中国人工智能系列白皮书

——人工智能驱动的生命科学

中国人工智能学会

二○二四年七月

《中国人工智能系列白皮书》编委会

主　　任：戴琼海

执行主任：王国胤

副 主 任：陈　杰　何　友 刘成林　刘　宏　孙富春　王恩东
王文博　赵春江　周志华　郑庆华

委　　员：班晓娟　曹　鹏　陈　纯　陈松灿　邓伟文　董振江
杜军平　付宜利　古天龙　桂卫华　何　清　胡国平
黄河燕　季向阳　贾英民　焦李成　李　斌　刘　民
刘庆峰　刘增良　鲁华祥　马华东　苗夺谦　潘　纲
朴松昊　钱　锋　乔俊飞　孙长银　孙茂松　陶建华
王卫宁　王熙照　王　轩　王蕴红　吾守尔·斯拉木
吴晓蓓　杨放春　于　剑　岳　东　张小川　张学工
张　毅　章　毅　周国栋　周鸿祎　周建设　周　杰
祝烈煌　庄越挺

《中国人工智能系列白皮书----人工智能驱动的生命科学》编写组

张世华　张学工　陈盛泉　李婷婷　刘红蕾
刘振栋　刘治平　王太峰　张　岳　郑旭彬

目　录

第 1 章 单细胞转录组预训练基础模型

1.1 单细胞基础模型概述

近年来，随着高通量单细胞测序技术的发展和普及，生物信息学领域内产生了以单细胞转录组为代表的数以亿计的单细胞数据，涵盖了上千种细胞类型、覆盖了不同的发育过程和细胞状态。国际上兴起的细胞图谱计划对这些海量单细胞数据进行了收集和组装，形成了HCA[1]、hECA[2]、CZ-cellxgene[3]等千万级别的大规模细胞图谱，扩展了单细胞组学数据的体量和多样性，为研究单细胞特性提供了宝贵的资源。而随着数据量的快速增长和数据异质性的提高，人们愈发意识到传统的单细胞算法难以有效捕捉大规模单细胞数据集中的生物规律和信息，这促使研究人员开始开发基于预训练人工智能的计算方法，通过构建单细胞转录组的基础模型学习大规模数据中蕴含的规律。

基础模型是一种在广泛数据上训练的机器学习模型，旨在通过大规模自监督学习进行训练，赋予其有效地适应广泛下游任务的能力。单细胞转录组数据中蕴含着丰富的生物学信息，构建单细胞转录组的基础模型能够学习基因表达中的调控规律，并将其与细胞类型识别、药物响应预测等多种下游任务建立关联，具有广阔的应用前景和价值。在自然语言、计算机视觉和语音处理等领域的基础模型构建中，Transformer 模型[4]已然成为了各种基础模型的骨干网络架构。Transformer 模型具有超群的长序列处理能力和扩展性，能够充分利用大规模数据并捕捉其中的数据特征，这使得 Transformer 模型在构建单细胞组学基础模型的过程中可以发挥关键作用。

目前，通过 Transformer 模型构建单细胞基础模型这一研究方向正处于早期探索阶段[5–12]，已有的预训练模型包括：scBERT、Geneformer、scGPT、scFoundation、tGPT、GeneCompass 和 scMulan 等。本章对已有的单细胞基础模型方法进行总结和归纳，分别对单细胞基础模型建模中的数据集、细胞表征、预训练任务建模、细胞和基

因嵌入、下游任务等内容进行概述，并对未来研究方向进行展望。

1.2 单细胞基础模型构建

通过 Transformer 模型构建单细胞基础模型的过程主要涉及数据预处理、数据编码和预训练任务构造三个步骤。单细胞基础模型的训练使用大规模单细胞数据集，并对数据特征维度等信息进行统一处理。数据编码过程主要包括对单细胞转录组数据的基因及其表达量进行编码；预训练过程则涉及预训练任务的构建和 Transformer 架构选择。经过编码的基因和表达量被输入 Transformer 中，经过自注意力机制进行长序列运算提取互作信息，并生成基因表征，进一步通过预训练任务的预测头进行自监督训练。

1.2.1 大规模单细胞数据集

目前的单细胞基础模型以基因为基本标识，以一个细胞为一个对象，在预训练阶段需要庞大的细胞数量以提供丰富的细胞多样性。高通量测序技术的飞速发展带来了大量的单细胞数据，hECA[2]、CZ-cellxgene[3]和 DISCO[13]等细胞图谱收集了千万级别的单细胞数据，涵盖了几百个数据集、百余种细胞类型、各年龄段的捐献者。这些细胞图谱不仅仅收集了数据集，还进行了一定的跨数据集统一处理。这样的数据规模和多样性，能够支持模型捕捉数据中的基因关系和分布特征。除了单细胞数据的收集之外，上述数据集还提供了丰富的元信息，包括细胞类型、所属器官、捐献者信息等。其中，hECA 对不同来源数据集中的元信息进行了系统的整理，使元信息字段在不同数据集中保持一致，例如，保证不同器官中相同细胞类型的细胞名称一致。这使得这些内容能够在基因表达之外，给模型提供更为宏观的信息；同时，这也需要模型进行专门设计进行兼容。

上述单细胞图谱都对不同来源的数据的基因列表进行了统一，使得图谱中所有细胞共享相同的基因列表。根据不同模型的设计，会通

过算法选取高变基因或使用完整基因。对基因表达矩阵的处理包括标准化和对数变换等步骤，旨在降低表达量中极端数据的影响，并使得特征范围可比。

1.2.2 单细胞数据编码嵌入表示

由于 Transformer 主要用于处理序列化的数据，而单细胞数据是表格类型的数据，因此需要将数据进行转化，从而能够支持 Transformer 模型进行处理。单细胞转录组基础模型通常将基因视为单词，细胞中的所有基因表达视为一个句子。由于每个基因天然是独立的单位，所以无需像自然语言处理那样对句子进行分词操作。而又由于与自然语言的词汇不同，在每一个单细胞的句子里，每个基因不仅由一个词汇（gene symbol）来表示，它还有对应的表达值。因此，需要对输入的基因名称和表达值分别进行编码，转为 Transformer 能够接收的格式。可以将基因和表达值分别使用不同的编码方式到相同维度的嵌入空间，然后通过相同位置编码逐元素求和得到最终输入 Transformer 模型的嵌入。目前对基因名称和对应的表达值存在不同的利用和编码方式。

1.2.2.1 基因名称的编码嵌入表示

为了让 Transformer 能够区分每一个输入的基因，需要对不同基因赋予不同的编码嵌入。大多数单细胞基础模型采用了自然语言处理中对 token 的编码方式，即通过 one-hot 编码和投影神经网络将词汇表中的每一个基因投影到一个高维嵌入空间。这使得每一个基因都通过编码成为相同维度的嵌入向量。这个投影过程具有可学习的参数，会随着 Transformer 的训练而进行更新，从而一定程度上能够捕捉基因之间的关系。

除了从数据中直接学习投影嵌入表示，GeneCompass[11]还通过引入外部知识，对基因赋予了其他的编码嵌入，包括启动子嵌入、共表达嵌入、基因族嵌入和基因调控网络嵌入。其中，启动子嵌入是使用

基因转录起始位点附近的碱基序列微调 DNABERT 模型[14]，并获取其对应的隐层表示来获得的。共表达嵌入、基因族嵌入和基因调控网络嵌入是通过 gene2vec 方法[15]获取的嵌入表示，即先将具有相似属性的基因构建基因对，再训练 gene2vec 模型使得相似基因可以获得相近的嵌入表示。这些编码具有相同的嵌入维度，从而经过聚合之后输入到 Transformer 模型之中。

1.2.2.2 基因表达值编码嵌入表示

基因表达值一方面可以用于给基因排序，通过位置编码的形式间接地提供表达水平的信息，另一方面也可以直接进行编码作为输入。本节介绍三种主要的表达值编码方式，可以将基因表达值的信息通过编码，叠加到基因编码上，作为 Transformer 的输入，包括排序编码、连续值投影编码和离散类别编码。

排序编码：根据基因表达量的高低可以对细胞中表达的基因由高到低排序，从而形成一个基因序列。由于 Transformer 对位置不敏感，可以通过跟自然语言中类似的位置编码对基因序列的位置进行编号，形成包含了表达量高低信息的位置编码。目前的 Geneformer 模型[8]采用了这种排序编码的方式。它的好处在于抹去了原始表达信息，从而更好地适应原始的 Transformer 架构，但随之而来的缺点是无法从排序后的序列中恢复原始表达。

连续值投影编码：经过标准化和对数变换等处理流程之后得到的基因表达量通常是一个连续的数值，为了将其映射到与基因编码相同的编码空间，需要对表达值进行投影。这个过程采用神经网络来完成，得到与基因编码相同维度的嵌入。这种编码形式理论上可以不经损失地使用原始的连续表达值，但是由于原空间维度过高，可能影响模型对有效信息的捕捉能力。scFoundation[6] 和 GeneCompass 模型中使用了连续值投影的编码。

离散类别编码：将编码空间离散化有助于模型的学习更为稳定，

也与基因编码的方式保持一致。因此，可以先将连续值进行离散化，得到诸多表达量区间，然后将表达量区间通过与基因编码相似的离散投影网络，将表达量投影到高维嵌入空间。离散类别编码也有多种实现方式，如 scMulan[5]通过动态分桶法，以每个细胞中表达值最高的基因为基准，划分多个区间；scGPT [7]通过分位数的方式来划分区间；BioFormers[9]提出可以通过非线性地对高表达、超高表达、低表达的基因采用不同的区间划分。

1.2.2.3 其他元素的编码

除了基因和表达值，其他元信息和特殊字符也可以被编码到 Transformer 之中。例如，scMulan 将以文本形式存在的细胞元信息以独立字符的方式进行编码，使得模型可以捕捉基因表达与元信息之间的关系，并且通过将不同的下游任务进行编码，使得模型能够通过接收不同的任务提示词来执行不同的功能。此外，包括批次信号、CLS、扰动信息等元素，也被应用于模型编码之中。这些特殊字符的编码可以给模型赋予额外的信息。

1.2.3 预训练任务建模

通过构建自监督学习任务的方式训练 Transformer 模型可以充分利用庞大的单细胞数据，从中学习调控规律和生物信息并应用于丰富的下游任务，从而在没有特定任务注释的情况下提高模型的泛化能力。这一自监督学习的范式已经在自然语言、计算机视觉等领域的基础模型构建过程中得到了广泛的印证。

在单细胞基础模型中，采用的预训练任务主要分为类似于 BERT 模型[16]使用的掩码预测（MLM）任务和类似于 GPT 模型[17]使用的因果逐个生成（CLM）任务。

1.2.3.1 基于 MLM 的预训练

MLM 是一种常见的自监督预训练方法，在自然语言处理中应用的典型代表为 BERT 及其变体[16,18–20]，目前的单细胞基础模型

scBERT[12]、Geneformer、scGPT 和 scFoundation 等，采取的是这种预训练任务。

具体而言，在单细胞的 MLM 任务中，某些基因表达量的值会被随机屏蔽（施加 Mask），然后模型通过自监督训练来预测这些被屏蔽的基因的基因表达水平。scFoundation 在这一基础上，还引入了恢复测序深度这一任务，进一步学习基因表达水平的信息。MLM 任务可以让模型学习到基因表达数据的分布和结构，同时还能捕捉到基因之间的潜在关系。

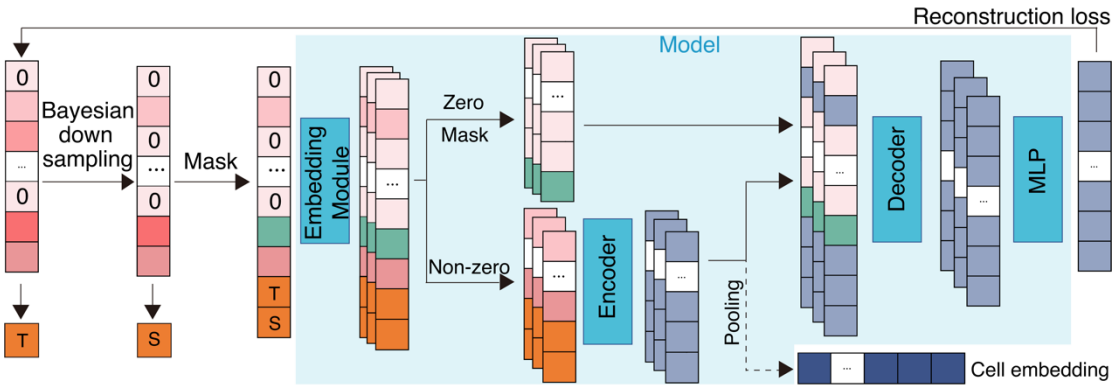


图 1-1 scFoundation 的建模方式

MLM 的预训练任务会选择使用 Transformer 的 Encoder 架构，它允许输入的所有元素通过双向的 Attention 机制获取全局信息，并得到每一个元素的高维嵌入表示。在预训练阶段，被屏蔽位置的元素的高维嵌入会被用于预测表达量，计算均方误差等损失，进行反向传播和梯度下降进行 Transformer 的参数更新。

通过 MLM 预训练后的基础模型捕捉到输入数据中的复杂结构和依赖关系，这对于理解单细胞组学数据中的基因表达模式和细胞状态具有重要作用。不过这一任务对屏蔽方式较为敏感，如何选择最佳的屏蔽策略，如屏蔽比例、屏蔽内容等，需要进行专门的测试和设计，不同的策略可能会对模型的训练产生显著影响。

1.2.3.2 基于 CLM 的预训练

目前，在自然语言处理领域最先进的大语言模型如 GPT 系列、Llama 系列[17,21–23]等均采用 CLM 方式构建生成式预训练任务。CLM 的任务是给定输入序列，预测下一个元素，在推理过程中可以通过生成完成任务。由于单细胞基因表达并没有天然的顺序，不同基础模型给出了各自的预训练任务构建方式。

tGPT[10]通过基因表达量的高低构造了基因的顺序，将预训练任务定义为给定某个位置之前的基因排序，预测下一个位置的基因，期望通过高表达基因逐渐预测所有低表达的基因。

scMulan 利用了注意力机制对位置不敏感的特点，没有对基因排序，而是通过随机打乱细胞中的基因顺序消除基因的排序，然后将预训练任务定义为给定某个位置之前的基因，预测细胞里其余基因和表达值，期望通过一部分基因预测其他基因。此外，scMulan 还加入了诸多元信息，如细胞类型、器官名、捐献者年龄、性别等。这些元信息可以作为输入序列的一部分，也可以作为预测对象。这使得在模型在预训练过程中构建了微观基因表达与宏观元信息之间的联系。通过设置诸多任务提示词，scMulan 可以在不同的下游任务中生成与之对应的内容，从而使用相同的预训练范式，能够同时进行多任务的预训练。

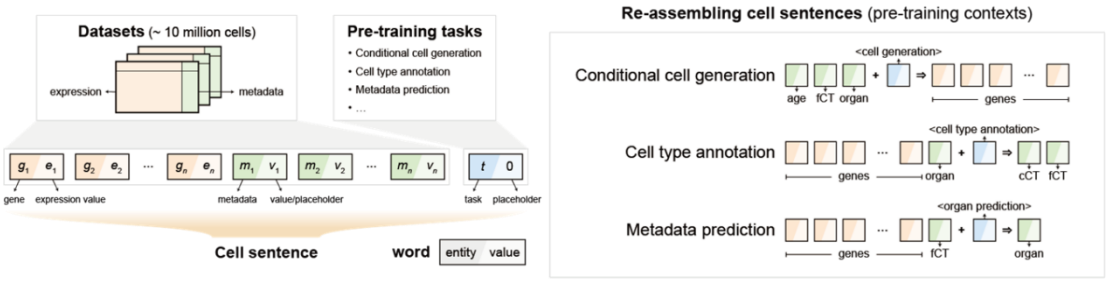


图 1-2 scMulan 对基因表达和元信息的使用范式

CLM 的预训练任务使用 Transformer 的解码器来进行训练。解码器通过特殊的因果注意力机制，使得每一个输入元素只能与它本身和它之前的元素产生注意力。在模型训练过程中，模型的一次前向和反向计算将会同时对所有输入元素进行训练，在单细胞转录组上的训练效率更高。

需要注意的是，CLM 方法得到的模型作为生成式模型，并不能显式地获取每一个输入基因经过 Transformer 之后的表征，其获取细胞表征的方式也有待进一步研究。

1.3 单细胞基础模型应用

在自监督预训练完成后，单细胞基础模型可被应用于多种下游任务，这充分展现了单细胞基础模型的可扩展性和通用性。目前的单细胞基础模型大多数通过在不同任务上进行微调执行对应任务，scMulan[5]由于在预训练阶段使用了部分元信息，可以在相关任务上无需微调执行多种下游任务。单细胞基础模型的应用主要包括：细胞嵌入表示、细胞类型注释、单细胞数据生成、推断调控网络和空间组学应用等。

1.3.1 基因嵌入表示和细胞嵌入表示

目前的单细胞基础模型，在经过预训练后都可以输出具有生物学含义的细胞嵌入表示。不同于输入 Transformer 之前对基因的嵌入表示，此处的嵌入表示是数据经过训练好的 Transformer 得到的。由于单细胞基础模型面对测试数据集具有良好的泛化能力，从而对新数据的细胞嵌入也可以保留基因之间和细胞之间的关系，具有较好的鲁棒性。细胞嵌入表示可以通过对所有基因嵌入平均的方式得到。

1.3.1.1 基因嵌入表示

基因的嵌入表示可以分为数据无关的嵌入和数据相关的嵌入表示。

应用于单细胞数据的 Transformer 在输入基因数据时，首先会生成某种维度的基因编码，如 1.2.2 节所述。这种基因编码通常在预训练过程中作为模型参数的一部分而进行更新。这类基因编码是模型参数的一部分，可以看做是与数据无关的基因嵌入。

数据相关的基因嵌入表示是将数据输入模型，然后从模型输出得到。一般而言这样的表示是从 Transformer 的最后一层输出层获取的，而在 Geneformer 模型中，使用的是 Transformer 输出的倒数第二层表示[8]。通过比较这些嵌入表示在不同细胞间的相似性得分，例如余弦相似性，可以为基因之间的共表达等关系提供新的见解。

1.3.1.2 细胞嵌入表示

低维空间中单个细胞的高质量表示是单细胞各种下游分析的关键组成部分。其中至关重要的是保存生物差异，如细胞类型和细胞状态，同时最大限度地减少技术混淆，如数据集之间的批次效应。在整合来自多个研究、组织甚至生物体的数据时，去除批次效应和相关协变量是极具挑战性的。Transformer 通过预训练任务在未知批次的情况下为细胞提供了一个有效嵌入表示，并且可以证明该表示对某些批处理效果稳健。

基于 Transformer 的细胞嵌入表示在许多方面与其他技术有所不同。基于变分自编码器的流行模型，如 scVI[24]和 scArches[25]，或最近提出的 SCimilarity[11]等模型明确地学习低维嵌入。Transformer 仅通过常用的自监督预训练任务并不显示产生低维的细胞嵌入，而是通过汇集单个细胞的 Transformer 输出的基因嵌入来实现细胞嵌入。例如将 Transformer 的每一个基因嵌入表示求均值得到细胞的嵌入表示，或者通过在输入中引入特殊的细胞标记，如 CLS，该标记的嵌入表示可以代表细胞的嵌入表示。此外，Transformer 输入标记的灵活性便于使用多模态特征进行细胞表示，例如 scGPT[7]可以将跨组学数据进行匹配和马赛克整合。

单细胞基础模型提取的细胞嵌入表示在多种下游任务表现更优异，被证实良好地去除批次效应的同时保留了生物差异。

1.3.2 单细胞类型注释

许多单细胞基础模型被设计用于单细胞类型注释这一下游任务，这也是评估单细胞基础模型的一个通用任务。单细胞基础模型已经显示出通过自监督预训练可以提高它们的细胞注释能力。

具体而言，在单细胞类型注释任务中，使用者可以将预训练得到的单细胞嵌入表示进行微调，从而实现对细胞类型注释。例如Geneformer[8]、scFoundation[26]等大多数模型都是通过微调实现细胞类型注释。而 scMulan[5]可以不经过微调实现细胞类型注释。由于细胞类型也是元信息的一部分，scMulan 得益于将细胞的元信息作为自回归学习的一部分这种特殊设计，可以无需微调直接进行细胞类型注释。Transformer 在泛化到未见数据集方面表现出了巨大的潜力，这对利用具有统一注释的参考单细胞数据图谱来注释新数据集至关重要。

1.3.3 单细胞数据生成

单细胞数据生成包括基因扰动数据生成、跨模态数据预测和基于元信息条件生成等。经过自监督预训练的单细胞基础模型可以通过模拟单个输入基因的扰动，在扰动条件下的预测其他的基因表达。例如，基因敲除或降低表达，或在细胞暴露于小分子等扰动条件下实现单细胞数据生成，这有利于进行虚拟药物试验从而实现药物的快速筛选等。跨模态预测是使用已知的模态来预测缺失的模态，例如 scMoFormer[27]和 scTranslator[28]利用基因组学数据预测蛋白组学数据。此外，生成性 Transformer 有可能直接模拟数据。例如，scMulan 使用指定的元信息条件作为输入，不需要任何组学特征即可生成单细胞数据，该模型可以用于在获取匹配对照组织具有挑战性的情况下进行对照组数据集的生成，并在一定程度上可以通过输入基因扰动在零样本条件下生成扰动后的细胞，进行虚拟扰动实验。

1.3.4 推断调控网络

单细胞预训练基础模型可以用于推断基因之间的相互作用和调控网络。细胞和基因组学特征标记之间的注意力分数可以用来识别细胞类型标记基因、与特定细胞表型相关的基因，以及与生物过程相关的基因，如发育调节因子，以及与特定细胞表型相关的基因。例如，Geneformer[8]和 GeneCompass[11]等模型通过分析基因嵌入之间的注意力分数来推断基因调控网络。

在传统方法中，识别与特定条件相关的组学特征，或者这些特征之间的相互作用，通常是通过特征与条件之间的相关性或通过分析特征嵌入的相似性来得出的。Transformer 引入了一种新颖的方法，即通过不同组学标记之间的注意力机制来学习多模态相互作用，生成可学习的特征关系。例如，结合 ATAC 和 RNA 数据可能揭示基于共结合转录因子的表达和染色质可及性的上下文特定的转录因子调控。

1.3.5 空间组学应用

单细胞基础模型在空间组学应用中也显示出了潜力。scGPT [7]、SpaFormer[29]和 CellPLM[30]直接应用于空间组学数据，在空间转录组基因表达插补任务上展示了有效结果。目前 SpaFormer 和 CellPLM 进行了空间信息的设计，将其整合到模型输入中，使用位置编码来编码细胞的空间坐标。空间转录组学的迅速发展以及 Transformer 在其他领域解析空间坐标的能力使得这些技术的整合成为一个有前景的新领域。

1.3.6 其他任务

单细胞基础模型由于各自的模型细节和侧重不同，设计了很多具有特色的下游任务，如单细胞药物响应预测、基因剂量敏感性预测实验等。例如，scFoundation 和 GeneCompass 可以结合 GEARs 等基因扰动预测模型，用于预测基因扰动的影响，Geneformer 和 GeneCompass 可以执行基因剂量敏感性预测任务，scFoundation 可以

增强输入数据的测序深度、并可在 bulk 数据上应用。

1.4 展望

目前在单细胞转录组数据上预训练得到的基础模型在零样本和微调场景下产生了优异的表现。如何结合单细胞多模态数据，如空间转录组、染色质开放性等信息构建基础模型，将是未来研究的一个重要方向。此外，对于单细胞基础模型中的涌现现象有待进一步探索。在应用方面，未来需要探索如何通过单细胞基础模型，实现疾病靶点发现和快速药物筛选，从而帮助更好解决更多的生命健康难题。

参考文献

[1] Science Forum: The Human Cell Atlas | eLife n.d. https://elifesciences.org/articles/27041 (accessed April 18, 2024).

[2] Chen S, Luo Y, Gao H, Li F, Chen Y, Li J, et al. hECA: The cell-centric assembly of a cell atlas. iScience 2022;25:104318. https://doi.org/10.1016/j.isci.2022.104318.

[3] Program CS-CB, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, et al. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data 2023:2023.10.30.563174. https://doi.org/10.1101/2023.10.30.563174.

[4] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Advances in Neural Information Processing Systems 2017;30.

[5] Bian H, Chen Y, Dong X, Li C, Hao M, Chen S, et al. scMulan: a multitask generative pre-trained language model for single-cell analysis 2024:2024.01.25.577152. https://doi.org/10.1101/2024.01.25.577152.

[6] Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large Scale Foundation Model on Single-cell Transcriptomics 2023:2023.05.29.542705. https://doi.org/10.1101/2023.05.29.542705.

[7] Cui H, Wang C, Maan H, Pang K, Luo F, Wang B. scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2023:2023.04.30.538439. https://doi.org/10.1101/2023.04.30.538439.

[8] Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature 2023:1–9. https://doi.org/10.1038/s41586-023-06139-9.

[9] Amara-Belgadi S, Li O, Zhang DY, Gopinath A. Bioformers: A

Scalable Framework for Exploring Biostates Using Transformers 2023:2023.11.29.569320. https://doi.org/10.1101/2023.11.29.569320.

[10] Shen H, Liu J, Hu J, Shen X, Zhang C, Wu D, et al. Generative pretraining from large-scale transcriptomes for single-cell deciphering. iScience 2023;26. https://doi.org/10.1016/j.isci.2023.106536.

[11] Yang X, Liu G, Feng G, Bu D, Wang P, Jiang J, et al. GeneCompass: Deciphering Universal Gene Regulatory Mechanisms with Knowledge-Informed Cross-Species Foundation Model. Bioinformatics; 2023. https://doi.org/10.1101/2023.09.26.559542.

[12] Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a Large-Scale Pretrained Deep Language Model for Cell Type Annotation of Single-Cell RNA-seq Data. Nature Machine Intelligence 2022;4:852–66.

[13] Li M, Zhang X, Ang KS, Ling J, Sethi R, Lee NYS, et al. DISCO: a database of Deeply Integrated human Single-Cell Omics data. Nucleic Acids Research 2021:gkab1020. https://doi.org/10.1093/nar/gkab1020.

[14] Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 2021;37:2112–20. https://doi.org/10.1093/bioinformatics/btab083.

[15] Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes based on co-expression. BMC Genomics 2019;20:82. https://doi.org/10.1186/s12864-018-5370-x.

[16] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv Preprint arXiv:181004805 2018.

[17] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I.

Language Models are Unsupervised Multitask Learners n.d.

[18] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach 2019. https://doi.org/10.48550/arXiv.1907.11692.

[19] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations 2020.

[20] Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics 2020;8:64–77. https://doi.org/10.1162/tacl_a_00300.

[21] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc.; 2020, p. 1877–901.

[22] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: Open and Efficient Foundation Language Models. arXiv Preprint arXiv:230213971 2023.

[23] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv Preprint arXiv:230709288 2023.

[24] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature Methods 2018;15:1053–8. https://doi.org/10.1038/s41592-018-0229-2.

[25] Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol 2022;40:121–30.

https://doi.org/10.1038/s41587-021-01001-7.

[26] Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large Scale Foundation Model on Single-cell Transcriptomics 2023:2023.05.29.542705. https://doi.org/10.1101/2023.05.29.542705.

[27] Heimberg G, Kuo T, DePianto D, Heigl T, Diamant N, Salem O, et al. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages 2023:2023.07.18.549537. https://doi.org/10.1101/2023.07.18.549537.

[28] Liu L, Li W, Wong K-C, Yang F, Yao J. A pre-trained large generative model for translating single-cell transcriptome to proteome 2023:2023.07.04.547619. https://doi.org/10.1101/2023.07.04.547619.

[29] Wen H, Tang W, Jin W, Ding J, Liu R, Dai X, et al. Single Cells Are Spatial Tokens: Transformers for Spatial Transcriptomic Data Imputation 2024. https://doi.org/10.48550/arXiv.2302.03038.

[30] Wen H, Tang W, Dai X, Ding J, Jin W, Xie Y, et al. CellPLM: Pre-training of Cell Language Model Beyond Single Cells 2023:2023.10.03.560734. https://doi.org/10.1101/2023.10.03.560734.

第 2 章 人工智能赋能细胞异质性刻画

2.1 概述

传统的基因组学研究通常只能提供细胞群体的信息，而单细胞测序技术的出现使得研究人员可以更加深入地了解细胞群体内部的细胞异质性，揭示不同细胞之间的转录表达模式、表型特征以及功能状态的差异，从而理解细胞群体中不同亚型的分布、相互作用以及在生理和病理过程中的作用。同时，随着测序技术的不断发展，包括单细胞转录组测序、单细胞表观组测序、单细胞蛋白质组测序等在内的不同组学单细胞测序技术的应用，也使得我们可以同时获得细胞多层次、多维度的数据，进一步丰富了对细胞异质性的理解。

尽管各类单细胞测序技术的发展为细胞异质性的刻画提供了强有力的工具，推动了细胞生物学领域的发展和深入研究，但单细胞测序数据的分析仍面临特征维度高、数据噪声大、稀疏程度高、批次效应强和技术差异大等多种挑战，如何有效地整合多源单细胞数据，准确地刻画细胞异质性，从而精准地辨识细胞类型并解析其基因调控规律，是亟待解决的关键科学问题。

随着人工智能技术的迅速发展，如何结合计算机算法与测序技术，更好地挖掘细胞异质性信息，是当今的重要研究热点之一。人工智能技术可以有效地应用于大规模生命组学数据的处理和分析，目前，针对细胞异质性刻画问题，研究人员提出了多个人工智能算法，这些方法涉及数据处理与分析的多个阶段：

降噪和数据清洗：人工智能方法可以应用在数据预处理环节，对数据进行降噪、校正和清洗，提高数据的质量和可靠性；

特征提取和降维：人工智能方法可以对数据进行特征提取和降维，挖掘数据中重要的模式和结构，减少高维度数据带来的问题；

聚类和分类：人工智能方法可以应用在细胞类型的识别和分类中，

帮助发现并定义不同的细胞类型，揭示细胞类型的特异性模式和机制；

数据整合和跨样本分析：人工智能方法可以整合不同来源的数据，消除批次效应和技术差异，实现跨样本的一致性分析和结果解释；

多组学联合析：人工智能方法可以将基因组学、转录组学、表观基因组学、蛋白质组学等多种不同组学的数据整合到一个框架中，提供更加全面且多维度的细胞信息。

以上人工智能方法按照对数据的需求程度，可以分为无监督学习、弱监督学习和有监督学习这三种主要类型，我们将依次介绍这三类方法的任务特点、数据需求、设计思路和代表性工作。

2.2 基于无监督学习的细胞异质性刻画

在许多实际应用中，获取带标注的数据通常代价高昂或不可行。无监督学习是一种不依赖标注数据，直接利用无标注的数据进行学习的人工智能方法，在没有数据标签的情况下分析和识别数据中的模式。无监督学习的目标通常是识别数据中的结构、关系或者数据的内在分布特性。在对单细胞各类组学数据进行下游分析之前，研究人员常常使用无监督学习方法进行数据预处理，比如特征提取和降维，以得到能够良好地表征细胞异质性的低维嵌入表示，从而用于进行后续各种下游分析。本节我们将探讨基于无监督学习的细胞异质性刻画方法（图 2-1），概述具有代表性的模型原理及此类方法中的代表性工作。

2.2.1 基于无监督机器学习的细胞异质性刻画方法

常用于细胞异质性刻画的传统机器学习方法包括主成分分析（Principal Component Analysis，PCA）、奇异值分解（Singular Value Decomposition，SVD）、非负矩阵分解（Non-negative Matrix Factorization，NMF）等降维方法，K-均值聚类（K-means Clustering）、K-中心点聚类（K-medoids clustering）、层次聚类（Hierarchical

Clustering）等聚类方法，以及基于贝叶斯框架的统计方法等。本节我们将重点关注上述人工智能方法在刻画细胞异质性方面的应用。

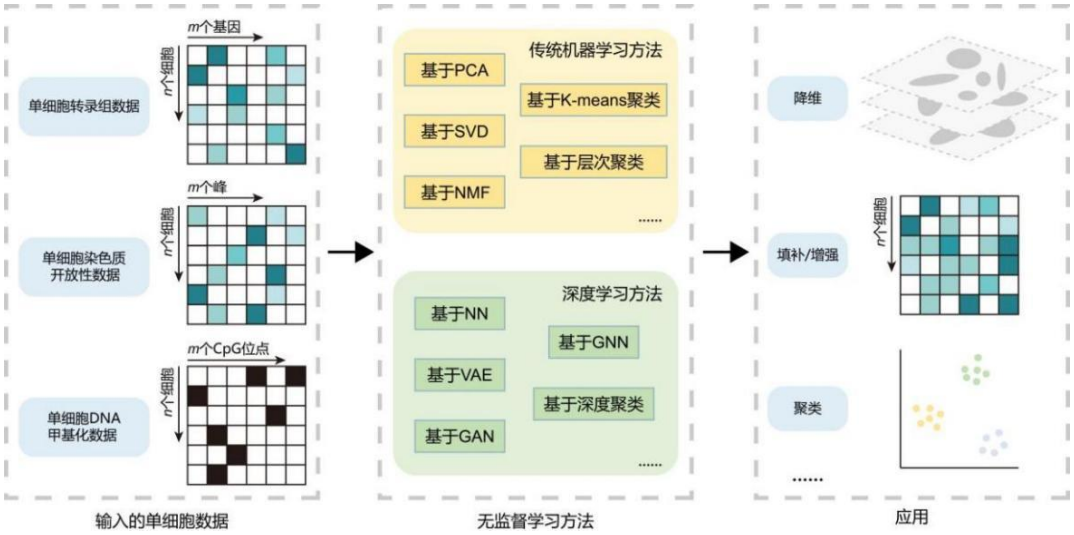


图 2-1 基于无监督学习的细胞异质性刻画方法

PCA 是最为广泛使用的降维方法之一[1]。基于 PCA 刻画细胞异质性的代表方法有 SC3，它首先对单细胞转录组数据的表达矩阵过滤基因和标准化，然后用欧几里得距离、皮尔逊系数和斯皮尔曼系数来分别度量细胞间的距离或相似性，再使用 PCA 或通过计算相关图拉普拉斯的特征向量来转换所有三种距离矩阵，得到转换过的 6 种距离矩阵，然后分别选取特征值最高的前d个特征向量得到$6*d$组低维表示，用于细胞聚类[2]。基于 PCA 得到单细胞数据低维表示的方法还有用于整合单细胞多组学数据的计算方法 Seurat v5[3]，适用于单细胞转录组数据的聚类方法 pcaReduce[4]，拟时序分析方法 TSCAN[5]、Monocle3[6]，以及适用于单细胞染色质开放性数据（Single-cell chromatin accessibility sequencing，scCAS）的计算方法 STREAM[7]、ChromSCape[8]和 EpiScanpy[9]等方法。

SVD 是一种广泛使用的基于矩阵分解的方法，同样可以用于对单细胞数据进行降维。代表性方法包括用于填补单细胞转录组数据的

ALRA[10]，用于分析 scCAS 数据的 ArchR[11]和 Signac[12]，以及用于整合单细胞多组学数据的 MultiMAP[13]等方法。

此外，基于矩阵分解的机器学习方法还包括非负矩阵分解。在 PCA 和 SVD 中，原始的矩阵**V**被近似分解为低秩的$\mathbf{V} \approx \mathbf{W} * \mathbf{H}$，分解出的两个因子矩阵**W**和**H**中往往含有负值元素。尽管从计算的角度来看，分解矩阵中的负值是允许的，但负值元素通常缺乏实际意义，难以进行模型解释。NMF 约束了原始矩阵**V**和分解矩阵**W**和**H**的非负性，其分解出的因子矩阵易于与细胞的类型及其表达模式对应起来。基于非负矩阵分解的代表性细胞异质性刻画方法包括用于单细胞多组学联合分析的 LIGER[14]、MOFA[15]和 MOFA+[16]，用于对单细胞转录组数据细胞类型识别的 NMFLRR[17]，用于填补 scCAS 数据的 scOpen[18]，以及用于增强 scCAS 数据的 scCASE[19]等方法。

而传统无监督机器学习中的聚类方法，如 K-means、K-medoids 和层次聚类，是用于识别细胞类型和状态，从而深入描述细胞间异质性的常用方法。例如，SC3 对上述$6 * d$组低维表示分别进行 K-means 聚类，得到$6 * d$组聚类结果，再对每组聚类结果计算相似性矩阵，对所有的相似性矩阵取均值得到一致性矩阵，再对其使用层次聚类以得到最终的聚类结果[2]。使用无监督聚类方法来刻画细胞间异质性从而辨识细胞类型的代表方法还有适用于单细胞转录组数据的 SIMLR[20]、SAME-clustering[21]，对 scCAS 数据进行聚类的 scABC[22]。传统的无监督学习方法还可以通过整合贝叶斯框架，提高模型对数据潜在结构的推断能力，此类代表性方法有 cisTopic[23]和 Melissa[24]。cisTopic 基于贝叶斯框架学习 scCAS 数据的低维嵌入，而 Melissa 则是通过概率图模型对单细胞 DNA 甲基化数据进行聚类和填补。

2.2.2 基于无监督深度学习的细胞异质性刻画方法

深度学习方法相较于传统机器学习方法的优势在于其能够通过多层次的非线性变换自动学习数据的复杂表示，这使得深度学习在处

理高维数据、图像识别、语音识别和自然语言处理等领域表现出色。深度学习能够自动提取和学习有用的特征，无需人工设计或选择特征，减少了对专业知识的依赖。本节我们将探讨基于无监督深度学习的细胞异质性刻画方法，重点关注基于神经网络（Neural Network，NN）、自编码器（Autoencoder，AE）及变分自编码器（Variational Autoencoder，VAE）、生成对抗网络（Generative Adversarial Network，GAN）、图神经网络（Graph Neural Network，GNN），以及深度聚类（Deep Clustering，DC）的方法。

基于常用的深度神经网络 NN，scVI 聚合单细胞转录组数据中相似细胞和基因的信息，并近似观察到基因表达值的分布[25]。而 scBasset 则基于卷积神经网络对 scCAS 数据的染色质开放峰区域对应的 DNA 序列进行建模，得到了高质量的 scCAS 数据低维嵌入表示，刻画细胞表观异质性[26]。

自编码器 AE 是一种通过神经网络进行数据编码和解码的模型，目的是学习数据的隐空间表示。变分自编码器 VAE 是自编码器的一种变体，对数据的隐空间分布进行约束，结合概率生成模型来模拟数据的生成，其中编码器学习数据分布的参数，解码器从这些分布中抽样生成数据。例如，scDHA 利用非负内核自动编码器和堆叠贝叶斯自动编码器实现单细胞转录组数据降维[27]；scVAE 基于 VAE 估计单细胞转录组数据预期基因表达水平和每个细胞的嵌入表示[28]；而基于 VAE 的方法也被广泛用于学习 scCAS 数据低维嵌入表示，包括 BAVARIA[29]、SCALE[30]、SCALEX[31]、uniPort[32]和 PeakVI[33]等。

生成对抗网络 GAN 通常由两个神经网络共同组成，一个是生成器（Generator），另一个是判别器（Discriminator）。生成器的目标是生成类似于真实数据的内容，而判别器的目标是判断给定的内容是否来自真实数据。这两个网络在互相竞争的过程中逐渐提高了生成器的生成能力，使得生成的内容更接近真实数据，而判别器则不断提高识

别真伪的能力。例如，DR-A 基于对抗变分自编码器的框架（生成对抗网络的一种变体），对单细胞转录组数据进行降维以刻画细胞异质性[34]；AGImpute 构建自编码器与生成对抗网络相结合的混合深度学习模型来估算已识别的丢失事件，以填补基因表达矩阵[35]；scDEC 针对 scCAS 数据构建耦合生成对抗网络，学习细胞嵌入表示的同时辨识细胞类型[29]。

相较于传统神经网络，图神经网络 GNN 能有效处理图结构数据，通过节点与其邻居之间的信息传递捕获图的拓扑关系，这使得 GNN 在节点分类、图分类和链接预测等任务中表现出色。例如，scGGAN 通过图卷积网络学习基因与基因的关系，并通过生成对抗网络学习全局单细胞转录组数据分布以对其进行填补，从而更好地刻画细胞异质性[36]；scGNN[37]和 scGNN 2.0[38]分别基于图自编码器和图注意力自编码器对单细胞转录组的基因表达矩阵进行填补；DeepTFni 则针对 scCAS 数据基于变分图自编码器来推断转录因子调控网络[39]。

进一步地，深度聚类 DC 方法通过结合深度学习和聚类算法，采用端到端的训练过程来优化细胞的嵌入表示和聚类质量，其基本思路是使用深度神经网络来提取和学习数据的特征，并结合常用的聚类技术进行聚类。深度聚类能够提升传统聚类方法在复杂数据集上的表现，已被成功用于单细胞组学数据的细胞异质性刻画和细胞类型辨识，例如 scDeepCluster[40]、DESC[41]和 scDAC[42]等。

2.3 基于弱监督学习的细胞异质性刻画

尽管无监督学习方法在细胞异质性刻画任务上表现出了良好的效果，但由于传统的无监督学习方法受限于所研究的目标数据本身，仍缺乏足够的精度。为此，许多现有的方法在刻画细胞异质性的过程中引入了弱监督学习策略，充分利用外部参考数据进行模型训练，以更多的有价值信息作为模型的参考，从而达到更准确的细胞异质性刻

画结果。

现有的弱监督细胞异质性刻画方法能够有效利用多种不同类型的数据作为参考（图 2-2）。首先，最常见的是利用 Bulk（细胞群）测序数据作为参考，与单细胞测序技术相比，Bulk 数据可能会丢失个体细胞的异质性信息，因为它提供的是细胞群体的整体平均信号。尽管如此，Bulk 数据仍能提供主要细胞类型的异质性信息以指导模型进行细胞异质性刻画。例如，Buenrostro 等人利用 Bulk 转录组数据和 Bulk 染色质开放性数据来验证单细胞测序结果可靠性，挖掘细胞整体基因表达变化并实现了细胞群体生物学过程分析[43]。通过将 Bulk 数据与单细胞数据相结合，能够提供更全面、多尺度的细胞分析视角，为深入理解细胞发育和功能提供更多线索和支持。其次，随着测序技术的发展和公共数据库的积累，公开数据库中已有海量单细胞数据。尽管不同实验条件下得到的不同数据集可能存在系统性差异，但是相同类型的细胞中仍存在一定的相似性。许多现有的方法能够结合其他单细胞数据集作为参考，以实现联合弱监督分析。最后，除测序数据外，多种已知的细胞类型特异性先验知识（如 Marker 基因信息）也可用于弱监督学习。

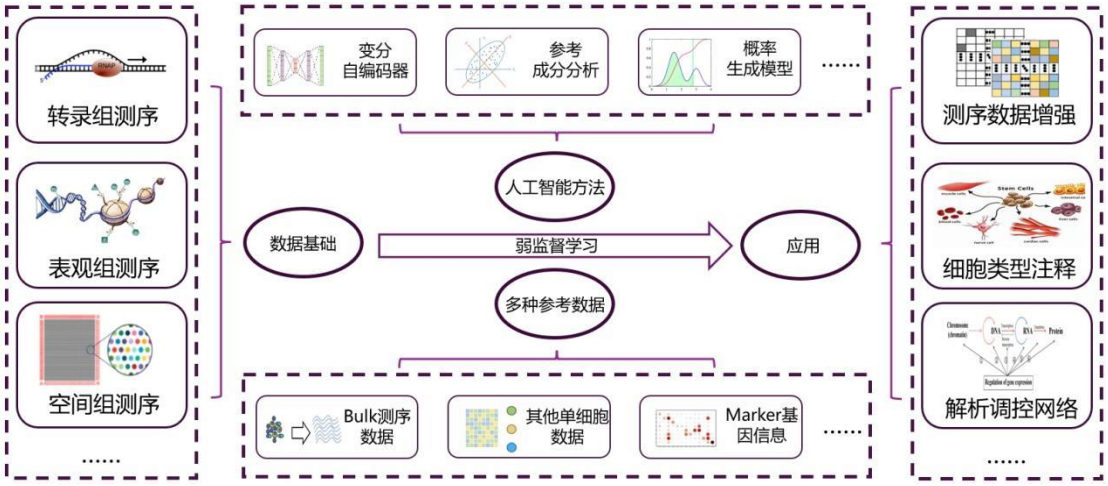


图 2-2 基于弱监督学习的细胞异质性刻画方法

2.3.1 刻画转录组数据细胞异质性的弱监督学习方法

在转录组方面，Li 等人开发了参考成分分析（RCA）方法，并刻画了人类结直肠肿瘤的细胞异质性[44]。RCA 从 BioGPS[45]下载了来自 Human U133A/GNF1H 基因图谱和原代细胞图谱的原始 Bulk 转录组数据作为参考数据，将单细胞转录组数据投影到由现有 Bulk 数据创建的全局参考面板上，并获取投影空间中的归一化坐标。结果表明，参考数据引导的聚类有较高的精度，能够有效降低数据中的技术差异和批次效应对下游分析的影响，而随着可用参考数据集的规模和多样性的扩大，参考数据引导的 RCA 的分辨率也将不断提高。CellAssign 是一种统计框架，可将单细胞转录组数据中的细胞分配给已知细胞类型[46]。CellAssign 通过计算每个细胞到细胞类型（由一组标记基因定义）或“新类”的概率来自动执行注释过程。这种唯一识别细胞类型的标记基因组合可以利用文献和数据库的专业知识建立，也可以直接从 PanglaoDB[47]等资源中获取。scINRB 则是在数据填补过程中引入了 Bulk RNA-seq 数据作为参考，即使在高缺失率和高维度的情况下，scINRB 也能准确填补缺失的基因表达值，改善细胞可视化、聚类和轨迹推断等下游分析效果[48]。

2.3.2 刻画表观组数据细胞异质性的弱监督学习方法

在表观组方面，Ji 等人开发了基于 scCAS 数据的顺式调节元件活性预测模型 SCATE[49]。SCATE 基于人和小鼠两个物种构建了参考 Bulk DNase-seq 数据库，该数据库由来自 ENCODE[50]项目生成的不同细胞类型的归一化 DNase-seq 样本组成。通过使用公开可用的 Bulk 数据，模型可以从中捕获稀疏的单细胞数据所无法捕获的宝贵信息。对于不同参考数据的需求，作者提供了接口，使得用户可以灵活地将自己的 Bulk 或伪 Bulk 数据扩充到已有的数据库中，以获取更精确的参考数据。Chen 等人开发的 RA3 是一种基于概率生成模型的 scCAS 数据分析方法[51]。RA3 可以使用 Bulk ATAC-seq 数据、Bulk

DNase-seq 数据和伪 Bulk 数据作为参考，实现对目标数据的整合分析。对于某些细胞群，特别是对于冷冻或固定组织中的细胞，可能很难获得 Bulk 测序样本，为此，RA3 提供了多种策略用于整合相同类型/聚类簇的单细胞数据来构建伪 Bulk 参考数据，这意味着其他单细胞数据集也可以有效地用于弱监督学习任务。

2.3.3 刻画空间转录组细胞异质性的弱监督学习方法

在空间转录组方面，同样发展了多种弱监督学习方法以刻画细胞的空间域异质性。例如，stPlus 是一种基于参考数据的方法，它利用单细胞转录组数据中的信息来增强空间转录组学[52]。stPlus 的输入是目标空间转录组数据和参考单细胞转录组数据，这些参考数据往往与空间数据相匹配或来自相似的组织。stPlus 可以充分利用参考数据中所有基因的整体信息，而不只局限于与空间转录组数据共享的基因。而 Li 等人开发的 PAST 方法是一种基于变分图卷积自编码器的空间转录组数据处理框架[53]。模型允许使用者从与目标空间转录组数据来自同一组织的外部空间转录组数据、相似组织的外部空间转录组数据、相似组织的外部单细胞转录组数据，或目标空间转录组数据本身作为自先验，四个方面来构建参考数据。结合参考数据，PAST 能够准确地刻画细胞的空间域异质性，有效促进空间模式域识别、空间轨迹推断等下游分析。

2.4 基于有监督学习的细胞异质性刻画

有监督的细胞异质性刻画是一种利用已知细胞标注信息指导模型识别和区分细胞类型或状态的方法。在此过程中，模型通过从带有细胞类型注释的数据集中学习特征，建立区分各种细胞类型的决策规则。相比于无监督和弱监督学习，有监督学习在刻画细胞异质性上展现出独特优势。首先，有监督学习利用细胞标注信息学习细胞类型的特异性模式，提供更为准确的细胞分类。其次，有监督学习在面对大

量高维数据时，往往能够找到更加鲁棒的特征表示。目前，基于有监督学习的细胞异质性刻画方法主要分为三大类：基于细胞间相似度的细胞异质性刻画、基于机器学习的细胞异质性刻画以及基于深度学习的细胞异质性刻画（图 2-3）。

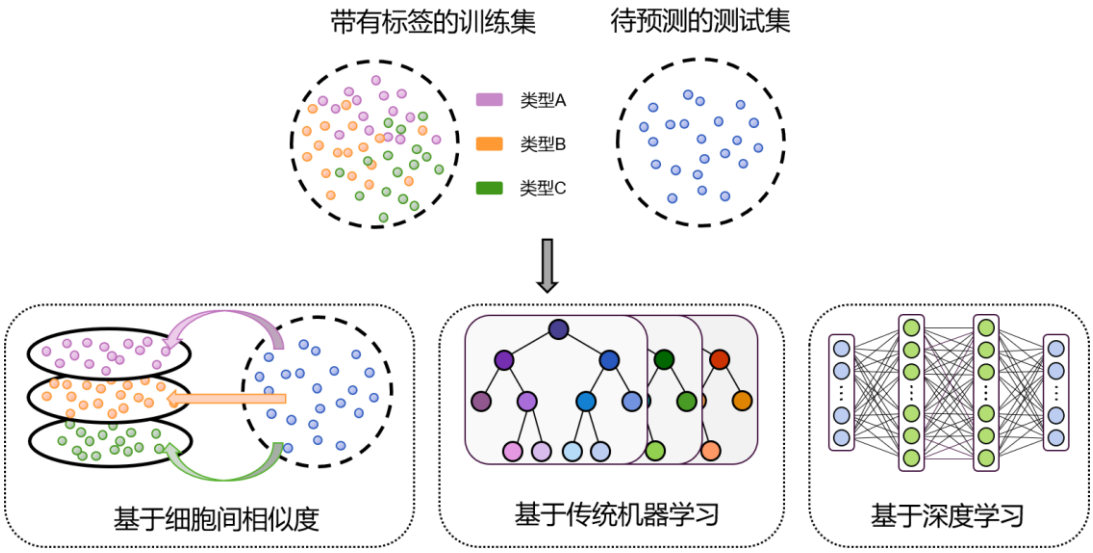


图 2-3 基于有监督学习的细胞异质性刻画方法

2.4.1 基于细胞间相似度的有监督学习方法

基于细胞间相似度的细胞异质性刻画本质上依赖于一个核心原则：属于相同类型的不同细胞在基因表达或表观修饰模式等方面具有显著的相似性。为了实现此类方法，首先需要有一个带细胞类型标注信息的数据集作为参考集。针对待标注数据集中的每一个细胞，通过皮尔逊相关系数、斯皮尔曼秩相关系数及余弦相似度等常用度量计算该细胞与参考数据集内各细胞之间的相似度。最终，每一细胞将被注释为参考集中与其最为相似的细胞所属的细胞类型。

目前，已有多种基于细胞间相似度的单细胞转录组数据注释方法。SingleR 通过选择高变基因，并计算待预测细胞与参考集中各个细胞类型的斯皮尔曼相关系数来实现对未知类型的细胞的标注[54]。CHETAH 算法则通过对参考集构建一个层次化分类树，实现对未标

注细胞的精确分类[55]。该过程首先基于参考单细胞转录组数据建立分类树，明确不同细胞类型之间的层次关系。随后，算法逐一处理输入细胞，通过遍历分类树，根据细胞的高变基因表达谱与参考集各个细胞类型的高变基因表达谱的相似度进行分类。如果一个细胞在分类过程中的任何阶段未能通过设定的阈值，其分类过程将终止，该细胞将被标记为未分配（位于树顶部）或中间状态（发生在分类树的内部）。通过这一方法，CHETAH 能够在维持高精确度的同时，有效避免对未知或未在参考数据中出现的细胞类型进行错误分类。不同于其他方法基于某个细胞和参考细胞表达谱的相似性这一原理，Cell-ID 使用的是另外一种思路：将某个细胞的特征基因集与表征细胞类型的参考基因集做富集分析，当在某个细胞类型的标记基因集上显著富集时，就将此细胞定义为该细胞类型[56]。然而，对于单细胞表观组数据而言，其维度更高、稀疏度更大，直接基于细胞间相似度进行细胞异质性刻画和细胞类型注释变得更加困难。因此，研究人员提出了 AtacAnnoR，旨在通过综合利用 scCAS 数据和单细胞转录组数据，精准地为待标注细胞分配细胞类型标签[57]。在第一轮注释中，AtacAnnoR 首先识别出细胞类型特异性的全局和邻近标记基因，通过计算待标注细胞与各参考细胞类型的基因表达之间的 Kendall’s tau 系数，识别出每个待标注细胞的候选细胞类型标签。第二轮注释中，通过清理和重新分配候选种子细胞的标签，使用加权 k-最近邻（WKNN）算法进一步精确注释细胞类型。

2.4.2 基于机器学习的有监督学习方法

然而，基于细胞间相似度的细胞异质性分析方法在处理高维数据时面临挑战，它们往往无法充分考虑变量间的非线性关系，也不具备自动提取复杂特征的能力。相对而言，基于机器学习的方法能够有效处理更为复杂的数据结构，不局限于线性关系，能够识别和学习到细胞异质性的深层次模式，从而构建出更加精准的细胞分类模型。

目前，已有多种基于机器学习的单细胞转录组数据细胞类型注释方法。例如，scmap 将待标注数据映射到参考数据集所在隐空间上，并利用 K 近邻算法实现细胞类型的注释[58]。scPred 则采用奇异值分解来识别具有高预测能力的基因，并使用这些基因训练支持向量机以分类细胞[59]。Garnett 利用单细胞转录组数据和预定义的细胞类型特异性标记基因来训练基于广义线性模型的分类器，从而注释细胞类型[60]。SciBet 通过 E-test 选取对分类重要的基因，并基于这些基因的平均表达值建立每个细胞类型的多项式模型[61]。在细胞类型分配过程中，SciBet 比较待标注细胞的基因表达谱和不同细胞类型模型的似然函数，以确定最匹配的细胞类型。devCellPy 则引入了 LayerObject 类来组织数据结构，使算法能学习数据集的注释层次，并在该层次结构中为每层训练一个 XGBoost 预测模型，这样可以自动地在正确的层次分支上对细胞亚型进行分类，从而精准地注释细胞类型[62]。

2.4.3 基于深度学习的有监督学习方法

尽管传统机器学习方法在单细胞数据的异质性刻画中取得了一定的成效，但这些机器学习模型通常需要手动选择特征，并且往往对高维数据和噪声敏感。相较于传统机器学习方法，基于深度学习的方法在表征细胞异质性时存在明显优势。深度学习方法通过自动特征学习减少了对先验知识的依赖，并且能从原始数据中直接提取复杂和非线性的特征，因此更适合处理高维与复杂的单细胞数据。

近年来，多个基于深度学习的单细胞转录组数据细胞异质性刻画方法相继发表。SuperCT 是第一个不依赖无监督聚类的单细胞转录组数据的深度学习细胞类型辨识方法，它基于全连接神经网络构建模型，并使用二进制信号表示基因表达水平来进行模型训练[63]。相较于 SuperCT 完全依赖于神经网络，Cell BLAST 额外引入了参考数据，通过采用一个基于神经网络的生成模型，实现了一种高度先进的单细胞转录组数据细胞异质性刻画方法[64]。该方法利用参考单细胞转录组

数据，自适应地学习从高维转录组空间到低维细胞嵌入空间的非线性映射，将待标注细胞映射到与参考细胞相同的低维空间中。接着，Cell BLAST 依赖于低维空间内的后验分布来精确地注释细胞类型。scDeepSort 则是一个基于加权图神经网络框架的预训练细胞类型注释方法[65]，模型由三个部分组成：用于存储图节点的嵌入层、学习图结构信息的加权图聚合层和最终输出细胞类型预测结果的线性分类层。通过在多个单细胞转录组数据中进行预训练，scDeepSort 能够实现稳健的细胞类型预测。scBERT 同样是一个预训练模型，受自然语言处理领域的 BERT（Bidirectional Encoder Representation from Transformers）模型的启发，scBERT 将这一基于 Transformer 的双向编码器表示模型应用于单细胞转录组数据[66]。通过在大量未标记的单细胞转录组数据上进行预训练，scBERT 获得了基因间交互作用的理解，然后将其转移到未训练和用户特定的单细胞转录组数据的细胞类型注释任务上进行监督微调，实现了稳健且准确的细胞类型注释。TOSICA 是一个基于 Transformer 的多头自注意力深度学习模型，能够使用生物学上的可解释对象（如通路或调控网络）进行可解释的细胞异质性刻画和细胞类型注释[67]。

在单细胞表观遗传组学方面，也有许多基于深度学习刻画细胞异质性的有监督方法。其中，EpiAnno 是针对 scCAS 数据提出的第一个细胞类型自动注释方法，是一个基于贝叶斯神经网络的概率生成模型，在 scCAS 数据的注释上有卓越性能[68]。RAINBOW 基于对比学习框架构建模型并融入参考数据，可以有效刻画细胞异质性并准确识别数据集中的新细胞类型[69]。CASCADE 则在全连接神经网络的基础上引入了仿真策略和基于 Masked Autoencoder 的去噪策略，在连续和不平衡的 scCAS 数据上的注释性能显著优于已有方法[70]。不同于上述方法，Cellcano 是一个两轮的有监督学习算法，它首先在参考数据集上训练多层感知机，并预测目标数据中的细胞类型，然后从预测结果中

选择一些被认为预测良好的目标细胞（称为锚点）组成新的训练集，使用这一带有伪标签的新训练集对知识蒸馏模型进行训练，以对剩余非锚点细胞进行注释，从而缓解了训练数据和目标数据之间的分布偏移问题[71]。

参考文献

[1] Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1-3): 37-52.

[2] Kiselev, V.Y., Kirschner, K., Schaub, M.T. et al. SC3: Consensus clustering of single-cell RNA-seq data. Nature Methods, 2017, 14(5): 483-486.

[3] Hao, Y., Stuart, T., Kowalski, M.H. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology, 2024, 42(2): 293-304.

[4] žurauskiene, J. & Yau, C. pcaReduce: Hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics, 2016, 17(1): 1-11.

[5] Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Research, 2016, 44(13): e117.

[6] Cao, J., Spielmann, M., Qiu, X. et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature, 2019, 566(7745): 496-502.

[7] Chen, H., Albergante, L., Hsu, J.Y. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nature Communications, 2019, 10(1): 1-14.

[8] Prompsy, P., Kirchmeier, P., Marsolier, J. et al. Interactive analysis of single-cell epigenomic landscapes with ChromSCape. Nature Communications, 2020, 11(1): 5702.

[9] Danese, A., Richter, M.L., Chaichoompu, K. et al. EpiScanpy: integrated single-cell epigenomic analysis. Nature Communications, 2021, 12(1): 5228.

[10] Linderman, G.C., Zhao, J., Roulis, M. et al. Zero-preserving

imputation of single-cell RNA-seq data. Nature Communications, 2022, 13(1): 192.

[11] Granja, J.M., Corces, M.R., Pierce, S.E. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature Genetics, 2021, 53(3): 403-411.

[12] Stuart, T., Srivastava, A., Madad, S. et al. Single-cell chromatin state analysis with Signac. Nature Methods, 2021, 18(11): 1333-1341.

[13] Jain, M.S., Polanski, K., Conde, C.D. et al. MultiMAP: dimensionality reduction and integration of multimodal data. Genome Biology, 2021, 22(1): 1-26.

[14] Welch, J.D., Kozareva, V., Ferreira, A. et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell, 2019, 177(7): 1873-1887.e1817.

[15] Argelaguet, R., Velten, B., Arnol, D. et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 2018, 14(6): e8124.

[16] Argelaguet, R., Arnol, D., Bredikhin, D. et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology, 2020, 21(1): 1-17.

[17] Zhang, W., Xue, X., Zheng, X. et al. NMFLRR: Clustering scRNA-Seq Data by Integrating Nonnegative Matrix Factorization with Low Rank Representation. IEEE Journal of Biomedical and Health Informatics, 2022, 26(3): 1394-1405.

[18] Li, Z., Kuppe, C., Ziegler, S. et al. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. Nature Communications, 2021, 12(1): 6386.

[19] Tang, S., Cui, X., Wang, R. et al. scCASE: accurate and

interpretable enhancement for single-cell chromatin accessibility sequencing data. Nature Communications, 2024, 15(1): 1629.

[20] Wang, B., Zhu, J., Pierson, E. et al. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. Nature Methods, 2017, 14(4): 414-416.

[21] Huh, R., Yang, Y., Jiang, Y. et al. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. Nucleic Acids Research, 2020, 48(1): 86-95.

[22] Zamanighomi, M., Lin, Z., Daley, T. et al. Unsupervised clustering and epigenetic classification of single cells. Nature Communications, 2018, 9(1): 2410.

[23] Bravo González-Blas, C., Minnoye, L., Papasokrati, D. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nature Methods, 2019, 16(5): 397-400.

[24] Kapourani, C.A. & Sanguinetti, G. Melissa: Bayesian clustering and imputation of single-cell methylomes. Genome Biology, 2019, 20(1): 61.

[25] Lopez, R., Regier, J., Cole, M.B. et al. Deep generative modeling for single-cell transcriptomics. Nature Methods, 2018, 15(12): 1053-1058.

[26] Yuan, H. & Kelley, D.R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nature Methods, 2022, 19(9): 1088-1096.

[27] Tran, D., Nguyen, H., Tran, B. et al. Fast and precise single-cell data analysis using a hierarchical autoencoder. Nature Communications, 2021, 12(1): 1029.

[28] Grønbech, C.H., Vording, M.F., Timshel, P.N. et al. ScVAE:

Variational auto-encoders for single-cell gene expression data. Bioinformatics, 2020, 36(16): 4415-4422.

[29] Liu, Q., Chen, S., Jiang, R. et al. Simultaneous deep generative modelling and clustering of single-cell genomic data. Nature Machine Intelligence, 2021, 3(6): 536-544.

[30] Xiong, L., Xu, K., Tian, K. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nature Communications, 2019, 10(1): 2410.

[31] Xiong, L., Tian, K., Li, Y. et al. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. Nature Communications, 2022, 13(1): 6118.

[32] Cao, K., Gong, Q., Hong, Y. et al. A unified computational framework for single-cell data integration with optimal transport. Nature Communications, 2022, 13(1): 7419.

[33] Ashuach, T., Reidenbach, D.A., Gayoso, A. et al. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. Cell Reports Methods, 2022, 2(3):

[34] Lin, E., Mukherjee, S. & Kannan, S. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. BMC Bioinformatics, 2020, 21(1): 1-11.

[35] Zhu, X., Meng, S., Li, G. et al. AGImpute: imputation of scRNA-seq data based on a hybrid GAN with dropouts identification. Bioinformatics, 2024, 40(2): btae068.

[36] Huang, Z., Wang, J., Lu, X. et al. scGGAN: single-cell RNA-seq imputation by graph-based generative adversarial network. Briefings in Bioinformatics, 2023, 24(2): bbad040.

[37] Wang, J., Ma, A., Chang, Y. et al. scGNN is a novel graph neural

network framework for single-cell RNA-Seq analyses. Nature Communications, 2021, 12(1): 1882.

[38] Gu, H., Cheng, H., Ma, A. et al. scGNN 2.0: a graph neural network tool for imputation and clustering of single-cell RNA-Seq data. Bioinformatics (Oxford, England), 2022, 38(23): 5322-5325.

[39] Li, H., Sun, Y., Hong, H. et al. Inferring transcription factor regulatory networks from single-cell ATAC-seq data based on graph neural networks. Nature Machine Intelligence, 2022, 4(4): 389-400.

[40] Tian, T., Wan, J., Song, Q. et al. Clustering single-cell RNA-seq data with a model-based deep learning approach. Nature Machine Intelligence, 2019, 1(4): 191-198.

[41] Li, X., Wang, K., Lyu, Y. et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nature Communications, 2020, 11(1): 2338.

[42] An, S., Shi, J., Liu, R. et al. scDAC: deep adaptive clustering of single-cell transcriptomic data with coupled autoencoder and dirichlet process mixture model. Bioinformatics, 2024, btae198.

[43] Buenrostro, J.D., Corces, M.R., Lareau, C.A. et al. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell, 2018, 173(6): 1535-1548.e1516.

[44] Li, H., Courtois, E.T., Sengupta, D. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nature Genetics, 2017, 49(5): 708-718.

[45] Wu, C., Orozco, C., Boyer, J. et al. BioGPS: An extensible and customizable portal for querying and organizing gene annotation resources. Genome Biology, 2009, 10(11): 1-8.

[46] Zhang, A.W., O'Flanagan, C., Chavez, E.A. et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nature Methods, 2019, 16(10): 1007-1015.

[47] Franzén, O., Gan, L.M. & Björkegren, J.L.M. PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. Database, 2019, 2019(1): baz046.

[48] Kang, Y., Zhang, H. & Guan, J. scINRB: single-cell gene expression imputation with network regularization and bulk RNA-seq data. Briefings in Bioinformatics, 2024, 25(3): bbae148.

[49] Ji, Z., Zhou, W., Hou, W. et al. Single-cell ATAC-seq signal extraction and enhancement with SCATE. Genome Biology, 2020, 21(1): 1-36.

[50] Luo, Y., Hitz, B.C., Gabdank, I. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Research, 2020, 48(D1): D882-D889.

[51] Chen, S., Yan, G., Zhang, W. et al. RA3 is a reference-guided approach for epigenetic characterization of single cells. Nature Communications, 2021, 12(1): 2177.

[52] Shengquan, C., Boheng, Z., Xiaoyang, C. et al. StPlus: A reference-based method for the accurate enhancement of spatial transcriptomics. Bioinformatics, 2021, 37(Supplement_1): I299-I307.

[53] Li, Z., Chen, X., Zhang, X. et al. Latent feature extraction with a prior-based self-attention framework for spatial transcriptomics. Genome Research, 2023, 33(10): 1757-1773.

[54] Aran, D., Looney, A.P., Liu, L. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology, 2019, 20(2): 163-172.

[55] de Kanter, J.K., Lijnzaad, P., Candelli, T. et al. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Research, 2019, 47(16): E95.

[56] Cortal, A., Martignetti, L., Six, E. et al. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. Nature Biotechnology, 2021, 39(9): 1095-1102.

[57] Tian, L., Xie, Y., Xie, Z. et al. AtacAnnoR: a reference-based annotation tool for single cell ATAC-seq data. Briefings in Bioinformatics, 2023, 24(5): bbad268.

[58] Kiselev, V.Y., Yiu, A. & Hemberg, M. Scmap: Projection of single-cell RNA-seq data across data sets. Nature Methods, 2018, 15(5): 359-362.

[59] Alquicira-Hernandez, J., Sathe, A., Ji, H.P. et al. ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biology, 2019, 20(1): 1-17.

[60] Pliner, H.A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. Nature Methods, 2019, 16(10): 983-986.

[61] Li, C., Liu, B., Kang, B. et al. SciBet as a portable and fast single cell type identifier. Nature Communications, 2020, 11(1): 1818.

[62] Galdos, F.X., Xu, S., Goodyer, W.R. et al. devCellPy is a machine learning-enabled pipeline for automated annotation of complex multilayered single-cell transcriptomic data. Nature Communications, 2022, 13(1): 5271.

[63] Xie, P., Gao, M., Wang, C. et al. SuperCT: A supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. Nucleic Acids Research, 2019, 47(8): e48-e48.

[64] Cao, Z.J., Wei, L., Lu, S. et al. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. Nature Communications, 2020, 11(1): 3458.

[65] Shao, X., Yang, H., Zhuang, X. et al. ScDeepSort: A pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. Nucleic Acids Research, 2021, 49(21): E122.

[66] 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. Nature Machine Intelligence, 2022, 4(10): 852-866.

[67] Chen, J., Xu, H., Tao, W. et al. Transformer for one stop interpretable cell type annotation. Nature Communications, 2023, 14(1): 223.

[68] Chen, X., Chen, S., Song, S. et al. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. Nature Machine Intelligence, 2022, 4(2): 116-126.

[69] Li, S., Tang, S., Wang, Y. et al. Accurate cell type annotation for single-cell chromatin accessibility data via contrastive learning and reference guidance. Quantitative Biology, 2024, 12(1): 85-99.

[70] Jia, Y., Li, S., Jiang, R. et al. Accurate Annotation for Differentiating and Imbalanced Cell Types in Single-cell Chromatin Accessibility Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024, 1-11.

[71] Ma, W., Lu, J. & Wu, H. Cellcano: supervised cell type identification for single cell ATAC-seq data. Nature Communications, 2023, 14(1): 1864.

第 3 章 人工智能赋能疾病诊疗

3.1 引言

随着全球人口的增长和老龄化趋势的加剧，医疗资源的短缺和医疗成本的上升成为各国面临的重大挑战。根据世界卫生组织（World Health Organization, WHO）的数据，预计到 2030 年，全球将有六分之一的人口超过 60 岁，这将对医疗系统的可持续性和效率带来巨大压力[1]。与此同时，复杂疾病如癌症、心血管疾病等的发病率不断上升，迫切要求医疗服务向更加精准和个性化的方向发展。如图 3-1 所示，传统的医疗模式已经难以满足日益增长的健康需求，特别是在大数据时代，传统的手工处理和分析方法已经无法有效处理海量的医疗数据。因此，医疗行业迫切需要新的技术和方法来提升诊疗效率和效果，同时降低成本，更好地满足人们对健康管理的需求。

人工智能（Artificial Intelligence, AI）凭借其强大的数据处理和分析能力，在医疗领域展现出巨大的潜力。AI 技术能够从多种数据源中提取、分析和利用信息，为医生和医疗机构提供决策支持和个性化治疗方案。研究表明，AI 在癌症早期诊断、药物研发、病理图像分析等领域取得了显著进展，为医疗行业带来新的希望和机遇[2]。

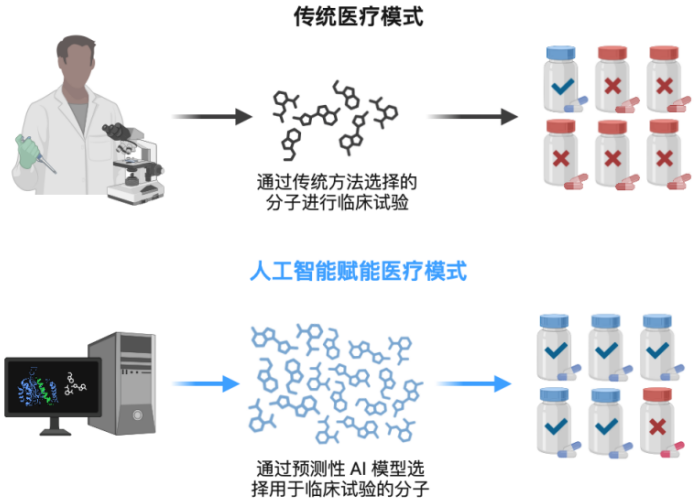


图 3-1 传统医疗模式与人工智能赋能医疗模式的比较（图片引自 biorender.com）

3.2 关键技术和应用

3.2.1 机器学习与深度学习

机器学习（Machine Learning，ML）是人工智能的核心技术之一，通过算法和模型从数据中学习规律和模式，从而进行预测和决策。传统机器学习算法有许多形式，大多数被设计用于处理表格数据，其中每个数据点都有一组明确的特征（例如，病人的年龄或基因突变状态），用于预测标签[3]。如图 3-2 所示，其中一种常见的算法称为随机森林（Random Forest, RF），它由一组决策树组成，每棵树基于训练数据构建，对输入特征进行一系列二进制决策，最终预测数据点的标签。另一个算法是支持向量机（Support Vector Machines, SVM），它在由输入特征定义的坐标系中学习一条直线（或多维空间中的超平面），将数据点分成两类。回归模型则通过学习输入特征的线性组合来预测连续标签（例如，线性回归（Linear Regression））或二元标签（例如，逻辑回归（Logistic Regression））。在医疗领域，机器学习可以应用于疾病预测、患者风险评估和个性化治疗方案的制定[4]。例如，一些基于机器学习的模型可以分析海量的临床数据和生物标志物，辅助医生精确预测患者患病风险，从而促进早期干预[5-7]。

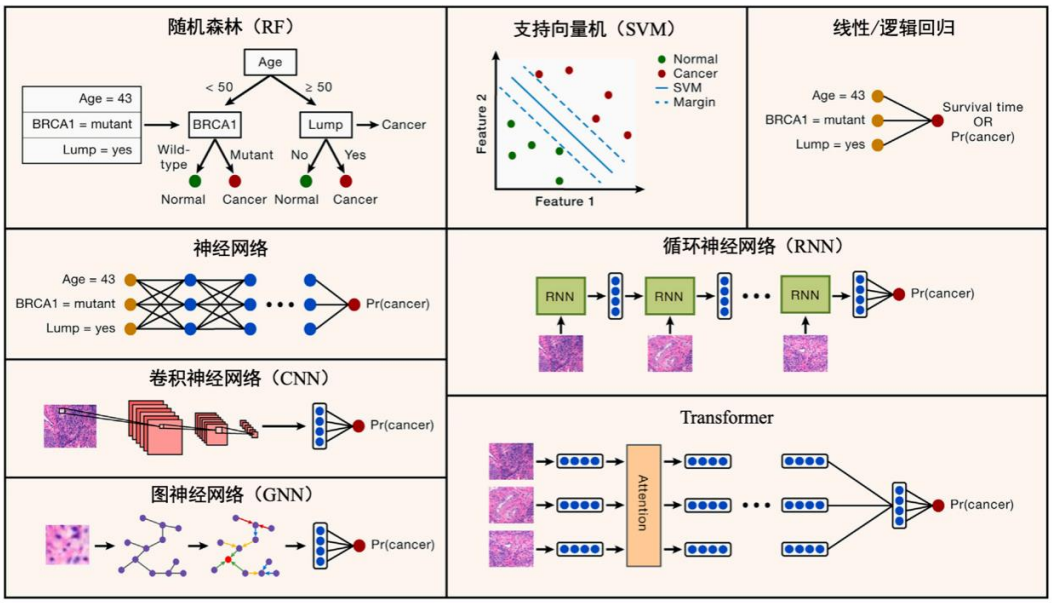


图 2 常见的机器学习模型（改编自[8]）

随着图形处理单元（Graphics Processing Unit, GPU）的广泛应用和性能提升，深度学习（Deep Learning, DL）作为机器学习的一个重要分支，已经在许多预测任务中逐步取代传统机器学习方法。DL 模型的核心组件是神经网络，它由一个或多个层次的单元组成，这些单元称为神经元，它们计算输入的加权和，然后应用非线性函数，生成一种称为嵌入（Embedding）的输入表示，最终用于预测输出。与传统机器学习模型相比，DL 模型具有更强的灵活性，并减少了对特征工程的依赖，因此它们擅长处理复杂的大数据和更广泛的非结构化数据类型，包括图像、文本和语音等[8]。然而，DL 模型通常需要更多的训练数据，这使得传统的机器学习模型在数据资源受限或处理表格数据的任务中仍然发挥着不可或缺的作用。为了处理非表格数据，神经网络的架构（例如，神经元或层次或神经元之间的连接数）被修改以适应所需的数据类型。如图 3-2 所示，卷积神经网络（Convolutional Neural Networks, CNN）主要用于提取图像特征。图神经网络（Graph Neural Networks, GNN）处理图数据，例如细胞-细胞相互作用图[9]或者药物分子结构[10]。递归神经网络（Recurrent Neural Networks, RNN）和 Transformer 网络则分析顺序数据，例如遗传序列或图像序列。这些模型类别中的每个都有许多特定的模型架构，例如基于 CNN 的 ResNet[11]或 U-Net[12]以及基于 RNN 的 LSTM[13]或 GRU[14]。综上所述，深度学习技术的快速发展和应用，正在为人工智能赋能疾病诊疗领域带来前所未有的机遇和挑战。

3.2.2 自然语言处理技术

自然语言处理（Natural Language Processing, NLP）技术使计算机能够理解、处理和生成自然语言文本。在生物医学领域，NLP 的应用尤为广泛，特别是在处理电子健康记录（Electronic Health Record, EHR）、医学文献和生物医学文本数据的分析中。例如，BioBERT[15]和 BlueBERT[16]模型都是基于 BERT[17]架构，专门为大规模生物医学

数据的预训练而设计。BioBERT 在命名实体识别（Named Entity Recognition，NER）、关系提取和问答系统等多种生物医学 NLP 任务中展现了显著的性能提升。BlueBERT 则通过在生物医学文献和临床记录的混合数据上进行训练，进一步强化了其处理临床和生物医学文本任务中的能力。此外，基于 GPT[19]架构的 BioGPT[18]专注于生物医学文本生成和理解。通过在广泛的生物医学语料库上的预训练，BioGPT 在生成相关领域文本和解答生物医学问题上表现出色。NLP 技术的应用使得医疗机构能够高效地从庞大复杂的医疗文本中自动提取关键信息，为临床决策和个性化治疗方案的制定提供重要支持[20]。

3.2.3 医疗图像分析技术

医疗图像分析技术借助深度学习算法实现了对医学影像的自动化分析和解读。在基于图像的癌症预测任务中，典型的机器学习工作流程如图 3-3 所示。这些算法能够精准识别各类医学影像（如 X 线摄影（X-ray）、超声影像（Ultrasound）、计算机断层成像（Computed Tomography, CT）、磁共振成像（Magnetic resonance imaging, MRI）以及正电子发射计算机断层显像（Positron Emission Tomography, PET）等）中的病变特征和异常，辅助医生进行更精确的诊断和治疗规划[21]。

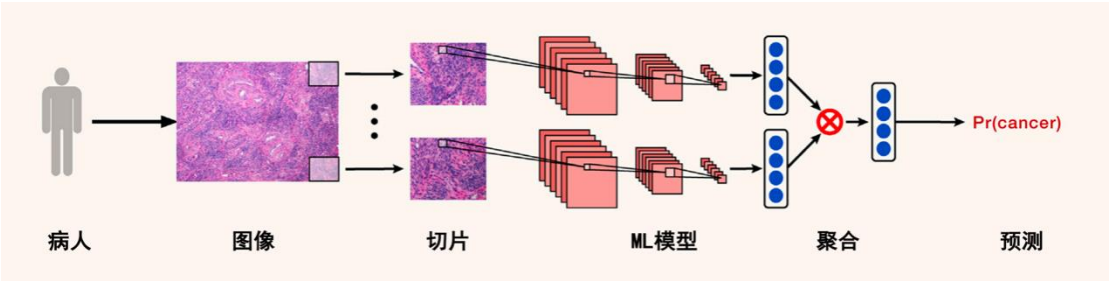


图 3-3 基于图像的癌症预测任务的通用机器学习模型工作流程（改编自[8]）

在这些技术中，深度学习模型的设计对医疗图像分割任务尤为重要。以 U-Net 为代表经典的 CNN 模型专为生物医学图像分割而设计，展现了极高的适用性和准确性。U-Net 以独特的 U 形结构而著称，该结构利用下采样路径捕捉图像的全局上下文信息，并通过上采样路径

实现细节的精细分割。U-Net 通过跳跃连接技术，将下采样路径中的特征图与上采样路径中的特征图融合，以保留高分辨率的特征信息[12]。例如，在肺部 CT 扫描中，U-Net 可以有效识别和分割出肺结节，为早期肺癌诊断提供支持[22]。与 U-Net 类似，V-Net 采用对称的编码器-解码器结构，其中编码器通过卷积和下采样提取图像特征，而解码器则通过反卷积和上采样生成分割结果。V-Net 特别使用 Dice 损失函数进行优化，使其在处理不平衡数据集时具有独特优势[23]。例如，在前列腺 MRI 图像中，V-Net 能够精确分割前列腺边界和内部结构[24]。此外，nnU-Net[25]和 Attention U-Net[26]在 U-Net 基础上分别引入了自适应模块和注意力机制。nnU-Net 通过自动化配置简化了参数调整过程，在多种任务中均展现出优秀的分割性能。在 BraTS2021 挑战中，nnU-Net 以卓越的脑肿瘤分割性能脱颖而出[27]。而 Attention U-Net 能够动态调整特征图的权重，专注于关键图像区域，进一步提高了分割的精度度。总体而言，基于深度学习的医疗图像分析系统在乳腺癌[28-29]、肺癌[30-31]等众多疾病的早期筛查中已经取得了显著进展，为疾病的精准诊断和治疗开辟了新路径。

3.2.4 知识图谱与数据整合技术

知识图谱是一种高效的结构化知识表示方法，能够精确捕捉并整合广泛的医学知识，从而帮助医生更好地理解疾病的复杂性和治疗选项。这种技术通过整合多源数据，包括基因组学数据、临床记录数据、病理报告数据等，将这些信息关联起来形成全面的疾病模型。例如，知识图谱可以将患者的遗传信息和临床症状相结合，揭示特定基因变异与疾病之间的关联，为个性化医疗提供科学依据，并指导制定针对性的治疗方案。此外，知识图谱在智能化医疗决策中扮演着关键角色。它利用自动化推理和先进的推荐系统，根据最新的医学研究和临床实践指南，为医生提供及时的、基于证据的诊疗建议。这种智能化支持不仅优化了治疗流程，还显著提升了医疗服务的整体质量[32-33]。除此

之外，如图 3-4 所示，知识图谱在医学研究和药物开发领域同样发挥着不可或缺的作用。通过对海量文献和临床试验数据的深入分析，知识图谱有助于识别新的疾病相关性、潜在的药物靶点以及创新的治疗策略。这种分析能力极大地加速了新药研发的进程，为医学界带来了前所未有的研究动力和创新潜力[34]。

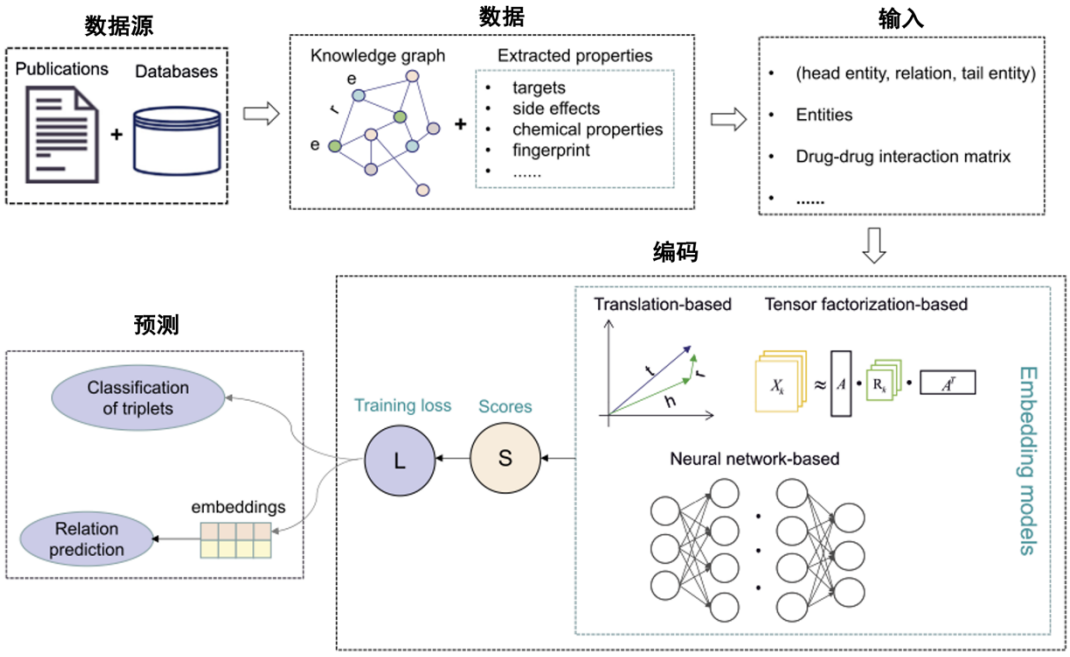


图 3-4 基于知识图谱的药物发现预测流程（改编自[35]）

3.2.5 生命科学领域的基础模型

生命科学领域大模型通常结合了多种生物医学数据（如基因组学、转录组学、蛋白质组学等），利用机器学习和深度学习技术进行综合分析。然而，与图像和文本数据相比，解码生命“语言”是一项更为复杂的任务，这一过程需要依赖大量生物学数据来建立数据间的联系。特别是在研究罕见疾病或在难以直接获取组织样本的情况下，这一任务的复杂性进一步增加。

迁移学习（Transfer Learning）为这一挑战提供了解决路径。通过在大规模通用数据集上预训练深度学习模型，并将其针对特定任务的有限数据进行微调，迁移学习策略使模型能够快速适应新的任务并支持多样化的下游应用。此外，基础模型（Foundation Models）已经在自然语言处理领域和计算机视觉领域取得了重大进展，并证明了其跨领域的适用性。如图 3-5 和图 3-6 所示，这些模型在蛋白质设计领域和单细胞转录组学领域显示出巨大发展潜力，为我们深入理解生命复杂性提供了全新视角和强有力的工具。

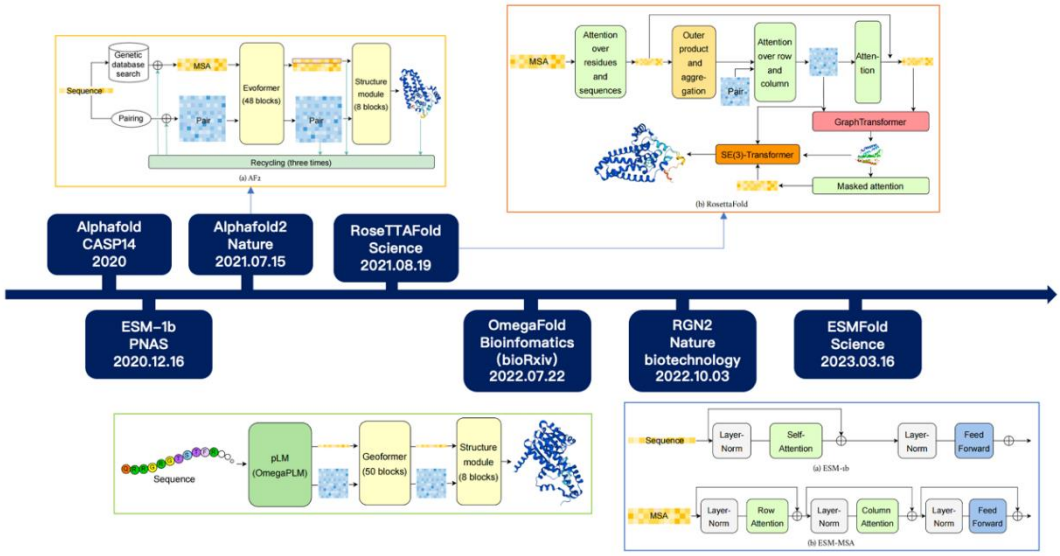


图 3-5 蛋白质设计领域相关模型的发展历程

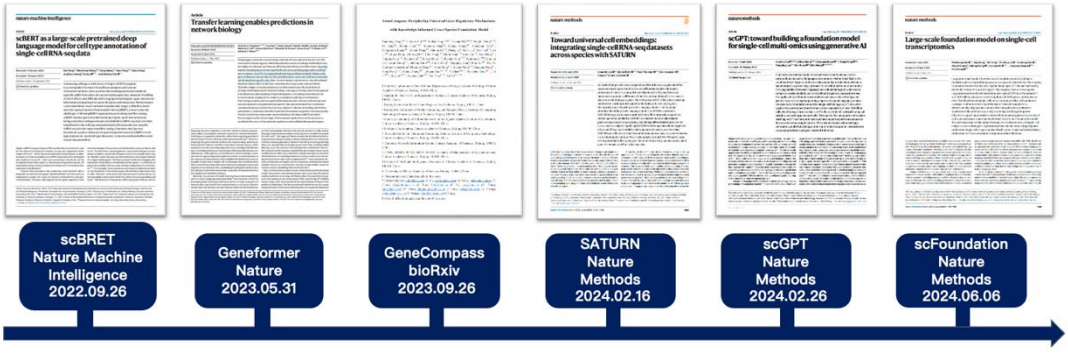


图 3-6 单细胞转录组学领域相关模型的发展历程

在蛋白质结构预测领域，DeepMind 开发的 AlphaFold 模型[36]通

过在大量已知蛋白质结构数据上的训练，利用深度学习技术显著提升了蛋白质三维结构预测的准确性。这一突破为理解蛋白质功能、药物设计和疾病机制提供了重要的工具。此外，AlphaFold[37]在罕见疾病研究中也展现了卓越的能力，通过预测与疾病相关的蛋白质结构，揭示了潜在的治疗靶点。例如，在遗传性疾病如囊性纤维化和亨廷顿舞蹈症的研究中，AlphaFold 的应用为揭示疾病的分子机制和寻找有效的药物候选分子提供了全新的视角。这种技术的进步不仅大幅提高了研究效率，还显著减少了传统实验方法的高昂成本，为全球生物医学研究带来了深远的影响。

在单细胞转录组学领域，一系列单细胞转录组大模型如 Geneformer[38]、scGPT[39]、scFoundation[40]、GeneCompass[41]等相继出现，预示着人工智能在生物医学领域的广泛应用和深远影响。这些模型通过在大规模的细胞转录组学数据上的预训练，具备强大的数据理解和处理能力，并在多种生物医学任务中实现出色的性能。例如，Geneformer 在有限患者数据的疾病建模中，成功识别出了心肌病的候选治疗靶点，加速研究人员发现关键网络调控因子和潜在治疗靶点。这一发现对于心脏病患者的精准诊断和个性化治疗至关重要，标志着 AI 技术在疾病预防和治疗中的潜力。scFoundation 结合了 xTrimoGene 架构和测序深度感知任务（Read-depth-aware, RDA），为细胞扰动响应预测、药物靶点发现等领域带来了创新的工具和方法，探索并推动了单细胞领域基础模型的边界（图 3-7）。这些模型的出现和应用标志着人工智能技术在生物医学研究中的日益成熟，为未来的科学探索和临床实践开辟了新的可能性。

图 3-7 scFoundation 模型的部分下游应用（摘自[40]）

3.3 展望

尽管人工智能在医疗领域展现出巨大的潜力，但仍面临诸多挑战。

首先，生物医学数据质量和隐私保护是亟需解决的核心问题。医疗数据涉及患者的敏感信息，任何数据泄露都可能导致严重的后果。因此，医疗机构在数据收集、存储和处理过程中必须严格遵守相关法律法规，如 GDPR（通用数据保护条例）和 HIPAA（健康保险携带与责任法案），并加强数据安全措施，确保患者隐私得到保护。

其次，人工智能模型的可解释性和在不同环境下的适应能力也限制了其在临床实践中的广泛应用。尽管 AI 在数据分析和预测方面展现了卓越的性能，但其“黑箱”特性使得医生和患者难以理解其决策过程。医生和患者需要对 AI 模型的诊断结果和治疗建议有充分的信任，而这种信任的建立依赖于对 AI 决策依据的透明化和可解释性。因此，提高 AI 模型的可解释性，采用如 LIME（局部可解释模型）和 SHAP（Shapley 值）等技术，提高模型的透明度，是当前研究的一个重要方向。

此外，AI 模型的适应能力也是一个关键问题，医疗环境的多样性和患者个体差异要求 AI 系统能够灵活适应不同的临床情境，这对 AI 模型的普适性和可靠性提出了更高的要求。

尽管如此，这些挑战背后也潜藏着巨大的机遇。人工智能可以通

过个性化治疗和预测，结合个体的基因组数据、生活方式特征和疾病历史，为每位患者量身定制最有效的治疗方案。这种精准医疗不仅可以提高治疗效果，还能减少不必要的治疗，降低医疗成本，同时帮助患者获得更好的治疗体验。同时，结合图像识别和自然语言处理技术，人工智能能够帮助医生在疾病的早期阶段进行迅速而准确的诊断。例如，通过智能影像分析，AI 可以帮助检测乳腺癌、肺癌等疾病的早期症状，提高早期筛查的效率和准确性，从而显著改善治疗效果和患者生存率。此外，通过远程医疗服务和智能化的健康管理系统，人工智能还能极大地扩展医疗服务的覆盖范围，从而打破时间和地域的限制，为偏远地区的患者提供高质量的医疗服务。这一技术不仅能够改善基层医疗服务的质量，还能够通过数据的集成和智能化管理，提高全球医疗资源的利用效率。

通过这些努力，我们有望在未来完成跨学科的合作和技术整合，促进生物医学数据的整合和共享，推动新技术的创新和应用实现更加智能化和个性化的医疗健康服务，为全球范围内的患者带来更好的健康成果和生活质量。

参考文献

[1] Wilmoth J R, Bas D, Mukherjee S, et al. World social report 2023: Leaving no one behind in an ageing world[M]. UN, 2023.

[2] Murali[1] N, Sivakumaran N. Artificial intelligence in healthcare–a review[J]. 2018.

[3] Boehm K M, Khosravi P, Vanguri R, et al. Harnessing multimodal data integration to advance precision oncology[J]. Nature Reviews Cancer, 2022, 22(2): 114-126.

[4] Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology[J]. Nature Reviews Cancer, 2018, 18(8): 500-510.

[5] Janssen B V, Verhoef S, Wesdorp N J, et al. Imaging-based machine-learning models to predict clinical outcomes and identify biomarkers in pancreatic cancer: a scoping review[J]. Annals of surgery, 2022, 275(3): 560-567.

[6] Jin T, Nguyen N D, Talos F, et al. ECMarker: interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages[J]. Bioinformatics, 2021, 37(8): 1115-1124.

[7] Jiang Y Z, Ma D, Jin X, et al. Integrated multiomic profiling of breast cancer in the Chinese population reveals patient stratification and therapeutic vulnerabilities[J]. Nature Cancer, 2024, 5(4): 673-690.

[8] Swanson K, Wu E, Zhang A, et al. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment[J]. Cell, 2023, 186(8): 1772-1791.

[9] Armingol E, Baghdassarian H M, Lewis N E. The diversification of methods for studying cell–cell interactions and communication[J]. Nature Reviews Genetics, 2024, 25(6): 381-400.

[10] Xiong J, Xiong Z, Chen K, et al. Graph neural networks for automated de novo drug design[J]. Drug discovery today, 2021, 26(6): 1382-1393.

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[12] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015: 234-241.

[13] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[14] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[15] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[16] Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets[J]. arXiv preprint arXiv:1906.05474, 2019.

[17] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[18] Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in

bioinformatics, 2022, 23(6): bbac409.

[19] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.

[20] Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review[J]. Computers in biology and medicine, 2023, 155: 106649.

[21] Zhou T, Dong Y, Huo B, et al. U-Net and its applications in medical image segmentation: a review[J]. Journal of Image and Graphics, 2021, 26(9): 2058-2077.

[22] 钟思华, 郭兴明, 郑伊能. 改进 U-Net 网络的肺结节分割方法[J]. Journal of Computer Engineering & Applications, 2020, 56(17).

[23] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). Ieee, 2016: 565-571.

[24] Aldoj N, Biavati F, Michallek F, et al. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net[J]. Scientific reports, 2020, 10(1): 14315.

[25] Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203-211.

[26] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.

[27] Luu H M, Park S H. Extending nn-UNet for brain tumor

segmentation[C]//International MICCAI brainlesion workshop. Cham: Springer International Publishing, 2021: 173-186.

[28] 王一凡, 刘静, 马金刚, 等. 深度学习在乳腺癌影像学检查中的应用进展[J]. Journal of Frontiers of Computer Science & Technology, 2024, 18(2).

[29] 王彤, 何萍, 苏畅, 等. 计算机辅助多模态融合超声诊断乳腺良恶性肿瘤[J]. 中国医学影像技术, 2021, 37(8): 1210-3.

[30] Alshmrani G M M, Ni Q, Jiang R, et al. A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images[J]. Alexandria Engineering Journal, 2023, 64: 923-935.

[31] Hroub N A, Alsannaa A N, Alowaifeer M, et al. Explainable deep learning diagnostic system for prediction of lung disease from medical images[J]. Computers in Biology and Medicine, 2024, 170: 108012.

[32] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine[J]. Scientific Data, 2023, 10(1): 67.

[33] Peng C, Xia F, Naseriparsa M, et al. Knowledge graphs: Opportunities and challenges[J]. Artificial Intelligence Review, 2023, 56(11): 13071-13102.

[34] Bonner S, Barrett I P, Ye C, et al. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective[J]. Briefings in Bioinformatics, 2022, 23(6): bbac404.

[35] Zeng X, Tu X, Liu Y, et al. Toward better drug discovery with knowledge graph[J]. Current opinion in structural biology, 2022, 72: 114-126.

[36] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein

structure prediction with AlphaFold[J]. nature, 2021, 596(7873): 583-589.

[37] Huang B, Kong L, Wang C, et al. Protein structure prediction: challenges, advances, and the shift of research paradigms[J]. Genomics, Proteomics & Bioinformatics, 2023, 21(5): 913-925.

[38] Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology[J]. Nature, 2023, 618(7965): 616-624.

[39] Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI[J]. Nature Methods, 2024: 1-11.

[40] Hao M, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics[J]. Nature Methods, 2024: 1-11.

[41] Yang X, Liu G, Feng G, et al. Genecompass: Deciphering universal gene regulatory mechanisms with knowledge-informed cross-species foundation model[J]. bioRxiv, 2023: 2023.09. 26.559542

第 4 章 人工智能助力医疗文本处理

4.1 医疗大数据简介及分类

医疗大数据是指在与人类健康相关的活动中产生的与生命健康和医疗相关的数据。随着医疗信息技术的快速发展，医疗数据的生成速度和数量呈指数级增长。

从数据的来源来说，可以分为临床大数据、健康大数据、生物大数据、运营大数据等。医疗大数据的形式常见的有 3 种：分类数据、序列数据、连续数据。分类数据比如病人的性别，是否使用某种治疗等，这类数据没有内在排序。序列数据是有明确排序的数据，比如血压值、血糖值等，除了排序性，这些数据之间可能还有距离，例如一个人过去 3 天的血压值，每一天就是等距。连续数据不仅是有序的，数据的自变量也是连续的，比如年龄、血压、体重等。

从数据格式上，医疗数据通常可以分为结构化数据和非结构化数据两类，它们在医疗信息管理中都扮演着重要角色。结构化数据是指可以直接存储和处理的数据，通常以表格或数据库的形式存在，具有明确定义的字段和格式，例如，患者基本信息、实验室检查数据等。非结构化数据则是指没有固定格式或字段，难以通过传统的数据库或表格进行处理和分析的数据类型。例如，电子病历、影像数据等。这些非结构化数据包含了丰富的临床信息，但要想进行系统化的分析和利用，需要借助自然语言处理（NLP）、图像分析等技术来提取和理解其中的内容。

医疗大数据不仅为医疗决策和研究提供了宝贵资源，同时也带来了数据处理和分析上的挑战。近年来，移动互联网、大数据、云计算等多项技术与各类医疗领域大数据不断跨界融合，相关的新技术应用于医疗行业的各个环节中，并且国家也出台了多项扶持政策。人工智能（AI）作为一种强大的技术工具，正在改变医疗大数据处理的方式

和效率。

4.2 医疗文本自然语言处理

影像学报告、电子病历、出院小结等都为重要的医疗健康大数据资源，不仅是医疗实践中的核心文档，也是连接医疗保健各个方面的桥梁。在国家推行使用电子病历和电子影像学报告的背景下，这些医疗文本中丰富的信息资源可以服务于临床实践、临床研究等[1,2]。但是目前大部分医疗文本为非结构化数据，给临床研究带来了困难。随着电子医疗文本的普及和医疗大数据时代的到来，将人工智能方法应用于非结构化医疗文本的自然语言处理问题，已成为当前的研究热点。

自然语言处理是从医疗文本中提取有用信息的关键技术。基于自然语言处理的医疗文本处理流程主要有句子边界识别、分词、共指消解、词性标记、句法分析、实体识别等。由于临床信息的复杂性和灵活性，影像学报告、电子病历、出院小结等医疗文本以自由文本（Free Text）的方式来记录，多为非结构化。通过自然语言处理，这些非结构化的医疗文本被转化为包含重要医学信息的结构化数据，后续可进行的病人聚类、临床辅助诊疗等研究分析[3,4]。

在美国，临床医学领域的自然语言处理研究可追溯到 20 世纪 60 年代，早期研究在有限的电子医疗文本中验证了可行性。自 20 世纪 80 年代以来，大量医学领域的知识库逐渐建立起来。例如 SNOMED CT 是被广泛应用的临床医学术语知识库之一[5]，UMLS（The Unified Medical Language System）是一体化的医学信息系统，它通过建立超级词表来统一医学术语概念，集成了 150 多种常用医学术语知识库[6]。随后，又出现了大量的临床医学自然语言系统，代表性的有 MedLEE、MetaMap、cTAKES、MedEx、KnowledgeMap 等。这些医学自然语言系统覆盖了医学信息抽取、医疗文本分类、医疗决策支持、信息管理、医疗信息问答、知识挖掘等诸多应用领域。与之相比，国内相关的医

学自然语言系统和知识库较为缺乏，限制了中文医疗文本自然语言处理研究的发展。

近年来，专业领域中文自然语言处理需求越来越大，而中文医学专业领域的语料资源较少。不同于以字母为基础的语言，中文是以字符为基础，学习算法目前也更受限制，中文知识库也较为受限。近年来，逐步出现针对于中文医疗文本的自然语言处理方法。

4.3 文本表示学习

文本表示学习是指将实际的文本内容转变成更易于计算机识别的信息，即对文本进行形式化处理，它依靠着高维空间向低维空间的转换，以将词来表示成一个低维的稠密实值向量[7]，进而表达文本词语的语义。常用方法有布尔模型、向量空间模型、概率模型等。这些向量随后可用于构建矩阵、拓扑结构或图数据，从而探索医疗实体和临床事件之间的复杂关系。使用向量空间模型方法需要对文本先进行分词，此时文本可看作一系列词的组合，之后对每个词加一个对应的权值，最初权值表示为 0 或 1，即当文本中出现该词，则值为 1，否则为 0，这种方法后续逐渐被更精确的词频代替。常用文本向量化方法有 BOW（词库、Bag of Words）模型、Mikolov 等人设计的 Word2Vec 模型[8]、以及 Quoc Le 等人提出的段落向量（Paragraph vector）法等[9]。BOW 方法中的 TF-IDF 向量表示法得到了广泛应用。TF-IDF 方法评估一个字词对于文件集或语料库中其中一个文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。将文本表示为 TF-IDF 向量后，向量间的 cos 角就可以用来测量文本间的相似度。针对 TF-IDF 方法，还有基于文本频率、信息增益、互信息、卡方检验的降维方法。然而，这类 BOW 方法存在一些不足：忽略了词的顺序以及词之间的语义联系，导致不同文本有可能会有同样的向量表示；实际问题中会计算出

较高的向量维数，给后续机器学习中带来维数灾难。

自 Word2Vec 到 Glove 再到 ELMO，词分布嵌入类模型由于其出色的词表示能力，可以在低维空间中高效的计算词的语义信息和词之间的语义联系，被广泛应用于医学文本的处理之中。根据模型的技术和应用场景，词分布嵌入类模型可以分为以下几类。各类方法的简介和举例如下：

1. 基于统计和分布假设的方法

通过统计分析词在文本中的共现关系来生成嵌入，常见方法为 LSA (Latent Semantic Analysis)，LDA (Latent Dirichlet Allocation)，PMI (Pointwise Mutual Information)。

2. 基于上下文窗口的词嵌入模型

通过训练一个浅层神经网络来生成固定维度的词向量，常见方法为 Word2Vec，GloVe (Global Vectors for Word Representation)，FastText 等。

3. 基于上下文动态生成的词嵌入模型

例如：ELMo (Embeddings from Language Models)，CoVe (Contextualized Word Vectors)

4. 基于 Transformer 的预训练语言模型

这类模型是上下文嵌入的高级演化，通过大规模预训练得到更强的语义表达能力，例如 BERT (Bidirectional Encoder Representations from Transformers)，GPT (Generative Pre-trained Transformer)，T5 (Text-to-Text Transfer Transformer)等。

5. 专为多模态或知识注入设计的词嵌入模型

例如，融入知识图谱信息的 ERNIE (Enhanced Representation through kNowledge Integration)，用于多模态任务，结合文本和图像对比学习的 CLIP (Contrastive Language–Image Pretraining)。

在医学文本的表示学习领域，Transformer[10]架构有效地解决了在

句子中捕获长距离依赖性的挑战，增强了模型理解上下文之间关系的能力。通过多头注意力机制学习到的上下文化词表示，以及在大规模语料库上的无监督预训练。基于 Transformer，如 BERT[11]，在文本表示学习任务中表现出了非常有前景的性能。然而，开放的医疗问题仍然具有挑战性，因为这些方法缺乏领域医疗知识来提升语义理解能力[12]。为了解决这个问题，一些工作[13-15]尝试将内部知识或外部知识整合到类似 BERT 的模型中。内部医疗知识主要包括语法知识、句法结构知识和语义知识等[12]。弱监督方法可以整合内部知识，然后设计基于知识的任务来学习文本中的医疗知识。例如，ERNIE[16]通过注释和掩码预训练数据中的短语和实体，融入了隐式的内部句法和语义知识。ERNIE-Health[17]使用医疗实体掩码算法来学习术语和其他医疗实体知识。CorefBERT[18]使用问答匹配任务来学习疾病描述与医生专业治疗之间的对应关系，从而获得了医疗实体知识之间的内在联系。与内部知识相比，外部医疗知识包含了医疗知识图谱、医疗领域特定数据和预训练数据的额外注释。根据格式的不同，它也可以分为结构化知识和非结构化知识。例如，BERT-MK[19]将医疗知识图谱中的子图视为一个整体，并对齐医疗文本以保留更多的结构信息。与结构化知识相比，非结构化知识（如医疗领域的数据）更加完整，但噪声也更多。K-ADAPTER[20]通过不同的适配器融入了医疗非结构化知识来学习词汇知识和语言知识。在生物医学文本训练的模型 BioBERT 基础上，UMLSBERT[21]利用 UMLS 去增强临床领域的知识表示，结果表明模型能更好的理解和表示医学文本中的语义信息。对于上述所有工作，知识都隐式地存储在其模型参数中。从知识增强方法中学习到的文本表示已经展示了其表达能力，并对下游任务的性能提升做出了贡献。

4.4 知识图谱

知识图谱是在自然语言处理的基础上发展而来，这个概念是谷歌

在 2012 年提出的，当时主要是为了将传统的基于关键字搜索的模型向基于语义的搜索升级。知识图谱本质上是一种揭示实体之间关系的语义网络，其节点代表实体（entity）或者概念（concept），边代表实体/概念之间的语义关系。相比于传统的机器学习算法，知识图谱能够从语义层面以结构化的形式表示知识，通过知识表示和推理，给人工智能系统提供可处理的先验知识，让其具有解决复杂任务的能力。随着智能信息处理技术，尤其是深度学习技术不断发展，知识图谱已广泛应用于智能搜索、智能问答、个性化推荐等领域。中文文本的知识图谱工作近年来在公开评测、领域扩展及上述的跨语料迁移方面也都取得了一些进展。

目前知识图谱也已经广泛应用于医疗领域[22,23]。医学知识图谱是在人工构建的专业知识库基础上，通过算法以及人工审核的方式不断扩充实体及关系来构建的，包括疾病、症状、药品、手术、非手术治疗等医学概念与多种医学关系。医学知识图谱的构建主要包括知识抽取、知识融合、知识应用等。医学知识图谱是疾病智能辅助决策工具的基石，使得计算机理解并做出智能的决策[24,25]。医学知识图谱在多项医学决策支持上都取得了成功的应用，例如预测药物点相互作用[26]、罕见病知识图谱辅助诊断模型[27,28]等。

中文医疗知识图谱领域近年来得到了业界广泛关注，目前在工业界主要有百度-灵医智惠、中国平安-平安好医生、阿里健康-医知鹿、腾讯-觅影等医学知识图谱，在学术界主要有 CMeKG（Chinese Medical Knowledge Graph，http://cmekg.pcl.ac.cn/ ）、BIOS 等。CMeKG 是基于大规模医学文本数据，利用文本挖掘技术研发的中文医学知识图谱。CMeKG 的构建参考了 MeSH、ICD、SNOMED 等权威的国际医学标准以及大规模多源异构的临床指南、诊疗规范等文本信息。CMeKG 涵盖疾病的临床症状、发病部位、药物治疗等 30 余种常见实体类型，100 余万概念关系及属性三元组。“生物医学信息学本体系统”BIOS

目前为全球最大开放生物医学知识图谱，BIOS 是首个完全由机器学习算法生成的大型开放生物医学知识图谱，其术语发现、语义分析、概念生成、关系发现、跨语言对齐完全由模型自动实现。对比美国开发几十年的“一体化医学语言系统”UMLS，BIOS 在短短几年的时间里，体量达到了 UMLS 的数倍，不仅扭转了中文领域缺乏大型开放生物医学知识图谱的困难局面，更充分证明了人工智能的巨大潜力。近年来，深度学习技术，尤其是图神经网络的发展，极大地推动了时序知识图谱的研究。常用的研究思路有以下两种：将动态图按照时间划分为每个时刻的图，然后进行处理，随着时间发展，每个时刻图中的边和节点可以变化；把时间 T 之前的所有边构造成一个图。常用算法中，DySAT 使用自注意力机制学习不同时刻的动态图表示[29]；EvolveGCN 思路便是对每个时刻 T 的图谱用 GCN 进行建模学习，用 RNN 去演化每个时刻 GCN 模型的参数[30]；TGAT 模型在处理时序知识图谱时，期望学习到邻域的时间拓扑信息，学习节点特征和时间之间的相互作用，将节点的嵌入表示看作为时间的函数[31]。dyngraph2vec 使用全连接层和递归层学习动态图嵌入的方法，并构建了动态模型库[32]。当前，已经有一些研究针对临床病历数据的时序知识图谱展开，Shang 等人[33]以患者疾病和药物为节点，考虑患者就诊期间的时序性，构建了一个患者的时序图模型，并在患者药物推荐上取得了良好的结果。

4.5 大语言模型在医疗文本中的应用

近年来，大语言模型在文本理解与生成方面展现出了卓越的能力，为文本分类、信息抽取等任务提供了新的解决方案。在医学领域，结合大数据技术和大语言模型的应用具有巨大的潜力和价值[34]。大语言模型通常采用 Transformer 架构，通过对大规模语料库进行自监督的预训练，学习文本的语法、语义和逻辑等特征，从而捕捉文本间复杂

的关系。通过对特定任务数据进行微调，适应不同的下游应用。GPT-4、PalM[35]和 LLaMA 等国际上的模型，以及国内的 ChatGLM、文心一言、通义千问、讯飞星火等，显示了大型语言模型在解决通用语言问题（如文本分类、问答、文档总结和文本生成等）方面的成功应用。此外，谷歌医疗团队最近发表了最新版本的医疗大模型 Med-PalM[36]，专注于医疗文本理解和信息抽取。BiomedGPT[37]是一个专为生物医学领域涉及的多模态通用基础模型，通过预训练和微调多种生物医学数据库，能够处理多样化的生物医学任务。

相对于传统的自然语言处理方法，大语言模型在电子病历信息抽取和结构化方面有着显著优势。电子病历通常包含大量的上下文信息、诊断过程和治疗方案等，传统方法常常难以充分考虑这些信息，而大语言模型能够更好地理解文本的语境和上下文信息，更准确地结构化这些复杂的信息。此外，大语言模型具备强大的泛化能力。能够处理各种类型和风格的医学文本，无需事先定义复杂的规则或特征工程。这种灵活性使得模型能够适应不同医疗实践中的各种数据格式和语言风格。

目前，大语言模型在电子病历信息抽取领域,尤其是中文电子病历方面的研究还在发展中，在支持临床决策和国际医疗数据标准化等应用中展示出巨大的潜力[38-40]。目前已有一些成功的应用，例如，山海医疗大模型可应用于门诊病历报告生成、手术记录撰写、商保管理；支付宝医疗大模型可应用于医疗问答、病历结构化和检索等。未来，它将为医疗信息管理和个性化医疗提供更为先进和有效的解决方案。

参考文献

[1] Giddings R, Joseph A, Callender T, et al. Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review. Lancet Digit Health. 2024;6(2):e131-e144.

[2] Montgomery-Csoban T, Kavanagh K, Murray P, et al. Machine learning-enabled maternal risk assessment for women with pre-eclampsia (the PIERS-ML model): a modelling study. Lancet Digit Health. 2024;6(4):e238-e250.

[3] Wang J, Zheng N, Wan H, et al. Deep learning models for thyroid nodules diagnosis of fine-needle aspiration biopsy: a retrospective, prospective, multicentre study in China. Lancet Digit Health. 2024;6(7):e458-e469.

[4] Daniel R, Jones H, Gregory JW. Predicting type 1 diabetes in children using electronic health records in primary care in the UK_ development and validation of a machine-learning algorithm. Lancet Digit Health. 2024;6:e386-95

[5] Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. Journal of the American Medical Informatics Association: JAMIA. 2014;21(e1):e11-19.

[6] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(Database issue):D267-270

[7] LOCKE S, BASHALL A, AL-ADELY S, et al. Natural language processing in medicine: A review[J]. Trends in Anaesthesia and Critical Care, 2021. https://doi.org/ 10.1016/j.tacc.2021.100233

[8] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word

representations in vector space. 2013 arXiv preprint arXiv:13013781.
[9] Le QV, Mikolov T. Distributed Representations of Sentences and Documents; 2014. pp.1188-1196.
[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need. Neural Information Processing Systems, 2017. https://doi.org/10.5555/3295222.3295349
[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North, 2019. https://doi.org/10.18653/v1/N19-1423
[12] Biawas Som S. "Role of chat gpt in public health." Annals of biomedical engineering 51.5 (2023): 868-869.
[13] Lewis, Mike, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.2019 arXiv preprint arXiv:1910.13461.
[14] Touvron, Hugo, et al. Llama 2: Open foundation and fine-tuned chat models. 2023 arXiv preprint arXiv:2307.09288.
[15] HAN X, ZHANG Z, DING N, et al. Pre-Trained Models: Past, Present and Future. AI Open, 2021. https://doi.org/10.1016/j.aiopen.2021.100080
[16] McIntosh, Timothy R., et al.. A culturally senstive test to evaluate nuanced gpt hallucination, IEEE Transactions on Artificial Intelligence (2023)
[17] CHEN Q, ZHU X, LING Z H, et al. Neural Natural Language Inference Models Enhanced with External Knowledge. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018. https://doi.org/

10.18653/v1/P18-1041

[18] MICHALOPOULOS G, WANG Y, KAKA H, et al. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus[C/OL]. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online. 2021.

[19] SHARMA S, SANTRA B, JANA A, et al. Incorporating Domain Knowledge into Medical NLI using Knowledge Graphs. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. https://doi.org/ 10.18653/v1/D19-1540

[20] LI Y, WEI B, LIU Y, et al. Incorporating knowledge into neural network for text representation. Expert Systems with Applications, 2018. https://doi.org/ 10.1016/j.eswa.2018.06.029

[21] SUN Y, SHUOHUAN W, YUKUN L, et al. ERNIE: Enhanced Representation through Knowledge Integration[J]. Cornell University - arXiv, 2019. https://doi.org/ 10.18653/v1/D19-1003

[22] Murali L, Gopakumar G, Viswanathan DM, Nedungadi P. Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study. J Biomed Inform. 2023;143:104403.

[23] Karthik Soman, Charlotte A. Nelson, Gabriel Cerono, Sergio E. Baranzini. Time-aware Embeddings of Clinical Data using a Knowledge Graph. Pac Symp Biocomput. 2023(28):97-108.

[24] Li T, Xiong Y, Wang X, Chen Q, Tang B. Document-level medical

relation extraction via edge-oriented graph neural network based on document structure and external knowledge. BMC Medical Informatics and Decision Making. 2021, 21 (Suppl 7): 368.

[25] Zhu Y, Che C, Jin B, Zhang N, Su C, Wang F. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. Health Informatics Journal. 2020, 26(4):2737-2750.

[26] Zhao D, Wang J, Sang S, Lin H, Wen J, Yang C. Relation path feature embedding based convolutional neural network method for drug discovery. BMC Medical Informatics and Decision Making. 2019, 19 (Suppl 2): 59.

[27] Latorre-Pellicer A, Ascaso A, Trujillano L, Gil-Salvador M, Arnedo M, Lucia-Campos C, et al. Evaluating Face2Gene as a Tool to Identify Cornelia de Lange Syndrome by Facial Phenotypes. Int J Mol Sci. 2020, 21 (3):1042.

[28] Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019, 47 (D1): D1018-D1027.

[29] Sankar, A., Wu, Y. , Gou, L., Zhang, W. , Yang, H. DySAT: Deep Neural Representation Learning on Dynamic Graphs via Self-Attention Networks. WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining. ACM 2020.

[30] SHANG C, TANG Y, HUANG J, et al. End-to-End Structure-Aware Convolutional Networks for Knowledge Base Completion[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 3060-3067.

[31] Pareja, A. , Domeniconi, G. , Chen, J. , Ma , T. , Leiserson, C.

Evolvegcn: evolving graph convolutional networks for dynamic graphs. Proceedings of the AAAI Conference on Artificial Intelligence,2020.

[32] Xu, D. , Ruan, C. , Korpeoglu, E. , Kumar, S. ,Achan, K. Inductive representation learning on temporal graphs. ICLR ,2020.

[33] Goyal P , Ch Hetri S R , Canedo A . dyngraph2vec: Capturing network dynamics using dynamic graph representation learning. Knowledge-Based Systems, 2019, 187.

[34] Murali L, Gopakumar G, Viswanathan DM, Nedungadi P. Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study. J Biomed Inform. 2023. 143:104403.

[35] Zhang, K., Zhou, R., Adhikarla, E. et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med. 2024. 30, 3129–3141

[36] PALM : Chowdhery, A. et al. PaLM: scaling language modeling with pathways. 2022 Preprint at 10.48550/arXiv.2204.02311.

[37] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023. 620(7972):172-180.

[38] de Hond A, Leeuwenberg T, Bartels R, et al. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health. 2024. 6(7):e441-e443.

第 5 章 人工智能助力 RNA 结构预测

5.1 背景

RNA 的研究被科学家称为永无止境的前沿。生命在于各种蛋白质，没有 RNA 就没有蛋白质。RNA 结构预测是相比于蛋白质结构预测来说是件更加困难的事情，RNA 的研究永无止境。

RNA 种类繁多、功能多样、不稳定，结构决定功能。许多烈性病毒就是 RNA 病毒，如肆虐全球的新冠病毒，就是 RNA 病毒。RNA 的结构预测公认比蛋白质结构预测更加困难。主要表现在以下几个方面：其一，RNA 可能随环境不同而存在多个稳定的不同结构态，其二，共进化信息有效提升了蛋白质结构预测精度，但对 RNA 结构预测帮助很小。其三，实验公布的 RNA 结构数量远小于蛋白质结构数量。尽管经过几十年的艰苦努力，相比预测蛋白质的三维结构，预测 RNA 三维结构仍然是一个非常巨大的挑战。

截至 2023 年 12 月，PDB 数据库中拥有超过 189000 个生物大分子结构可用，含有 RNA 的结构仅占总结构数的 0.86%，其中，包括与其他分子复合的 RNA 结构。PDB 每年新发布的 RNA 结构数量(深色)及数据库中累计的 RNA 结构数量，RNA 结构数量增长缓慢。这表明 RNA 三级结构测定的效率极其低下，RNA 结构数量还远不能满足研究人员对结构和功能探索的需求。

Science 封面：新型 AI 技术有望破解 RNA 结构预测难题。

RNA 三级结构预测的主要困难在于其构象采样和打分函数的构建。对于构象采样的问题，Rosetta 框架的出现为 RNA 构象采样提供了新的思路，在 Rosetta 框架下基于枚举采样和随机抽样方案的 RNA 三级结构预测算法有效地提高了构象采样能力。而对于打分函数而言，机器学习相关方法克服了传统打分函数打分不准确的弊端，基于三维卷积神经网络的 RNA 结构打分函数不仅提高了结构打分的质量，还在一定程度上提高了 RNA 三级结构预测的精度。

人类基因组计划的实施使得大量生物分子序列、结构及功能的相关数据呈几何倍数增长的趋势出现。生物信息学是一个跨多学科的研究领域，该领域主要基于生物计算方法来对大量的生物大分子数据进行分析，旨在发现其中隐藏的生物模式及相关信息，此外，通过对相关信息的进一步分析可以促进对生物运行机制的研究。生物信息学和高通量测序技术的快速发展显著地提高了我们探索人类微生物组的能力，并为各种疾病的研究提供了理论基础和解决方案。在近期的研究报告中，专家和学者利用生物信息学方法研究了肿瘤突变、乳腺癌、宫颈癌、鼻咽癌、Ig A 肾病等疾病，并从基因水平对这些疾病进行了更深入的研究。生物信息学的本质就是处理大量的生物数据，并从中获得想要的信息。

蛋白质、多糖及核糖是生命系统中必不可少的生物大分子，生物大分子的结构预测仍然是生物信息学领域的一项重大挑战，特别是 RNA 三级结构的预测。RNA 是一种由核糖核苷酸组成的多功能生物大分子。RNA 在疾病分析领域发挥着重要作用，如研究口腔鳞状细胞癌需要了解 microRNAs，而研究食管癌需要先研究 lncRNAs，这表明对 RNA 的研究将为疾病研究提供坚实的理论基础。此外，对 RNA 结构的探索是研究活细胞中低丰度 pre-mRNA 与 RNA-蛋白质相互作用的基础，此项研究能够帮助研究人员进一步理解细胞生命活动中 RNA 的功能，这使得 RNA 的相关研究成为一大热点。

RNA 在生物体内有多种功能，其主要功能是将存储在 DNA 里面的遗传信息转化为蛋白质，并引导蛋白质分子的合成。RNA 的功能逐渐受到关注，在最近的研究中，研究人员发现了 RNA 的一些新功能，有些部分 DNA 分子片段转录成 mRNA，进一步翻译成蛋白质，而另一部分 DNA 分子片段只转录成 RNA，不能进一步翻译，无法翻译成蛋白质大分子的 RNA 是非编码 RNA(non-coding RNA)。非编码 RNA 能够控制蛋白质合成、调节转录过程并进行翻译，除此之外非

编码 RNA 还具有一些更加复杂的生物学功能，如剂量补偿、染色质调控、基因组印记、核组织及基于代谢物浓度变化来进行基因表达调控等。

总部位于美国马萨诸塞州剑桥市的克雷数学研究所(Clay Mathematics Institute，CMI)，在 2000 年提出了世界 7 大数学难题，而 NP 完全问题1(non- deterministic polynomial complete problem)是世界 7 大数学难题之一，近似算法是处理 NP 完全问题(NP 难问题)的一种本质方法。新型冠状病毒是 RNA 病毒，冠状病毒(coronavirus，CoV)的 RNA 结构通常包含 H 型假结(pseudoknot)，包含假结的 RNA 结构预测问题是 NP 完全问题[1]。有关 RNA 的研究已经多年被 Science 列入世界主要科技进展，1986 年，Science 上刊发了诺贝尔奖获得者 Dulbecco[2]关于人类基因组测序的有关论文，相关论文的发表极大地推动了 20 世纪人类基因组计划(Human Genome Project，HGP)的实施，也催生了生物信息学/计算生物学学科的发展。

从 2019 年底开始在全球肆虐的新型冠状病毒(COVID-19)给人类带来了巨大灾难，新型冠状病毒属于 RNA 病毒，RNA 多为单链结构，该结构不稳定、易变异，这为疫苗的研制增加了难度。冠状病毒是有包膜的正股单链 RNA 病毒，直径为 80～120nm，约由 3 万个碱基组成，其遗传物质是已知 RNA 病毒中最大的。目前已经发现至少 7 种致病性冠状病毒，其中，严重急性呼吸综合征冠状病毒(severe acute respiratory syndrome coronavirus，SARS-CoV)、中东呼吸综合征冠状病毒(Middle East respiratory syndrome coronavirus，MERS-CoV)曾在人群中大范围传播流行，证明了冠状病毒在动物间、人与人之间传播的可能性。研究表明，蝙蝠身上能携带超过 100 多种病毒，是许多高致病性病毒的天然宿主，对人类社会造成巨大威胁的 SARS-CoV 正是来自中华菊头蝠。2019 年发现的 SARS-CoV-2 就属于蝙蝠 SARS

冠状病毒和中东呼吸综合征冠状病毒的病毒群。

遗传物质决定生命体的性状，结构决定功能，冠状病毒拥有目前几乎已知所有 RNA 病毒中最长的 RNA 碱基序列，RNA 结构预测问题来源于 RNA 编码的秘密，也来源于病毒疫苗药物研制的困难性。用实验来测定指数级的数量庞大的 RNA 结构代价太大，不现实也不可能。除 RNA 的一级结构能用实验的方法来测定测序外，RNA 二级结构、三级结构甚至四级结构，用实验的方法测定十分困难，因而用计算方法与复杂性理论来分析预测 RNA 结构成为不可缺少的选择。

结构决定功能，想要探究 RNA 的功能，特别是 RNA 有些复杂的生物学功能，就必须要先了解 RNA 的结构。目前国内外的 RNA 三级结构测定方法主要有两种。第一种方法是利用 X 射线、核磁共振及冷冻电镜等实验测定方法，采用实验的方法测得的结果比较精确且可靠，但是构象数量随着 RNA 长度的增加呈指数增长，导致成本太高，也不可能穷举。第二种算法是基于生物计算的结构预测方法，当前的 RNA 三级结构预测算法主要有基于知识挖掘的预测方法和基于物理的预测方法。基于知识挖掘的三级结构预测方法依赖已知的 RNA 模板数据库，基于物理的预测方法减少了对数据库的依赖，但是仍存在结构建模精度不够高的问题，无法满足当前的结构预测需要。因此针对这个现状，需要对现有方法进行改进创新。

由于 RNA 分子和蛋白质具有不同的折叠方式，所以将蛋白质的研究方法应用到 RNA 的研究中得到的结果不佳。在蛋白质领域，存在一个假设，假设大分子的原生构象具有最低自由能，并且自由能函数近似为氢键、范德瓦耳斯力、静电力和溶剂化项之和。本书针对现有技术的缺陷，假设大分子原生构象具有最低自由能，但不同的 RNA 分子的三级结构中，根据碱基相互作用的不同类型，分配不同的权值，通过线性加和后得到相应自由能。此外，针对单线程构象能力受限制问题，可以采用并行机制，同时对建模结果进行了多重判断，得到一

个专门用于 RNA 三级结构预测的算法—逐步蒙特卡罗(Monte Carlo，MC)并行化算法[3]。

21 世纪初，随着由中国和美国、英国、法国、德国、日本科学家共同参与的人类基因组计划的全部完成，人类进入后基因时代—人类细胞图谱计划时代。根据基因表达的分子信息，对所有人类细胞种类进行定义，而 RNA 在细胞中的转录和表达起着非常重要的作用。近年来，全球有关 RNA 的研究，特别是冠状病毒 RNA 的研究，引起了全球众多学者的极大关注。RNA 是单链折叠结构，RNA 在遗传信息从 DNA 表达为蛋白质的过程中起转录作用。RNA 结构预测，特别是 RNA 三级结构预测甚至四级结构预测是当今学术界研究的热点，但普遍存在预测准确度不高、特异性和敏感性不理想、预测算法时空复杂度高等问题。冠状病毒的 RNA 结构往往包含 H 型假结，包含假结的 RNA 结构预测问题被证明是 NP 完全问题，而作为世界 7 大数学难题之一的 NP 完全问题的研究给我们带来了极大的困难。为了获取 RNA 结构功能信息，获知生物分子的生物学功能，寻找非编码 RNA 基因，利用机器学习、深度学习、层次聚类、蒙特卡罗方法等人工智能的典型技术，结合 RNA 病毒结构特性，特别是现在全球大流行的新型冠状病毒结构，结合最大 k-补割、稠密 k-子图问题等典型的 NP 难的问题，以及困难性未知的最小结构熵问题，有望解决 RNA 结构预测算法与复杂性中存在的世界前沿问题，探索生命起源和进化，揭开 RNA 编码秘密，为研究冠状 RNA 病毒机理和靶向核酸药物研制提供理论和技术指导。

不同于 DNA 的双螺旋结构，RNA 是单链结构，RNA 碱基序列中包含 A、C、G、U 四种碱基。由于碱基是平面结构，其边缘的氢原子供/受体可近似地划分为三个配对边：Watson-Crick(W)边，Hoogsteen(H)边，以及 Sugar(S)边。配对边影响 RNA 折叠结构的稳定性，稳定性也可以用碱基配对所需要的自由能量来衡量，并且自由能

量越小，RNA 结构越稳定。

RNA 能量模型包括结构单元间的近邻相互作用模型、独立结构单元模型等。最邻近邻居模型可以看作一种独立结构单元模型的特殊情况，其结构单元中堆叠结构与环结构是由最邻近碱基对决定的，RNA 分子的自由能量主要是堆叠结构和环结构的贡献。环结构对 RNA 折叠结构的稳定性有非常重要的作用，但对环结构的热动力学研究相对较少，其结构的稳定性可以由自由能量参数来衡量[4]。AU、CG 基对是 RNA 碱基序列中常见的茎环结构，RNA 茎环结构的邻位基对可能有十余种的组合数，预测 RNA 结构的本质是找出 RNA 碱基序列的各位点之间的配对关系。然而 GU 错配现象在 RNA 碱基序列中也经常发现，包含 GU 错配的情况大约有十几种邻位关系的组合。利用寡核苷酸合成技术，我们可以合成大量用于实验的寡核苷酸链，进一步提高了自由能量参数的正确率，Mathews 和 Turner[5]改进的自由能量参数成为目前普遍采用的参数。

许多 RNA 病毒中含有假结结构，如冠状病毒中通常含有 H 型假结。假结是 RNA 分子中最广泛的三级结构单元，假结的存在使 RNA 结构更加复杂化，假结在不同的 RNA 分子中有催化、调节、构造等非常重要的功能，在探索生命科学的现象、规律中具有十分重要的意义[6,7]。假结是非常复杂和稳定的 RNA 结构，包含假结的 RNA 结构预测是目前 RNA 结构预测研究的难点和关键点。1985 年，Pleij 等成功地预测了几种毒菌 RNA 的假结结构[6]，Kolk 等在 1998 年予以证实了假结结构的存在性[7]。有关含假结的 RNA 结构预测算法近似理论与技术的研究是近似算法领域研究中的热点之一。在多项式时间可解的问题得到研究之后，包含假结 RNA 折叠结构预测的 NP 难问题的近似算法研究成为算法理论设计与分析经典领域中的活跃分支。

通过 RNA 结构分析，本书抽象设计出有效的精确确定性算法来预测三级结构甚至四级结构，利用近似算法来求解包含假结的 RNA

结构预测这一理论上是被证明的 NP 完全问题，利用近似算法分析设计中提出的新思想、新观点来预测 RNA 结构，提高预测的精度、特异性、敏感性。本书的研究有助于 RNA 结构预测近似算法与复杂性，以及算法不可近似性的发展；也有助于 RNA 结构预测理论在生物医药产业实践中的指导，特别是在加快生物制药、冠状病毒药物研制和疫苗研制进度角度，具有极其重要的意义。

生物信息学/计算生物学从 20 世纪 80 年代开始逐渐形成一门学科，南加利福尼亚大学 Waterman 开创了生物信息学和计算生物学的先河，1981 年，Smith 与 Waterman 提出了著名的序列比对的 Smith-Waterman 算法，该算法改进了 Needleman-Wunsch 算法的不足。美国的 Pipas 和 McMahon 最先提出如何运用计算机技术预测 RNA 二级结构。1994 年，Walter 和 Turner 对同轴堆叠在 RNA 折叠中的作用进行了研究，研究主要包括嵌套结构，但许多 RNA 结构中还包含非嵌套结构—假结，假结破坏了动态规划算法依赖的 RNA 折叠结构的嵌套子结构的性质，假结还使 RNA 结构预测问题变为 NP 难问题，增加了问题的困难性 [8,9]。Zuker 等[10]提出了 Mfold 算法，将动态规划算法引入最邻近邻居热力学模型。Rivas 和 Eddy[11]提出了关于 RNA 二级结构预测的 Pknots 算法，可以预测任意的平面假结和部分非平面假结，但其时间复杂度为 O(n6)，空间复杂度为 O(n4)，时空复杂度太高，该算法通过限制假结的类型来预测含假结的 RNA 的二级结构，太高的时间复杂度和空间复杂度严重制约了该算法所能计算的问题规模，使带假结的 RNA 结构预测变得异常困难。含假结的 RNA 结构预测在国际上受到高度重视，是 RNA 结构预测领域中的典型问题和热点。关于假结参数可以用非假结参数乘以系数 g(0.83)作为补偿[12]，这些参数值一部分为理论估计值，另外一部分参数由实验结果计算得到。Nixon 等[13]对 mRNA 假结结构加以研究，提出移码突变的 mRNA 解决方案。Ieong 等[14]于 2003 年提出了最大堆叠基对数问

题，并成功地设计了该类问题近似性能比为 3 的近似算法。Lyngsø[15]设计了时间复杂度高达 O(n81)的最大堆叠基对数问题的精确算法，该算法难以理解更不实用，同时，Lyngsø 提出了最大堆叠数问题，证明该最大堆叠数问题属于 NP 难问题，并设计了多项式时间近似方案。Ruan 等[16]和 Ren 等[17]也对 RNA 假结进行了研究，分别提出了包含假结的启发式算法和环匹配算法，Huang 和 Ali[18]对 RNA 假结结构的预测敏感性进行了研究，Han 等[19]提出了包含假结的 RNA 结构比对算法。

20 世纪末，清华大学自动化系李衍达院士和张学工教授在国内率先致力于生物信息学/计算生物学的研究，清华大学自动化系汪小我、李梢也在基因调控分析与建模、复杂疾病计算分析等方面取得了若干研究成果。吉林大学徐鹰长期致力于癌症生物信息学、微生物信息学和结构生物信息学等相关领域的研究，在生物通路与网络的计算方法和模型研究、比较基因组分析、蛋白质结构预测与建模等方面做出了重要的和公认的贡献。中南大学王建新、李敏利用参数化算法等理论与技术在生物信息计算领域进行了深入系统的研究，在长非编码疾病关联竞争性内源预测等方面取得了具有领先水平的一批理论成果。近年来，国内许多学者开展了 RNA 结构预测的研究，特别是 RNA 二级结构预测。中国科学院计算技术研究所徐琳等[20]提出一种对动态规划矩阵采用分块技术的细粒度并行算法，对面向现场可编程门阵列(field programmable gate array，FPGA)的 RNA 二级结构进行预测，提高了算法效率。陈翔等[21]根据 RNA 折叠的特点，提出了一种启发式搜索算法来预测带假结的 RNA 二级结构，该算法以 RNA 的茎区为基本单元，采用启发式搜索策略在茎区的组合空间中搜索自由能最小并且出现频率最高的 RNA 二级结构，该算法能降低搜索 RNA 二级结构的时间复杂度。吉林大学刘元宁等[22]提出 14 种类型的 RNA 假结结构，并使用一种改进的 RNA 平面结构表示法—弧图，利用相容

矩阵与迭代矩阵来求出具有全局最大最优能的 RNA 茎区组合。近年来在癌症基因驱动检测、识别 RNA 内源性模块等方面，西安电子科技大学 Li 等[23]和 Wen 等[24]取得了丰硕的成果。Yue 等[25]利用贝叶斯网络结合不同算法来预测小 RNA，提高了预测的敏感性和特异性。2011 年美国罗切斯特大学的 Ellaousov 提出了包含假结的 RNA 二级结构快速预测算法，该算法的时间复杂度为 O(n2)，预测准确度为 69.3%，但长度超过 700 的核苷酸的预测精度不理想。2015 年，山东大学李国君联合吉林大学、美国阿肯色州立大学、佐治亚大学等的研究人员共同提出了一种新的 RNA 转录组组装工具 Bridger，其研究成果发表在国际著名学术杂志 Genome Biology 上。Gupta 等[26,27]在求解 Rent-or-Buy 问题时，把博弈论的费用分摊方法应用到近似算法的设计与分析中，成果分别发表在理论计算机科学国际顶会(IEEE Annual Symposium on Foundations of Computer Science)和国际著名期刊 Journal of the ACM[27]上。近似算法的不可近似性成为近年来近似算法领域中的一个新的热点[28]，近似算法及随机算法的去随机化技术为包含假结和冠状病毒的 RNA 结构预测提供了新思路、新方法[29,30]。若把 RNA 序列碱基(核苷酸)看作图的顶点，两碱基(核苷酸)若配对，则在它们之间画一条线段，若途中线段之间存在交叉，则说明 RNA 结构中存在假结，可以把 RNA 结构优化问题转化为图问题，利用深度学习、近似算法和随机算法理论与技术，设计 NP 难包含假结的 RNA 结构预测近似算法，证明问题的可近似性或近似难度。如果一个茎区的形成能使 RNA 结构更稳定，那么表明该结构更有可能先形成，用自由能来衡量 RNA 结构的稳定性，因而本书提出的预测算法可以采用自由能作为评估和衡量候选茎区的标准，设计相关 RNA 假结结构预测近似算法，相关研究论文可以参考文献[31]和[32]。香港大学的 Wong 等[33,34]对含复杂假结的 RNA 折叠结构加以研究，设计了效果不错的 RNA 结构比对方法，主要来判断 ncRNAs (non-coding RNAs)，并且

在超过 350 个 ncRNA 家族中进行了实验。2012 年，Wong 等[35]设计了包含简单假结的 RNA 结构比对算法，其时间复杂度为 O(mn3)，并设计了 RNA 结构比对算法，该算法能处理假结，时间复杂度为 O(mn4)。刘振栋等[36,37]提出了含假结的 RNA 结构近似算法及启发式算法。2013 年，麦吉尔大学的 Reinharz 等[38]利用加权样本和抽样方法设计了加权样本算法，对 RNA 二级结构加以预测，取得了良好的效果。刘振栋等[39]深入分析了含假结的 RNA 折叠结构内部特性，基于堆叠数最大化和能量最小化原理，提出了含假结的 RNA 结构预测算法。华盛顿大学的 Andronescu 等[40,41]对具有最邻近邻居的参数的 RNA 折叠结构进行研究，提出了利用 RNA 序列数据库来确定参数值的方法。芝加哥大学的 Babai[42]针对图同构问题找到了一个拟多项式时间的算法，该算法可以同时对两个网络系统计算加以优化，使生物计算网络更加简单。2015 年，Keane 等[43]研究了含包装信号的 HIV2-1 的 RNA 折叠结构，对 HIV-1 的研究有独到的见解。2016 年 Kucharík 等[44]详细阐述了假结在 RNA 折叠结构中的特性，对假结的理解更为深刻。近年来对单细胞的研究如火如荼，2017 年，Gomez-Schiavon 等[45]对单细胞 RNA 分子中的 BayFish 机理进行详细研究，加深了对单细胞的理解。

在对各类疾病进行分析时，与 RNA 的关联性研究必不可少，如研究乳腺癌需要了解 microRNAs 的结构与功能[46]，研究 Autophagy-related lncRNAs 的结构与功能对研究食管癌至关重要[47]，这表明对 RNA 的研究可以为疾病研究提供坚实的理论基础。RNA 通常会形成复杂的空间结构，其线性核苷酸序列经过碱基配对组成二级结构，二级结构通过折叠决定其三维空间中的结构[48]。RNA 的功能取决于其三级结构及与其他分子在细胞中的相互作用，RNA 二级结构已经提供了 RNA 分子的碱基序列蓝图，我们仍然需要进一步探索

RNA 的三级结构[49]。

目前用于 RNA 三级结构采集的生物学实验方法有冷冻电镜法[50]、核磁共振法[51]等，但是由于 RNA 三级结构极不稳定，容易受到环境的影响而发生突变，同时由于基因的进化，很难获取 RNA 的第三级接触信息，所以获取一段连续的、完整的 RNA 片段是非常困难的。因此，需要利用生物信息学的方法和技术，结合已知的生物分子结构及其功能特点，利用计算机技术来预测 RNA 的三级结构[52]。

目前在生物大分子的三级结构预测领域，蛋白质的结构预测方法已经取得显著进展，但是该方法却难以用于预测 RNA 的三级结构，其原因是目前预测蛋白质结构的方法主要利用了相关已知蛋白质的结构，通过机器学习的手段进行训练，提取相关蛋白质的特征，建立数学模型[53]。但是通过实验测得的 RNA 结构数目远远少于蛋白质，不足以提供大量有效的训练集数据，因此预测蛋白质结构的方法并不适用于 RNA，需要发展更有效的生物计算方法来进行 RNA 三级结构的预测。

5.2 研究现状

近年来，研究人员发现 RNA 具有剂量补偿等复杂的生物学功能，RNA 结构研究引起了广泛重视。然而，RNA 三级结构预测相关研究仍处在起步阶段，与蛋白质结构预测相关研究成效相差甚远。RNA 三级结构预测相关研究一直落后于蛋白质结构预测的相关研究，主要有三个原因。第一，与蛋白质结构相比，RNA 分子结构上有更多的自由度，因此 RNA 结构数量更多，结构预测计算量大。第二，非沃森-克里克碱基对是 RNA 分子折叠结构的核心，虽然其数量有限但是却难以识别，这为 RNA 的三级结构预测增加了难度。第三，RNA 构象空间比蛋白质构象空间要大得多。综合 RNA 与蛋白质的自由度和分子量分析，100nt(核苷酸，nucleotide)的 RNA 三级结构预测与 200～

300aa(amino acids，氨基酸)蛋白质结构预测的建模难度相当[19,20]，这足以证明 RNA 三级结构预测的困难性。正是由于 RNA 三级结构预测比蛋白质结构预测更困难，所以 RNA 三级结构预测的相关研究发展缓慢。

RNA 分子一般是线状单链结构，然而 RNA 分子的某些区域可自身回折，进行碱基互补配对并形成局部双螺旋结构。RNA 双螺旋中，一般是 A 与 U 配对、G 与 C 配对，但存在非标准配对，如 G 与 U 错配对。RNA 分子中的双螺旋与 A 型 DNA 双螺旋相似，而非互补区则膨胀形成前面介绍的凸出(bulge)或者环(loop)，短的双螺旋区域和环可以形成发夹结构，发夹结构是 RNA 中最普通的二级结构形式，二级结构进一步折叠形成三级结构，RNA 分子只有在具有三级结构时才有活性。RNA 能与蛋白质形成核蛋白复合物，RNA 的四级结构是 RNA 与蛋白质的相互作用形成的，RNA 结构预测是计算生物学与生物信息学的典型问题。

致力于发展一种新的 RNA 三级结构预测工具来预测出更多的 RNA 三级结构。生物计算领域出现了很多 RNA 三级结构预测算法，典型的 RNA 三级结构预测算法主要包括两类：一类是基于知识的 RNA 三级结构预测算法，另一类是基于物理的 RNA 三级结构预测算法。基于知识的 RNA 三级结构预测算法主要包括 MANIP 算法、ModeRNA 算法、RNABuilder 算法、3dRNA 算法等。ModeRNA 算法和 RNABuilder 算法是基于同源建模的 RNA 三级结构预测算法，通过基于片段的插入方法对没有模板的区域进行建模，并利用力场进行集合优化，获得物理上合理的构象。

基于物理的 RNA 三级结构预测算法是根据生物物理的原则，通过搜索 RNA 三级结构的构象空间，寻找自由能最低的构象，采样方法都是动态的，且基于蒙特卡罗算法或者分子动力学方法进行构象空间搜索采样，典型算法有 FARNA 算法、FARFAR 算法、SWA 算法、

SWM 算法等。

截至 2023 年 12 月，PDB 数据库中拥有超过 189000 个生物大分子结构可用，含有 RNA 的结构仅占总结构数的 0.86%，其中，包括与其他分子复合的 RNA 结构。PDB 每年新发布的 RNA 结构数量(深色)及数据库中累计的 RNA 结构数量，RNA 结构数量增长缓慢。这表明 RNA 三级结构测定的效率极其低下，RNA 结构数量还远不能满足研究人员对结构和功能探索的需求。图 5-1 为 DNA、RNA 与蛋白质关系的中心法则。

图 5-1　DNA、RNA 与蛋白质关系的中心法则

A-U 碱基的 W/W 顺式配对，G-C 碱基的 W/W 顺式配对，以及 G-U 碱基的 W/W 顺式配对是 RNA 标准碱基配对(canonical base pairs)。然而研究发现，目前观察到的 RNA 分子中，标准碱基配对占据了约 80%。虽然非标准碱基配对(noncanonical base pairs)仅占 20%，但是对于提高 RNA 三级结构预测精度至关重要，非标准碱基配对的精准预测是 RNA 三级结构预测的重点和难点。

RNA 三级结构预测关键有两个方面：一方面，利用构象采样方法生成候选结构；另一方面，利用合适的打分函数来评估生成的这些候选结构。通常 RNA 三级结构预测算法中采用的评估标准是基于具有最低能量的结构最稳定、最接近原生构象的原理；打分函数的优劣很大程度上会影响 RNA 结构预测结果的好坏，当前已经开发出了一些比较好的打分函数，如 RASP、RNAKB potentia、3dRNAscore 和 Rosetta 等打分函数。对于 RNA 结构预测的进一步研究需要从这两个方面进行。此外，RNA 三级结构预测的关键组成还包括分子表示方

式和自由度。

近年来，研究人员基于生物计算提出了一系列 RNA 三级结构预测算法，包括 ModeRNA[54]、3dRNA[55]、FARFAR[56]、MANIP[57]等，这些算法主要基于 RNA 的碱基序列及其二级结构，已在 RNA 的三级结构预测领域取得了一定的进展。此外，Rosetta 的出现也为进一步实现 RNA 三级结构的精确预测创造了可能。Rosetta[58]是一项用于模拟生物大分子结构的综合性框架模型，作为一套用途广泛、灵活性强的框架，它涵盖了大量有关 RNA 及蛋白质三级结构预测的设计、组装工具与算法，通过对 Rosetta 套件中性能的不断改进，其结构预测效果得到进一步提高，如抗体和抗原建模的对接与设计[59]，研究人员利用 Rosetta 套件可以有效地预测 RNA 三级结构。

RNA 三级结构预测的主要影响因素有自由度、采样方法、能量函数、分子表示方式。在 Rosetta 框架中，生物计算方法通常受两方面影响。一方面，通过各种抽样方法生成大量候选结构。另一方面，使用一个评估这些候选结构的鉴别器。对于 RNA 或者蛋白质结构预测而言，鉴别器通常是指能量函数[60]，例如，最近更新的 Rosetta 能量函数[61]。而低效的采样方法一直是 RNA 高分辨率建模的瓶颈。如果不对构象空间进行有效采样，那么就不可能实现精确的建模和严格的高分辨率能量函数测试。

为了提高构象采样能力，Sripakdeevong 等[62]提出了一种假设，通过每次添加一个残基递归地构建模型，枚举出单个 RNA 数百万种构象，并覆盖所有构建路径。Watkins 等[63]进一步指出，用随机抽样代替确定性枚举抽样将降低计算成本，提高建模精度。为了进一步降低计算成本，提高建模精度和建模完整度，在采样时采用并行机制，并对建模结果进行进一步判断和处理。

2018 年，Liu 等[64]对包含假结的 RNA 折叠结构加以研究，降低了时间复杂度，改进了预测精度、特异性和敏感性。2019 年，Meng

等[65]针对 RNA 结构预测设计了 RAG-Web 方法，对 RNA 结构有了更深的认识。2020 年，Rivas 等[66]在研究 RNA 结构时计算了 RNA 碱基序列的变化，阐述了碱基序列的配对规律。2020 年，Menden 等[67]利用深度学习技术对 RNA 结构相关的组织表达加以深入分析，其成果发表在 Science 上。2020 年，Liu 等[68]对 RNA 折叠结构的盆跳图(basin hopping graph，BHG)与障碍树进行深入解析，提出了基于扩展结构的 RNA 预测算法。Guo 等[69]采用降维技术来研究蛋白质与蛋白质之间，以及 RNA 与蛋白质的关系。2020 年，山东大学 Zheng 和 Liu[70]进行了最大 k-补割问题和稠密 k-子图问题的研究。2021 年，斯坦福大学的 Townshend 等[71]采用 18 个已知的 RNA 结构设计了一个几何深度学习方法来预测 RNA 结构精确模型，在 blind RNA 预测方面取得了非常好的效果。2021 年，Park 等[72]对 RNA 介导的 DNA 转座系统和靶向选择的基础结构加以研究，加深了对 RNA 介导功能的理解。2021 年，Niu 等[73]用深度学习和降维技术来研究 RNA 与蛋白质之间的相互关系。2022 年，Rasmussen 等[74]在 Nature 上发表了用 RNA 结构揭示疾病和健康关系的论文。2021 年 11 月 9 日在南非首次检测到奥密克戎(英文名：Omicron，编号：B.1.1.529)新型冠状病毒变种，对冠状病毒的 RNA 结构研究迫在眉睫。2022 年，Garcia-Beltran 等[75]在 Cell 上提出了基于 mRNA COVID-19 的疫苗增强剂对 SARS-CoV-2 奥密克戎变种的中和免疫方法，给奥密克戎变种的防治提供了有效途径。2022 年，Liu 等[76]提出了基于蒙特卡罗策略和原子精度的 RNA 三级结构的预测算法，从原子精度对 RNA 的三级结构进行深入研究。至今为止，RNA 结构中特别是 RNA 冠状病毒的 RNA 结构分析预测还存在许多需要研究的问题，期待我们来探索其中的秘密。

Liu 等分别在 2018 年、2020 年对 RNA 折叠结构的 BHG 与障碍树进行深入解析，提出了基于扩展结构的 RNA 预测算法。2020 年、

2021 年 Liu 等用深度学习和降维技术来研究蛋白质之间、蛋白质和 RNA 之间的相互关系，从而进一步加深了对 RNA 结构的理解，2022 年 1～9 月，Liu 等[76-78]发表了有关基于蒙特卡罗策略和原子精度的 RNA 三级结构的预测算法、细胞组织单细胞 RNA 预测算法、基于组合优化策略的 attC 结合位点预测算法。冠状病毒的 RNA 结构预测 NP 完全问题近似算法、近似难度的分析证明等工作具有挑战性，这些挑战性的工作会激发我们极大的研究热情。

RNA 结构中特别是 RNA 冠状病毒的 RNA 结构分析预测还存在众多需要研究解决的问题，其中，有些多项式确定性精确算法、绑定蛋白质问题、NP 完全问题近似算法仍有改进的余地[79-82]，如求解含任意假结最大结构数问题是否是 NP 难的，是否存在该问题的最大 k-补割问题近似算法？病毒 RNA 最大茎区问题如何转换为最小结构熵问题？如何提高 RNA 结构预测近似算法中预测特异性和敏感性？NP 难问题的不可近似性的证明也极具挑战性。

Artem Nemudryi 等人将 CRISPR 核糖核酸酶的序列特异性 RNA 切割与可编程的 RNA 修复相结合，在 RNA 中进行精确的删除和插入，建立了一种重组 RNA 技术直接应用于 RNA 病毒的简易工程[83]。McCauley 等人发现自然修饰有利于 RNA 的天然折叠，表明共价 RNA 修饰可能在生命起源的过程中代谢发挥了关键作用[84]。2024 年 2 月，在科学出版社出版的学术专著中，利用深度学习技术对带权多粒度扫描策略的转录因子结合位点，RNA 结构预测及其复杂性领域加以详细说明[85]。许多生物分子凝聚体依赖于 RNA 和 RNA 结合蛋白，2024 年 3 月的《Science Advances》发表的论文中，Tebbe 等人提供了一种获取 RNA-蛋白质结构信息的方法，生物分子凝聚物中的配合物可能对生物的整体结构建模至关重要[86]。2024 年 3 月，Elizabeth Pennis 等人在《Science》发表的论文中利用 RNA 结构特性，可以在动物身上绘制彩色图案，也为探索 RNA 结构机理提出了有趣科学问题[87]。

RNA 结构决定 RNA 功能、RNA 结构预测算法和人工智能技术的改进，为寻找非编码 RNA 基因，以及为 RNA 病毒和靶向核糖体药物研制提供了新思路、新方法。

5.3 机器学习与深度学习

机器学习的核心是设计和分析一些算法，这些算法旨在让机器自动学习数据信息。经典的机器学习方法已经在多个领域取得了巨大的成功，然而语音等数据具有多维度特点，传统的机器学习方法难以对如此高维度的数据进行处理。深度学习(deep learning，DL)的出现为该问题的解决提供了可能。深度神经网络可被视为由多个隐含层组成的神经网络结构模型，属于机器学习的一个分支。调整神经元的连接方式、改变激活函数、增加网络模型深度等方式可以有效地优化深层神经网络。

5.3.1 卷积神经网络

卷积神经网络(convolutional neural network，CNN)是一种基于视觉感受野机制的具有卷积结构的前馈神经网络，神经元感受野是指视觉神经系统中的视网膜上的一块区域，仅刺激这块区域时才可以激活该神经元，很多感受野交错重叠在一起，最终覆盖整个视线域。

卷积神经网络的基本结构单元主要有池化层、卷积层及全连接层，且卷积神经网络具有池化、共享权值及局部感受野等结构特性。与全连接网络相比，卷积神经网络能够进行空间平移、旋转等操作，这样既能保留其数据内部的关联性，还能够有效地减少网络模型中的相关参数，卷积结构可以有效地降低模型出现过拟合现象的概率。

5.3.1.1 最新进展

近年来，研究人员基于机器学习和深度学习提出了一系列卷积神经网络（CNN）改进算法，算法主要基于 CNN 的基础结构及其特性，已在视觉任务领域取得了一定的进展。此外，FlashInternImage 和 ViT

等架构的出现也为进一步实现 CNN 性能提升创造了可能。

CNN 性能提升的主要影响因素有网络结构、优化方法、损失函数、模型表示方式。在 FlashInternImage 和 ViT 框架中，深度学习方法通常受两方面影响。一方面，通过各种优化方法改进模型性能。另一方面，使用一个评估这些模型性能的指标。对于 CNN 或者其他深度学习模型而言，指标通常是指损失函数，例如，最常用的交叉熵损失函数。而低效的优化方法一直是 CNN 性能提升的瓶颈。如果不对模型进行有效优化，那么就不可能实现精确的预测和严格的性能提升。

2024 年初提出的一种高效的变形卷积网络 DCNv4，重新思考了动态和稀疏操作在视觉应用中的使用。枚举出单个 CNN 数百万种可能的操作，并覆盖所有构建路径。DCNv4 是一种高效的动态和稀疏操作符，它重新思考了可变形卷积的动态特性，并简化了内存访问，运行速度和性能都有显著提升。相较于前一版本 DCNv3，DCNv4 使用一个线程处理同一组中的多个通道，这些通道共享采样偏移和聚合权重。这样可以减少内存读取和双线性插值系数计算等工作负载，并且可以合并多个内存访问指令，具体如图 5-2 所示。

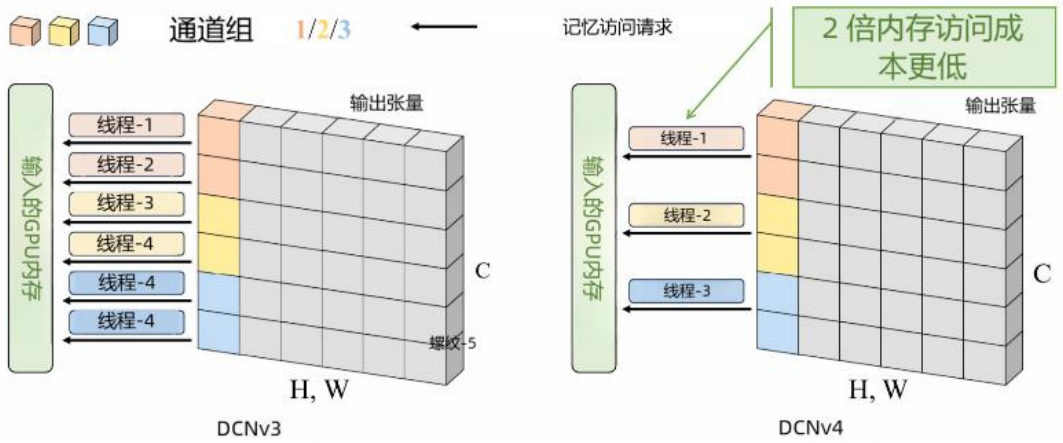


图 5-2　DCNv4 相对于 DCNv3 的线程改进

配备 FlashInternImage 骨干网络的 DCNv4 不仅提高了运行速度，还改善了各种视觉任务的性能。值得注意的是，DCNv4 还展示了其

作为通用操作符的多功能性和有效性。通过将其集成到 ConvNeXt 和 ViT 等先进的架构中，DCNv4 进一步提高了吞吐量和准确性。此外，DCNv4 在潜在扩散模型中也有出色的表现，展示了其在增强生成模型方面的潜力。

可以用空间位置选择性地放大或衰减滤波器将降低计算成本以提高模型精度。为了进一步降低计算成本，提高模型精度和完整度，在训练时采用并行机制，并对模型结果进行进一步判断和处理。在卷积之后，通过一个编码器网络将坐标传递，并通过乘法门应用于卷积后的数据，实现了根据空间位置选择性地放大或衰减滤波器的功能，具体流程如图 5-3 所示。

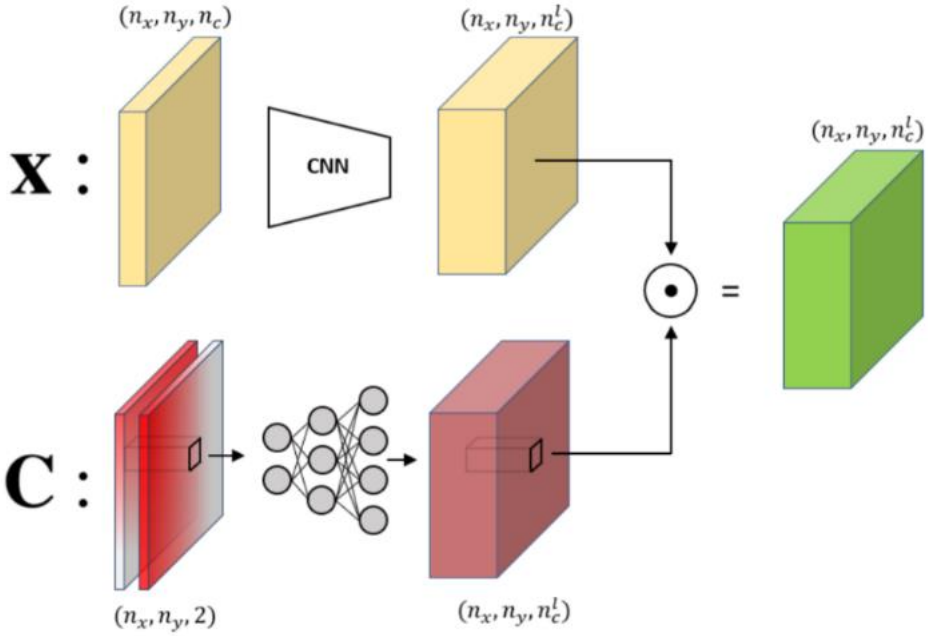


图 5-3 卷积 CoordGate 模块

在卷积 CoordGate 模块中，数据 X 和坐标 C 分别通过卷积神经网络（CNN）和多层感知机（MLP）进行处理，然后对得到的张量进行哈达玛乘积（Hadamard product）。这种技术为 CNN 提供了一种新的、高效的计算空间变化卷积的方法。实验证明，CoordGate 在 U-Net 中的应用能够在图像去模糊等任务中取得比传统方法更好的效

果，为计算机视觉应用提供了更强大和空间感知的解决方案。

通过引入了大卷积核可产生三个效果：扩大感受野，增加空间模式的抽象层次，通过增加深度改进模型的一般表示能力。UniRepLKNet 则是一种通用大卷积核 ConvNet 架构。它将 3×3 卷积添加到小卷积核 ConvNet 中，期望通过扩大感受野、增加空间模式的抽象层次和通过增加深度改进模型的一般表示能力，来提升 CNN 的性能。提出了一种稀疏重参数块（Dilated Reparam Block），该块使用非稀疏的小卷积核和多个稀疏的小卷积核层来增强非稀疏的大卷积核层，它的超参数包括大卷积核的大小 K、并行卷积层的大小 k 和膨胀率 r，具体流程如图 5-4 所示。

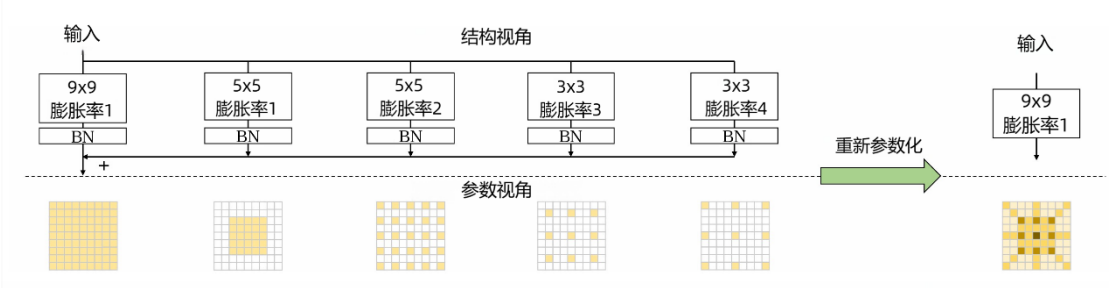


图 5-4　稀疏重参数块

图中包含四个并行层，UniRepLKNet 还引入了一种基于块设计的架构指导原则，既能进行通道间通信又能进行空间聚合的高效结构来增加深度。实验结果显示，经过 ImageNet-22K 预训练后的 UniRepLKNet-S 具有很高的准确性，并且运行速度比 RepLKNet-31L 快 3 倍。

CNN 在基因组学中的应用也日益增多。研究人员利用 CNN 分析基因组序列，识别与特定疾病相关的突变和调控元件，帮助理解复杂的遗传机制。此外，CNN 还被用于单细胞 RNA 测序数据的分析，帮助识别细胞类型和状态，揭示细胞异质性。在 2024 年 7 月的《Scientific Reports》发表的研究中，通过整合单细胞 RNA 测序和卷积神经网络，

揭示了阿尔茨海默病中小胶质细胞的异质性及其复杂的细胞间相互作用，展示了深度学习在基因组研究中的应用前景[88]。随着技术的不断进步，卷积神经网络的应用范围将进一步扩大，可能在更多领域带来突破性的成果，尤其是在生物信息学的深层次研究中。

5.3.2 三维卷积神经网络

三维卷积神经网络是由二维神经网络改进而来的。由于二维卷积神经网络不能很好地捕获视频资源中的时空信息，因此产生了三维卷积神经网络。二维卷积的输出为二维特征图，多用于单通道，而在多通道时图像的多通道信息都被压缩了。三维卷积神经网络可以很好地解决该问题，因为其输出仍是三维特征图，能够捕获视频中的空间和时间特征信息。

随着机器学习和深度学习方法的发展，卷积神经网络方法开始被广泛地应用。一维卷积神经网络（1D CNN）一般用来学习和处理一维的序列类数据；二维卷积神经网络（2D CNN）通常用于目标监测、自然语言处理及图像处理等领域，典型的 2D CNN 算法有 AlexNet、VGG-Net、GoogLeNet、LeNet-5 等；而三维卷积神经网络（3D CNN）则广泛应用于医学领域及视频处理领域。

近年来，三维卷积神经网络逐渐被应用到了生物大分子结构预测领域。例如，在蛋白质结构预测领域，一种端到端优化的可微模型通过优化全局的几何结构并且不违反局部共价化学的几何三元来耦合局部与全局的蛋白质结构，该模型能够在没有预先获取共同进化数据的条件下预测出新的蛋白质折叠结构。基于神经网络来预测碱基对之间距离的 AlphaFold 算法，通过简单的梯度下降算法实现了无须复杂的采样程序即可生成蛋白质结构。AlphaFold2 仍然是一种基于三维卷积神经网络的蛋白质建模方法，该算法利用多序列比对手段，将有关蛋白质结构的物理和生物学知识整合到深度学习算法的设计与实现中。三维卷积神经网络在蛋白质结构预测领域的应用提高了蛋白质的

结构预测准确度，并且能够在无法明确同源蛋白质结构的条件下进一步研究蛋白质的功能。2024 年 5 月 8 日，Google DeepMind 发布了新一代 AlphaFold3, 用于预测蛋白质、DNA、RNA、小分子等的几乎所有生物分子结构和相互作用, AlphaFold3 相较于前版本，能够在与其他分子共同作用时建模蛋白质。

在 RNA 结构预测领域，三维卷积神经网络也得到了应用，基于三维卷积神经网络对 RNA 三级结构预测进行评估，即 RNA 3D CNN，该算法使用结构的三维网格表示作为输入，无须人工提取特征，而是在隐藏层内部直接进行特征处理。3D CNN 的主要优势在于其能够处理三维特征图，直接提取空间和时间信息。这种能力使得它在视频处理、医学影像以及生物信息学等领域表现突出。在视频处理方面，3D CNN 能够同时分析帧之间的变化以及每一帧的细节，增强了运动分析的效果。在蛋白质结构预测领域，3D CNN 的应用也取得了革命性的进展。一种端到端优化的可微模型通过全局几何结构的优化与局部共价化学几何的耦合，能够有效整合局部与全局的蛋白质结构信息。这种方法在没有共同进化数据的情况下，实现了对新蛋白质折叠的预测。

Townshend 和 Eismann 提出了一个基于三维卷积神经网络的结构模型 ARES，该模型不需要任何有关结构模型的相关概念及与评估其准确性相关的假设，具有较强的灵活性。此外，ARES 模型不仅可以针对 RNA 结构预测，还可以应用到其他类型分子系统的结构预测。ARES 模型是一种基于 3D CNN 的结构预测模型，展示了 3D CNN 在生物分子研究中的广泛应用潜力。随着深度学习技术的不断进步，3D CNN 在生物信息学领域的应用将继续扩展，未来的研究可能集中在模型集成、数据增强、多模态学习以及增强模型的可解释性等方向。

5.3.3 基于 ResNet 的三维卷积神经网络

残差网络(ResNet)也是卷积神经网络，在保持卷积核大小不变的

情况下，增加网络的宽度及深度能有效地提升网络模型的性能，然而当网络深度过深时，将会出现梯度爆炸或梯度弥散问题，该问题可以通过正则化初始化来解决。然而，退化问题无法通过上述方法解决，仍然会出现随着网络深度增加，模型训练效果可能接近饱和甚至下降的现象。因此，神经网络不能够简单地通过增加深度来进行优化，ResNet 的出现是为了解决网络深度增加带来的网络退化和梯度弥散问题。ResNet 内有多个残差学习单元，ResNet 残差单元可以表示为

$$y_l = h(x_l) + F(x_l, W_l) \tag{5.1}$$

$$x_{l+1} = f(y_l) \tag{5.2}$$

$$h(x_l) = x_l \tag{5.3}$$

式中，l 表示第 l 个残差单元；x_l 与 x_{l+1} 分别表示其输入和输出；$F()$表示残差函数；$f()$表示 ReLU 型激活函数。ReLU 函数有很多种，具体如图 5-5 所示。

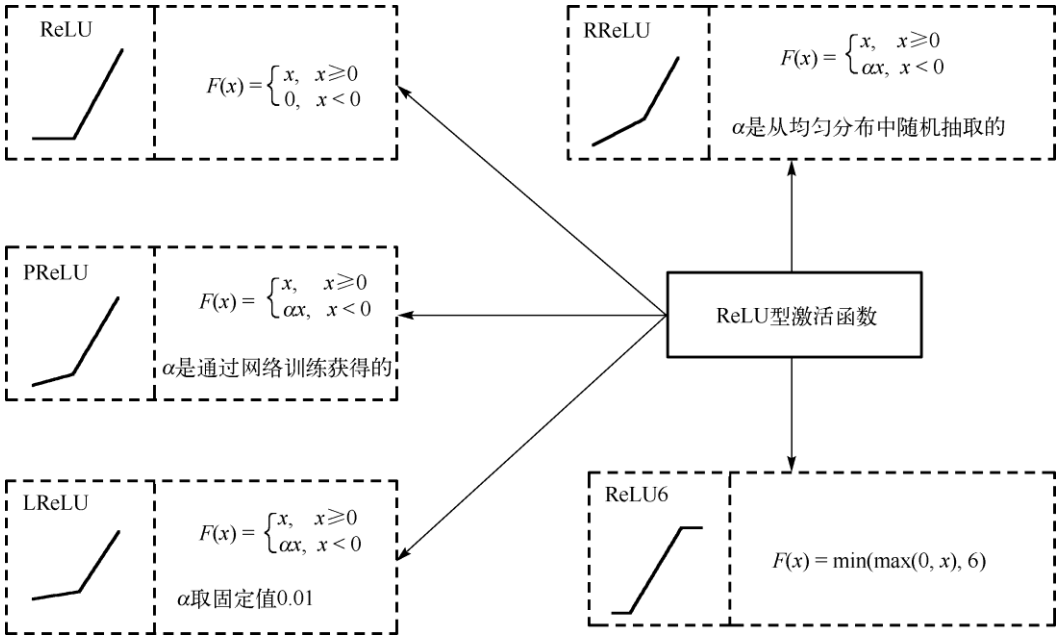


图 5-5　ReLU 型激活函数

ResNet 从其浅层 l 到深层 L 的学习特征为

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{5.4}$$

ResNet 目前广泛地应用于医学图像分类、超分辨率、重建、合

成、疾病检测等医学图像分析领域，并取得了很大进展，因此，本书期望用 ResNet 来对 RNA 三级结构打分函数进行改进和优化。机器学习、深度学习与算法及计算复杂性理论助力 RNA 结构方面的研究。假设 RNA 片段由 15 个碱基(核苷酸)组成，理论上其结构数为 13 万亿个，这是一个天文数字。冠状病毒约由 3 万个碱基组成，其遗传物质是已知 RNA 病毒中最长的，理论上其结构数更是天文数字，并且病毒在不停地变种，可能的 RNA 三级结构数更是天文数字，不可能逐一用实验来测定，只能用计算的方法，特别是通过设计人工智能近似算法来计算其可能的结构，会得到意想不到的结果。

参考文献

[1] Lyngsø R B, Christian N S. Pseudoknots in RNA pseudoknotted structure[C]. Proceedings of Recomb, Tokyo, 2000.

[2] Dulbecco R. A turning point in cancer research: Sequencing the human genome[J]. Science, 1986, 231: 1055-1056.

[3] Yang Y R, Liu Z D. A comprehensive review of predicting method of RNA tertiary structure[J]. Computational Biology and Bioinformatics, 2021, 9(1): 15-20.

[4] Turner D H, Sugimoto N, Freier S M. Improved parameters for prediction of RNA structure[J]. Biophysics Chemistry, 1988, 17(2): 167-192.

[5] Mathews D H, Turner D H. Prediction of RNA secondary structure by free energy minimization[J]. Current Opinion in Structural Biology, 2006, 16(5): 270-278.

[6] Walter A E, Turner D H, Kim J, et al. Coaxial stacking of helixes enhances binding of oligo onucleotides and improves predictions of RNA folding[J]. Proceedings of the National Academy of Sciences, 1994, 91(2): 9218-9222.

[7] Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history[J]. Bioinformatics, 1999, 15(6): 446-454.

[8] Hochbaum D S. Approximation algorithms for NP-hard problems[J]. ACM SIGACT News, 1997, 28(2): 40-52.

[9] Vazirani V. Approximation Algorithms. Berlin: Springer, 2001.

[10] Zuker M, Mathews D H, Turner D H．Algorithms and Thermodynamics for RNA Secondary Structure Prediction：A Practical Guide in RNA Biochemistry and Biotechnology. Den Haag City: Kluwer

Academic Publishers, 1999: 11-43.

[11] Rivas E, Eddy S R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of Molecular Biology, 1999, 285(5): 2053-2068.

[12] van Batenburg F H, Gultyaev A P, Pleij C W, et al. PseudoBase: A database mRNA pseudoknots[J]. Nucleic Acids Research, 2000, 28(1): 201-204.

[13] Nixon P L, Rangan A, Kim Y G, et al. Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot[J]. Journal of Molecular Biology, 2002, 322(3): 621-633.

[14] Ieong S, Kao M Y, Lam T W, et al. Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs[J]. Journal of Computational Biology, 2003, 10(6): 981-995.

[15] Lyngsø R B. Complexity of Pseudoknot Prediction in Simple Models[M]. Berlin: Springer, 2004: 919-931.

[16] Ruan J, Stormo G D, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots[J]. Bioinformatics, 2004, 20(1): 58-66.

[17] Ren J, Rastegari B, Condon A, et al. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots[J]. RNA, 2005, 11(10): 1494-1504.

[18] Huang X, Ali H. High sensitivity RNA pseudoknot prediction[J]. Nucleic Acids Research, 2007, 35(2): 656-663.

[19] Han B, Dost B, Bafna V. Structural alignment of pseudoknotted RNA[J]. Journal of Computational Biology, 2008, 15(7): 489-504.

[20] 徐琳，李晓民，谭光明，等．面向FPGA的RNA二级结构预测并

行算法研究[J]. 计算机学报, 2006, 2(29): 233-238.
[21] 陈翔, 卜东波, 张法, 等. 基于局部茎搜索的RNA二级结构预测算法[J]. 生物化学与生物物理学进展，2009, 36(1): 115-121.
[22] 刘元宁, 张浩, 李誌, 等. RNA假结结构分析[J]. 吉林大学学报(工学版), 2009, (S1I): 265-269.
[23] Li F, Gao L, Wang B B. Detection of driver modules with rarely mutated genes in cancers[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 17(2): 390-401.
[24] Wen X, Gao L, Hu Y X. LAceModule: Identification of competing endogenous RNA modules by integrating dynamic correlation[J]. Frontiers in Genetics, 2020, 11(3): 235-241.
[25] Yue D, Guo M Z, Chen Y D, et al. A Bayesian decision fusion approach for microRNA target prediction[J]. BMC Genomics, 2012, 13(S8): S13.
[26] Gupta A, Kumar A, Pál M, et al. Approximation via cost-sharing: A simple approximation algorithm for the multicommodity rent-or-buy problem[C]. Proceedings of the 44th IEEE Annual Symposium on Foundations of Computer Science, Washington, 2003: 606-615.
[27] Gupta A, Kumar A, Pál M, et al. Approximation via cost sharing: Simpler and better approximation algorithms for network design[J]. Journal of the ACM, 2007, 54(3): 1-38.
[28] Hassin R, Monnot J, Segev D. Approximation algorithms and hardness results for labeled connectivity problems[J]. Journal of Combinatorial Optimization, 2007, 14(4): 437-453.
[29] Williamson D, van Zuylen A. A simpler and better derandomization for an approximation algorithm for single-source rent-or-buy[J]. Operations Research Letters, 2007, 35(6): 707-712.

[30] Lau L C M. Singh: Additive approximation for bounded degree survivable network design[C]. Proceedings of the 40th ACM Symposium on Theory of Computing, New York, 2008: 759-768.

[31] Liu Z D, Li H W, Zhu D M. A predicting algorithm of RNA secondary structure based on stems[J]. Kybernetes, 2010, 39(6): 1050-1057.

[32] Liu Z D, Xia C L, Zhu D M. Improved algorithm for RNA secondary structure prediction including pseudoknots[J]. Advances in Systems Science and Applications, 2010, 10(4): 710-716.

[33] Wong T K F, Lam T W, Sung W K, et al. Structural alignment of RNA with complex pseudoknot structure[J]. Lecture Notes in Computer Science, 2009, 5724(6): 403-414.

[34] Wong T K F, Wan K L, Hsu B Y, et al. RNASAlign: RNA structural alignment system[J]. BMC Bioinformatics, 2011, 27(15): 2151-2152.

[35] Wong T K F, Chiu Y S, Lam T W, et al. Memory efficient algorithms for structural alignment of RNAs with pseudoknots[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(1): 161-168.

[36] Liu Z D. Approximation algorithm of RNA folding including pseudoknots[J]. International Review on Computers and Software, 2012, 7(6): 2942-2946.

[37] Liu Z D, Zhu D M. New heuristic algorithm of RNA structure prediction including pseudoknots[J]. Journal of Computers, 2013, 8(2): 279-283.

[38] Reinharz V, Ponty Y, Waldispühl J. A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution[J]. Bioinformatics, 2013, 29(13): 308-315.

[39] Liu Z D, Zhu D M, Ma H W. Predicting scheme of RNA folding structure including pseudoknots[J]. International Journal of Sensor Networks, 2014, 16(4): 229-235.

[40] Andronescu M, Condon A, Hoos H H, et al. Computational approaches for RNA energy parameter estimation[J]. RNA, 2010, 16(12): 2304-2318.

[41] Andronescu M, Condon A, Turner D H, et al. Determination of RNA folding nearest neighbor parameters[J]. Methods Molecular Biology, 2014, 1097: 45-70.

[42] Babai L. Graph isomorphism in quasipolynomial time[J]. Combinatorics and Theoretical Computer Science Seminar, 2015, 13(2): 18-26.

[43] Keane S C, Heng X, Lu K, et al. Structure of the HIV-1 RNA packaging signal[J]. Science, 2015, 348(6237): 917-921.

[44] Kucharík M, Hofacker I L, Stadler P F, et al. Pseudoknots in RNA folding landscapes[J]. Bioinformatics, 2016, 32(2): 187-194.

[45] Gomez-Schiavon M, Chen L F, West A E, et al. BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells[J]. Genome Biology, 2017, 18(2): 164.

[46] Nuoroozi G, Mirmotalebisohi S A, Sameni M, et al. Deregulation of microRNAs in oral squamous cell carcinoma, a bioinformatics analysis[J]. Gene Reports, 2021, 11(3): 101241.

[47] Wu D, Ding Y, Fan J B. Bioinformatics analysis of autophagy-related lncRNAs in esophageal carcinoma[J]. Combinatorial Chemistry and High Throughput Screening, 2021, 24(4): 101241.

[48] Tang L. A path to predict RNA tertiary structures[J]. Nature Methods,

2018, 15(7): 650.

[49] Weeks K M. Piercing the fog of the RNA structure-ome[J]. Science, 2021, 373(6558): 964-965.

[50] Kappel K, Zhang K, Su Z, et al. Accelerated cryo-EM-guided determination of three- dimensional RNA-only structures[J]. Nature Methods, 2020, 17(10): 699-707.

[51] Fan X, Wang J, Zhang X, et al. Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution[J]. Nature Communications, 2019, 10(4): 2386.

[52] Yang Y, Liu Z. A comprehensive review of predicting method of RNA tertiary structure[J]. Computational Biology and Bioinformatics, 2021, 9(3): 9-15.

[53] Perez A, Morrone J A, Brini E, et al. Blind protein structure prediction using accelerated free-energy simulations[J]. Science Advances, 2016, 2(11): e1601274.

[54] Magdalena R, Kristian R, Tomasz P, et al. ModeRNA: A tool for comparative modeling of RNA 3D structure[J]. Nucleic Acids Research, 2011, 39(2): 13-22.

[55] Zhao Y, Huang Y, Gong Z, et al. Automated and fast building of three-dimensional RNA structures[J]. Scientific Reports, 2012, 2(5): 727-734.

[56] Das R, Karanicolas J, Baker D. Atomic accuracy in predicting and designing noncanonical RNA structure[J]. Nature Methods, 2010, 7(6): 291-294.

[57] Massire C, Westhof E. MANIP: An interactive tool for modelling RNA[J]. Journal of Molecular Graphics and Modelling, 1998, 16(2): 197-205.

[58] Das R, Baker D. Macromolecular modeling with rosetta[J]. Annual Review of Biochemistry, 2008, 77(8): 363-382.
[59] Schoeder C T, Schmitz S, Adolf-Bryfogle J, et al. Modeling immunity with rosetta: Methods for antibody and antigen design[J]. Biochemistry, 2021, 60(6): 825-846.
[60] Li J, Zhu W, Wang J, et al. RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks[J]. PLoS Computational Biology, 2018, 14(2): 1-18.
[61] Bradley P, Misura K, Baker D. Toward high-resolution de novo structure prediction for small proteins[J]. Science, 2010, 309(11): 1868-1871.
[62] Sripakdeevong P, Kladwang W, Das R. An enumerative stepwise ansatz enables atomic- accuracy RNA loop modeling[C]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 10(9): 20573-20578.
[63] Watkins A M, Geniesse C, Kladwang W, et al. Blind prediction of noncanonical RNA structure at atomic accuracy[C]. Science Advances, 2018, 4(5): eaar5316.
[64] Liu Z D, Zhu D M, Dai Q H. Predicting model and algorithm in RNA folding structure including pseudoknots[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2018, 32(10): 1-17.
[65] Meng G, Tariq M, Jain S. RAG-Web: RNA structure prediction/design using RNA-As- Graphs[J]. Bioinformatics, 2019, 13(5): 647-648.
[66] Rivas E, Clements J, Eddy R S. Estimating the power of sequence covariation for detecting conserved RNA structure[J]. Bioinformatics, 2020, 11(9): 3072-3076.

[67] Menden K, Marouf M, Oller S. Deep learning-based cell composition analysis from tissue expression profiles[J]. Science, 2020, 6(28): 51-59.

[68] Liu Z D, Li G, Liu J S. New algorithms in RNA structure prediction based on BHG[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2020, 34(13): 1-14.

[69] Guo Z F, Wang P P, Liu Z D, et al. Discrimination of thermophilic proteins and non- thermophilic proteins using feature dimension reduction[J]. Frontiers in Bioengineering and Biotechnology, 2020, 8: 1-10.

[70] Zhang P, Liu Z D. Approximating max k-uncut via LP-rounding plus greed, with applications to densest k-subgraph[J]. Theoretical Computer Science, 2020, 849(14): 173-183.

[71] Townshend R, Eismann S, Watkins A M, et al. Geometric deep learning of RNA structure[J]. Science, 2021, 373(6531): 1047-1051.

[72] Park J U, Tsai A W L, Mehrotra1 E, et al. Structural basis for target site selection in RNA-guided DNA transposition systems[J]. Science, 2021, 373(2): 768-774.

[73] Niu M T, Wu J, Zou Q, et al. Predicting RNA-binding proteins using deep learning[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(9): 3668-3676.

[74] Rasmussen M, Reddy M, Nolan R, et al. RNA profiles reveal signatures of future health and disease in pregnancy[J]. Nature, 2022, 601(15): 422-427.

[75] Garcia-Beltran W F, Denis K J S, Hoelzemer A, et al. mRNA-based COVID-19 vaccine boosters induce neutralizing immunity against SARS-CoV-2 Omicron variant[J]. Cell, 2022, 185: 457-466.

[76] Liu Z D, Yang Y R, Li D Y, et al. Prediction of RNA tertiary structure based on random sampling strategy and parallel mechanism[J]. Frontiers in Genetics, Section Computational Genomics, 2022, 12(8): 1-10.

[77] Liu Z D, Lv X R, Chen X, et al. Predicting algorithm of tissue cell ratio based on deep learning using single-cell RNA sequencing[J]. Applied Sciences, 2022, 12(5790): 1-14.

[78] Liu Z D, Chen X, Li D Y, et al. Predicting algorithm of attC site based on combination optimization strategy[J]. Connection Science, 2022, 34(1): 1895-1912.

[79] Ito T M, Ogawa S, Ashida K, et al. Accurate magnetic field imaging using nanodiamond quantum sensors enhanced by machine learning[J]. Scientific Reports, 2022, 12: 13942.

[80] Nguyen L, van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features[J]. Nature Communications, 2022, 13: 4013.

[81] Kong J H, Ha D, Lee J, et al. Network-based machine learning approach to predict immunotherapy response in cancer patients[J]. Nature Communications, 2022, 13: 3703.

[82] Szczerba M, Johnson B, Acciai F, et al. Canonical cellular stress granules are required for arsenite-induced necroptosis mediated by Z-DNA-binding protein[J]. Science, 2023, 16(12): 776.

[83] Artem Nemudryi, Anna Nemudraia, Joseph E. Nichols,et al, CRISPR-based engineering of RNA viruses, Science Advances,2023, eadj8277 (2023):1-9.

[84] McCauley O. Meyer, Ryota Yamagami, Saehyun Choi, Christine D.

Keating, Philip C. Bevilacqua,RNA folding studies inside peptide-rich droplets reveal roles of modified nucleosides at the origin of life, Science Advances,2023, eadh5152 (2023):1-15.

[85] 刘振栋、肖传乐、邹权、张博峰.生物信息学中RNA结构预测算法与复杂性，北京：科学出版社，2024年2月.

[86] Tebbe de Vries, Mihajlo Novakovic, Yinan Ni, Izabela Smok, Specific protein-RNA interactions are mostly preserved in biomolecular condensates, Science Advances,2024, eadm7435 (2024):1-12.

[87] Elizabeth Pennisi, Surprise RNA paints colorful patterns on butterfly wings, Science,2024,383(6687):1039-1040.

[88] Wu, X., Liu, M., Zhang, X.et al. Elucidating Microglial Heterogeneity and Functions in Alzheimer's Disease Using Single-cell Analysis and Convolutional Neural Network Disease Model Construction. Scientific Reports 14, 17271 (2024).

第 6 章 人工智能识别组学生物标志物

6.1 背景

生物标志物在医学和生物学领域具有重要作用，它们是客观测量和评估的生物特征，能够指示生理或病理过程以及药物对体内生物过程的反应。生物标志物在疾病的早期诊断、预测和预防、个性化治疗、临床试验以及疾病进展和预后评估中发挥着关键作用，已经成为现代生物学和医学中不可或缺的一部分。例如，癌症中的肿瘤标志物（如 PSA 用于前列腺癌，CA-125 用于卵巢癌）有助于早期发现和监测；心血管疾病中的心肌损伤标志物（如肌钙蛋白）用于诊断心肌梗死；糖尿病患者的 HbA1c 水平用于长期血糖控制评估。通过检测这些生物标志物，可以制定个性化治疗方案，提高疗效，减少副作用，同时在新药开发过程中，生物标志物用于评估药物的疗效和安全性。未来，随着技术进步和对生物系统理解的深入，生物标志物的应用将更加广泛和精准，尤其是多重标志物组合、液体活检以及人工智能和大数据分析的结合，将显著提升医学诊断和治疗的效果。

人工智能在识别生物标志物的应用日益增多，尤其在处理和分析高通量组学数据时表现出极大的潜力。组学数据（包括转录组、蛋白质组等）通常包含数千到数万个变量，其中只有少数特征与生理或病理状态密切相关，因此识别生物标志物的过程旨在从高维的组学数据中提取出具有较强预测能力的标志物，其本质为机器学习中的特征选择问题。早期选择生物标志物的方法仅依靠单一组学并结合一些先验信息，如基因之间的调控关系。而随着测序技术的日益成熟，结合多种不同组学选择生物标志物的方法应运而生。

6.2 常见的单组学方法

高通量组学数据描述了生物体内各个分子层面上的信息，反应了

生物体在正常或疾病状态下的复杂生物学过程。而高通量组学数据中通常仅有少数特征与特定的生理或病理状态密切相关。特征选择的目的就在于从这些高维数据中筛选出尽可能少的特征，同时尽可能提高模型性能。通常，特征选择方法主要可分为过滤式、嵌入式和包裹式三种类型。

6.2.1 过滤式

过滤式方法通常被用作特征选择过程中的数据预处理步骤，以减少数据集中的特征数量。虽然过滤式方法可以单独用于特征选择，但它们不足以完全捕捉特征与目标之间的复杂关系，特别是在涉及非线性复杂模式或交互作用的任务中。工具包 Caret（Classification And REgression Training）[1]提供了一个全面的机器学习框架，支持多种模型的训练、参数调优和特征选择。Boruta[2]是一个基于随机森林的特征选择方法，通过创建“阴影特征”（即随机打乱的真实特征）来测试每个特征与响应变量之间的相关性是否显著高于随机噪声。由于这种特征选择方式没有与分类器结合，所以选择出的特征通常不能达到最优的分类性能。

6.2.2 包裹式

包裹式特征选择方法是对不同的特征子集进行评估以获得最优集合。这种选择方式将分类器的性能作为最终的评价标准，其目的就是为给定的分类器“量身定做”特征子集。最常见的包裹式方法包括递归特征消除（Recursive Feature Elimination, RFE）等。例如，Guyon 等人[3]提出了一种支持向量机递归特征消除（SVM-RFE）方法，该方法使用 SVM 分类器来评估特征的重要性，并通过递归的方式逐步删除重要性最低的特征。Li 和 Liu[4]已通过该方法从自发性早产基因表达数据有效识别出 54 个生物标志物。Kursa 等人[5]提出了基于随机森林的递归特征消除方法（RF-RFE），该方法利用 RF 分类器度量变量的重要性来进行特征选择。Fortino 等人[6]提出的一种新型的多岛自适

应遗传算法 GARBO。它通过调整遗传操作符的概率和特征的初始排名，有效地优化了特征选择过程。

6.2.3 嵌入式

嵌入式特征选择是将特征选择和分类器的训练过程融为一体，即在训练分类器的同时自动地选择特征子集，这与过滤式和包裹式的特征选择方法有明显区别。例如，Feng 等人[7]采用最小绝对收缩和选择算子（LASSO）回归分析方法，筛选出 14 个差异表达免疫相关基因用于晚期冠状动脉疾病（CAD）的诊断，进而构建了一个基于这些生物标志物的晚期 CAD 的诊断模型。Huang 等人[8]通过整合 $L_{1/2}$ 正则化的稀疏性和 L_2 正则化的群组效应，提出了一种混合 $L_{1/2+2}$ 正则化（Hybrid $L_{1/2+2}$ Regularization, HLR）方法，并采用坐标下降算法优化带有 HLR 惩罚的逻辑回归模型。这一方法有效应对了基因数量远超样本量的挑战，并克服了传统逻辑回归在高维小样本数据分析中的过拟合问题，并能够自主选择有利特征。Díaz-Uriarte 等人[9]采用随机森林算法进行基因选择和分类，并开发了 R 包 varSelRF，该方法旨在从微阵列数据中提取最尽可能小的基因子集，以实现对样本的精确分类预测。

6.3 从网络中发展生物标志物

虽然这些特征选择方法在特定环境下能够找到部分生物标志物，然而它们并没有考虑到生物系统的复杂性以及基因之间的关联性，这导致了所有的基因都是以孤立节点存在于子集中，忽略了基因间的相互作用和协同效应。实际上，复杂疾病往往不仅由单个分子的异常引起，而是在多个信号通路和分子网络的交互作用共同影响下的结果。为了深入探索疾病的分子机制并提升诊断准确率，研究者们开始探求考虑分子间相互作用的方法，即通过分析分子组或更大的分子集合的相互作用，揭示复杂的分子相互作用和信号传导路径，从而识别出能

够反映分子间相互作用的模块生物标志物或网络生物标志物。

目前已有一些基于网络的特征选择方法。例如 Horvath 等人[10]提出了加权基因共表达网络分析（Weighted Gene Co-expression Network Analysis, WGCNA）方法，通过分析基因表达模式的相似性来识别共表达的基因模块，并将这些模块与外部表型关联，从而识别与疾病密切相关的基因集。然而，该方法并未考虑针对特定研究问题的先验特征，且其构建的网络主要是基于数据驱动，侧重于揭示基因之间的共表达关系，而不涉及基因间的因果关系或调控机制。另一方面，基因调控网络（Gene Regulatory Network, GRN）可以很好地缓解这一问题。GRN 将基因、转录因子等生物分子作为节点，通过分子间的调控关系作为连接节点的边，以网络图的形式直观地反应生物分子间相互作用关系。在 GRN 中，每个节点都充当信息处理的单元，接受来自其他基因的信号调控，并据此调整其自身的表达水平。连线则代表互动的性质与强度，既可以是正向的促进作用，也可以是负向的抑制作用。基于高通量测序技术收集的大量生物数据，结合已知 GRN 与特征选择方法，能够显著提升预测模型的准确性及生物标志物的可解释性。为此，Li 等人[11]将基因网络的连接结构作为约束条件纳入支持向量机模型中，提出了嵌入式连通网络约束支持向量机方法（CNet-SVM），用于在保持基因间固有图形结构的同时，从高通量组学数据中识别和分类癌症生物标志物。Zhang 等人[12]提出一种可解释基于网络的博弈论方法，将基因到模型选择的过程视为一个合作博弈，每个特征的组合贡献通过合作博弈理论度量，即 Shapley 值来评估，并通过赤池信息准则（AIC）在模型选择中进行了统计验证，有效区分了肝细胞癌和健康样本。Wang 等人[13]提出了一种基于不同状态下重构基因调控网络识别生物标志物的生物信息学新方法。Shang 等人[14]采用由特定网络中的表型状态指导的迭代监督模块检测方法，并通过网络拓扑中心性在局部和全局进行基于块的模块排名，以检测

可靠的生物标志物模块。

6.4 单组学研究的局限性

尽管单组学研究在其特定领域内提供了宝贵的见解，但它们各自都存在局限性。生物过程是复杂和多层次的，单一组学数据往往无法全面揭示生物体的复杂性。例如，基因组学数据可能显示某个基因的突变，但这种突变如何影响蛋白质的功能和细胞代谢则需要蛋白质组学和代谢组学的数据来解释。不仅如此，单组学方法通常不能提供足够的信息来解析生物过程中各组分之间的相互作用和网络。此外，单组学研究通常只能在静态条件下捕捉到一时的状态，而生物过程是动态变化的。例如，一个基因的表达可能因为环境变化或发育阶段的不同而波动，单一时间点的转录组分析可能无法完全捕捉这种动态变化。这种静态的视角限制了我们对生物调控网络动态性的理解。

为了克服这些局限，研究者们逐渐趋向于采用多组学整合分析方法，通过综合考虑多种组学数据，形成对生物体状态更全面的理解。多组学研究通过整合不同层面的分子信息，能够揭示更为复杂的生物机制，如基因表达如何被转录后修饰控制，以及这些过程如何影响细胞的代谢路径。这种整合分析不仅增强了特征的筛选能力，还提高了疾病预测的精度，对精准医疗等领域具有重要的应用价值。

6.5 多组学的研究的优势

在当今生物科学和医学领域，单组学研究的局限性促使科学家探索更为全面的研究方法，即多组学研究。多组学研究是一种整合基因组学、转录组学、蛋白质组学、代谢组学等多种生物信息的方法，以获取关于生物体系统的全面视图。这种整合的方法不仅能提供单一数据层面无法揭示的信息，还能更有效地解析复杂的生物过程和疾病机制。其优势可总结为以下几点：

全面的生物系统视图：通过整合不同类型的组学数据，多组学研

究能够提供一个全面的系统生物学视角，揭示不同分子层次之间的相互作用和网络。例如，基因变异如何影响蛋白质表达，以及这些蛋白质如何影响代谢路径，这些信息对于理解复杂疾病的生物学基础至关重要。

改进疾病诊断和预测：多组学数据可以用来识别新的疾病标志物和治疗靶点，这些标志物和靶点可能在单一组学研究中无法检测到。此外，多组学分析通过整合多层次数据，能够提高疾病预测模型的准确性，对于精准医疗和个体化治疗策略的制定具有重要价值。

揭示复杂疾病机制：多组学分析有助于解析复杂疾病的多因素起因，包括遗传、表型和环境因素。这种整合视角为疾病提供了一个更为复杂的因果框架，有助于科学家和临床医生更好地理解疾病的多样性和个体差异。

比如，2024 年 Jiang 等人[15]进行了一项综合多组学分析，由于在公开可用的大规模研究中，存在亚洲患者的代表性不足的问题，研究者建立了一个包括 773 名中国乳腺癌患者的综合多组学队列，专门针对中国乳腺癌患者进行了深入研究，揭示了该疾病患者的不同分层以及治疗上的脆弱性。该研究发现，与西方人群相比，中国患者中存在更多可靶向的 AKT1 基因突变，以及更高比例的 HER2 富集亚型，这为个性化抗 HER2 治疗提供了潜在机会。此外，研究还明确了铁死亡作为基底样乳腺癌的治疗靶点，并通过整合临床、基因组和代谢组数据，有效地对患者进行了分层，为亚洲人群的精准医疗带来了新的启示。此外，多项文献研究表明，多组学数据可以提高患者临床结果的预测准确性[16-22]。

6.6 多组学数据的整合策略

在生物信息学和精准医学的领域中，多组学数据整合是一个关键的研究方向。随着组学技术的发展，研究者们可以从基因组、转录组、

蛋白质组、代谢组等多个层面获取数据。多组学数据整合可以揭示这些不同层面之间的相互作用，从而为疾病的诊断、治疗和预后提供更全面的视角。多组学数据整合策略主要分为：前融合、中融合和后融合，如图 6-1 所示。前融合一般是学习特征数据的简单表示，而中融合和后融合大都是将这些数据组合成更加抽象的表示，因此可以选择在任意位置进行模态融合操作。三种融合方式各有优劣，并没有证据表明哪个融合方式是最好的，选择合适的融合方式应取决于任务的性质和数据的特点。以下是三种策略的介绍。

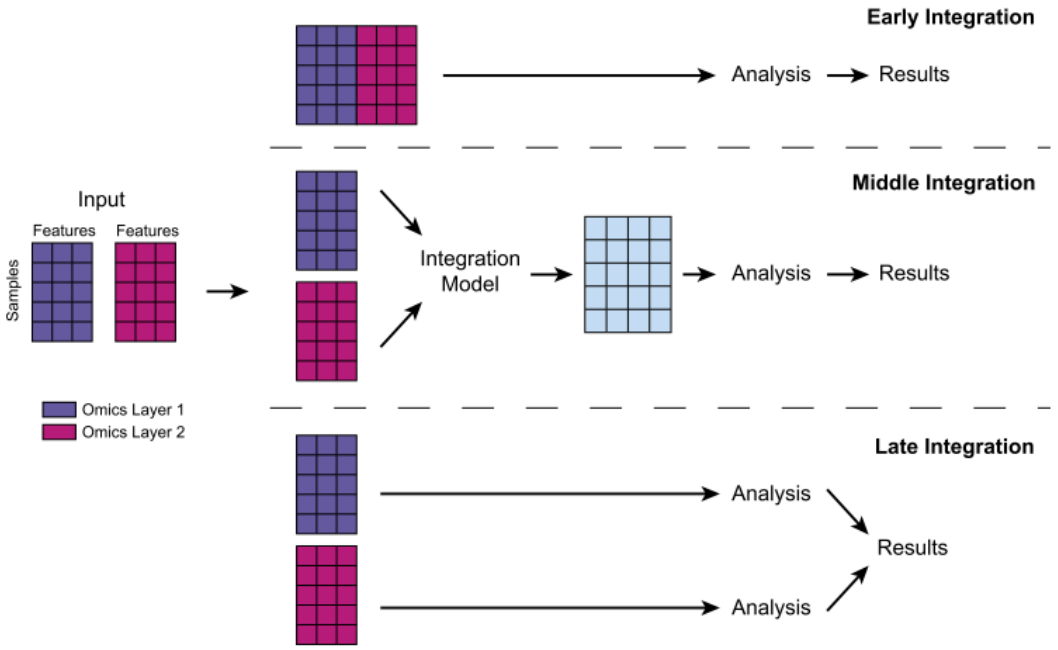


图 6-1 前融合、中融合和后融合策略对比图

6.6.1 前融合

前融合是指在数据分析的早期阶段就开始整合不同来源的组学数据。在这种策略中，原始数据（如 DNA 序列、RNA 表达数据和蛋白质丰度数据）会在进行任何单独的分析之前被合并。这种方法通常涉及到创建一个综合的数据集，其中每个样本的所有组学数据都被视为一个整体来进行分析。方法介绍：（1）特征串联：特征串联是前融

合中最直接的方法之一，它涉及将来自不同组学的特征直接串联到一个长向量中。例如，一个样本的基因表达数据、蛋白质丰度和代谢物浓度可以直接合并成一个特征向量。这样的方法便于使用传统的机器学习算法进行后续分析，如分类和聚类。（2）多维缩放融合：由于前融合的原始数据通常包含大量的冗余信息，因此，多模态前融合方法常常与特征提取方法相结合以剔除冗余信息。在多维缩放融合中，首先使用降维技术（如 PCA 或 t-SNE）处理每种组学数据，以减少每个数据集的维度。然后，这些降维后的数据集被合并为一个综合数据集。这种方法有助于保留每种组学数据中的主要变异信息，同时使得整合后的数据更易于管理和分析。

优点：前融合的优点在于其能够最大限度地利用多种组学数据中的信息，从而提供一个全面的生物学视角。（1）全面性：前融合通过在分析开始前整合数据，确保了对生物样本的全面观察，使得从基因到蛋白质再到代谢产物的每个层面的数据都被考虑在内。这种全面的数据视角是理解复杂生物过程中不同分子层面如何相互作用的关键。（2）减少偏差：在数据处理的早期阶段进行整合可以减少因分离分析而引入的偏差。单独分析不同的组学数据可能会导致重要信号的丢失或误解，因为独立分析可能忽略了不同数据类型之间的相互作用。（3）提高数据整合的一致性：通过前融合，不同来源的数据在分析前被统一处理，从而提高了数据处理过程的标准化和一致性。这对于需要精确生物标记物检测和复杂疾病机理研究的情况尤其重要。

实际案例：Liu 等人[23]将 mRNA、miRNA 和蛋白质丰度的数据堆叠至同一矩阵 X，并利用因子分析（Factor Analysis, FA）对其进行建模。对高级别浆液性卵巢腺癌患者数据集的分析表明，整合多组学数据可以识别与临床结果相关的重要分子特征，如对铂类治疗的反应和患者生存状态。此外，发现了关键分子（如 HOX 基因家族和 RUNX3 转录因子）在癌症发生和治疗抵抗机制中发挥重要作用。

6.6.2 中融合

中融合是在数据处理的中间阶段整合来自不同组学的数据。在这种策略中，各组学数据首先独立进行预处理和特征提取，然后将这些处理后的数据或特征先转化为高维特征，再传入模型的中间层进行融合。这种方法允许更加灵活地处理和分析各种数据，同时保持了数据原始特征的部分独立性。方法介绍（1）特征整合：特征整合是中融合中常见的方法之一，它涉及将从各个组学数据中提取的特征进行整合。例如，可以将基因表达数据中的表达量特征、蛋白质组数据中的蛋白质丰度特征、以及代谢组数据中的代谢物浓度特征合并，形成一个综合特征集，用于后续的分析。（2）基于模型的融合：在这种方法中，各组学数据先分别通过模型（如支持向量机、随机森林等）处理，提取有用的信息或学习到的特征，然后再将这些特征整合到一个模型中进行最终的分析。这种方式有助于从每种数据中挖掘最有价值的信息，同时减少不同数据源之间的噪音干扰。（3）联合学习：联合学习策略通过在特征提取阶段应用多视图学习算法，例如典型相关分析（Canonical Correlation Analysis, CCA）或偏最小二乘法（Partial Least Squares, PLS），来同时处理多种类型的组学数据。这种方法试图找到不同数据集之间的最大相关性，以增强预测模型的准确性和鲁棒性。

优点：（1）灵活性：中融合方法的一大优势就是可以灵活地选择融合的位置。允许研究者在保持各自数据特性的同时整合多源信息，便于使用特定的统计模型和机器学习算法。（2）精确性：通过整合多个数据层面的特征，可以更精确地预测疾病的发展。

实际案例：Wang 等人[24]提出的 MOGONET 方法即采用了中融合策略。通过使用 MOGONET 方法，预测不同的肿瘤亚型并识别关键的生物标记物。此研究的核心是利用来自 mRNA 表达数据、DNA 甲基化数据和 microRNA 表达数据来增强疾病亚型的分类准确性，并揭示可能的治疗靶点。MOGONET 使用图卷积网络（Graph Convolutional

Network, GCN）来处理多组学数据。在此框架中，每个节点通过聚合其邻居节点的信息来更新其特征表示。通过多层图卷积，MOGONET能够从局部到全局层面整合节点特征，提取与疾病相关的深层次生物学信号。经过图卷积层的特征提取后，使用全局池化操作（例如全局平均池化）来获取图的整体表示，这一表示反映了所有生物分子间复杂的相互作用和整体的生物学状态。随后，MOGONET 针对各组学数据的初始预测结果，构建跨组学发现张量，从而充分表征异质组学数据之间标签的相关性。最终，这些数据被输入到视图相关性发现网络（View Correlation Discovery Network, VCDN）方法中进行最终的多分类预测，从而得到癌症亚型分类预测结果。

6.6.3 后融合

后融合是指在数据分析的最后阶段整合来自不同组学的结果。在这种策略中，各种组学数据如基因组学、转录组学、蛋白质组学和代谢组学等首先独立进行深入分析，每种数据都输出自己的分析结果或生物标志物，然后再将这些结果进行综合分析或决策支持。后融合特别适用于各种数据需要不同专业工具或模型独立处理，而最终结果需要整合所有数据视角的场景。方法介绍：（1）结果级整合：后融合通过整合每种组学分析后的结果来提供全面的生物学见解。例如，可以独立分析基因表达数据和蛋白质组数据，每个分析可以识别不同的疾病生物标志物，然后通过统计方法如投票系统、决策树或元分析，将这些结果汇总来提出最终的疾病预测或分类。（2）网络融合：在网络融合策略中，独立分析的结果被用来构建网络模型，如基因调控网络或蛋白质互作网络。然后，这些网络基于其相互作用或功能联系被整合，以揭示跨组学层面的复杂生物过程或疾病机制。（3）模型集成：模型集成方法涉及将不同组学数据分析的预测模型或分类器的输出进行集成。这可以通过集成学习技术实现，如随机森林的集成、堆叠泛化或提升方法，以提高预测的准确性和鲁棒性。

优点：（1）专业性强：后融合允许每种组学数据使用最适合其特性的分析方法，保证了数据处理的专业性和深入性。（2）灵活性高：后融合提供了高度的灵活性，在数据分析的最后阶段整合结果，可以根据研究需要选择不同的整合策略、保证了各组学数据的最佳处理方式，最大限度地发掘了每种数据的独立价值。（3）综合性好：后融合在分析的最后阶段整合结果，能够从多个角度全面评估研究对象，提供更全面的生物学或疾病理解。

实际案例：一种采用后融合策略的多组学数据整合方法是MOFNet（Multi-Omics data Fusion Network）[25]，这种多组学融合策略，是通过整合 mRNA 表达、DNA 甲基化、microRNA 表达等多类型异质组学数据，对乳腺癌、低级别脑胶质瘤、胃腺癌三种发病率较高的癌症进行分型研究。MOFNet 主要由组学特异性学习方法 SGO（Similarity Graph pOoling with structure learning）与多组学数据整合的方法 VCDN 两个模块组成。MOFNet 仅用至多 25%的特征进行预测，便可超越当前已有的癌症分型方法的性能，实现癌症亚型的精准分类。

在组学特异性学习方面，MOFNet 使用了 SGO 方法。SGO 方法是一种基于改进图池化的冗余特征消除编码器，主要包含图卷积、图池化和图结构学习三部分。SGO 方法将图卷积神经网络作为组学特异性学习的基础模块。在此基础上，SGO 提出了一种改进的图池化网络，以实现特征维度的压缩，解决了传统图池化方法不能有效优化组学调控关系一致性的问题。同时，SGO 通过引入多图对应节点得分先汇聚再分发的机制，有效增强了组学调控关系的可解释性。此外，SGO 还引入了图结构学习方法 GSL（Graph Structure Learning）。该方法通过评估孤立节点与其他节点的相似度，将孤立结点与最佳节点相连，解决了孤立结点阻碍信息有效流动的问题。

在多组学数据整合方面，MOFNet 构建了一个跨组学发现张量，

并输入给 VCDN 以完成最终预测。该跨组学发现张量由异质组学数据的初步预测结果构成，反映了跨组学标签的相关性。VCDN 通过探索深层标签空间中异质组学数据类型之间的潜在相关性，有效整合来自每种组学数据的初始预测。通过这种方式，MOFNet 能够深入挖掘并整合各组学层次的互补信息，从而提高分类任务的精确度和可靠性。

另一种使用了后融合的方法是 iRANK 算法[14]，iRANK 算法通过先分别计算各层网络中的节点 PR 值，然后将这些不同层次的 PR 值进行融合，最终得到节点的 iPR 值，这一过程正是后融合的典型应用。iRANK 算法的输入为多组学数据和对应的多层分子网络，输出为节点的 iPR 值。iRANK 先计算每个单层网络的 PR 值，然后将多层网络的 PR 特征值进行融合，得到节点的 iPR 值。最后以 iPR 值作为特征衡量网络中节点的重要性。

6.7 临床中的应用

近年来，部分通过分析多组学数据发现的生物标志物已被用于临床实验，且取得较好效果。2002 年，Veer 等人开发了 Mammaprint 试剂盒，通过分析 70 个基因的表达情况，将 ER 阳性乳腺癌患者划分为高风险和低风险预后组，并指出部分高风险 ER 阳性患者可能从辅助化疗中获益[26]。然而，对于 ER 阴性乳腺癌患者，MammaPrint 大多将其评定为高风险，展现出在预后评估和治疗反应性预测方面的局限。2004 年，Paik 等人[27]开发了 Oncotype DX 试剂盒，利用 21 个关键基因的表达分析来对乳腺癌患者的复发风险进行三级分类，识别出对化疗有显著反应的高风险患者群体。EndoPredict 检测 11 个基因的表达评估乳腺癌患者在仅接受内分泌辅助治疗情况下的复发风险[28]。PAM50 通过定量检测 50 个基因的 mRNA 表达评估特定乳腺癌患者在手术后接受内分泌治疗的远处复发风险[29]。

6.8 总结

随着人工智能技术的不断进步，其在组学数据分析和生物标志物识别方面的应用潜力被越来越多地挖掘。尤其是在癌症诊断和治疗中，通过从高通量组学数据中筛选和识别生物标志物，人工智能正逐步展现出其独特的优势和广阔的应用前景。研究人员通过综合运用基于网络的方法、监督模块检测、特征选择技术以及机器学习模型，不仅能够更精确地识别与疾病相关的关键生物标志物，还能深入理解它们在疾病进程中的作用机制。尽管单组学分析取得了一定的成果，但每种单组学方法都存在各自的局限性。为了解决这些问题，多组学研究通过整合不同层面的分子信息，揭示了更为复杂的生物机制。这种整合分析不仅增强了特征筛选的能力，还提高了疾病预测的精度，对精准医疗具有重要作用。通过多组学数据的协同分析，研究人员可以获得更加全面和深入的生物学见解，从而推动疾病诊断、预后评估和个性化治疗的发展，最终实现精准医学的目标。

参考文献

[1] KUHN M. Caret: classification and regression training [J]. Astrophysics Source Code Library, 2015: ascl: 1505.003.

[2] KURSA M B, RUDNICKI W R. Feature selection with the Boruta package [J]. Journal of statistical software, 2010, 36: 1-13.

[3] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines [J]. Machine learning, 2002, 46: 389-422.

[4]李苓玉, 刘治平. 基于机器学习的自发性早产生物标记物发现 [J]. 南京大学学报, 2021, 57(5): 767-74.

[5] KURSA M B. Robustness of Random Forest-based gene selection methods [J]. BMC bioinformatics, 2014, 15: 1-8.

[6] FORTINO V, SCALA G, GRECO D. Feature set optimization in biomarker discovery from genome-scale data [J]. Bioinformatics, 2020, 36(11): 3393-400.

[7] FENG X, ZHANG Y, DU M, et al. Identification of diagnostic biomarkers and therapeutic targets in peripheral immune landscape from coronary artery disease [J]. Journal of Translational Medicine, 2022, 20(1): 399.

[8] HUANG H-H, LIU X-Y, LIANG Y. Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2+ 2 regularization [J]. PloS one, 2016, 11(5): e0149675.

[9] DíAZ-URIARTE R, ALVAREZ DE ANDRéS S. Gene selection and classification of microarray data using random forest [J]. BMC bioinformatics, 2006, 7: 1-13.

[10] LANGFELDER P, HORVATH S. WGCNA: an R package for weighted correlation network analysis [J]. BMC bioinformatics, 2008, 9:

1-13.

[11] LI L, LIU Z-P. Biomarker discovery from high-throughput data by connected network-constrained support vector machine [J]. Expert Systems with Applications, 2023, 226: 120179.

[12] ZHANG Z, SUN C, LIU Z-P. Discovering biomarkers of hepatocellular carcinoma from single-cell RNA sequencing data by cooperative games on gene regulatory network [J]. Journal of Computational Science, 2022, 65: 101881.

[13] WANG Y, LIU Z-P. Identifying biomarkers for breast cancer by gene regulatory network rewiring [J]. BMC bioinformatics, 2022, 22(Suppl 12): 308.

[14] SHANG H, LIU Z-P. Network-based prioritization of cancer biomarkers by phenotype-driven module detection and ranking [J]. Computational Structural Biotechnology Journal, 2022, 20: 206-17.

[15] JIANG Y-Z, MA D, JIN X, et al. Integrated multiomic profiling of breast cancer in the Chinese population reveals patient stratification and therapeutic vulnerabilities [J]. Nature Cancer, 2024, 5(4): 673-90.

[16] GüNTHER O P, CHEN V, FREUE G C, et al. A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers [J]. BMC bioinformatics, 2012, 13: 1-18.

[17] HUANG Z, ZHAN X, XIANG S, et al. SALMON: survival analysis learning with multi-omics neural networks on breast cancer [J]. Frontiers in genetics, 2019, 10: 166.

[18] KIM D, LI R, DUDEK S M, et al. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network [J]. BioData mining, 2013, 6: 1-14.

[19] SINGH A, SHANNON C P, GAUTIER B, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays [J]. Bioinformatics, 2019, 35(17): 3055-62.

[20] SUN Y, GOODISON S, LI J, et al. Improved breast cancer prognosis through the combination of clinical and genetic markers [J]. Bioinformatics, 2007, 23(1): 30-7.

[21] VAN DE WIEL M A, LIEN T G, VERLAAT W, et al. Better prediction by use of co-data: adaptive group-regularized ridge regression [J]. Statistics in medicine, 2016, 35(3): 368-81.

[22] WANG B, MEZLINI A M, DEMIR F, et al. Similarity network fusion for aggregating data types on a genomic scale [J]. Nature methods, 2014, 11(3): 333-7.

[23] LIU Y, DEVESCOVI V, CHEN S, et al. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties [J]. BMC systems biology, 2013, 7: 1-13.

[24] WANG T, SHAO W, HUANG Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification [J]. Nature communications, 2021, 12(1): 3445.

[25] ZHANG C, LI P, SUN D, et al. MOFNet: A Deep Learning Framework of Integrating Multi-omics Data for Breast Cancer Diagnosis; proceedings of the International Conference on Intelligent Computing, F, 2023 [C]. Springer.

[26] KNAUER M, MOOK S, RUTGERS E J, et al. The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer [J]. Breast cancer research treatment, 2010, 120: 655-61.

[27] PAIK S, SHAK S, TANG G, et al. A multigene assay to predict

recurrence of tamoxifen-treated, node-negative breast cancer [J]. New England Journal of Medicine, 2004, 351(27): 2817-26.

[28] MüLLER B M, KEIL E, LEHMANN A, et al. The EndoPredict gene-expression assay in clinical practice-performance and impact on clinical decisions [J]. PloS one, 2013, 8(6): e68252.

[29] MA X-J, SALUNGA R, DAHIYA S, et al. A five-gene molecular grade index and HOXB13: IL17BR are complementary prognostic factors in early stage breast cancer [J]. Clinical cancer research, 2008, 14(9): 2601-8.

第7章 蛋白质语言大模型的前沿探索和展望

7.1 从通用语言大模型到蛋白质语言大模型

近年来，自然语言处理（NLP）领域取得了显著进展，尤其是基于 Transformer 架构的自然语言大模型，如 BERT 和 GPT-3 等。这些模型通过在大规模文本数据上进行预训练，能够捕捉语言中的复杂模式和关系，从而在各种语言任务中表现出色。它们的成功不仅在于强大的计算能力和深度学习算法，还在于自监督学习方法的应用，如掩码语言模型（MLM）和自回归模型（ARM），使得模型能够在大规模无标签数据上学习，捕捉到语言中的深层次模式和关系，从而为理解和生成语言提供了强大的基础。

自然语言大模型的核心是 Transformer 架构，它通过自注意力机制和并行计算，能够高效处理长序列数据。海量数据的预训练使得表征具有很强的泛化能力，微调策略进一步增强了模型的在不同任务数据上的表现。

然而，语言模型的应用不仅限于自然语言处理。从中心法则（The central dogma of molecular biology[18]）出发，我们认为生命过程中有其独特的语言模式，从 DNA 转录到 RNA，再翻译成蛋白，而蛋白作为细胞对外的通信单元，与其他的蛋白进行相互作用，从而在生命体重发挥作用。可以说蛋白质是其中对话的语言。随着生物信息学和计算生物学的发展，研究人员开始将自然语言中先进的语言模型技术应用于蛋白质序列的分析和预测。蛋白质语言模型（Protein Language Models, PLMs）应运而生，旨在通过学习大量蛋白质序列数据，捕捉蛋白质中的复杂模式和关系，从而在蛋白质结构预测、功能预测和设计等任务中发挥作用。

蛋白质序列由 20 种氨基酸组成，其序列和结构信息对生物功能有重要影响。那么以氨基酸为基础 token，形成的蛋白序列就很像我们在自然语言中的句子。蛋白质之间的相互作用，那就形成了段落或

者说是对话（如图 7-1）。

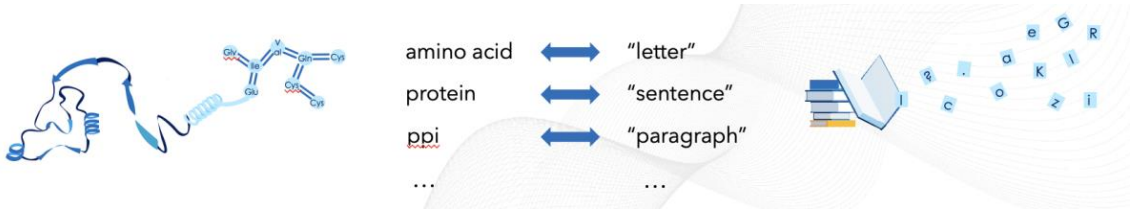


图 7-1 将蛋白质的组成和相互作用类比自然语言

然而，蛋白质语言又有他独特的地方，自然语言中有比较明确的短语，而且多数具有局部连续的特性。而蛋白质序列中的氨基酸更像是单字，虽然具有功能性的片段或者 domain 可以认为是短语，但是这些短语的定义比较模糊，不像自然语言有固定的词库。另外蛋白的折叠也使得在序列中不连续的氨基酸组成了特定结构或者功能片段。这些都让蛋白质序列中的语义建模比较自然语言场景更有挑战[1]。

同时，蛋白质数据不仅包括序列信息，还包括结构、功能等多模态信息。因此，蛋白质语言模型也可以结合多尺度建模、多模态数据整合和生物学知识，以提高模型的准确性和功能性。

总的来说，从自然语言大模型到蛋白质语言模型的演变，既展示了语言模型技术在不同领域的广泛应用和巨大潜力，也需要我们能够考虑到蛋白质氨基酸序列的特点，通过结合自然语言处理的先进技术和生物信息学的特殊需求，这样才能够在生物应用领域的任务上有所进展。

7.2 蛋白质语言大模型的前沿探索与尝试

7.2.1 数据的来源和整理

进行蛋白质语言模型的预训练最基础的就是单序列数据，蛋白质序列数据的规模可以非常庞大。例如，UniProt[8]数据库包含数亿条蛋白质的序列信息，以及丰富的蛋白家族标签信息，而 ColabFoldDB[17]

中更是拥有数以十亿计的宏基因组的数据，记录了种类更为丰富的蛋白质序列数据。在基于这类数据进行预训练的时候，通常会对序列进行聚类，并按照聚类的类别进行训练时的数据采样，保证学习的过程中序列的多样性足够高，从而能够让训练比较平稳的进行。

当我们对标自然语言中的以对话或者以段落作为输入的预训练场景时，发现蛋白语言预训练也可以考虑使用蛋白质对话的数据，比较直接的数据就是蛋白质的相互作用数据，也就是 PPI（protein-protein interaction）数据，例如 STRING 数据库中就收录了超过 14000 个物种、6 千多万种蛋白、200 多亿个相互作用的信息。这些数据可以为预训练提供更加丰富的数据上下文。但 STRING 中的数据也有其自身的缺陷，比如有大量的由规则推断的相互作用，其正确性比较难以保证，推断的基于往往是基于家族信息，这样使得这些相互作用的蛋白对相对比较聚集，这样就更加需要利用聚类去重，规则过滤等手段来清洗出利于预训练学习的样本。

除开纯序列的数据之外，也会利用其他模态（特别是结构数据）来进行预训练学习的，这里主要先重点阐述基于氨基酸序列的语言大模型，其他的预训练模型在后面展望的环节会加以阐述。

7.2.2 训练范式

目前进行预训练语言模型在应用中使用的场景主要是两类，一类是需要对给定蛋白序列进行某种属性判断或者打分，这类主要在于提取最合适蛋白序列表征特征。另一类是需要生成满足特定条件的蛋白序列，按照顺序生成组成蛋白的氨基酸序列。因此这两类的场景对应的训练目标是不一样的，参照下图可以总结为：

掩码语言模型（MLM）目标：这项任务涉及预测序列中被随机掩码的标记。这些标记由特殊标记[MASK]指示。这项任务与 ESM2[6]，proTrans[3],ankh[4]等的训练方法一致，专注于双向上下文理解。

广义语言模型（GLM）目标：这项任务涉及预测序列中的后续

标记，包括短掩码跨度（由[sMASK]指示）和序列末尾的较长跨度（由[gMASK]标记）。GLM 目标考虑了预测后续单词的单向上下文，前缀编码部分仍然是双向的。典型的代表是 proGen[5]的工作。

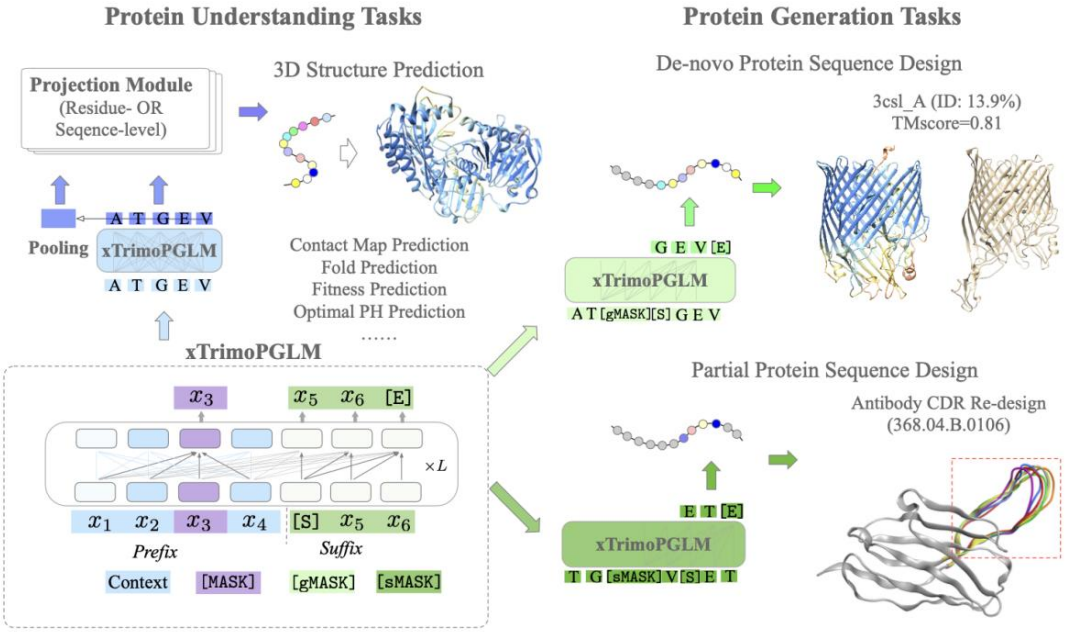


图 7-2 在 xTrimoPGLM 中融合了两类自然语言预训练的任务，即掩码语言模型（MLM）和广义语言模型（GLM）

训练范式结合了偏向于语义理解的 mask language model 以及偏向于生成的自回归模式，从而服务于蛋白的属性理解优化任务和蛋白生成设计任务。

在训练的模式方面，百图生科研发的 xTrimoPGLM，受到课程学习思路的启发，预训练阶段分为两个不同的阶段：

初始预训练：使用 MLM 目标，专注于在大约 4000 亿个标记上快速最小化损失。这个阶段旨在增强模型的理解能力。

后续训练：采用统一的方法，将 MLM 和 GLM 目标以特定比例（20% MLM，80% GLM）合并。这个阶段使用额外的 6000 亿个标记，致力于完善模型的表征和生成能力。

通过这样的学习过程，有效地结合了两方面训练目标的逻辑，兼

顾了不同的应用场景。也让 xTrimoPGLM 模型能够在应用落地中具有更多的优势。近些年有很多的基于序列的蛋白语言模型被以各种形式发表出来，这里也将这些工作在模型训练的各方面的信息做一个对比，如下图，这里可以看到，目前为止，由百图生科和清华一起研发的 xTrimoPGLM 模型是参数规模最大，同时也是训练数据最丰富的蛋白质序列预训练语言模型。

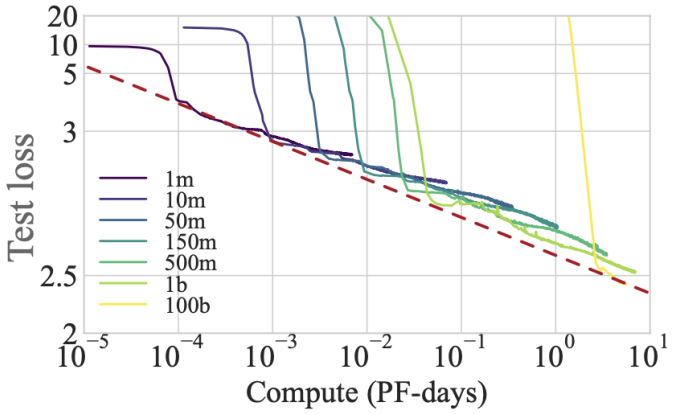
	Pre-trained Models				
Model	Parameters	Layers	Datasets	Embedding Dim	Machine Learning Tech
xTrimoPGLM-100B	BioMap **~100B**	**72**	**Uniref+ Colabfold DB**	**10240**	**Mixture Obj Masked Language Modeling + Causal Language modeling**
ESM-15B	Meta 15B	48	UR50/S 2018_03	5120	Masked Language Modeling
ESM2(3B)	Meta 3B	36	UR50/D 2021_04	2560	Masked Language Modeling
ESM2(650M)	Meta 650M	33	UR50/D 2021_04	1280	Masked Language Modeling
ESM2(150M)	Meta 150M	30	UR50/D 2021_04	640	Masked Language Modeling
OmegaPLM	Helixon 670M	66(head=1)	Uniref50 2021_04	1280	Masked Language Modeling
ProtBERT	aws 420M	/	/	/	Masked Language Modeling
Prot-T5-XL-Ur50 (3B)	NVIDIA 3B	/	/	/	Masked Language Modeling
ProGen	salesforce 1.2B	36	280M	1028	Casual Language Modeling

图 7-3 各类蛋白质语言模型的不同参数对比

7.2.3 蛋白质语言模型的 Scaling Law

当我们讨论大预言模型的时候，Scaling Law（尺度定律）往往被认为是模型是否可以继续发展和成功的保障，在数据保证以及算力保证的情况下，加大模型参数规模在对未知新数据的理解上能够获得不断提升的效果。参考下图，在 xtrimoPGLM 的训练中，当我们将模型的规模从 1M 不断提升到 100B 的过程中 L 的下降与计算资源之间的关系呈现 power-law 的分布。这也说明了模型规模不断的扩大的过程

中，模型对数据的学习能还在呈现上升的趋势。



$$--- \; L = 2.63 * C^{-0.017}$$

$$\text{One PF-day} = 8.64 \times 10^{19}(\text{FLOPs})$$

图 7-4 xTrimoPGLM 训练过程中发现的尺度定律（scaling law）

而另一方面，当用新的测试数据集来验证训练后的语言模型的预测能力的时候，也可以看到拥有 100B 参数的 xTrimoGPLM 模型的对于每个 token 的 PPL 得出的判断效果要明显由于参数规模更小的 ESM2(15B)和 proGen2（6.4B）。这也从侧面证明了蛋白语言模型中也存在着 scaling law，并且目前模型的规模下仍然有提升的空间，当我们可以有更丰富的蛋白数据的情况，更大规模的参数或许可以得到更加多有意义的表征层特征。

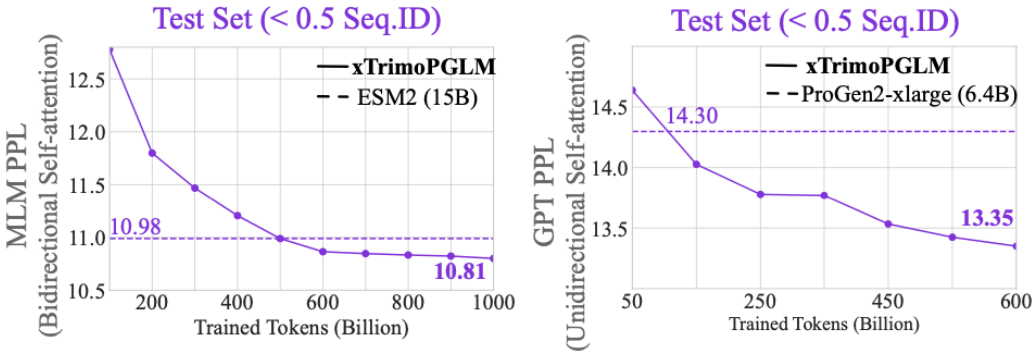


图 7-5 对比 xTrimoPGLM 与其他预训练模型在 PPL 上的表现，xTrimoPGLM 凭借更大的模型规模获得了更低的 PPL

测试集和挑选了 10000 个蛋白序列，并保证这部分蛋白与训练集中蛋白序列相似性低于 0.5. MLM PPL 可以理解为当前位置在多少个氨基酸的范围内选择有不确定性。从 random 角度来看，每个位置的氨基酸选择都有 20 种，所以随机情况下 PPL 应为 20。

7.2.4 语言模型应用落地

7.2.4.1 基于蛋白理解的一系列的 Benchmark 任务

前面提到蛋白质预训练语言模型在很大程度上需要解决蛋白质理解的问题，经典的蛋白应用相关的场景大多都落在蛋白质理解的范畴中，下图中给出了四大类不同的蛋白相关的应用任务，包括蛋白结构预测，蛋白功能预测，蛋白相互作用预测，以及蛋白质属性预测。在这些应用中，我们用预训练语言模型拿到的表征，来测试在这些应用任务上的能力。

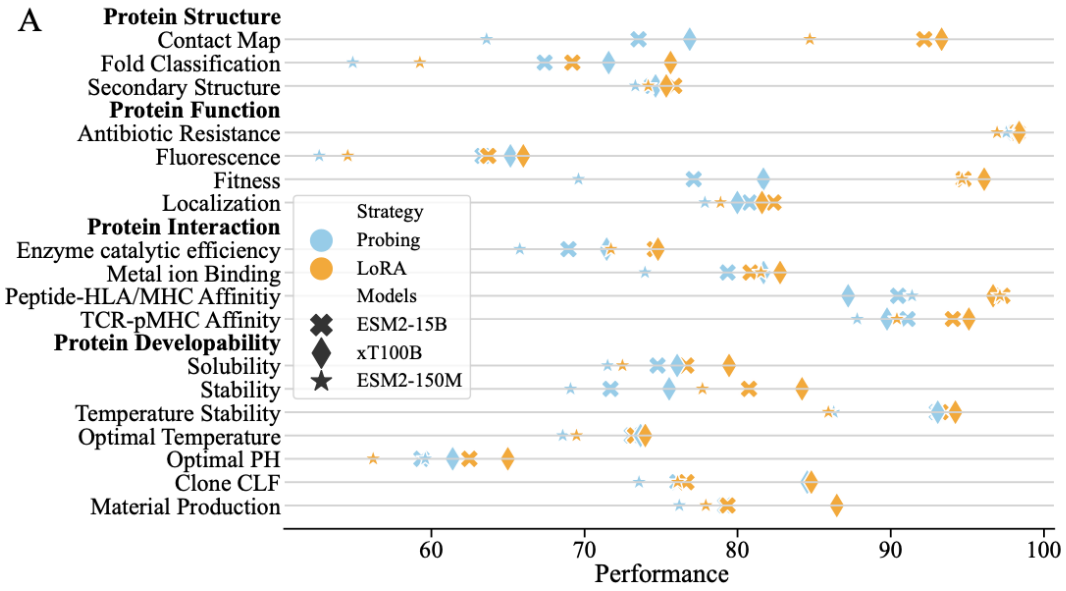


图 7-6 xTrimoPGLM 在一系列蛋白相关的计算预测任务重获得了更好的表现，优于 ESM2-15B，同时微调模型会获得更大的提升。

这是典型的基础大模型在下游场景的应用，从常规做法来说，有固定大模型参数的 Probing 的使用方式，也有利用大模型和应用数据进行微调（finetuning）的使用方式，由于模型参数规模比较大，微调

的方案选择了 LoRA[9]的方法。通过测试，对比不同大小的蛋白质序列预训练模型，更大的模型通过微调的方式会获得很大的效果提升。这说明结合大模型和应用场景的数据，可以获得出色的应用算法能力的提升。

7.2.4.2 蛋白语言模型辅助结构预测

蛋白结构预测任务一直是生命科学领域皇冠上的明珠，由于结构决定了功能，所以当我们可以利用 AI 将结构有效的解析出来时，后续很多的工作都可以找到更好的解决方案。在 ESM 和 xTrimoPGLM 的研究中可以看到，当我们拥有了基于蛋白质语言模型的表征，可以去除 MSA（同源序列对齐）的依赖。预训练大模型从大量的蛋白序列中学习到了氨基酸序列中各种关联信息，其中也包含了同源序列。百图 xtrimoPGLM 采用了这样的策略（如图 7-7），从而在单体蛋白以及蛋白质复合物的结构预测的速度达到了 AF2 的数十倍。同时，也为一些 MSA 缺失的蛋白序列的结构预测带来了增益，特别像抗体蛋白这类突变比较多的蛋白，不论是抗体单体的结构预测能力，还是抗原抗体复合物的结构预测，相比于 AF2 都得到不错的提升。

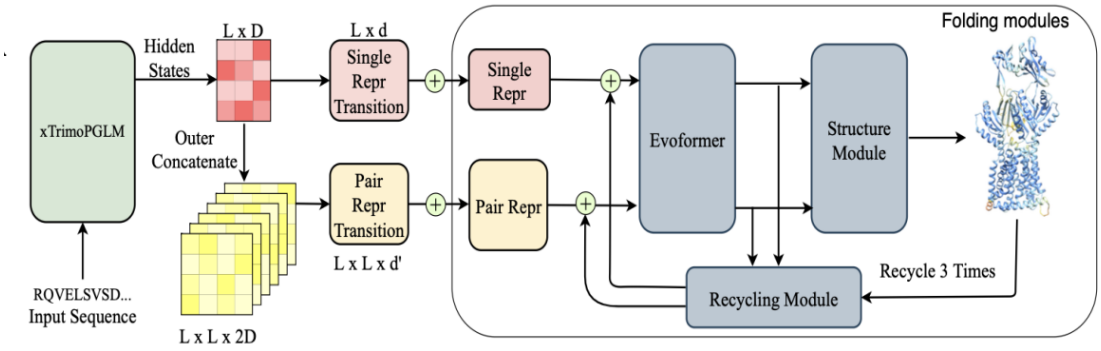


图 7-7 借助 xTrimoPGLM 底座，帮助蛋白质结构预测模型提升整体效率和效果

7.3 对于蛋白质语言模型以及 AI 进行蛋白质设计的展望

7.3.1 多模态融合的蛋白质预训练

前面我们讲了很多基于氨基酸序列的蛋白质预训练语言模型，从某种程度上说序列是当前最容易活动也最值得信赖的信息，但我们也

知道 3D 结构才是更加接近蛋白功能的表征，因此当前很多蛋白质相关的预训练的工作都会利用包括电镜解析的真实 3D 结构或者 AlphaFold2 等算法预测的蛋白结构数据来提取结构模态的信息。

从结构模态的角度来进行预训练，可以捕捉包括主链骨架结构（键长、键角），侧链转角，结构表面形态（法向量）等诸多方面的特征，以及在这些方面特征之间的高阶关联信息，在这个角度也有不少有代表性的工作，包括 GearNet[14]，Masif[15]，PIsToN[16]等等。这些模态的预训练一方面可以得到更加高效的结构表征，从而为蛋白结构生成方面提供基础，另一方面，也能够通过融入序列和结构的对齐数据，实现从结构到序列的生成，如 proteinMPNN[11]。这使得一些序列设计的任务可以综合两方面的信息也就是已知的一些结构的信息和序列信息，来共同控制位置序列的生成，使设计更加合理和可用，如 LM-DESIGN[12], DyMean[13], ESM3[10]。

除蛋白质序列和结构角度，我们可以再扩宽一些思路，在蛋白质序列信息和我们可以更大规模获取的细胞的基因表征之间进行融合学习。当我们把每个蛋白质的基本表征代入到以一定的基因或蛋白表达为数据形态的单细胞测序数据的预训练中时，就可以让单细胞数据中的基因间的联系来帮助我们调节蛋白的氨基酸序列表征。这也是百图生科提出 xTrimo 系列多层次预训练概念时想要做到的（如下图）。

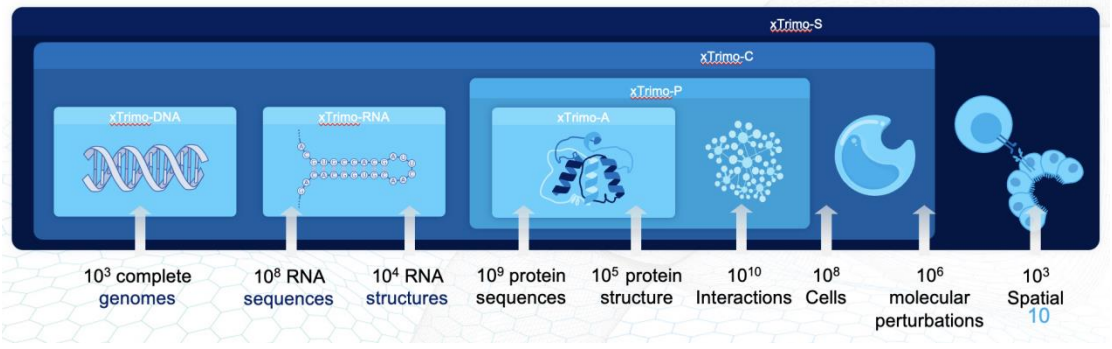


图 7-8 不同尺度的蛋白相关的数据，及其数据规模。从单体的角度看有 DNA，RNA，蛋白质的数据，进而在细胞层面有蛋白质相互作用，以及在更大的系统层面包含了细胞之间的相互作用。

7.3.2 对数据的期待

数据是蛋白质语言模型走向更加通用的基础，前面提到很多数据带来的局限性，那么如果我们展望数据方面的发展，什么样的数据能够更好的帮助蛋白质语言模型发展呢？

首先，反映物理规律的数据，如精准力场以及精准力场下的动力学模拟，这类能够反应第一性定律，可以用计算能力换去海量数据样本，一方面能够然后让 AI 来学习数据其中不能显式计算的能量规律，从而可以让 AI 模型某种程度具有还原第一性原理的能力，从而完成进一步的推断。另一方面，补充当前蛋白结构数据不完善的问题，因为从 PDB 的角度来看大部分的结构信息只是蛋白结构的一个采样，让物理规律帮助我们采样更加完整的数据，进而让 AI 能够表征更加完整的蛋白质结构分布的数据，让模型表达能力本身更加强。

其次，除开获取价格高昂的结构数据，更多高通量的实验数据，比如蛋白属性测量，蛋白-蛋白相互作用的数据实际是可以更加同时也更加经济的帮我们探索足够多样的蛋白空间的（比如 Ribosome Display 等）。我们期待着有更好的实验技术，能够让实验产生更多高质量的测量数据，为蛋白质表征的学习带来切实的指导信息。

7.3.3 语言模型与 AI 蛋白质设计的思路

蛋白质设计是一个多目标优化的问题，通常我们需要满足从结构功能，到序列合理性（如表达能力）、蛋白属性（如稳定性、多聚、免疫原性等）等多方面的需求。基于语言模型进行蛋白质的生成式设计是其中一种蛋白设计的手段。

生成式模型的训练方式决定了他们可以满足的设计要求，换句话说，每一个生成式的模型所具有的知识是特定模态下来的数据分布，当训练样本足够的情况下，数据分布完整且精细，那么可以根据条件进行分布抽样的能力就会比较好，进而达到设计的需求。在我们前面提到的基于不同模态的预训练、生成式预训练中，所用到的数据包括

蛋白质序列数据，蛋白质 3D 结构数据等，蛋白质序列的数据总量虽然很大，但是由于自然界有太多的蛋白家族，他们都具有自己特殊的序列片段，在这类特定的特征空间下的数据量往往是比较小的，能否建立完整的特征分布是存在问号的。同样的，面对仅仅存在的几十万实验测定的蛋白结构数据，我们想建立一个比较完善的结构特征+序列特征的数据分布刻画，是很困难的。

数据不足的问题困扰着大家向更加通用的设计方法发展，但是同时也启发了大家，从知识经验和特定种类数据的角度来思考，更高效的融合入设计模型中，减少对数据的需求。比如，profluent 公司在进行 crispr 酶设计的时候，就大量了结合了结构生物学的知识，对于酶蛋白中功能区域进行保守设计，而将蛋白质语言模型用来对非保守部分进行生成式处理，提升其设计的多样性。在减少模型压力的同时，也获得了很好的应用能力。

因此，在当前的数据场景下，预训练的模型的成功需要从两个方面寻找机会，一方面我们等待数据的，另一方面也是积极寻找融入少量知识的方案，起到事半功倍的效果。

参考文献

[1] Ofer, Dan, Nadav Brandes, and Michal Linial. "The language of proteins: NLP, machine learning & protein sequences." Computational and Structural Biotechnology Journal 19 (2021): 1750-1758.

[2] Chen, Bo, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei et al. "xTrimoPGLM: unified 100B-scale pre-trained transformer for deciphering the language of protein." arXiv preprint arXiv:2401.06199 (2024).

[3] Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs et al. "Prottrans: Toward understanding the language of life through self-supervised learning." IEEE transactions on pattern analysis and machine intelligence 44, no. 10 (2021): 7112-7127.

[4] Elnaggar, Ahmed, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. "Ankh: Optimized protein language model unlocks general-purpose modelling." arXiv preprint arXiv:2301.06568 (2023).

[5] Nijkamp, Erik, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. "Progen2: exploring the boundaries of protein language models." Cell systems 14, no. 11 (2023): 968-978.

[6] Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa et al. "Language models of protein sequences at the scale of evolution enable accurate structure prediction." BioRxiv 2022 (2022): 500902.

[7] Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." arXiv

preprint arXiv:2001.08361 (2020).

[8] Apweiler, Rolf, et al. "UniProt: the universal protein knowledgebase." Nucleic acids research 45, no. D1 (2017): D158-D169.

[9] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

[10] Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science 379, no. 6637 (2023): 1123-1130.

[11] Dauparas, Justas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile IM Wicky et al. "Robust deep learning–based protein sequence design using ProteinMPNN." Science 378, no. 6615 (2022): 49-56.

[12] Kong, Xiangzhe, Wenbing Huang, and Yang Liu. "End-to-end full-atom antibody design." arXiv preprint arXiv:2302.00203 (2023).

[13] Zheng, Zaixiang, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. "Structure-informed language models are protein designers." In International conference on machine learning, pp. 42317-42338. PMLR, 2023.

[14] Zhang, Zuobai, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. "Protein representation learning by geometric structure pretraining." arXiv preprint arXiv:2203.06125 (2022).

[15] Gainza, Pablo, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael M. Bronstein, and Bruno E. Correia.

"Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning." Nature Methods 17, no. 2 (2020): 184-192.

[16] Stebliankin, Vitalii, Azam Shirali, Prabin Baral, Jimeng Shi, Prem Chapagain, Kalai Mathee, and Giri Narasimhan. "Evaluating protein binding interfaces with transformer networks." Nature Machine Intelligence 5, no. 9 (2023): 1042-1053.

[17] Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. "ColabFold: making protein folding accessible to all." Nature methods 19, no. 6 (2022): 679-682.

[18] Crick, Francis. "Central dogma of molecular biology." Nature 227, no. 5258 (1970): 561-563..

第 8 章 人工智能基因调控

8.1 基因调控概述

基因调控是指生物体内控制基因表达的机制，基本内容是调控蛋白与其靶 DNA 或 RNA 分子之间的相互作用，包括发生于 DNA 水平上的调控、转录控制和翻译控制，微生物水平调控和多细胞生物的基因调控三种水平。微生物通过基因调控改变其代谢的方式以适应周围环境的变化，这类调控是短暂和可逆的。多细胞生物的基因调控是细胞分化、形态变化、个体发育的基础，这种调控一般是长期、不可逆的。

调控基因组学研究具有重要的生物学意义。基因调控可以调节微生物的氨基酸、核苷酸之类物质的合成，受调控的菌种应用于发酵工业，可使产量大幅增加，例如植物苯丙氨酸解氨酶（PLA）应用于农作物生产。随着生产、生活环境的需要，植物抗病育种、生物活性成分成为现代生物技术焦点之一，能够提高植物抗性，减少经济损失和农药对人类健康及生态环境的影响，另外，其次生代谢产物富集在疾病的治疗和医疗方面具有不可替代的作用。除此之外，基因表达调控在环境保护、食品工艺、养殖等很多行业中也有重要的应用。

细胞调节基因转录以协调细胞活动对细胞内和细胞外信号的反应。转录是主要受转录因子（TF）调节，转录因子是结合靶基因并且可以对靶基因的转录速率产生积极或消极的影响。基因组 DNA 与结构蛋白紧密结合形成称为核小体的复合物，核小体是染色质的基本单位，使得大多数基因无法被转录机制访问。基因转录起始位点附近的区域被称为启动子，需要通过紧密置换来暴露堆积的核小体。通过先锋 TF 的结合，可以触发 DNA 可及性的变化。其他 TF 可以结合到远端 DNA 的顺式调控元件（CRE）与辅因子和其他蛋白质协同作用，用于调节和稳定 RNA 聚合酶-蛋白质复合物。染色质、转录因子和基因之间的相互作用产生了复杂的调控回路，可以表示为基因调控网络

GRN。GRN 的研究有助于了解细胞身份在疾病中是如何建立、维持和破坏的。GRN 可以从实验数据（历史上的大量组学数据）和文献中推断出来。单细胞多组学技术的出现导致了新的 AI 计算方法的发展，这些方法利用基因组、转录组和染色质可及性信息以前所未有的准确率推断 GRN。

8.2 基序检测的人工智能算法

转录因子（TF）是关键的调节蛋白，通过结合转录因子结合位点（TFBS）或基序的短 DNA 序列来控制细胞的转录速率。识别和表征 TFBS 对于理解控制细胞转录状态的调控机制至关重要。基序（motif）是基因组中重复出现的序列模式，是转录因子的结合位点，对调节细胞内蛋白质的产生至关重要，基序分析对促进医学治疗和理解细胞过程具有重要意义[1]。识别基因组中转录因子结合位点或基序是破译基因调控机制的关键之一。在过去的几十年里，已经开发了几种实验方法来恢复含有 TFBS 的 DNA 序列。

转录因子结合特定的 DNA 序列，并控制 DNA 转录成 mRNA。转录因子结合位点可以通过几种高通量检测方法来测定，包括 PBM、SELEX、ChIP-和 CLIP-seq 技术。与此同时，已经提出了基于这些 DNA 序列发现和鉴定 TFBS 基序的计算方法。这是生物信息学中研究最广泛的问题之一，被称为基序发现问题。近年来，基序识别算法主要分为基于统计策略和基于人工智能学习两大类。基于统计的模型仍有较大的局限性，特别是对于识别序列组成和二级（或三级）结构组成的 RNA 结合蛋白。深度学习非常适用于基因组学，多个学习层可以在神经元内捕获多个级别的处理和抽象信息，DeepBind 和 DeepFinder 是其中两种应用较广的基序识别算法。Alipanahi 等人[2]将来自深度学习的最先进技术融入到 DeepBind 的开发中。该方法通过两个步骤预测蛋白质与 DNA 或 RNA 序列的结合亲和力，包括应

用卷积模块进行表征学习和应用特征组合预测模块。Deepbind 从原始序列中捕获结合特异性，通过发现新的序列 motif 并将他们组合起来预测绑定分数。图形处理单元（GPU）用于自动训练高质量模型，不依赖于专家调优。DeepBind 的优点是能够自动选择模型参数和复杂度，深度学习在训练方式上的进步使得 DeepBind 更加实用。DeepBind 对挖掘更大的数据集非常有用，如 ENCODE 和 Roadmap Epigenomics。

Lee 等人[3]提出的 DeepFinder 使用具有与结合位点相关特征的深度学习神经网络来构建基序识别模型，其采用一种改进的三阶段 DNA 基序预测方法，这种方法具有两个新特征：第一，采用一组基序发现工具，用于从输入序列子集中初步预测候选结合位点。第二，提取与最具潜力的候选结合位点相关的特征，用于深度神经网络学习。DeepFinder 计算框架有三个连续的步骤：(1) 数据集划分为五个非重叠子集。(2) 四种新的基序发现工具应用于其中一个分区子集，以预测假定的基序和各自的结合位点，使用聚类算法对每个工具返回的前三个基元进行合并和分割，提取并集中与候选绑定位点相关的 76 个特征，并将其用于堆叠式自编码神经网络学习。(3) 学习神经网络用于预测相关的结合位点。DeepLIFT 将每个神经元的激活与参考值进行比较，选择哪些内核对 TFBS 定义贡献最大，从而减少基序冗余[4]。TF MoDISco[5]通过使用分配给内核的重要性得分对发现的基序进行聚类和聚合，扩展了这一想法。然而，在不丢失 DNN 学习到的一些信息的情况下构建可解释模型仍然是一个悬而未决的挑战。

TF-DNA 相互作用不仅涉及 TF 和 DNA 之间的直接结合，还涉及多个结合亚区（长期相互作用）和具有 TF 高阶结构的核苷酸（短期相互作用）之间的相互作用。长短期记忆网络（LSTM）和双向 LSTM（BLSTMs）可以有效地捕获序列信号的长期和短期依赖性。LSTM 和 BLSTMs 非常适合模拟 TF-DNA 相互作用，因为基因组序列可以被视为具有长期和短期依赖性的序列信号。深度神经网络在仅从

DNA 序列预测顺式调控活性的基因组训练模型方面也取得了一定的突破，已被证明足够复杂和通用，可以捕获顺式调节的参数，比如Sei[6]等算法。更精细分辨率的功能基因组学检测的发展，如染色质免疫沉淀，使我们能够学习更详细的模型，甚至捕捉到基因与其他转录因子之间的相互作用，这些相互作用遵循与 DNA 螺旋周期一致的周期依赖性。Enformer 模型经过训练，可以预测 200kb 序列中的数千个染色质谱，也被证明在预测扰动方面非常有用[7]。此外，在人类基因组序列上训练的几个神经网络模型表明，它们可以预测变异如何影响基因表达[8]。然而，这些模型对于单个变异通常是不可靠的，对于启动子比增强子变异更可靠。尽管这些模型捕获了有助于基因表达的主要特征，并且对许多应用非常有用，但它们与可靠检测所有较弱效应和可靠预测变异功能所需的模型还有一定距离。鉴于捕获顺式调控所需的大量参数，建立可靠的定量模型将需要更多的数据。

8.3 基因调控网络构建的人工智能算法

基因调控网络（GRN）自基因表达数据产生以来，一直是生物信息学研究的热点。利用计算方法揭示基因表达调控的复杂结构是近几十年来出现的一项具有挑战性的任务。识别基因的相互作用有助于理解 GRN 的拓扑结构和每个基因的作用，这对研究细胞在环境中行为背后的复杂机制至关重要[9]。

根据基因表达构建基因调控网络一直是研究人员关注的焦点。在估计基因表达的因果关系时，加权基因共表达网络分析（WGCNA）是最简单、最受欢迎的方法[10]。它计算整个转录组成对的相关性，以识别共表达基因。由此产生的网络通常被称为基因共表达网络。由于相关性的对称性，基因共表达网络及其相互作用是无方向的，缺乏对基因模块因果关系的估计，增加了潜在的假阳性发现。当前几种主要的构建基因调控网络的方法框架包括无监督、半监督和监督学习。无

监督模型主要分为布尔模型，贝叶斯模型，微分方程模型和信息理论模型四种类别。微分方程是最广泛使用的一类动力学模型，Jiguo 等人[11]提出的微分方差模型考虑了代谢物浓度随时间的变化。半监督学习有仅从样本中学习和从正样本和未标记数据中学习两种方法。Patel 等人[12]提出了一种基于随机森林和 SVM 的迭代方法，通过自训练来预测每个转录因子的调节。Jisha 等人[13]使用聚类方法从未标记的数据中提取可靠的反例， 提出的监督学习主要有 SVM 和深度神经网络。监督学习需要基因表达数据和已知的基因间调控，但比无监督和半监督方法更准确。Mordelet 等人[14]将基因调控网络推断的问题分解为大量的二元分类问题，每个子问题都与转录因子相关联，而 SVM 用于预测 GRN。compareSVM[15]可以用来比较线性，高斯， Sigmoid 和多项式内核四种 SVM 内核函数，其包括优化，比较和预测三个步骤。深度神经网络是一个受大脑神经网络启发的强大模型，其性能在广泛的应用中已经大幅度的提升。Mandal 等人[16]利用杂交杜鹃搜索花传粉算法（FPA）训练一个递归神经网络来选择基因的最佳组合。

为了获得基因之间的调控关系，多个模态的组学信息被用来进行整合估计。GENIE3[17]及 GRNBoost2[18]等方法首先根据以前报道的调控机制将调控因子 TF 与靶基因区分开来，然后训练预测靶基因的表达，这显著减少了要考虑的相互作用的数量。通过这样做，无向交互被转化为有向连接，从而引入了假定的因果关系。由于编码 TF 的 mRNA 转录物需要许多过程才能变成作为一种功能性蛋白质，仅靠转录水平可能无法提供足够的信息进行推理。因此，Buenrostro 等人[19]使用染色质可及性数据推断 TF 可能靶向的基因调控元件。染色质可及性数据将 GRN 推断分为两个步骤：第一，将转录因子分配给基因调控元件，即开放染色质地区；第二，分配这些调控元件对基因的影响。此类推理方法的一些示例包括 ATAC2GRN[20]、LISA[21]以及 SPIDER[22]。

随着单细胞 RNA 测序技术的出现，GRN 方法已被用于推断细胞类型特异性 TF-基因相互作用，以及这些 GRN 中发生的动态变化发展[23]。Aibar 等人[24]针对 scRNA-seq 数据提出了量身定制的方法 SCENIC，是 GRNBoost2 方法的扩展。该方法通过利用 TF-基因共表达模式，生成细胞类型特异性 GRN，并且通过启动子区域 TF 与基序结合的信息来优化 GRN 的边。单细胞分辨率的提高可以识别动态细胞状态以及在不同事件进展中的作用，如发育、细胞分化和疾病进展等。伪时间方法可以用来表征这样的连续变化，由此产生的 GRN 可以为细胞命运中的复杂过程提供有价值的线索。LEAP[25]和 SINCERITIES[26]是基于伪时间信息推断带有方向性的 GRN。单细胞染色质可及性与单细胞转录组学一起，让 GRN 的估计精度得到了大幅度改进[27, 28]。一些早期的研究从未配对的多组学数据中推断出 GRN，研究人类髓系细胞分化[29]、小鼠胚胎发育[30]和树突状细胞的 HIV 感染[31]。现在 GRN 推理的新方法会同时利用 scRNA-seq 和 scATAC-seq，用于 GRN 推理的多模态数据可以来自相同或者不同的细胞。如果多模态数据来源于不同的细胞，可以单独构建每个模态的 GRN，然后进行合并。如果多模态数据来源于相通的细胞，可以在同一细胞单元中对这两种模式进行建模，比如 DeepMAPS（参考文献 62）[32]、FigR[33]、GLUE[34]、scAI[35]和 SOMatic[36]等方法。如果两种模态都使用集成方法进行匹配，未配对的数据仍然可以建模。多模态 GRN 推理方法使用的扩展框架与单模态有所不同。具体来说，多模态方法根据 TF 表达预测基因表达，使用基序信息将 TF 分配给可访问的顺式调控元件 CRE，将 CRE 与有距离限制的基因组靶基因相关联。

即使 GRN 推理方法使用类似的建模策略，他们之间的结果可能不同，因为对于 TF 结合事件的预测，不同方法使用不同的 TF 结合基序数据库预测算法。TF 的覆盖范围不同，预测算法的模型绑定也

不同。大多数方法允许使用与其默认值不同的 TF 结合基序数据库，修复所使用的基序匹配器算法，比如 CENIC+[37]，集成了三种算法，cisTarget、DEM 和 HOMER。此外，GRN 推断方法使用不同的基因组距离截断，将开放的染色质区域分配给靶基因，比如近距离 10kb，中等距离 100kb，远端效应高达 1000kb。由于功能基因在最近的距离处被极大地富集，和阈值距离的差异可能会影响由此推断出 GRN。多模态 GRN 推理方法会生成一个候选支架网络，该网络由与 CRE 相关的 TF 三联体组成，其中 CRE 与靶基因相连。为了生成最终的 GRN 结构，不同的方法通常假设 TF、CRE 和基因等因素之间存在线性关系。线性建模假设一个变量，与另一个变量成正比的变化，例如 TF 转录本或 CRE 开放性。相比之下，非线性建模可以适应更复杂的交互协同效应等变量[38]。尽管人们普遍认为基因表达是一个非线性过程，但线性建模由于简单实用，通常更受欢迎。独立于所使用的建模策略，评估相互调控作用可以使用频率论或贝叶斯概率统计框架。频率论方法定义了事件的概率由于事件发生的次数比例很大相同的实验，而贝叶斯概率将其定义为基于以下因素对所述事件发生的置信度测量观测数据和先验信息。多模态 GRN 推理方法可以根据其建模策略与输入类型进行分组。FigR 和 GRaNIE[39]等使用了线性回归；DIRECT-NET[40]和 SCENIC+使用非线性回归（随机森林）；PECA[41]和 Symphony[42]使用贝叶斯建模。相比之下，CellOracle[43]、Inferelator 3.0[44]和 Pando[45]为用户提供多种建模策略。在细胞发育中，往往没有连续性的数据来定义不同的组，scMEGA[46]和 IReNA[47]利用推断出来的轨迹信息分别线性和非线性地推断 GRN。此外，Dictys[48]、scMTNI[49]和 TimeReg[50]使用组合细胞类型和轨迹数据为 GRN 建模提供信息，而 CellOracle 和 SCENIC+使用轨迹数据进行建模下游分析。ANANSE[51]、sc-compReg[52]和 SCENIC+等方法构建了细胞特异性的 GRN。

为转录生成全基因组结合转录因子（TF）需要大量的实验，从

而 GRN 的推理方法中预测 TF 结合可以基于开放基因组区域上的先验信息。这些信息来自大量的 TF-DNA 结合检测，如染色质免疫沉淀，ChIP-seq 实验等。和 TF 特异性结合的基因组序列，通常称为 TF 结合型基序。有几个数据库收集此类检测结果，并生成 TF 结合基序集合。在 GRN 推理过程中，如果数据库之间的覆盖范围不同，可以合并它们以增加 TF 数量。此外，一些计算已经开发出利用 TF 结合基序的算法以预测结合事件，称为基序匹配器算法。因为不同的方法对 TF 结合预测的建模不同，结果可能会有所不同。

高通量染色体构象捕获（Hi-C）是研究染色体和基因组三维构象的关键技术，应用下一代测序技术对空间上彼此靠近（即接触）的染色体区域进行测序。因此，Hi-C 数据捕获了基因组染色体区域之间的相互作用，以构建基因组的 3D 构象，并研究长程基因增强子相互作用。这种方法通常需要将数据转换为 2D 染色体接触矩阵，其中存储了染色体区域 i 与染色体区域 j 相互作用的频率，其中 i 和 j 是染色体区域的索引。因此，Hi-C 接触矩阵可以被视为图像。然而，Hi-C 数据，特别是单细胞 Hi-C 数据通常是嘈杂和不完整的，因此染色体接触基质中的染色体相互作用可能是假阳性，或者基质中可能缺少相互作用。深度学习方法（例如，GAN）可用于对 Hi-C 数据进行去噪[53]；此外，扩散模型（例如 DDPM）可以实现 Hi-C 染色体接触矩阵的去噪，以改善 3D 基因组构象建模，并研究基因和调控元件（例如增强子）之间的空间相互作用[54]。然而，DDPM 的深度学习架构通常是 U-Net，它可能不如 Hi-C 数据去噪方法 ScHiCEDRN[55]中使用的深度残差网络强大。因此，如果应用于 Hi-C 数据去噪，DDPM 的架构可以更新为深度残差网络，以提高其去噪能力。

参考文献

[1] Maston, G.A., S.K. Evans, and M.R. Green, Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **2006**. 7: p. 29-59.

[2] Alipanahi, B., A. Delong, M.T. Weirauch, and B.J. Frey, Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, **2015**. 33(8): p. 831-838.

[3] Lee, N.K., F.L. Azizan, Y.S. Wong, and N. Omar, DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery. *Biotechnology & Biotechnological Equipment*, **2018**. 32(3): p. 759-768.

[4] Shrikumar, A., P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. in International conference on machine learning. 2017. PMlR.

[5] Avsec, Ž., M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, and A. Kundaje, Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, **2021**. 53(3): p. 354-366.

[6] Chen, K.M., A.K. Wong, O.G. Troyanskaya, and J. Zhou, A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, **2022**. 54(7): p. 940-949.

[7] Karollus, A., T. Mauermeier, and J. Gagneur, Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome biology*, **2023**. 24(1): p. 56.

[8] Sasse, A., B. Ng, A. Spiro, S. Tasaki, D.A. Bennett, C. Gaiteri, P.L. De Jager, M. Chikina, and S. Mostafavi, How far are we from personalized gene expression prediction using sequence-to-expression

deep neural networks? *Biorxiv: the Preprint Server for Biology*, **2023**.

[9] Huang, Y., I.M. Tienda-Luna, and Y. Wang, Reverse engineering gene regulatory networks. *IEEE Signal Processing Magazine*, **2009**. 26(1): p. 76-97.

[10] Langfelder, P. and S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, **2008**. 9: p. 1-13.

[11] Cao, J., X. Qi, and H. Zhao, Modeling gene regulation networks using ordinary differential equations. *Next generation microarray bioinformatics: methods and protocols*, **2012**: p. 185-197.

[12] Patel, N. and J.T. Wang, Semi-supervised prediction of gene regulatory networks using machine learning algorithms. *Journal of biosciences*, **2015**. 40: p. 731-740.

[13] Jisha, A. and A. Jereech. Gene regulatory network: a semi supervised approach. in International Conference on Electronics Communication and Aerospace Technology ICECA. 2017.

[14] Mordelet, F. and J.-P. Vert, SIRENE: supervised inference of regulatory networks. *Bioinformatics*, **2008**. 24(16): p. i76-i82.

[15] Gillani, Z., M.S.H. Akash, M.M. Rahaman, and M. Chen, CompareSVM: supervised, Support Vector Machine (SVM) inference of gene regularity networks. *BMC bioinformatics*, **2014**. 15: p. 1-7.

[16] Mandal, S., A. Khan, G. Saha, and R.K. Pal, Large-scale recurrent neural network based modelling of gene regulatory network using cuckoo search-flower pollination algorithm. *Advances in bioinformatics*, **2016**. 2016.

[17] Huynh-Thu, V.A., A. Irrthum, L. Wehenkel, and P. Geurts, Inferring regulatory networks from expression data using tree-based

methods. *PloS one*, **2010**. 5(9): p. e12776.

[18] Moerman, T., S. Aibar Santos, C. Bravo González-Blas, J. Simm, Y. Moreau, J. Aerts, and S. Aerts, GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, **2019**. 35(12): p. 2159-2161.

[19] Buenrostro, J.D., P.G. Giresi, L.C. Zaba, H.Y. Chang, and W.J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, **2013**. 10(12): p. 1213-1218.

[20] Pranzatelli, T.J., D.G. Michael, and J.A. Chiorini, ATAC2GRN: optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference. *BMC genomics*, **2018**. 19: p. 1-13.

[21] Qin, Q., J. Fan, R. Zheng, C. Wan, S. Mei, Q. Wu, H. Sun, M. Brown, J. Zhang, and C.A. Meyer, Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome biology*, **2020**. 21: p. 1-14.

[22] Sonawane, A.R., D.L. DeMeo, J. Quackenbush, and K. Glass, Constructing gene regulatory networks using epigenetic data. *npj Systems Biology and Applications*, **2021**. 7(1): p. 45.

[23] Consortium*, T.T.S., R.C. Jones, J. Karkanias, M.A. Krasnow, A.O. Pisco, S.R. Quake, J. Salzman, N. Yosef, B. Bulthaup, and P. Brown, The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, **2022**. 376(6594): p. eabl4896.

[24] Aibar, S., C.B. González-Blas, T. Moerman, V.A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, and J. Aerts, SCENIC: single-cell regulatory network inference and clustering. *Nature methods*, **2017**. 14(11): p. 1083-1086.

[25] Specht, A.T. and J. Li, LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics*, **2017**. 33(5): p. 764-766.

[26] Papili Gao, N., S.M. Ud-Dean, O. Gandrillon, and R. Gunawan, SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, **2018**. 34(2): p. 258-266.

[27] Chen, S., B.B. Lake, and K. Zhang, High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, **2019**. 37(12): p. 1452-1457.

[28] Ma, S., B. Zhang, L.M. LaFave, A.S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V.K. Kartha, and T. Tay, Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, **2020**. 183(4): p. 1103-1116. e20.

[29] Ramirez, R.N., N.C. El-Ali, M.A. Mager, D. Wyman, A. Conesa, and A. Mortazavi, Dynamic gene regulatory networks of human myeloid differentiation. *Cell systems*, **2017**. 4(4): p. 416-429. e3.

[30] Starks, R.R., A. Biswas, A. Jain, and G. Tuteja, Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics & chromatin*, **2019**. 12: p. 1-16.

[31] Johnson, J.S., N. De Veaux, A.W. Rives, X. Lahaye, S.Y. Lucas, B.P. Perot, M. Luka, V. Garcia-Paredes, L.M. Amon, and A. Watters, A comprehensive map of the monocyte-derived dendritic cell transcriptional network engaged upon innate sensing of HIV. *Cell reports*, **2020**. 30(3): p. 914-931. e9.

[32] Ma, A., X. Wang, J. Li, C. Wang, T. Xiao, Y. Liu, H. Cheng, J.

Wang, Y. Li, and Y. Chang, Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications*, **2023**. 14(1): p. 964.

[33] Kartha, V.K., F.M. Duarte, Y. Hu, S. Ma, J.G. Chew, C.A. Lareau, A. Earl, Z.D. Burkett, A.S. Kohlway, and R. Lebofsky, Functional inference of gene regulation using single-cell multi-omics. *Cell genomics*, **2022**. 2(9).

[34] Cao, Z.-J. and G. Gao, Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, **2022**. 40(10): p. 1458-1466.

[35] Jin, S., L. Zhang, and Q. Nie, scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome biology*, **2020**. 21: p. 1-19.

[36] Jansen, C., R.N. Ramirez, N.C. El-Ali, D. Gomez-Cabrero, J. Tegner, M. Merkenschlager, A. Conesa, and A. Mortazavi, Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. *PLoS computational biology*, **2019**. 15(11): p. e1006555.

[37] Bravo González-Blas, C., S. De Winter, G. Hulselmans, N. Hecker, I. Matetovici, V. Christiaens, S. Poovathingal, J. Wouters, S. Aibar, and S. Aerts, SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature methods*, **2023**. 20(9): p. 1355-1367.

[38] Zuin, J., G. Roth, Y. Zhan, J. Cramard, J. Redolfi, E. Piskadlo, P. Mach, M. Kryzhanovska, G. Tihanyi, and H. Kohler, Nonlinear control of transcription through enhancer–promoter interactions. *Nature*, **2022**. 604(7906): p. 571-577.

[39] Kamal, A., C. Arnold, A. Claringbould, R. Moussa, N.H. Servaas,

M. Kholmatov, N. Daga, D. Nogina, S. Mueller-Dott, and A. Reyes-Palomares, GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Molecular Systems Biology*, **2023**. 19(6): p. e11627.

[40] Zhang, L., J. Zhang, and Q. Nie, DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Science Advances*, **2022**. 8.

[41] Duren, Z., X. Chen, R. Jiang, Y. Wang, and W.H. Wong, Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences*, **2017**. 114: p. E4914 - E4923.

[42] Bachireddy, P., E. Azizi, C. Burdziak, V.N. Nguyen, C.S. Ennis, K. Maurer, C.Y. Park, Z.-N. Choo, S. Li, S.H. Gohil, N. Ruthen, Z. Ge, D.B. Keskin, N. Cieri, K.J. Livak, H.T. Kim, D.S. Neuberg, R.J. Soiffer, J. Ritz, E.P. Alyea, D. Pe'er, and C.J. Wu, Mapping the evolution of T cell states during response and resistance to adoptive cellular therapy. *Cell reports*, **2021**. 37 6: p. 109992.

[43] Kamimoto, K., B. Stringa, C.M. Hoffmann, K. Jindal, L. Solnica-Krezel, and S.A. Morris, Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, **2023**. 614: p. 742 - 751.

[44] Gibbs, C.S., C.A. Jackson, G.-A. Saldi, A. Tjärnberg, A. Shah, A. Watters, N. De Veaux, K. Tchourine, R. Yi, and T. Hamamsy, High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. *Bioinformatics*, **2022**. 38(9): p. 2519.

[45] Fleck, J.S., S.M.J. Jansen, D. Wollny, F. Zenk, M. Seimiya, A. Jain, R. Okamoto, M. Santel, Z. He, and J.G. Camp, Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, **2023**.

621(7978): p. 365-372.

[46] Li, Z., J.S. Nagai, C. Kuppe, R. Kramann, and I.G. Costa, scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference. *Bioinformatics Advances*, **2023**. 3(1): p. vbad003.

[47] Jiang, J., P. Lyu, J. Li, S. Huang, J. Tao, S. Blackshaw, J. Qian, and J. Wang, IReNA: Integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles. *Iscience*, **2022**. 25(11).

[48] Wang, L., N. Trasanidis, T. Wu, G. Dong, M. Hu, D.E. Bauer, and L. Pinello, Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nature Methods*, **2023**. 20(9): p. 1368-1378.

[49] Zhang, S., S. Pyne, S. Pietrzak, S. Halberg, S.G. McCalla, A.F. Siahpirani, R. Sridharan, and S. Roy, Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nature Communications*, **2023**. 14(1): p. 3064.

[50] Duren, Z., X. Chen, J. Xin, Y. Wang, and W.H. Wong, Time course regulatory analysis based on paired expression and chromatin accessibility data. *Genome research*, **2020**. 30(4): p. 622-634.

[51] Xu, Q., G. Georgiou, S. Frölich, M. van der Sande, G.J.C. Veenstra, H. Zhou, and S.J. van Heeringen, ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic acids research*, **2021**. 49(14): p. 7966-7985.

[52] Duren, Z., W.S. Lu, J.G. Arthur, P. Shah, J. Xin, F. Meschi, M.L. Li, C.M. Nemec, Y. Yin, and W.H. Wong, Sc-compReg enables the comparison of gene regulatory networks between conditions using

single-cell data. *Nature Communications*, **2021**. 12(1): p. 4763.

[53] Highsmith, M. and J. Cheng, VEHiCLE: a variationally encoded Hi-C loss enhancement algorithm for improving and generating Hi-C data. *Scientific Reports*, **2021**. 11(1): p. 8880.

[54] Ho, J., A. Jain, and P. Abbeel, Denoising diffusion probabilistic models. *Advances in neural information processing systems*, **2020**. 33: p. 6840-6851.

[55] Wang, Y., Z. Guo, and J. Cheng, Single-cell Hi-C data enhancement with deep residual and generative adversarial networks. *Bioinformatics*, **2023**. 39(8): p. btad458.

第 9 章 人工智能赋能多组学融合

9.1 人工智能与多组学融合概述

随着基因测序技术的快速发展，产生了大量的基因组学、转录组学、表观遗传学、蛋白质组学数据，以及成对的多种组学数据，即针对同一样本同时测量其两至三种组学特征。不同的组学描述了遗传过程中，从基因 DNA 编码到转录翻译表达调控过程中的多个环节。但是，随之而来的多种组学数据的对齐与融合成为了多组学数据分析中的一大难点问题。

人工智能技术与深度神经网络技术已经广泛地应用于图像与文字、视频与音频等多模态数据融合处理中，并取得了不错的效果，也为多组学数据融合提供了先进的策略和方法。目前，基于人工智能的多组学融合方法主要可以分为四种类型：一是基于深度神经网络的方法（图 9-1A），由于深度神经网络所需的训练数据量较大，因此多用于单细胞多组学中；二是基于非负矩阵分解的方法（图 9-1B），通过非负矩阵分解将多组学数据分解为共有的因子矩阵和各组学固有特征的矩阵，并基于这些分解后的矩阵进行下游任务；三是基于贝叶斯统计方法（图 9-1C），将不同组学投射到隐空间，拟合其分布，并将其中一个组学拟合为另一个组学的条件概率分布，利用贝叶斯公式进行推算；四是基于图（网络）的方法（图 9-1D），因不同组学在基因和生物学功能上有关联，基于这种关联构建调控关系图（网络），应用图论相关算法对数据进行融合。不同方法之间会有交叉融合，例如，目前多数贝叶斯方法和图方法都基于自编码神经网络将数据投射到隐空间，在隐空间中使用贝叶斯统计和图方法，或者使用图神经网络对多组学数据进行卷积处理。在本章节中，基于贝叶斯的深度学习和使用图神经网络的方法都被归类为贝叶斯与图方法。

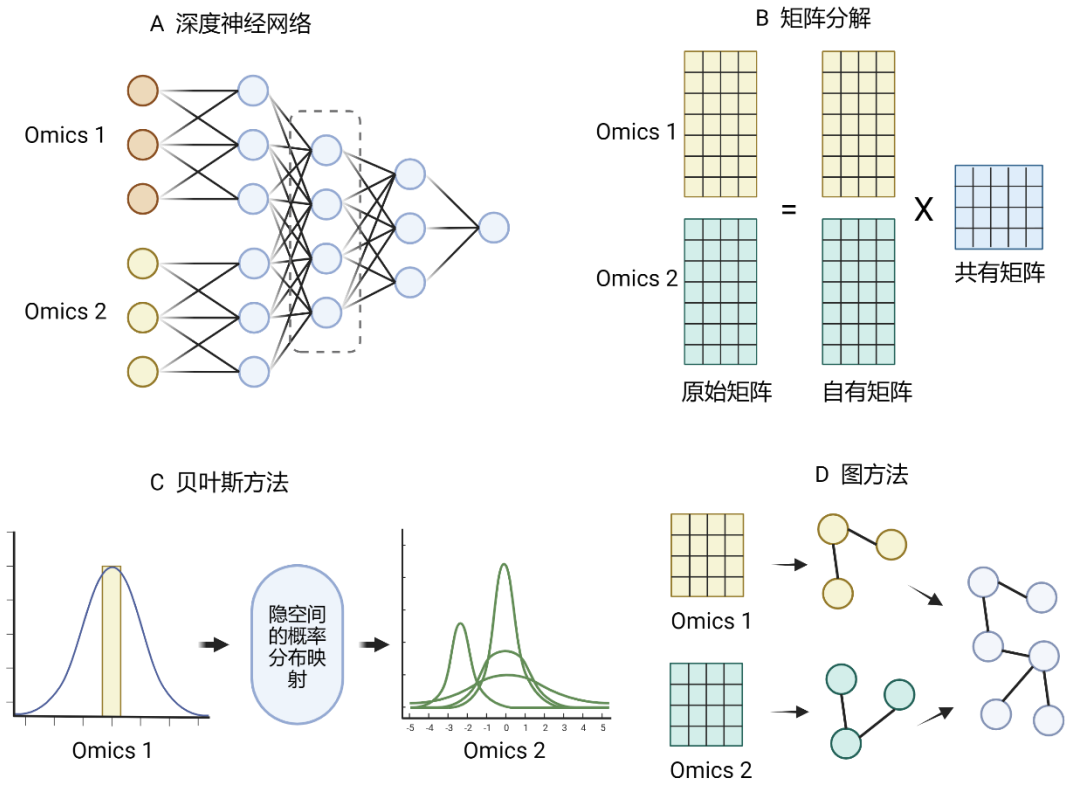


图 9-1 基于人工智能的多组学融合方法

目前主要的组学数据有基因组学、转录组学、表观遗传学、蛋白质组学数据，而近几年的研究主要集中在单细胞转录组学、表观遗传学、蛋白质组学数据的融合。因此，从组学的角度，可以将多组学融合方法的应用分为三类：一是单细胞转录组与表观遗传组数据融合，大部分方法都针对单细胞 RNA 测序（scRNA-seq）与单细胞染色质可及性（scATAC-seq）数据的融合而开发；二是单细胞转录组与蛋白质组数据融合，主要针对单细胞 RNA 测序（scRNA-seq）与蛋白质表达（ADT）数据的融合；第三类为三种组学的融合，包括单细胞 RNA 测序、单细胞染色质可及性与蛋白质表达（RNA-ATAC-ADT）数据融合。由于 RNA 和 ATAC 的配对数据成本较低、数据量较大，因此大多数方法都是针对 RNA-ATAC 多组学融合而开发。还有一些方法既可以用于两种组学、又可以应用于三种组学数据，在本章中都归为三种组学融合的方法。另外，目前热点的空间组学属于组学信息

与空间信息的融合，会在其他章节详细描述，在本章中不再赘述。

人工智能在多组学数据融合的应用中存在不同的融合策略，可以在不同阶段对两种或多种数据进行融合。主要包括早期阶段融合、中间阶段融合、最终阶段融合（图 9-2）。早期阶段融合指的是，在数据输入到模型之前，先进行数据对齐，然后统一输入到人工智能模型，模型根据组学配对的信息统一进行处理。这种方式的优势在于，能够融入生物学信息，具备较好的可解释性，但并非每一种组学都能够很好的与另一种组学进行对齐。中间阶段融合多见于深度神经网络的架构，前端先对每一种组学使用各自不同的嵌入（embedding）将不同组学投射到隐空间中，然后在隐空间中进行融合。这种方式的优势在于，无需人工对组学进行对齐就可以将不同组学映射到相同的隐空间，提供更好的抽象解析，融合效果较好，是目前使用较多的方式。最终阶段融合主要针对特定的下游任务，分别使用机器学习模型对各自组学进行处理得到结果，最后根据不同组学的最终结果进行整合判断，是一种集成学习的方法。其优势在于，不需要对不同组学进行交互处理，每一种组学有一套独立的模型，相对比较简单，但融合效果较弱，且只能适用于特定下游任务，泛化性不足。

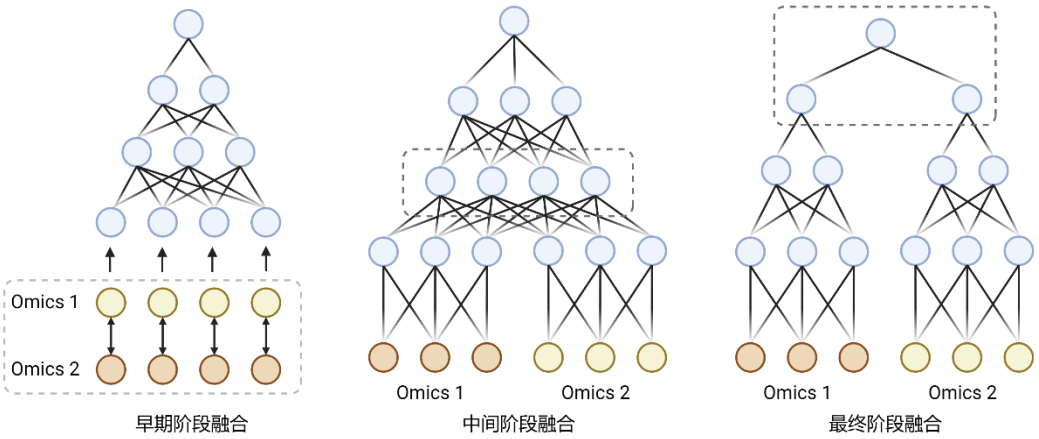


图 9-2 基于人工智能的多组学数据融合策略

本章节将首先介绍基因组学、转录组学、表观遗传学、蛋白质组学数据的来源，即不同组学的测序技术，以及配对多组学数据的测序技术。然后，将从多组学数据的维度，即单细胞转录组与表观遗传组数据融合、单细胞转录组与蛋白质组数据融合、三种组学（RNA-ATAC-ADT）数据融合三个方面，以及人工智能方法的维度，即深度神经网络、矩阵分解、贝叶斯统计、图方法四种方法，介绍人工智能赋能多组学融合的最新成果。

9.2 多组学测序技术

本章将介绍多种组学数据的获取方式，即测序技术，以及多组学的测序技术。一方面，近年来生命科学技术高通量测序技术已经发展到单细胞尺度，另一方面，机器学习和深度学习模型在训练的过程中需要使用大量的数据，而单细胞测序的出现能够满足其数据需求。因此，本章主要解释单细胞尺度的多种组学的测序技术（图 9-3）。

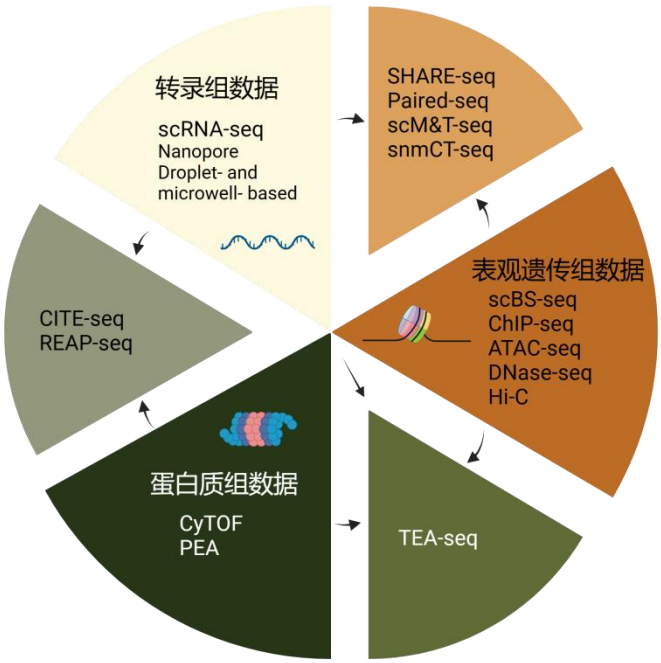


图 9-3 多组学测序技术（CITE-seq，细胞转录组和表位的索引；DR-seq，基因

组 DNA-信使 RNA 测序；FISH，荧光原位杂交；G&T-seq，基因组和转录组测序；MERFISH，多重化错误鲁棒 FISH；Paired-Tag，单个细胞 RNA 表达和靶向测序的 DNA 的平行分析；PEA，近距离扩展测定法；REAP-seq，RNA 表达和蛋白质测序；scBS-seq，单细胞亚硫酸氢钠测序；scM&T-seq，单细胞甲基组和转录组测序；scRNA-seq，单细胞 RNA 测序；SHARE-seq，转座酶可接近染色质测定和 RNA 表达的同时高通量测序；smFISH，单分子 FISH；TEA-seq，转录本、表位和染色质可及性的同时三模单细胞测量。）

9.2.1 单细胞基因组学

单细胞基因组学指的是在以细胞为尺度在 DNA 层面上进行的研究，主要技术为单细胞 DNA 测序 (scDNA-seq)，它能够在单细胞分辨率上测量遗传变异来阐明遗传异质性。单细胞 DNA 测序 (scDNA-seq) 已被证明可以有效识别拷贝数畸变、体细胞突变和追踪细胞谱系。它在癌症研究中得到了广泛的应用，有助于跟踪不同细胞克隆的生长并了解肿瘤的发展。将转录组和基因组单细胞测序配对可以帮助揭示基因调控机制和基因型-表型关联。由于基因组和转录组的对应比较直接、且数据量较少，人工智能在单细胞转录组和基因组融合的应用较少，因此本章节不作过多介绍。

9.2.2 单细胞转录组学

单细胞转录组学是指在单细胞水平上进行 RNA 测序，研究细胞之间基因的转录水平，单细胞 RNA 测序技术（scRNA-seq）已被广泛应用于基础和转化研究中。为了在单细胞水平上测量转录，必须将每个细胞从其起源组织中分离出来，这可以使用多种技术实现，包括荧光激活细胞分选（FACS），激光捕获显微解剖和微流控技术。基于液滴或微阱的方法，如 Drop-seq 和 10× Genomics Chromium，生成完整的互补 DNA（cDNA）池，从而实现对成千上万个细胞的无偏分析，随后使用 Illumina 短读测序进行。利用纳米孔长读测序的技术可以生

成完整的序列以及有关序列多样性、剪接和嵌合转录本的信息。新技术，如分割池技术，通过使用单细胞的组合索引，提供了一种无需细胞分离的替代方法。

9.2.3 单细胞表观遗传学

表观遗传学研究的是在不改变 DNA 序列的前提下，通过某些机制引起可遗传的基因表达或细胞表现型的变化，表观遗传组数据使我们能够研究将基因组内容转化为多种功能和稳定细胞状态的机制。染色质可及性指示，对 DNA 的物理接触是建立和维持细胞身份的基本调控机制。

胞嘧啶甲基化是一种关键的表观遗传层，指示基因组 DNA 的转录潜力，可以使用亚硫酸氢钠测序来检测。单细胞胞嘧啶甲基化测序（scBS-seq）可以映射甲基化位置。将 scBS-seq 与 scRNA-seq 结合使用 scM&T-seq，可以同时获得单细胞甲基组和转录组测序数据。另一种技术，scNMT-seq（单细胞核小体、甲基化和转录组测序），将 scM&T-seq 与 NOMe-seq（核小体占据和甲基化测序）相结合，以捕获单细胞水平上的转录组、甲基组和染色质可及性。

9.2.4 单细胞蛋白质组学

蛋白质组学是对不同的蛋白质的表达水平进行测量，研究它们在生物体内发挥的作用。尽管蛋白质是由 mRNA 翻译而成，但是 mRNA 和蛋白质水平不一定相关，部分原因是由于转录后的调控机制。许多单细胞蛋白质组学的测量方法主要使用与可检测分子结合的抗体，如高维度流式细胞术（FACS）和飞行时间流式细胞术（CyTOF；金属同位素）。另一种方法是在 RNA 分析中并行进行蛋白质测量的近距离扩展测定法（PEA）。在 PEA 中，抗体特异性地与目标蛋白结合后，抗体上的探针对进行近距离杂交，形成 PCR 模板，继而利用引物对样本进行扩增，最后通过 qPCR 或 NGS 实现定量检测。

9.2.5 单细胞多组学

新的高通量测序技术的发展使得对多种组学的同时测量成为可能，从而产生联合多模态的数据，为组织和器官功能中涉及的分子和细胞过程提供了更全面的理解。多种技术使研究人员能够在单个细胞中同时测量 DNA 甲基化和转录组数据，例如 scM&T-seq、scMT-seq、scTrio-seq 和 snmCT-seq。Perturbseq 和 CRISP-seq 是测量基于 CRISPR 的转录干扰和高通量单细胞 RNA 测序的技术。类似的技术，如 Paired-seq 和 SNARE-seq，在单个细胞或核中研究转录组和染色质可及性。这些技术产生的数据允许在单细胞水平上进行多模态组学分析[1]。CITE-seq 和 REAP-seq 允许同时测量转录组和蛋白质组的表达水平。TEA-seq 和 scNMT-seq 能够同时测量三种组学。

9.3 转录组学与表观遗传学数据融合

9.3.1 基于深度神经网络方法

本小节将介绍近几年由华人提出的使用深度神经网络融合转录组学与表观遗传学的方法。中国科学院上海生物化学与细胞生物学研究所的陈洛南团队提出了单细胞多模态变分自动编码器 scMVAE[2]和 DCCA[3]。 scMVAE 模型利用随机优化和多模态编码器，首先通过聚合相似细胞和特征上的双组学数据来近似 GMM 先验下的联合潜在特征，然后通过每种组学数据的解码器重构观察到的表达值，同时考虑到每种类型数据的归一化，该模型适用于从同一单个细胞中测量的 scRNA-seq 和 scATAC-seq 数据的整合分析。DCCA 通过估计一定分布的每个 VAE 模型分别对每个组学数据进行建模，然后通过在彼此之间循环转移关于低维表示的注意力图来协调多组学数据，用于对同一细胞的 scRNA-seq 和单细胞表观基因组学 (scEpigenomics) 数据进行综合分析。

同济大学的李高阳团队提出了一种非对称深度生成模型，即单细

胞多视图分析器 scMVP[4]，它通过聚类一致性自动学习 scRNA-seq 和 scATAC-seq 数据的常见潜在表示-约束多视图变分自动编码器模型（VAE），并通过特定于层的数据生成过程（包括 Transformer 的基于自注意力的 scATAC）从多组学数据的公共潜在嵌入中估算每个单层数据。scMVP 提供了一种高效的深度生成模型，用于对同一单细胞的多个组学测量进行联合分析，并实现同时多模态分析联合分析测序数据的数据标准化、聚类、联合嵌入、可视化、轨迹推断和 CRE 预测。

香港科技大学的杨燦和吴若昊团队创建了一个名为Portal的工具[5]，能够快速而准确地整合来自样本、技术平台、数据模态和物种的图谱级别数据集。它将来自不同研究的数据集视为具有特定领域效应（包括技术变异和其他来源的非期望变异）的不同领域，通过整合对抗机制的统一域转换框架实现了良好的数据对齐性能。它在潜在空间中学习细胞的表示，其中移除了特定领域的效应。域转换网络的非线性使得 Portal 能够考虑复杂的特定领域效应。

清华大学的张强锋团队开发了名为 SCALEX 的方法[6]，该方法基于变分自编码器（VAE）框架，用于在线整合异质单细胞数据。SCALEX 的编码器被设计为一个数据投影函数，当投影单个细胞时，它仅保留批次不变的生物数据组分。重要的是，该投影函数是通用的，不需要在新数据上重新训练，因此使得 SCALEX 能够以在线方式整合单细胞数据。通过与大量基准数据集的合作，他们证明了 SCALEX 在整合精度、可扩展性和计算效率等方面显著优于在线 iNMF 以及非在线单细胞数据整合工具。这些优势使得 SCALEX 特别适用于整合和研究利用当今随着生物学和医学领域中不断增加的单细胞研究而不断增长的数据集。

香港中文大学的李煜团队提出了一种基于对比循环对抗自动编码器（Con-AAE）的框架[7]，利用两个自动编码器将两种模态的数据

映射到两个低维流形中，并在对抗性损失的约束下进行训练，旨在为每种模态开发出分离的表示，这些表示在协调子空间中无法被对抗网络识别。然而，仅使用对抗性损失可能会导致模型崩溃。为了避免这个问题，他们进一步引入了一种新的循环一致性损失。此外，他们还训练了一个没有成对信息的模型用于数据对齐任务，并通过自监督对比学习明确考虑数据噪声。

四川大学的彭玺团队院发现细胞异质性不仅不是干扰，反而可以利用这一特性来促进数据整合。具体而言，他们发现 scATAC-seq 细胞的染色质可及性与 scRNA-seq 基因表达呈现可变的相关性。具有更高正相关性的 scATAC-seq 细胞表现出更小的组学差异，这些差异较容易进行整合，并能够弥合两种组学之间的差距。根据这一观察结果，他们设计了一种名为 scBridge 的异质转移学习方法[8]，用于多组学数据的整合。scBridge 首先使用带注释的 scRNA-seq 数据预热一个深度神经分类器，然后通过可靠性建模识别出具有较小组学差异的 scATAC-seq 细胞。然后，通过跨组学原型对齐，将可靠的 scATAC-seq 细胞与 scRNA-seq 细胞进行整合。最后，scBridge 从可靠的 scATAC-seq 细胞中选择并合并到带注释的 scRNA-seq 数据中，以缩小组学差距。通过重复上述过程，组学差异逐渐减小，整合的细胞数量增加，最终实现了数据的整合结果。

北京大学的邓明华团队提出了 MoClust 作为一种用于聚类单细胞多组学数据的模型[9]。MoClust 可应用于转录组和蛋白组数据，以及转录组和表观基因组数据。首先，它通过不同的自编码器分别对每个组学数据集进行建模，并学习低维的组学特定潜在表示。其次，在预训练阶段引入了一种新颖的自动双重子发现模块，以净化输入数据并提升聚类性能。第三，MoClust 包含一个对齐表示的对比模块。通过余弦角而不是欧氏距离来估计表示之间的相似性，以调整不同组学中潜在不一致的度量方式。

山东大学-南洋理工大学人工智能联合研究中心的王峻团队提出了一种名为 scMCs 的方法[10]。他们解决方案的主要思想是设计一个信息提取和融合模块，以精细处理从异质组学中学到的个性和共性，并构建更全面、更丰富的单细胞多组学数据融合、聚类和多聚类表示。

加州大学的谢晓晖团队提出了一个名为 SAILERX 的深度学习框架，以改进多组学或单模态和多模态单细胞测序数据集的分析[11]。他们使用变分自动编码器（VAE）对 scATAC-seq 数据进行建模，而 scRNA-seq 数据的嵌入是预先训练的，不会在训练时显式进行建模。他们通过最小化两个模态之间嵌入空间中成对相似性的距离来进一步进行规范化，这鼓励细胞的局部结构与参考模态相似，同时适应不同模态之间的技术噪声差异较大。SAILERX 的建模选择允许混合整合具有 scATAC-seq 测量的数据集和具有配对的 scRNA-seq 和 scATAC-seq 的数据集，有效利用高质量多模态数据的信息来改进单模态数据集的分析。

佐治亚理工学院的张秀苇团队提出了 scDART，用于 ATAC-Seq 和 RNA-Seq 轨迹集成的单细胞深度学习模型[12]。这是一种可扩展的深度学习框架，它将数据模态嵌入到共享的低维潜在空间中，保留原始数据集中的细胞轨迹结构。

加拿大多伦多大学的王波团队通过对超过 3300 万个细胞进行预训练，提出了单细胞基础模型 scGPT[13]。研究人员建立了一个专门针对非顺序组学数据的统一生成预训练工作流，并将 Transformer 架构进行了适应，同时学习细胞和基因表示。此外，研究人员还提供了具有任务特定目标的微调流程，旨在促进在各种不同任务中应用预训练模型。预训练基础模型的采用将极大地拓展我们对细胞生物学的认识，并为未来的发现奠定坚实基础。

9.3.2 基于矩阵分解方法

本小节将介绍近几年由华人提出的基于矩阵分解方法融合转录

组学与表观遗传学的方法。加利福尼亚大学的聂清团队提出了一种单细胞聚集和整合方法（scAI）来整合来自同一细胞的转录组和表观基因组谱（即染色质可及性或 DNA 甲基化）[14]。scAI 考虑到单细胞表观基因组数据的极其稀疏和近二元性质。通过无监督的迭代学习，scAI 在表现出相似基因表达和表观基因组谱的细胞亚组中聚集表观基因组数据。这些相似的细胞是通过同时从转录组和聚集的表观基因组数据中学习细胞-细胞相似性矩阵来计算的。因此，scAI 代表了具有生物学意义的低秩矩阵的转录组和表观基因组谱，允许识别细胞亚群；在共享的二维空间中同时可视化细胞、基因和位点；以及转录调控关系的推断。

为了解决收集参考知识的问题，浙江大学陈华钧团队提出了一个能够丰富模式和文本实例的模式感知参考存储方案 SCHEMA[15]，将人工注释和弱监督文本的实例与结构化模式进行对齐，使符号知识和文本语料库在表示学习中处于同一空间。然后，构建一个包含来自符号模式和训练实例衍生的知识的统一参考存储。为了解决利用参考知识的问题，提出了基于检索的参考集成方法来选择信息丰富的知识作为提示。由于并非所有外部知识都是有利的，利用基于检索的方法动态地从模式感知的参考存储中选择最相关于输入序列的知识作为提示。通过这种方式，每个样本都可以获得多样化和合适的知识提示，在资源有限的环境中提供丰富的符号指导。

9.3.3 基于图/网络方法

北京大学高歌团队提出了一种基于图连接的统一嵌入方式 GLUE[16]，用于同时整合非配对的单细胞多组学数据并推断不同组学间的调控相互作用。通过显式地建模各组学层次之间的调控相互作用，GLUE 以一种生物学直观的方式弥合了各种组学特征空间之间的差距。系统化的基准测试和案例研究表明，GLUE 对于异质单细胞多组学数据是准确、稳健且可扩展的。此外，GLUE 被设计为一个通用的

框架，可以以模块化的方式轻松扩展和快速应用于特定情境。

清华大学的陈恳和陈睿团队开发了一种新的方法称为双标准相关分析（双阶标准相关分析）和相关的计算工具称为 Bi-CCA[17]，从两个不同实验生成的两个数据矩阵中学习行和列之间的最佳对齐（即单元对应和特征交互）。从 bi-CCA 导出的排列矩阵可以用于导出多组学对齐的细胞，可以用作下游调节网络推断的输入。bi-CCA 利用完整的特征信息，实现了 RNA 和 ATAC 数据之间双极细胞亚型的准确对齐。它还通过整合 RNA 和质谱细胞术数据，使能够发现新的细胞类型区分基因-蛋白质链接。

西安电子科技大学的马小科团队提出了一种基于网络的整合聚类算法 NIC[18]，用于通过融合单细胞转录组（scRNA-seq）和表观基因组谱（scATAC-seq 或 DNA 甲基化）来识别细胞类型。NIC 包括自适应图学习、整合分析和细胞聚类这三个主要组成部分。图学习过程通过利用细胞的轮廓自动构建细胞的相似性网络，而整合分析过程通过探索细胞的相似性网络的结构来提取细胞的特征。根据提取的细胞特征来识别细胞类型。因此，NIC 的目标函数也由对应于前两个程序对应的两个子成本组成。

山东大学刘炳强和康钦马团队开发了 MarsGT（使用单细胞图变换器用于罕见群体推断的多组学分析）[19]。这是一种用于从单细胞多组学数据中识别罕见细胞群体的端到端深度学习模型。内部工具 DeepMAPS16 显示了异构图转换器（HGT）的优越性能，这是一种强大的图神经网络架构，可以处理大规模异构和动态图，通过联合分析 mumi 数据的生物网络推理和细胞聚类。在这样的基础上，MarsGT 引入了一个基于概率的 HGT 框架来分析来自异质图的单细胞多组学数据，包括细胞、基因和峰，这可以建立峰-基因调控关系，并利用这种关系来表征罕见的细胞群。

南京医科大学的郭雪江团队提出了一个称为 scapGNN 的统一框

架[20]。这是一个基于图神经网络（GNN）的框架，可以推断和重建基因-细胞、基因-基因和细胞-细胞关联关系，将稀疏的单细胞谱数据转化为稳定的基因-细胞关联网络。此外，scapGNN 整合了单细胞多组学数据，计算了单细胞通路活性评分，并通过量化网络信息识别了细胞表型相关的基因模块。

9.4 转录组学与蛋白质组学数据融合

9.4.1 基于神经网络方法

新泽西理工学院的 Zhi Wei 团队开发了一种多模态深度学习模型 scMDC[21]，用于多模态单细胞数据的聚类分析。scMDC 采用了多模态自动编码器，用于处理来自不同模态的连接数据，并使用两个解码器分别解码每个模态的数据。为了进一步提高潜在特征学习，scMDC 引入了基于 Kullback-Leibler 散度的损失（KL 损失），以吸引相似的细胞并分离不相似的细胞。整个模型，包括自动编码器、KL 损失和深度 K 均值聚类，同时进行优化。scMDC 是一种端到端的多模态深度学习聚类方法，用于建模不同的多组学数据。

中国科学院计算技术研究所的吴杨和赵屹团队设计了一种包含 RNA 序列和体内 DMS-seq 结构特征的的深度双模信息融合网络名为 DeepFusion[22]，通过结合从 DMS-seq 数据中获得的结构特征来揭示蛋白质与 RNA 的相互作用。DeepFusion 融合了两个基于卷积神经网络和长短期记忆网络的子模型，以提取局部主题信息和长期上下文信息。

深圳大学的欧阳乐团队提出了一种新颖的深度嵌入式多组学聚类与协作训练模型 DEMOC[23]，用于整合转录组学和蛋白组学数据，以联合识别细胞簇。DEMOC 模型首先将 CITE-seq 数据中的转录组学和蛋白组学数据视为描述细胞不同方面的两个视角，然后在转录组学数据上应用基于集成学习的 scRNA-seq 填补方法 EnImpute，生成填

补后的转录组学数据作为第三个描述细胞的视角。针对这三种数据视角预训练了三种不同类型的自编码器框架，获得了它们的低维表示和初始聚类中心。随后引入协作训练框架，对这三个数据视角进行联合聚类。

北京大学的邓明华团队提出了 scCTClust 用于对 CITE-seq 数据进行聚类[24]。为了解决 CITE-seq 数据组学之间的维度差异，难以量化组学对聚类对象的贡献，以及组学表示之间可能的低相关性等问题，scCTClust 首先利用了两个针对转录组学和蛋白质组学的神经网络，分别从两个维度提取特征。然后，利用深度典范相关分析方法，找到了两个维度之间的最大相关性，以便将它们合并到一个共享的低维空间中。在此基础上，引入了一种新的分散聚类方法，用于在低维空间中对细胞进行聚类。

宾夕法尼亚大学的 Justin Lakkis 和 Mingyao Li 团队提出了 sciPENN（单细胞蛋白质嵌入神经网络）[25]，用于预测和插补蛋白质表达、量化不确定性、在低维嵌入中集成数据集，以及合并多个 CITE-seq 数据集。sciPENN 可以使用截断损失方法集成多个 CITE-seq 数据集，即使它们的蛋白质面板并不完全重叠。sciPENN 的优势具有高度可扩展性和计算效率。随着多模态数据集规模的持续增长，准确且高效的计算方法对于扩展其应用在实践中至关重要。

9.4.2 基于矩阵分解方法

厦门大学的帅建伟团队设计了一个灵活的单细胞多模态分析框架 CITEMO[26]。CITEMO 使用主成分分析（PCA）分别获取转录组和 ADT 的低维表示，然后再次利用 PCA 来整合这些低维多模态数据进行下游分析，CITEMO 框架涵盖了一系列为多模态数据设计的过程，并同时输出转录组、ADT 和多模态组学分析结果。研究证明，CITEMO 可以轻松应用于大样本分析，并具有出色的稳健性。

9.4.3 基于贝叶斯统计学方法

匹兹堡大学的 Wei Chen 团队开发了一种新颖的贝叶斯随机效应混合模型 BREM-SC[27]，该模型可对配对的单细胞转录组和蛋白质组数据进行联合聚类。作为一种基于概率模型的方法，BREM-SC 能量化每个单细胞的聚类不确定性。在处理来自不同数据源的数据时，集成聚类方法能够将单独的聚类整合起来，确定与每个数据源最匹配的细胞总体分区。然而，大多数集成聚类方法假设单独的聚类已知，并且忽略了不确定性，而忽略数据的异质性可能会导致遗漏每个数据源的重要特征。BREM-SC 概率模型提供了一个统一的框架，能够联合分析多源数据，并考虑到不同数据源之间的相关性。

匹兹堡大学的 Wei Chen 团队还开发了 SECANT[28]，一种生物学指导的 SEmi 监督方法，用于单细胞多组学的聚类、分类和注释。它可以分析 CITE-seq 数据，也可以联合分析 CITE-seq 和 scRNA-seq 数据。SECANT 的创新之处包括：（1）利用从表面蛋白数据中识别出的可信细胞类型标签作为细胞聚类的指导；（2）为每个细胞聚类提供可信细胞类型的一般注释；（3）利用细胞类型标签不确定或缺失的细胞来提高性能；（4）准确预测 scRNA-seq 数据的可信细胞类型。作为一种基于模型的方法，SECANT 可以通过易于解释的后验概率量化结果的不确定性，而且该框架可以扩展到处理其他类型的多组学数据。

9.4.4 基于图/网络方法

悉尼大学的 Hani Jieun Kim 和 Yingxin Lin 团队介绍了一个实现了一系列从双重检测到综合分析的 CITE-seq 数据的方法和工具的计算框架，CiteFuse[29]。该软件包用于双重检测、模式整合、聚类、RNA 和蛋白质表达差异分析、抗体衍生标签评估、配体-受体相互作用分析以及 CITE-seq 数据的交互式网络可视化。团队利用模拟和实际 CITE-seq 数据证明了 CiteFuse 整合两种数据模态 RNA 和蛋白表达的能力，以及与单模式剖析生成的数据相比的相对优势。CiteFuse 代表

了第一个专门设计用于系统地整合 CITE-seq 数据中单细胞 RNA 和 ADT 模式的方法。

9.5 转录组学、蛋白组学与表观遗传学数据融合

9.5.1 基于神经网络方法

美国斯坦福大学的 Howard Y. Chang 和 James Zou 团队提出了 BABEL[30]，可以在单个细胞的转录组和染色质配置文件之间进行转换。借助一种新颖的可互操作的神经网络模型，BABEL 可以直接从细胞的 scATAC-seq 生成 scRNA-seq，反之亦然。这使得在只有一种模态实验可用时，可以计算合成配对的多组学测量。该团队进一步展示了 BABEL 可以整合额外的单细胞数据模态，例如 CITE-seq，从而实现染色质、RNA 和蛋白质之间的转换。BABEL 为数据探索和假设生成提供了一个强大的方法。

澳大利亚悉尼大学 Y. X. Rachel Wang、美国斯坦福大学 Wing H. Wong 等研究人员提出了 scJoint[31]，用于整合规模化的、异质的单细胞 RNA 测序(scRNA-seq)和单细胞 ATAC 测序(scATAC-seq)数据集。scJoint 利用注释的 scRNA-seq 数据中的信息，在半监督框架下使用神经网络同时训练标记和未标记数据，实现标签传递和联合可视化。使用图谱数据以及由 ASAP-seq 和 CITE-seq 生成的多模态数据集，研究人员证明了 scJoint 在计算效率上的优势，并且始终达到比现有方法更高的细胞类型标签准确性，同时提供有意义的联合可视化。

军事医学研究院应晓敏团队和伯晓晨团队提出了一种用于单细胞多模态数据的拼图式整合和知识传递的深度概率框架 MIDAS[32]，通过使用自监督模态对齐和信息论潜在解缠方法，同时实现了拼图数据的降维、插补和批次校正。通过评估其在三模态和拼图整合任务中的性能，展示了其优越性以及可靠性。此外，该团队还构建了一个人类外周血单核细胞的单细胞三模态图谱，并制定了定制的迁移学习和

相互参考映射方案，以实现从图谱到新数据的灵活和准确的知识传递。在骨髓拼图数据集上的拼图整合、伪时分析和跨组织知识传递等应用展示了 MIDAS 的多功能性和优越性。

9.5.2 基于矩阵分解方法

香港城市大学汪建平和李帅成团队提出了一个概率张量分解框架 SCOIT[33]，用于从单细胞多组学数据中提取嵌入。SCOIT 包含各种分布，包括高斯、泊松和负二项分布，以处理稀疏、嘈杂和异质的单细胞数据。该框架可以将多组学张量分解为细胞嵌入矩阵、基因嵌入矩阵和组学嵌入矩阵，从而实现各种下游分析。通过基因嵌入，SCOIT 实现了跨组学基因表达分析和整合基因调控网络研究。此外，嵌入还允许同时进行跨组学插补，可用于仅具有一种组学剖面的细胞子集的情景。

9.5.3 基于图/网络方法

山东大学刘丙强、密苏里大学 Dong Xu 和俄亥俄州立大学 Qin Ma 团队提出了 DeepMAPS[34]，用于从 scMulti-omics 中推断生物网络。DeepMAPS 将 scMulti-omics 建模为一个异质图，并使用多头图变换器在局部和全局上下文中稳健地学习细胞和基因之间的关系。它展示了在肺部肿瘤白血病细胞 CITE-seq 数据和匹配的弥漫性小淋巴细胞淋巴瘤 scRNA-seq 和 scATAC-seq 数据中推导出细胞类型特异性生物网络的竞争能力。此外，该团队部署了一个带有多功能和可视化功能的 DeepMAPS Web 服务器，以提高 scMulti-omics 数据分析的可用性和可重复性。

美国密西根州立大学 Jiliang Tang 教授团队针对单细胞的三个关键任务：模态预测、模态匹配和联合嵌入，提出了一种通用图神经网络 scMoGNN[35]。该方法对单细胞的不同模态分别进行建模，根据单细胞测序数据构建 GNN 网络，将不同组学的生物知识添加到图网络中作为额外的结构性信息，从而捕捉细胞和模态之间的高阶结构关系。

同时，该方法表现出高度灵活性，可在不同模式的单细胞任务进行扩展使用，有效解决传统的单细胞数据整合技术的局限性。

南开大学张瀚教授和腾讯 AI Lab 姚建华团队提出了 scMHNN[36]，该模型基于超图神经网络来整合单细胞多组学数据开发。在对各种模态之间的复杂数据关联建模后，scMHNN 在多组学超图上执行消息传递过程，可以捕获高阶数据关系并整合多种异质特征。随后，scMHNN 通过自监督方式中的双对比损失学习可辨别的细胞表示。基于预训练的超图编码器，该团队进一步引入了预训练和微调范式，只需少量标记的细胞作为参考，即可实现更准确的细胞类型注释。

中南大学李敏团队提出了一种单细胞多模式 Louvain 聚类框架 scMLC[37]。scMLC 构建了多重单模态和跨模态的细胞网络，以捕获模态特异性和模态间的一致信息，然后采用稳健的多重社区检测方法来获得可靠的细胞簇。此外，scMLC 具有灵活性，可以扩展到具有两种以上模态的单细胞测序数据。

青岛科技大学于彬教授团队提出了一个基于图卷积网络的单细胞多组学数据整合通用框架 GCN-SC[38]。在多个单细胞数据中，GCN-SC 通常选择具有最多细胞数的数据作为参考数据集，其余数据作为查询数据集。它利用相互最近邻算法识别细胞对，为参考和查询数据集内部和之间的细胞提供连接。进一步采用 GCN 算法，利用这些细胞对构建的混合图来调整查询数据集的计数矩阵。最后，在可视化之前使用非负矩阵分解进行降维。

参考文献

[1] Athaya T, Ripan R C, Li X, 等. Multimodal deep learning approaches for single-cell multi-omics data integration[J]. Briefings in Bioinformatics, 2023, 24(5): bbad313.

[2] Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data[J]. Briefings in Bioinformatics, 2021, 22(4): bbaa287.

[3] Zuo C, Dai H, Chen L. Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data[J]. A. Mathelier. Bioinformatics, 2021, 37(22): 4091–4099.

[4] Li G, Fu S, Wang S, 等. A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data[J]. Genome Biology, 2022, 23(1): 20.

[5] Zhao J, Wang G, Ming J, 等. Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets[J]. Nature Computational Science, 2022, 2(5): 317–330.

[6] Xiong L, Tian K, Li Y, 等. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space[J]. Nature Communications, 2022, 13(1): 6118.

[7] Wang X, Hu Z, Yu T, 等. Con-AAE: contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration[J]. A. Mathelier. Bioinformatics, 2023, 39(4): btad162.

[8] Li Y, Zhang D, Yang M, 等. scBridge embraces cell heterogeneity in single-cell RNA-seq and ATAC-seq data integration[J]. Nature Communications, 2023, 14(1): 6045.

[9] Yuan M, Chen L, Deng M. Clustering single-cell multi-omics data with MoClust[J]. A. Mathelier. Bioinformatics, 2023, 39(1): btac736.

[10] Ren L, Wang J, Li Z, 等. scMCs: a framework for single-cell multi-omics data integration and multiple clusterings[J]. J. Wren. Bioinformatics, 2023, 39(4): btad133.

[11] Cao Y, Fu L, Wu J, 等. Integrated analysis of multimodal single-cell data with structural similarity[J]. Nucleic Acids Research, 2022, 50(21): e121–e121.

[12] Zhang Z, Yang C, Zhang X. scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously[J]. Genome Biology, 2022, 23(1): 139.

[13] Cui H, Wang C, Maan H, 等. scGPT: toward building a foundation model for single-cell multi-omics using generative AI[J]. Nature Methods, 2024.

[14] Jin S, Zhang L, Nie Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles[J]. Genome Biology, 2020, 21(1): 25.

[15] Yao Y, Mao S, Zhang N, 等. Schema-aware Reference as Prompt Improves Data-Efficient Knowledge Graph Construction[A]. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Taipei Taiwan: ACM, 2023: 911–921.

[16] Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding[J]. Nature Biotechnology, 2022, 40(10): 1458–1466.

[17] Dou J, Liang S, Mohanty V, 等. Bi-order multimodal integration of single-cell data[J]. Genome Biology, 2022, 23(1): 112.

[18] Wu W, Zhang W, Ma X. Network-based integrative analysis of single-cell transcriptomic and epigenomic data for cell types[J]. Briefings

in Bioinformatics, 2022, 23(2): bbab546.

[19] Wang X, Duan M, Li J, 等. MarsGT: Multi-omics analysis for rare population inference using single-cell graph transformer[J]. Nature Communications, 2024, 15(1): 338.

[20] Han X, Wang B, Situ C, 等. scapGNN: A graph neural network–based framework for active pathway and gene module inference from single-cell multi-omics data[J]. S. Huang. PLOS Biology, 2023, 21(11): e3002369.

[21] Lin X, Tian T, Wei Z, 等. Clustering of single-cell multi-omics data with a multimodal deep learning method[J]. Nature Communications, 2022, 13(1): 7705.

[22] Qiao Y, Yang R, Liu Y, 等. DeepFusion: A deep bimodal information fusion network for unraveling protein-RNA interactions using in vivo RNA structures[J]. Computational and Structural Biotechnology Journal, 2024, 23: 617–625.

[23] Zou G, Lin Y, Han T, 等. DEMOC: a deep embedded multi-omics learning approach for clustering single-cell CITE-seq data[J]. Briefings in Bioinformatics, 2022, 23(5): bbac347.

[24] Yuan M, Chen L, Deng M. Clustering CITE-seq data with a canonical correlation-based deep learning method[J]. Frontiers in Genetics, 2022, 13: 977968.

[25] Lakkis J, Schroeder A, Su K, 等. A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation[J]. Nature Machine Intelligence, 2022, 4(11): 940–952.

[26] Hu H, Liu R, Zhao C, 等. CITEMO ^{XMBD} : A flexible single-cell multimodal omics analysis framework to reveal the heterogeneity of

immune cells[J]. RNA Biology, 2022, 19(1): 290–304.

[27] Wang X, Sun Z, Zhang Y, 等. BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data[J]. Nucleic Acids Research, 2020, 48(11): 5814–5824.

[28] Wang X, Xu Z, Hu H, 等. SECANT: a biology-guided semi-supervised method for clustering, classification, and annotation of single-cell multi-omics[J]. S. Yooseph. PNAS Nexus, 2022, 1(4): pgac165.

[29] Kim H J, Lin Y, Geddes T A, 等. CiteFuse enables multi-modal analysis of CITE-seq data[J]. A. Mathelier. Bioinformatics, 2020, 36(14): 4137–4143.

[30] Wu K E, Yost K E, Chang H Y, 等. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution[J]. Proceedings of the National Academy of Sciences, 2021, 118(15): e2023070118.

[31] Lin Y, Wu T-Y, Wan S, 等. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning[J]. Nature Biotechnology, 2022, 40(5): 703–710.

[32] He Z, Hu S, Chen Y, 等. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS[J]. Nature Biotechnology, 2024.

[33] Wang R H, Wang J, Li S C. Probabilistic tensor decomposition extracts better latent embeddings from single-cell multiomic data[J]. Nucleic Acids Research, 2023, 51(15): e81–e81.

[34] Ma A, Wang X, Li J, 等. Single-cell biological network inference using a heterogeneous graph transformer[J]. Nature Communications, 2023, 14(1): 964.

[35] Wen H, Ding J, Jin W, 等. Graph Neural Networks for Multimodal Single-Cell Data Integration[A]. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining[C]. Washington DC USA: ACM, 2022: 4153–4163.

[36] Li W, Xiang B, Yang F, 等. scMHNN: a novel hypergraph neural network for integrative analysis of single-cell epigenomic, transcriptomic and proteomic data[J]. Briefings in Bioinformatics, 2023, 24(6): bbad391.

[37] Chen Y, Zheng R, Liu J, 等. scMLC: an accurate and robust multiplex community detection method for single-cell multi-omics data[J]. Briefings in Bioinformatics, 2024, 25(2): bbae101.

[38] Gao H, Zhang B, Liu L, 等. A universal framework for single-cell multi-omics data integration with graph convolutional networks[J]. Briefings in Bioinformatics, 2023, 24(3): bbad081.