释或关联，这些节点在KG中相当于孤立点，使模型难以对其进行推断 [51]。例如，一个尚无任何已知靶点或类似结构的全新化合物，在药物-基因-疾病图谱中是孤立的，模型无法可靠预测其与疾病的关系 [52]。为提高图谱质量，研究者利用机器学习和NLP自动扩充图谱，如从文献中预测新的基因-疾病关联或药物-靶标关系 [53]。但这些**自动化推理**又可能引入错误，即知识图谱中充斥**错误或不确定的关系**，甚至文献里的某些结论不可复现 [53]。当前针对KG中**虚假关系**的处理研究较少，这可能影响模型的可靠性。Wu等人强调，需要将**生物物理性质**、**情景上下文**和**多尺度信息**引入知识图谱，构建更精细、更可靠的网络模型以提高预测的**因果合理性**和**泛化性** [54] [55]（下文将讨论）。

跨物种迁移学习

跨物种迁移学习利用模式生物的数据和知识来提升人类生物医学预测能力，是当前应对数据不足和外推的重要策略之一 [56] [2]。由于实验伦理和技术限制，我们常在模式生物（如小鼠、果蝇、斑马鱼等）中研究基因功能和药物效应，然后将这些发现应用于人类。然而，不同物种间存在生物差异，如何**传递知识**是个挑战。AI技术通过挖掘不同物种间**保守的生物模式**，可以实现跨物种的表型预测建模 [56]。例如，前述**GeneCompass**模型正是通过将**人和小鼠**海量单细胞数据融合训练，构建了一个跨物种的细胞表达基础模型，使得在人-鼠数据之间进行**迁移学习**成为可能 [26]。它将人类和小鼠的基因表达共同投影到统一空间，依赖于基因的正交对应关系和对生物过程的共同编码，从而在下游任务中，可以通过在小鼠数据上微调模型，来帮助预测人类对应情形的结果 [26]。再如**SATURN**模型，这是首个将**大规模蛋白质语言模型**（例如ESM2）提取的进化信息与单细胞转录组数据相结合的跨物种方法 [57]。SATURN认识到不同物种基因并非一一对应，为克服缺乏直接同源的问题，它采用**软聚类**将功能相似的基因分组为“宏基因”概念，然后同时将这些宏基因在两物种中的表达模式输入模型进行训练 [58]。具体而言，SATURN使用条件自编码器结合ZINB损失来联合学**人类与小鼠**的单细胞表达，并辅以弱监督信号使同一物种数据内的细胞在潜在空间拉开，不同物种中相似细胞靠近 [59]。最终，SATURN能学到**跨实验、跨物种**的**通用细胞嵌入**，在没有严格一一对应基因的情况下，把两个物种的细胞进行对齐 [59]。这些跨物种模型表明，利用**进化保守性**和大规模跨物种数据预训练，可以让模型掌握基本的生物规律，从而对**分布外（不同物种）的数据**作出合理预测。这对于药物从动物模型到人类的疗效安全性预测，未知人类基因功能的推断等，都有重大意义。

当然，跨物种迁移也有其局限。一方面，不同物种的生物网络可能出现**新颖功能**或**重要差异**，模型需要能捕获**非保守**部分，否则只能预测共有模式而忽略物种特异性现象。另一方面，训练跨物种模型常需**高度匹配**的多物种数据（例如相似条件下的人和动物单细胞测序数据），获取和构建这样的数据集并不容易 [60]。一些研究尝试借助**知识图谱**将物种对齐（如通过基因同源关系连接物种网络），或引入**域适应**算法减少物种间分布差异。这些都是当前跨物种AI建模的活跃方向。

解决数据稀缺、OOD泛化与异构数据挑战

多组学整合和基因型-表型预测面临的核心困难可以归结为：**有标签的数据稀缺且含糊、分布移位/外推困难（OOD问题）**、以及**高维异构数据的融合与因果解释** [61] [62]。前文介绍的各种技术各有所长，以下总结它们应对这些挑战的策略：

- **数据稀缺与标签不足**：深度学习模型往往需要大量标记样本才能表现良好，但在生物医学中，高质量**表型标签**获取不易，且不同组学的**配对数据**更是有限 [63]。针对这一问题，无监督和自监督学习大显身手——自动编码器、对比学习、Transformer预训练等方法可以充分利用**未标注数据**提取有信息量的表征 [3]。例如，scVI/scANVI等VAE模型只需原始计数数据就能训练，从中学习批次校正和生物变异的潜在表示 [4]；对比学习通过配对组学自身的一致性训练，无需显式表型标签即可对齐多模态 [31]；大型基础模型（scGPT等）更是通过预训练**跨越了有标签数据的限制**，在成百上千万细胞数据上自我监督获取知识，然后微调时仅需少量标签即可达到出色性能 [22] [24]。此外，知识图谱提供的**先验联系**相当于引入“额外信息”：即使某些关系没有直接数据支撑，模型也能通过图谱中的邻接关系**推断**，相当于变相增加了训练信息。综上，这些方法从不同角度**缓解了标注匮乏**：要么充分利用无标数据，要么整合已有知识，从而降低对大规模标注的依赖。值得一提的是，标签的**定义模糊**也是问题之一，例如复杂疾病表型往往定义不统一。对此有研究在引入**表型本体（如PATO）**以标准化描述，提高跨研究的一致性 [64]。未来把这些本体学知识融入模型，有望进一步解决标签不准确的问题。

- **分布外泛化（OOD）**：即模型应对训练时未见过的新情况（新基因、新药物、新环境等）的能力。生物领域的“宇宙”极其庞大，而当前数据覆盖范围狭窄，例如数万基因中只有少数有已知配体，海量可能的分子里只有很小一部分被测试过 [65]；再如体外细胞系的药物效应不代表人体真实反应，这都是显著的**域偏移** [66]。提升OOD泛化，本文作者提出了多管齐下的策略。其中一大思路是**生物学先验约束**：通过**生物启发的端到端模型**和**物理信息融合的知识图谱**，让模型基于机制而非相关性进行预测 [67] [54]。具体而言，端到端多模态模型强制按照生物系统的层次信息流（DNA→RNA→蛋白→表型）建模 [67]。这样的模型内在地尊重了**因果级联关系**，比起直接在基因和表型间做黑箱映射，更有望在新情况下保持合理性。这类模型还可以通过**转移学习**学习不同层级间的映射（如TransPro模型从转录组预测蛋白质组 [68]），当遇到OOD输入时，模型已掌握更一般的机制（如基因影响蛋白进而影响表型的链条），因此更能**外推** [69]。**跨物种训练**也是扩大分布覆盖的一种方式——将人类以外的生物数据加入训练，相当于扩充了数据多样性，使模型学到更普适的特征，从而对人类未见过的基因/药物也有一定经验 [56] [26]。此外，模型需要对自己的**不确定度**有所估计 [70]。在药物发现等高风险应用中，让模型输出对新样本预测的置信度，可以帮助研究者判断何时不应相信模型、需进一步实验验证 [70]。诸如深度贝叶斯方法、集成学习都可用于量化不确定度，从而在模型面对OOD情形时提供安全保障。

- **高维异构数据融合**：多组学数据类型繁多（DNA序列、RNA表达矩阵、显微图像、临床数据等），直接拼接往往适得其反，需要考虑每种数据的**特性和尺度**。为此，**多模态模型架构**设计至关重要。生物学的中央法则和层次结构为我们提供了指导：可以针对每一层级或每一模态，开发**专门的子模型（子网络）**来提取其特征，然后在更高层结合。例如Yang等人的模型就对不同模态采用不同网络（图像用CNN、序列表达用全连接网络、三维基因组相互作用Hi-C用图卷积），再将这些局部表征映射到共享的潜在空间，实现各模态间的**翻译与补全** [71]。类似地，在作者提出的框架中，每种数据模态都可以先训练一个**该模态的基础模型**（如分别有DNA序列模型、转录组模型等） [72]。当存在跨层级的配对数据时，比如同一批样本测了基因组和转录组，则通过**对比学习或迁移学习**将这两个基础模型连接起来，使DNA表征和RNA表征对齐 [72]。以此类推，多层级的数据通过**局部成对对齐**最终串联成**端到端**网络。训练完成后，即便某些中间层的数据缺失（例如患者没有脑组织的蛋白质组数据），模型也能利用学到的**跨层映射**和其它来源的数据推断出缺失层的信息，从而实现从基因直接到临床表型的预测 [73]。这种逐层预训练、逐层对接的策略，避免了一开始就要求所有模态同时配对，可充分利用不完整的数据集，解决了真实研究中**多组学数据不齐全**的难题 [74]。另外，在知识图谱一侧，作者也建议构建**多尺度、多上下文**的网络模型：例如将全局的基因-疾病网络细分为不同疾病亚型或不同患者的子网，每个子网针对特定情境调整节点和连接 [55]。再如引入**物理相互作用细节**（如突变是否破坏蛋白结合，结合力变化多少）为边的权重和方向 [75]。这些做法使知识图谱能够更精细地表示**异质性**：不同患者、不同细胞类型有各自的

网络拓扑，从而模型可以针对具体环境进行预测 [55] [76]。总之，无论神经网络还是知识图谱路线，核心都是针对异构生物数据的特点设计融合方式，**最大程度保留关键生物信息**并**避免无意义的噪声干扰**，以提高模型在实际复杂系统中的适用性和解释性。

图2：**端到端多模态深度学习框架示意**。【左】传统方法往往要求所有模态数据同时成对获取，并在模型中直接拼合不同模态特征，这对数据收集要求高且难以利用未配对数据。【右】生物启发的端到端深度学习模型按生物层级构建，每个模态可用本模态的大模型预训练，在有配对数据的两层之间，通过对比学习或迁移学习连通模型。一旦完成全链路训练，即使缺少某层数据（如缺失某组学），模型也能通过内部学到的关联予以弥补，实现从基因型经由中间表型到最终表型的预测 [74] [77]。

实际生物医学场景中的应用组合

上述技术的最终目的，是在真实世界的生物医学问题中发挥作用。以下通过**药物发现**和**疾病表型预测**两个场景，说明多组学AI技术的组合应用如何带来突破。

- **药物发现与精准用药**：传统药物研发以靶点为中心，但代价高且成功率低，而基于**表型**的药物发现重新受到重视 [78] [79]。AI驱动的多组学模型可以极大加速这一过程。在细胞或模型生物中进行**扰动组学**实验（如加入候选化合物，测定处理前后的转录组、蛋白质组、代谢组变化），通过深度模型连接这些**端表型（endophenotype）**变化与临床表型，可用于**筛选和优化药物** [80] [81]。例如前述BioLORD模型利用多组学数据**预测细胞对新药的响应**，ChemCPA则将**化合物结构和基因表达变化联系**起来以推断单细胞级别的药物效应 [17] [19]。这些模型能够在**未有实验数据的新药物**上作出预测，大大拓展了虚拟筛选范围。此外，多组学模型还能**揭示药物作用机制**：通过知识图谱将药物-靶点-通路-表型串联，找出药物引发特定分子变化进而导致疗效或毒副作用的因果链。近年来的研究表明，药物导致的多组学变化可作为一种全面的“指纹”，如果能与患者的疾病多组学特征匹配，就有望实现**个性化用药**和**老药新用（药物重定位）** [82] [83]。例如，有工作将患者的转录/蛋白特征与药物在细胞中诱导的转录响应进行匹配，从而筛选出**潜在治疗阿尔茨海默症**的候选药物 [84]。再如**DeepReal**框架通过引入多尺度信息，成功预测了GPCR靶点针对**OOD全新化合物**的活性 [85]。总的来说，在药物发现场景中，**端到端多组学模型+知识图谱**的组合能够把**体外实验**和**临床结果**桥接起来，显著提高新药发现的效率和准确性 [86] [87]。

- **疾病亚型与表型预测**：复杂疾病（癌症、神经退行性疾病等）往往存在不同分子亚型，对治疗反应各异。多组学整合模型可用于**疾病亚型识别、预后预测和诊断分类**等方面，以实现精准医疗。例如在癌症中，将基因突变、表观遗传、转录、蛋白等多层数据融合，可以更准确地区分肿瘤亚型或预测患者生存。**DSIR**和**DLSF**等模型利用深度子空间学习和循环自编码器，从多组学数据中提取共同特征来进行癌症**分型** [88] [89]。它们通过构建样本间的一致性矩阵或共享自表达层，识别出了在多模态下稳定存在的患者子群 [90] [89]。实践证明，这种多组学方法得到的亚型与临床预后密切相关，优于单一组学的分类 [90]。又如Faisal等人的研究，把**病理组织图像**与**基因组变异、转录表达**等数据融合，用深度网络预测癌症患者生存和分层风险 [91]。该模型不仅预测准确度高，还能通过分析多模态特征的重要性找到潜在生物标志物 [92]。在更基础的层面，**CLEIT**模型证明，引入一个中间的转录组表型层（即**内表型**）来连接基因突变和细胞表型，可以显著提升从基因预测表型的性能 [93]。类似地，**TransPro**利用转录组预测蛋白组，再用预测的蛋白组去预测有无药物处理的细胞表型，结果**比直接用实验转录数据预测更准**，因为蛋白组作为更贴近表型的层次起到了提升作用 [69]。这些研究表明，将**多层次生物数据**有机串联，模型可以更好地区分疾病机制差异，提高表型预测的鲁棒性和精准性。

综上，AI驱动的多组学整合在生物医学多个领域展现了强大的应用前景。从**发现新药**、**寻找疾病标志物**，到**个性化治疗决策**，多模态深度学习与知识图谱、迁移学习等技术的结合，正在帮助我们从海量生物数据中提炼有用信息，**揭示基因-环境互作对表型的因果影响** [94]。这一多尺度建模框架有望加速**未解疾病**机制的阐明，发现新的**分子靶标和生物标记物**，并推动**精准医学**的发展，为目前尚无有效疗法的疾病提供创新的解决方案 [94]。

代表性技术和模型一览

下表总结了本文涉及的主要AI技术类别及其代表模型，以及各自的应用特点：

技术方法	代表模型（示例）	应用场景与特点
自动编码器类	ScVI/ScANVI [4] [6] TotalVI [7] Cobolt/MultiVI [9] [11] scMVP [14] GLUE [16]	单细胞RNA等高维数据的降维与批次效应消除；扩展到多组学联合表示（RNA-ATAC-蛋白等）；需要部分配对数据指导融合，面临数据稀疏和噪声挑战。
Transformer基础模型	scGPT [22] [24] Geneformer、scFoundation [22] GeneCompass（跨物种）[26]	大规模预训练的单细胞基础模型，基于Transformer自注意力捕捉基因交互和细胞特征；可微调用于多任务；缓解标注不足但需海量数据和算力；GeneCompass实现人-鼠跨物种建模。
对比学习整合	scCLIP [31] MatchCLOT [32]	自监督对齐多模态数据（如转录组-染色质可及性）至统一空间；利用配对样本对训练，能融合大规模数据集；需要高质量成对数据且模型设计需避免过拟合技术噪声。
图神经网络（GNN）	GEARS [37] Lee等人细胞通信图模型 [42]	在基因、细胞等生物网络上执行学习；用于预测基因组合扰动效应、细胞-细胞相互作用对疗效的影响等；可整合复杂关系提高泛化，但受限于图谱质量，需应对组合干扰的混杂因素。
知识图谱融合	BioBridge [44] OFA [46]	利用知识图谱连接多模态基础模型或多领域图数据，实现跨模态检索、语义推理；仅训练桥接模块高效灵活，但需高质量KG且部分任务效果仍待提升；图谱不完备和错误关系需进一步处理。
跨物种迁移学习	SATURN [58] GeneCompass [26]	融合不同物种数据训练统一模型，借助进化保守信号实现人-鼠等跨物种细胞表征对齐；有助于知识从模式生物迁移到人类，提高OOD预测；挑战在于物种差异和配对数据获取。

上述各类模型各擅胜场，且并非相互孤立。在实际应用中，它们往往**相互结合**：例如，先用**无监督模型**预训练提取表征，再在**知识图谱**上用GNN整合，再通过**迁移学习**跨域应用；或者利用**基础模型**提供的表示作为对比学习的起点等等。这种多技术协同，将是未来生物医学AI发展的重要趋势。展望未来，随着更多高质量多组学数据积累和生物知识的充实，我们有望构建更加精准、可解释的**AI驱动生物系统模型**，深入揭示生命复杂性并加速新疗法的发现 [94] 。

---

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] [48] [49] [50] [51] [52] [53] [54] [55] [56] [57] [58] [59] [60] [61] [62] [63] [64] [65] [66] [67] [68] [69] [70] [71] [72] [73] [74] [75] [76] [77] [78] [79] [80] [81] [82] [83] [84] [85] [86] [87] [88] [89] [90] [91] [92] [93] [94] [2407.06405] AI-driven multi-omics integration for multi-scale predictive modeling of causal genotype-environment-phenotype relationships
https://ar5iv.labs.arxiv.org/html/2407.06405v1