



单细胞扰动建模研究进展

齐晓宁^{1,2}, 赵连鹤¹, 陈润生³, 赵屹^{1*}

1. 中国科学院计算技术研究所, 泛在系统中心实验室, 北京 100190

2. 中国科学院大学, 北京 100049

3. 中国科学院生物物理研究所, 北京 100101

* 联系人, E-mail: biozy@ict.ac.cn

2025-05-15 收稿, 2025-07-04 修回, 2025-08-26 接受, 2025-08-26 网络版发表

国家重点研发计划(2022YFF1203303)、国家自然科学基金(92474204, 32341019, 32070670)、宁波市医疗卫生高端团队重大攻坚项目(2023030615, 2024020919)、北京市自然科学基金(L222007)、宁波市科技计划(2023Z226, 2024Z229)、广州国家实验室专项(GZNL2023A03001)和肿瘤系统医学全国重点实验室资助课题(KF2422-93)资助

摘要 单细胞扰动技术为解析细胞在单细胞分辨率下对扰动的响应机制提供了强有力的手段, 广泛应用于基因功能研究、疾病机制解析及新药研发等领域. 扰动建模的核心目标在于揭示扰动对细胞状态的影响, 其任务涵盖扰动效应解析; 扰动响应预测; 扰动组合预测和扰动生成与设计等方面. 近年来, 人工智能驱动的建模方法不断涌现, 极大推动了单细胞扰动建模的发展. 本研究聚焦该研究领域, 系统梳理并综述了单细胞扰动建模的核心任务与研究进展. 在数据方面, 汇总了32种单细胞扰动技术(涵盖遗传扰动和化学扰动等)及54个包含多组学、多模态信息的数据集. 在方法方面, 根据扰动建模的任务分类, 回顾了代表性的计算模型, 涵盖从传统统计推断方法到基于机器学习的预测模型, 再到近年来广泛应用的图神经网络模型、深度生成模型, 包括变分自编码器、最优传输模型以及受到大语言模型启发而发展的通用基础模型等前沿方法. 最后, 本文对当前单细胞扰动建模所面临的挑战进行了批判性评估, 并展望了未来的发展方向与研究前景, 如构建系统性的扰动图谱、发展通用型扰动大模型、融合多模态扰动建模, 以及虚拟细胞等前沿探索.

关键词 单细胞扰动, 扰动建模, 单细胞组学, 扰动响应, 人工智能

人体由约37万亿个细胞构成^[1,2], 这些细胞在应对内源性调控或外部刺激(如基因突变、药物作用、环境变化等)时展现出复杂而动态的响应机制. 深入理解细胞在扰动条件下的适应性变化, 不仅有助于揭示基因调控网络和疾病发生机制, 还为药物研发和精准治疗提供理论基础和技术支撑. 扰动实验旨在建立扰动与响应之间的因果联系, 探究多尺度生命系统如何应对外部条件变化, 通常通过遗传扰动、小分子处理或环境调控等方式开展系统性筛选研究.

传统扰动筛选研究受限于技术通量, 通常仅在群体水平进行, 难以捕捉细胞异质性及扰动效应的细粒

度特征^[3-5]. 近年来, 单细胞和空间组学技术的迅速发展显著提升了生物研究的分辨率, 使得单细胞水平上对细胞的表型进行深度和广度的多模态高内涵表征成为可能^[6,7]. 2016年, 以单细胞CRISPR(single-cell clustered regularly interspaced short palindromic repeat, scCRISPR)为代表的技术将CRISPR(clustered regularly interspaced short palindromic repeat)筛选与单细胞RNA测序(single cell RNA sequencing, scRNA-seq)相结合, 以实现大规模单细胞分辨率的混合遗传扰动筛选^[6,8,9]. 此外, 2019年, Srivatsan等人^[10]提出了sci-Plex, 它将化学筛选与scRNA-seq相结合, 经济高效地同时定量筛选

引用格式: 齐晓宁, 赵连鹤, 陈润生, 等. 单细胞扰动建模研究进展. 科学通报

Qi X, Zhao L, Chen R, et al. Advances in single-cell perturbation modeling (in Chinese). Chin Sci Bull, doi: 10.1360/CSB-2025-0682

上百种化学物扰动三种细胞系的转录响应。单细胞扰动筛选技术将扰动(主要包括遗传扰动和化学扰动)与单细胞测序(single-cell sequencing)结合,能够捕捉扰动细胞的染色质可及性^[11]、基因表达^[6]、蛋白水平^[12]和形态学^[13]等高内涵表征(图1),系统描绘细胞状态及其扰动响应。这些实验的逐步规模化促进了大规模扰动及其组合的扰动响应表征,描绘了细胞状态在扰动条件下的适应性变化的全景图谱。

单细胞扰动技术的发展与数据积累使得数据驱动的扰动建模迅速发展。扰动建模的核心任务涵盖扰动效应解析、扰动响应预测、扰动组合预测及扰动生成与设计等方面。具体而言,扰动效应解析有助于揭示扰动对基因^[6,14]、基因模块^[15,16]到细胞水平^[17,18]的影响;扰动响应预测可以预测未观测扰动条件下的组学^[19,20]或表型响应;扰动组合预测推断基因或药物组合的协同或拮抗作用^[21,22];扰动生成与设计通过优选和生成优化实验设计。近年来,单细胞扰动建模工具呈现爆发式增长,推动研究范式由实验驱动向数据驱动转变。多种建模范式相继涌现,包括统计方法^[16,23]、变分自编码器^[19,24]、图神经网络^[20,21]、最优传输模型^[25,26],以及通用基础模型^[27,28]等,这些方法为解析和预测复杂扰动效应、揭示基因调控网络与新药开发和疾病治疗策略提供了有力支持。

单细胞扰动建模涉及药理学、生物学、数学和计算机科学等多学科的交叉融合、数据繁杂、建模任务多样但缺乏基准,阻碍了科学界广泛地研究和工具使用。基于以上需求和发展趋势,本研究聚焦于单细胞扰动建模领域(图1),综述了单细胞扰动建模的核心任务,系统汇总单细胞扰动技术以及数据集,回顾代表性建模方法,最后对扰动建模进行总结,并展望了未来的发展方向与研究前景。

1 扰动建模关键任务

扰动建模旨在揭示扰动对细胞状态和功能的影响,进而服务于基因功能解析、疾病机制研究与新药发现等生物医学核心问题。为系统梳理该领域的研究任务,本文在整合Ji等人^[29]、Gavrilidis等人^[30]和Rood等人^[31]对扰动建模目标的分析基础上,结合当前单细胞扰动实验技术、数据类型与建模方法的发展现状,将扰动建模归纳为以下四个关键任务:(1)扰动效应解析;(2)扰动响应预测;(3)扰动组合预测;(4)扰动生成与设计。

1.1 扰动效应解析

扰动效应解析是单细胞扰动实验中的核心任务之一,旨在多层次、系统性地揭示扰动对细胞系统所产

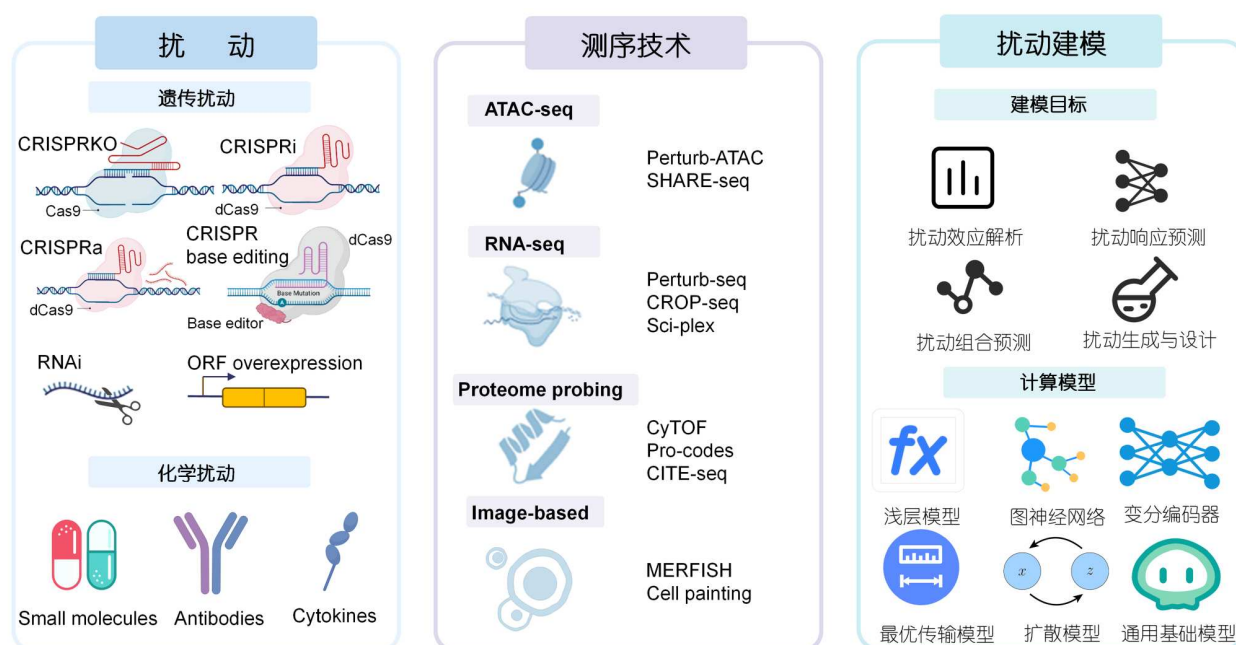


图1 单细胞扰动技术及扰动建模示意图。部分图标来源于BioRender.com和Iconfont

Figure 1 Schematic diagram of single cell perturbation technology and perturbation modeling. Some icons are sourced from BioRender.com and Iconfont

生的影响,系统地量化并理解扰动对基因、基因模块到细胞水平所产生的影响,进行因果推断.通过比较扰动组与对照组在转录组水平的差异,结合算法建模,研究者可以揭示基因编辑或药物处理带来的分子变化,进而阐明其生物学功能与作用机制.在基因和基因模块层面,解析的重点是识别扰动引起的显著差异表达基因、直接或间接作用的靶点基因,以及关键的转录调控因子,挖掘涉及的生物通路和信号转导机制,推断潜在的调控路径,并识别能逆转疾病相关异常表达图谱的干预手段^[6,15,18];在细胞层面,则关注细胞的异质性,解析扰动对细胞群体组成和状态的影响,包括不同细胞亚群的丰度变化、命运转变,以及特定亚群对扰动的选择性响应等^[18].

1.2 扰动响应预测

扰动响应预测是当前扰动建模领域中最为核心且研究广泛的任务之一,旨在预测细胞在受到特定扰动(如药物处理、基因编辑)后所发生的组学变化(如转录组、蛋白组)以及表型变化(如药物敏感性、细胞状态转变等)^[22,32].其中组学层面包括:(1)组学响应重建,即给定扰动条件,重建完整的扰动响应谱^[25,33]; (2)扰动响应表型预测,即预测扰动诱导下的细胞状态、分类、药物敏感性标签等表型特征^[22,32]; (3)分布外(out-of-distribution, OOD)组学响应预测,即预测泛化到未观测扰动条件(如时间、基因、化合物、化合物剂量等)或细胞类型的细胞响应^[19,34].该任务不仅有助于解析扰动对细胞命运的影响,是理解药物作用机制和筛选潜在治疗靶点的重要途径,同时也具备显著的实用价值,可在无需大规模实验的前提下,预测未观测扰动或细胞类型下的反应,提升研究效率并降低成本.

1.3 扰动组合预测

扰动组合预测关注多个扰动因素(如药物-药物、基因-基因)之间的协同或拮抗作用预测,是理解复杂调控网络与实现组合扰动优化的关键步骤.该任务通常被建模为分类问题(如协同vs非协同)或回归问题(预测组合干预的敏感性评分),广泛应用于联合用药策略优化、抗药性机制研究以及基因调控图谱构建等场景^[21,35].扰动组合效应建模的具体任务有预测药物组合的协同或拮抗效应^[36],以及推断基因对之间的功能性相互作用类型,如协同、上位作用、冗余或抑制^[21]等.

1.4 扰动生成与设计

扰动生成与设计任务旨在庞大且高维的扰动空间中,基于特定的生物学目标或期望的细胞表型,智能地筛选、生成或优化扰动条件(如药物、基因或其组合),从而提升实验效率并实现更具干预价值的生物效应.与传统的穷举式实验不同,该任务依托于主动学习或生成建模等策略,结合已有扰动响应数据,迭代更新模型、引导实验设计或直接生成符合目标的候选扰动,为药物发现、个性化治疗和细胞命运调控等应用提供数据驱动的决策支持.该方向涵盖典型子任务如:在有限实验预算下的扰动优选(active intervention selection)^[37,38]、基于目标表型的逆向分子设计^[39,40](转录特征驱动的药物生成)、组合扰动优化^[41,42]等,是实现从建模到可控干预闭环的关键环节.

2 单细胞扰动技术及数据

高质量且大规模的单细胞扰动数据是单细胞扰动建模研究的基础保障.单细胞扰动技术通过在单细胞分辨率下施加基因或化学扰动,记录扰动条件下细胞的响应,从而以高通量、高精度的方式揭示基因功能、作用机制及药物响应特征,为功能基因组学与药物研发提供了关键的数据支持.根据所施加扰动类型的不同,现有的单细胞扰动实验技术及数据集可大致分为两大类:单细胞遗传扰动与单细胞化学扰动,分别对应基因编辑与小分子药物处理所引发的细胞状态变化.这两类扰动在技术路径、数据特征及应用目标上各具特色(表1),共同构成了单细胞尺度下扰动研究的核心基础.

2.1 单细胞遗传扰动技术

CRISPR技术是功能基因组学研究不可或缺的工具^[69].2016年,单细胞CRISPR筛选(scCRISPR sequencing)实现了大规模单细胞分辨率的混合遗传扰动筛选^[6,8,9].其核心技术创新在于单向导RNA(single guide RNA, sgRNA)的设计,通过引入特异性条形码或可检测序列的sgRNA,实现对每个细胞扰动的高效注释.根据扰动的机制,遗传扰动主要被分类为敲除(knockout, CRISPRKO)、敲减(knockdown, CRISPRi)、激活(activation, CRISPRa)、位点突变(mutation, CRISPR base editing)和组合扰动(combinatorial perturbation),实现对特定基因序列的高效编辑.除CRISPR技术外,开放阅

表 1 代表性单细胞扰动技术汇总

Table 1 Indicative examples of single-cell perturbation methods

扰动类别	扰动技术	组学模态	扰动机制	扰动鉴别方式	单细胞平台	发布年份	参考文献
遗传扰动	Perturb-seq	RNA	CRISPRi、CRISPRKO	Barcode	10×Chromium	2016	[6]
	CRISP-seq	RNA	CRISPRKO	Barcode	MARS-seq	2016	[9]
	Mosaic-seq	RNA	CRISPRi	Barcode	Drop-seq	2017	[43]
	CROP-seq	RNA	CRISPRKO	Polyadenylated sgRNA	Drop-seq	2017	[44]
	POKI-seq	RNA	CRISPR knock-in	Barcode	10×Chromium	2020	[45]
	TAP-seq	RNA	CRISPRi	Polyadenylated sgRNA	10×Chromium	2020	[46]
	Direct-seq	RNA	CRISPRKO、CRISPRa	sgRNA	10×Chromium/ Fluidigm C1	2020	[47]
	Direct-capture Perturb-seq	RNA	CRISPRKO、CRISPRi、 CRISPRa	sgRNA	10×Chromium	2020	[48]
	scifi-RNA-seq	RNA	CRISPRKO	Polyadenylated sgRNA	10×Chromium	2021	[49]
	FiCS Perturb-seq	RNA	CRISPRi	Barcode	10×Chromium	2025	[50]
	SEUSS	RNA	ORF overexpression	ORF barcode	Droplet-based scRNA-seq	2018	[51]
	Perturb-ATAC	ATAC	CRISPRi、CRISPRKO	Barcode	Fluidigm C1	2019	[11]
	Spear-ATAC	ATAC	CRISPRi、CRISPRKO	sgRNA in genome	10×Chromium	2019	[52]
	CRISPR-sciATAC	ATAC	CRISPRKO	Polyadenylated sgRNA	Combinatorial indexing	2021	[53]
	Pro-Codes	Protein	CRISPRKO	Barcode	CyTOF	2018	[54]
	Imaging-based CRISPR screen	Imaging	CRISPRKO	Polyadenylated sgRNA	MERFISH	2019	[55]
	CRAFT-ID	Imaging	CRISPRKO	sgRNA in genome	Microscopy	2020	[56]
	Raft-seq	Imaging	CRISPRKO	cDNA mutation	CellRaft air system, microscope	2022	[57]
	SHAREseq	RNA+ATAC	ORF overexpression	ORF barcode	10×Chromium	2023	[58]
	ECCITE-seq	RNA+Protein	CRISPRKO	sgRNA	10×Chromium	2019	[59]
	Perturb-CITE-seq	RNA+Protein	CRISPRKO	Polyadenylated sgRNA	Drop-seq, flow cytometer	2019	[12]
	OverCITE-seq	RNA+Protein	ORF overexpression	ORF barcode	10×Chromium	2022	[60]
	Perturb-FISH	RNA+Imaging	CRISPRi、CRISPRKO	Amplified sgRNA-based optical barcode	MERFISH	2025	[61]
	Perturb-map	RNA + Protein + Imaging	CRISPRKO	Barcode	CyTOF, MICSSS, MIBI, 10×Visium	2022	[62]
	NIS-seq	RNA + Protein + Imaging	CRISPRKO、CRISPRa、 cDNA overexpression	sgRNA	Custom High-throughput imaging and <i>in situ</i> sequencing platform	2024	[63]
化学扰动	MIX-seq	RNA	Drug treatment	SNP+Cell Hashing	10×Chromium	2020	[64]
	sci-Plex	RNA	Drug treatment	Nuclear Hashing	10×Chromium	2020	[10]
	96-plex scRNA-seq	RNA	Drug treatment	DNA oligonucleotide tagging of cellular proteins	10×Chromium	2020	[65]
	Microwell-based scRNA	RNA	Drug treatment	Barcode	Microwell-based scRNA	2021	[66]
	TraCe-seq	RNA	Drug treatment	Barcode	10×Chromium	2022	[67]
	Mosaic	RNA	Drug treatment	GigaLab	Mosaic	2025	[68]
	JUMP Cell Painting	Imaging	Drug treatment	Unique JUMP identifier	Five types of microscope	2024	[13]

读框过表达(open reading frame overexpression, ORF overexpression)也是常用的遗传扰动手段之一，可用于

激活或过表达目标基因或调控序列，有助于系统性地解析基因序列功能。

近年来,单细胞遗传扰动技术迅速发展,形成了以不同组学层次为表型的多种技术路线。根据整合单细胞组学手段的不同,当前遗传扰动主要分为五类:(1) 基于RNA-seq的扰动技术;(2) 基于ATAC-seq(assay for transposase-accessible chromatin using sequencing)的扰动技术;(3) 基于蛋白质表达探测(proteome probing)的扰动技术;(4) 基于多组学的扰动技术;(5) 基于成像(image-based)的扰动技术。接下来,我们根据单细胞技术的分类简明地介绍了该领域的基本技术(图2(a)),关于扰动测序更详细的技术细节可以参考Cheng等人^[70]的一篇优秀综述。

2.1.1 基于RNA-seq的单细胞遗传扰动技术

基于RNA-seq的scCRISPR-seq是单细胞遗传扰动研究中最主流的策略之一。其核心在于利用慢病毒载体递送包含sgRNA、细胞条形码(cell barcode, CBC)、sgRNA条形码(guide barcode, GBC)及分子标识符(unique molecular identifier, UMI)等模块的构建体。其中,sgRNA用于引导核酸酶靶向目标基因组区域,CBC用于标记每个细胞,而GBC用于标记每个sgRNA,UMI用于标记每个转录本。根据sgRNA的类型对已有方法进行分类,主要分为3种策略:(1) sgRNA特异性条码;(2) 多聚腺苷酸化sgRNA;(3) 带有捕获序列的sgRNA。

早期的开创性技术,如Perturb-seq^[6]和CRISP-seq^[9],首次将CRISPR扰动与scRNA-seq相结合,通过在sgRNA构建中引入GBC,结合单细胞捕获中的CBC,实现了扰动与单细胞转录组的准确关联。然而,该策略依赖病毒共包装,易发生模板切换(template switching),导致sgRNA与条形码错误配对,影响扰动注释准确性,需通过阵列包装或同源转移等额外技术手段加以缓解。X-Atlas/Orion^[50]通过检测剂量依赖性遗传效应推出了“固定-冻存-单细胞测序”FiCS(fix-cryopreserve-scRNA-seq)Perturb-seq平台,揭示了基因活性如何随着特定干预措施的强度而变化。

为解决病毒共包装的依赖问题,CROP-seq^[44]对载体设计进行了优化,创新性地sgRNA的表达置于RNA聚合酶II(Pol II)启动子下,并引入多聚腺苷酸化(polyadenylation)信号,使得sgRNA随mRNA进入标准的scRNA-seq流程中,无需额外GBC设计,从而提升了通用性与平台兼容性。Mosaic-seq^[43]通过引入索引CRISPR构建体,实现对增强子功能扰动及其在单细胞水平下的转录响应测序,拓展了对非编码调控元件功

能的研究能力。

面对全转录组测序在低表达基因或微弱扰动效应下信噪比有限的问题,TAP-seq^[46]在Perturb-seq框架下引入目标基因的特异性富集扩增策略,通过优化扩增引物、降低引物二聚体生成和非特异扩增,显著提高了灵敏度和检测可靠性。为进一步提高通量、降低成本,scifi-RNA-seq^[71]开发了单细胞组合流体索引技术,在细胞透化后先进行预索引,再通过微滴分离进行测序。预索引允许为每个液滴加载多个细胞,并通过计算解复用单个表达谱,使单实验可处理数千个样本,大幅提升了样本通量和成本效益。

为进一步提升sgRNA捕获效率,direct capture Perturb-seq^[48]通过在sgRNA构建体中嵌入可识别的捕获序列(capture sequence, CS),结合互补引物实现sgRNA的特异性扩增与直接捕获。但该方法依赖10× Genomics平台特有的探针和引物设计,难以兼容其他测序平台。为提升方法的通用性,Direct-seq^[47]进一步将原有CS优化为可与poly(dT) RT引物退火的A/G混合序列,从而支持不同测序平台下的sgRNA捕获,显著增强了方法的适应性与可扩展性。

除CRISPR技术外,SEUSS^[51]基于发育关键转录因子(transcription factors, TFs)ORF过表达筛选,scRNA-seq和适应度评价,系统解析关键TF对人类多能干细胞(human pluripotent stem cells, hPSCs)转录组与功能状态的影响,拓展了对发育调控网络的理解。

总体来看,基于单细胞遗传扰动技术体系不断演进,围绕检测灵敏度、通量、平台兼容性 & 文库构建便捷性等关键问题持续优化,已成为高通量功能基因组学研究的关键支撑技术。

2.1.2 基于ATAC-seq的单细胞遗传扰动技术

为探索遗传扰动对染色质可及性的影响,研究者开发了多种整合ATAC-seq的scCRISPR-seq方法,实现了单细胞分辨率的染色质可及性分析。Perturb-ATAC^[11]首次将多重CRISPRKO或CRISPRi与scATAC-seq相结合,能够同时检测sgRNA信息与全基因组开放染色质区域。但是Perturb-ATAC存在通量低、PCR(polymerase chain reaction)中片段未分离及背景噪音高等问题。为提升实用性,Spear-ATAC^[52]对sgRNA结构进行了优化,引入预整合接头和生物素标记引物,使sgRNA片段能被特异性富集,有效降低了背景噪音,同时显著提升了细胞通量并降低了成本。另一种方法CRISPR-sciATAC^[53]则采用条形码组合索引策略“细胞

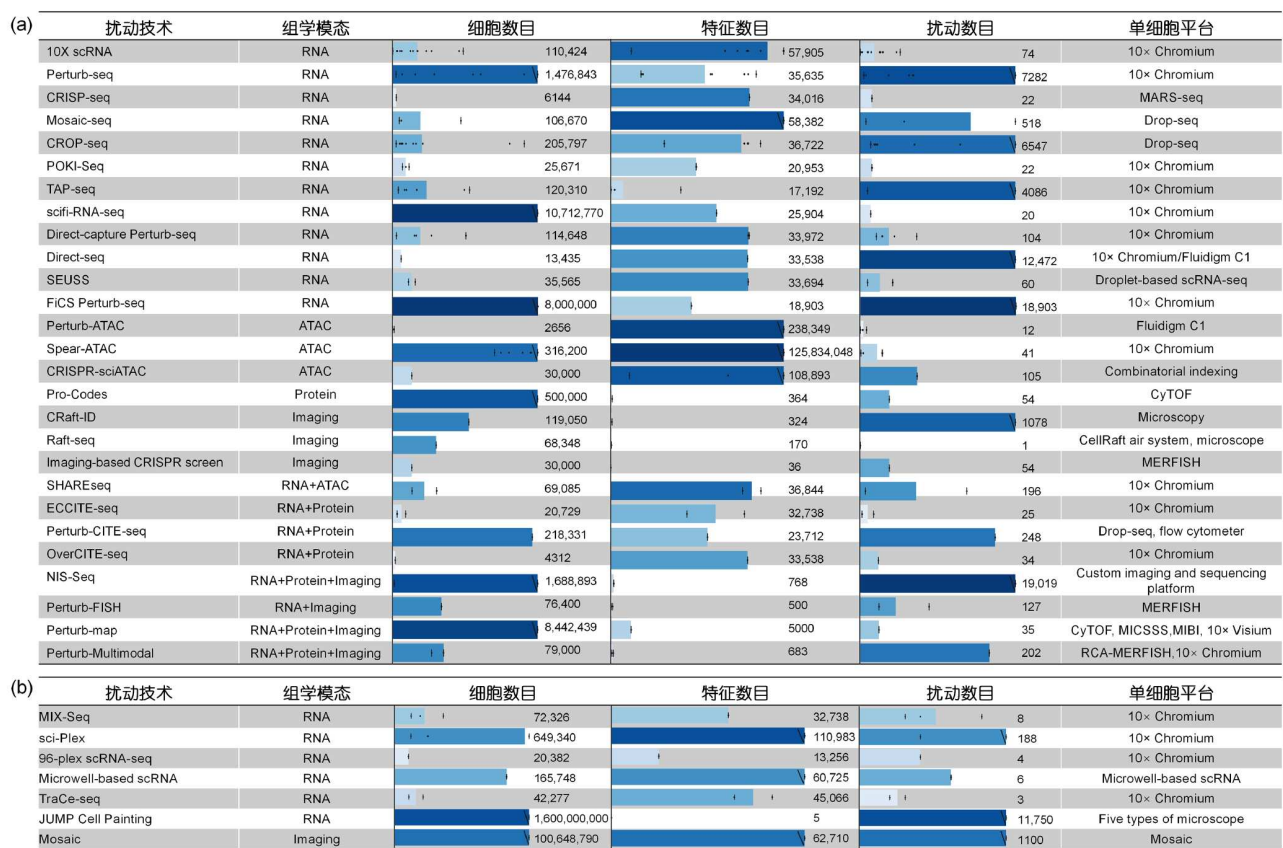


图2 单细胞扰动技术对比图。(a) 单细胞遗传扰动技术对比图。(b) 化学扰动技术对比图。条形图的长度代表各项技术对应数据的均值，竖线表示最小值和最大值，散点表示每个样本的数值。为便于显示，过长的条形图被斜线截断，标注的数字代表该技术对应数据的最大值

Figure 2 Comparison of single-cell perturbation techniques. (a) Comparison of single-cell genetic perturbation technologies. (b) Comparison of single-cell chemical perturbation technologies. The length of the bar represents the average value of each metric for each technique. The vertical lines represent the minimum and maximum values. The scattered points represent individual sample values. For clarity, bars that are too long are truncated with a slash, and the maximum value for that technique is marked with a number

条形码(cell barcode)”，能够同时标记sgRNA和ATAC片段，实现了高通量sgRNA捕获与单细胞表观组信息联合测定。该策略在通量与成本上与Spear-ATAC相当，但在解析微弱染色质变化方面存在一定局限，需结合液滴平台或更高分辨率算法进一步优化。整体而言，该类技术显著拓展了scCRISPR在表观遗传调控机制研究中的应用潜力。

2.1.3 基于Proteome probing的单细胞遗传扰动技术

基于Proteome probing的scCRISPR技术通过检测细胞表面或胞内蛋白水平的变化，弥补了传统仅依赖转录组数据的局限，尤其适用于解析信号转导、免疫应答等复杂表型。目前常见的策略主要包括两类：一种是通过串联表位构建蛋白条形码(Pro-Codes)，在CyTOF(cytometry by time of flight)中与细胞表面蛋白共同检测；另一种是通过DNA偶联抗体标记细胞膜蛋白

后进行微流控单细胞分离并进行联合测序。

Pro-Codes^[54]是首个支持大规模单细胞遗传扰动的蛋白条码系统，通过合成编码10种不同表位的DNA序列进行三元组合，共生成120种可区分的蛋白条码，并融合dNGFR(a truncated receptor without an intracellular domain)以作为Pro-Codes。该系统与CRISPR gRNA共表达，借助质谱流式细胞仪CyTOF实现扰动信息与蛋白表型的单细胞关联，具备高扩展性和高通量优势。Perturb-map^[62]进一步整合空间蛋白组学和空间转录组，实现对肿瘤微环境中扰动效应的空间解析。然而，Pro-Codes的检测通量受限，不适用于超大规模的基因筛选实验。

为提高检测通量和通用性，研究者提出基于抗体条码测序的策略。Perturb-CITE-seq^[12]利用带有寡核苷酸条码的抗体标记细胞表面蛋白，结合sgRNA信息构

建联合文库, 实现扰动与蛋白表型的同时检测. 细胞哈希(cell hashing)^[72]则通过标记普遍表达的表面蛋白以引入样本特异性条码, 支持多样本混合测序及双细胞识别. 这两种策略常被联合用于提高蛋白水平扰动效应的解析精度与样本通量.

2.1.4 基于多组学的单细胞遗传扰动技术

在CITE-seq(cellular indexing of transcriptomes and epitopes)基础上, ECCITE-seq^[59]整合了转录组、T细胞受体(T-cell receptor, TCR)、蛋白质、样本标签和sgRNA, 实现多模态数据的联合检测; Perturb-CITE-seq^[12]结合Perturb-seq与CITE-seq, 实现sgRNA、信使核糖核酸(messenger RNA, mRNA)与表面蛋白的同时检测. 这些多组学方法为解析遗传扰动对蛋白表达和功能状态的影响提供了强有力工具, 拓展了scCRISPR的应用深度与广度. OverCITE-seq^[60]为兼容ORF过表达筛选而设计, 可在单细胞层面同时读取外源ORF的表达信息及细胞的转录组和表面蛋白特征. Perturb-Multimodal^[73]通过整合CRISPR混合筛选与单细胞RNA测序及RNA、蛋白质和亚细胞形态的多重成像实现多模态数据的联合检测.

Joung等人^[58]设计了一个包含所有注释的人转录因子亚型的条形码ORF文库, 利用SHARE-seq技术^[74]实现对ORF驱动的染色质可及性与基因表达联动的单细胞图谱绘制. SHARE-seq^[74]建立在单细胞多组学技术Paired-seq^[75]基础上, 利用多轮添加条形码的方式测量同一单细胞中的染色质可及性和基因表达, 具备高度可扩展性.

2.1.5 基于成像的单细胞遗传扰动技术

基于成像的scCRISPR技术通过结合CRISPR扰动与高分辨率光学成像, 突破了传统测序技术在形态与空间维度上的限制, 支持对细胞形态、空间定位及动态行为的精细表征. 这类方法大致可分为两类: 一类基于光学条码识别技术, 包括原位测序(*in situ* sequencing, ISS)、多轮荧光原位杂交(multiplexed error-robust fluorescence *in situ* hybridization, MERFISH), 以及荧光原位杂交(fluorescence *in situ* hybridization, FISH); 另一类结合物理分选与成像, 如microraft平台对目标细胞进行成像后筛选并测序.

在光学条码策略中, Feldman等人^[76,77]开发的ISS技术利用原位逆转录、padlock探针连接与滚环扩增(rolling circle amplification, RCA), 实现了图像与扰动身份的同步读取. Wang等人^[55]基于MERFISH提出ima-

ging-based CRISPR screen结合构建的可匹配sgRNA的条形码编码, 可实现sgRNA身份判别及空间表达模式分析. Perturb-FISH^[61]结合MERFISH与T7启动子驱动的sgRNA条形码转录扩增, 在细胞内实现sgRNA与转录组的共成像, 增强了短序列检测灵敏度与空间解析分辨率. NIS-seq^[63]则在sgRNA载体中插入T7启动子, 使得sgRNA可被大量转录, 生成的RNA拷贝经逆转录、锁式探针(padlock probe)延伸与滚环扩增后, 在原位实现高灵敏度信号放大, 从而精准定位扰动信息并结合空间转录图谱.

基于物理分选与成像的策略如CRaft-ID^[56]和Raft-seq^[57]通过microraft平台对固定或活细胞成像, 并物理分选感兴趣细胞进行后续分析. Raft-seq^[57]特别优化了活细胞成像的载体结构, 从而扩展了对动态细胞表型的记录能力. 然而, 这类方法大多依赖单细胞互补脱氧核糖核酸(complementary DNA, cDNA)的逐一扩增与测序, 通量受限, 扰动识别需依赖基因突变位点而非标准sgRNA条码.

2.2 单细胞化学扰动技术

随着单细胞技术的发展, 研究者将小分子化学扰动与单细胞测序结合, 实现了在高通量条件下对细胞药物响应的系统性解析, 尤其适用于刻画细胞间对药物及其他分子的异质性响应. 与遗传扰动相比, 化学扰动实验通常更侧重于扰动条件组合的规模化设计与细胞身份的高效溯源, 常采用如标签技术、组合索引、自然多态性识别和图像分析等手段, 以实现大规模药物筛选和扰动反应解析(图2(b)).

MIX-seq^[64]是一种利用单核苷酸多态性(single-nucleotide polymorphism, SNP)进行细胞溯源的方法, 能够在无需引入外源标签的前提下识别混合细胞来源. 该方法通过将不同细胞系混合处理, 结合SNP分析进行解复用, 实现了对数百种癌细胞系在单细胞分辨率下的药物响应表征. 结合细胞哈希技术, 还可扩展至处理时间和剂量等变量, 显著提升实验通量. sci-Plex^[10]采用细胞核散列(nuclear hashing)标记策略, 将短单链DNA标签引入细胞核中, 以标记不同扰动条件下的细胞, 从而在单次实验中实现对数千种药物或剂量扰动条件的组合索引, 可在单次实验中同时解析成千上万扰动条件下的转录响应. 该方法极大提升了扰动筛选效率, 适用于大规模药物筛选. Zhang等人^[68]开发了高通量Mosaic平台, 由高度多样化且优化平衡的“细胞村(cell

village)”组成,有效减少批次效应,并以前所未有的规模实现对数千种条件进行单细胞分辨率的并行分析,构建了覆盖50个癌细胞系与1100多种小分子的Tahoe-100M数据集,包含超1亿个单细胞表达谱,显著提升了高通量化学扰动研究的能力。

在组合化学扰动研究中, Gehring等人^[65]通过将识别性寡核苷酸连接至细胞表面蛋白,在固定细胞上实现化学标记,构建了高多路复用的单细胞RNA测序平台。该策略可在一次实验中解析多种药物及剂量组合的扰动效应,为组合药物研究提供了新的手段。

此外, Drop-seq和基于微孔的scRNA-seq平台虽最初并非为化学扰动设计,但凭借其高通量与稳定性被广泛应用于药物处理后的单细胞分析。例如, Aissa等人^[78]利用Drop-seq进行肿瘤药物处理的时间序列研究,揭示了转录动态与耐药性、联合治疗之间的关联。Zhao等人^[66]则将药物刺激与微孔平台整合,分析新鲜肿瘤组织切片的个体化药物响应模式。

TraCe-seq^[67]是一种将可表达条形码与3'端scRNA-seq兼容整合的谱系追踪技术。该方法通过慢病毒将高复杂性条形码整合入细胞基因组,实现克隆水平的追踪与扰动响应监测,适用于研究药物处理过程中异质性群体的适应性响应与命运转换。

在成像方向, Cell Painting技术采用多通道荧光染料标记细胞核、线粒体、内质网等亚细胞结构,结合高内涵成像与图像分析算法对细胞形态特征进行高通量量化。尽管不依赖组学数据,该方法在药物作用机制推断、表型聚类及药物重定位中发挥了重要作用,且常与转录组数据整合使用。为推进该领域标准化, Broad Institute发起JUMP Cell Painting项目^[13],联合多方力量建立覆盖11万种小分子和2万余个遗传扰动条件的标准化细胞形态数据库,已采集超16亿张细胞的表型图像,为后续机制研究与药物筛选提供了宝贵资源。

2.3 单细胞扰动数据集

近年来,伴随单细胞扰动技术的快速发展,多个规模庞大、用途广泛的单细胞扰动数据集相继建立,为研究基因功能及其调控网络、药物作用机制及药物研发提供了丰富资源。以下分别对经典的遗传扰动和化学扰动数据集做简要描述(表2)。

2.3.1 遗传扰动数据集

遗传扰动类数据集中, CRISPR技术驱动的大规模

单细胞筛选数据集最具代表性。Dixit等人^[6]和Adamson等人^[8]基于Perturb-seq分别在K562和BMDC细胞系中进行了近百个基因的CRISPRi实验,绘制了扰动响应的转录图谱,其中BMDC数据集还包含跨时间点的动态数据。此外, Son等人^[86]、Jin等人^[89]和Ursu等人^[90]基于Perturb-seq技术,分别研究了CRISPRKO对2型糖尿病胰岛细胞、大脑细胞以及位点突变对A549肺癌细胞的功能影响。Replogle等人^[102]在超过250万个细胞中进行了全基因组尺度的CRISPRi Perturb-seq,系统描绘了K562和RPE1细胞中数千个基因功能丧失后的复杂表型。Norman等人^[87]则构建了25万个细胞的CRISPRa组合扰动数据集,是基因互作研究的重要资源。

Datlinger等人^[44]提出了CROP-seq技术,在Jurkat T细胞中实现了多基因CRISPRKO筛选与转录响应的联合分析。Shifrut等人^[93]将CROP-seq技术应用于原代人类T细胞的基因CRISPRKO筛选。Tian等人^[92,94]进一步基于CROP-seq在人类诱导多能干细胞来源神经元(human induced pluripotent stem cells (iPSCs)-derived neurons)中开展CRISPRi与CRISPRa筛选,生成包含逾20万细胞的数据集。Gasperini等人^[91]和Hill等人^[95]则实现了K562和MCF10a细胞系的组合基因CRISPRa筛选。

在更高通量的场景中, TAP-seq被用于在K562细胞系中实现基因组规模CRISPRi筛选定量测量,例如Schraivogel等人^[46]利用TAP-seq技术构建了人类增强子扰动-靶基因图谱,涵盖1778个增强子区域,以进行功能验证与靶点筛选。X-Atlas/Orion^[50]基于FiCS Perturb-seq平台发布当前全球最大规模的公共扰动数据库,包含8000000个深度测序细胞,涵盖HCT116和HEK293T两种细胞系基因组规模的剂量分层CRISPRi筛选。另一方面, ORF的过表达筛选被广泛应用于原代细胞^[60],人多能干细胞(human pluripotent stem cells, hPSC)^[51]和胚胎干细胞^[58,80]等具备分化功能的细胞基因过表达筛选中,借助OverCITE-seq^[60]、SHAREseq^[58]、SEUSS^[51]和10×scRNA等平台构建了多组学数据集已覆盖近百万细胞,支持蛋白表型与转录反应的联合检测。此外,基于ECCITE-seq^[7]、CaRPool-seq^[98]、Perturb-CITE-seq^[12]和ECCITE-seq^[59]等多组学技术的应用,进一步拓展了遗传扰动数据在蛋白水平和免疫特征层面的表征能力。

2.3.2 化学扰动数据集

在化学扰动数据集中, sci-Plex^[10]实现了对数千种药物及剂量条件下的单细胞转录组高通量测量。其中sci-Plex3涵盖了A549、K562和MCF7三种细胞系,对

表 2 代表性单细胞扰动数据

Table 2 Indicative examples of single-cell perturbation datasets

扰动类别	扰动技术	组学模态	扰动模型	扰动机制	数据量	多时间点	多剂量	扰动组合	参考文献
遗传扰动	10×scRNA + perturbation	RNA	Embryonic stem cells	ORF overexpression	60997	—	—	—	[58]
		RNA	T cell	CRISPRKO	6932	—	—	—	[79]
		RNA	Embryonic stem cells	ORF overexpression	174942	2 Rounds	—	Yes	[80]
		RNA	T cell	CRISPRa	103805	—	—	Yes	[81]
		RNA	Multiple teratoma cell types	CRISPRKO	98958	—	—	—	[82]
		RNA	LUHMES	CRISPRi	14473	—	—	—	[83]
		RNA	T cell	CRISPRKO	8095	—	—	—	[84]
		RNA	T cell	CRISPRKO	10018	—	—	—	[85]
		RNA	K562	CRISPRi	86111	—	—	Yes (part)	[8]
		RNA	T2D islet cell	CRISPRKO	19665	—	—	Yes	[86]
	Perturb-seq	RNA	K562	CRISPRa	111668	—	—	Yes	[87]
		RNA	human foreskin fibroblasts	CRISPRi	35601	—	—	—	[88]
		RNA	Brain	CRISPRKO	34939	—	—	—	[89]
		RNA	RPE1、K562	CRISPRi	1960923	—	—	—	[89]
		RNA	A549	CRISPR base editing	263423	—	—	—	[90]
		RNA	K562、BMDC	CRISPRi	195990	BMDC: 0 h, 3 h	—	—	[6]
		RNA	K562	CRISPRi	289959	—	—	Yes	[91]
		RNA	iPSC derived neurons	CRISPRi	182790	—	—	—	[92]
	CROP-seq	RNA	T cell	CRISPRKO	52236	—	—	—	[93]
		RNA	iPSC derived neurons	CRISPRi、CRISPRa	53493	—	—	—	[94]
		RNA	MCF10a	CRISPRi	12007	—	—	Yes	[95]
		RNA	Jurkat cells	CRISPRKO	5905	—	—	—	[44]
	Mosaic-seq	RNA	K562	CRISPRi	106670	—	—	Yes	[96]
		RNA	K562	CRISPRi	24142	—	—	Yes	[43]
	TAP-seq	RNA	K562	CRISPRi	320503	—	—	Yes (part)	[46]
	POKI-seq	RNA	T cell	CRISPR knock in	62116	—	—	Yes	[45]
	scifi-RNA-seq	RNA	Jurkat cells	CRISPRKO	20228191	—	—	—	[49]
	Direct-capture Perturb-seq	RNA	K562	CRISPRi、CRISPRa	259673	—	—	Yes (part)	[48]
	Quant-seq	RNA	iPSC	CRISPRi	275708	D14, D21, D28	—	Yes	[92]
	scGOF-seq	RNA	ND islet cell	CRISPRKO	19792	—	—	Yes	[86]
	SEUSS	RNA	hPSC	ORF overexpression	60237	—	—	Yes	[51]
	scRNA-seq CRISPRi	RNA	hESC	CRISPRi	17234	—	—	—	[97]
	FiCS Perturb-seq	RNA	HCT116、HEK293T	CRISPRi	8000000	—	Dose-dependent	—	[50]
	Perturb-ATAC	ATAC	B cell、epidermal keratinocytes	CRISPRi、CRISPRKO	4190	—	—	Yes	[11]

(续表2)

扰动类别	扰动技术	组学模态	扰动模型	扰动机制	数据量	多时间点	多剂量	扰动组合	参考文献
遗传扰动	Spear-ATAC	ATAC	K562、GM12878、MCF7	CRISPRi	4423481	K562: D3, D6, D9, D21	—	Yes	[52]
	CaRPool-seq	RNA+Protein	THP-1	CRISPR RNAi	30707	—	—	Yes	[98]
	ECCITE-seq	RNA+Protein	THP-1	CRISPRKO	20729	—	—	—	[7]
		RNA+Protein	K562	CRISPRKO	7343	—	—	—	[59]
	OverCITE-seq	RNA+Protein	T cell	ORF overexpression	4312	—	—	—	[60]
	Perturb-CITE-seq	RNA+Protein	melanoma cell	CRISPRKO	218331	—	—	Yes	[12]
	SHAREseq	RNA+ATAC	Embryonic stem cells	ORF overexpression	791791	—	—	—	[58]
	ASAP-seq	RNA+ATAC+Protein	T cell	CRISPRKO	10018	—	—	Yes	[99]
	Cell Painting	Imaging	U2OS	CRISPRKO, ORF overexpression	1600000000	—	—	—	[13]
	Perturb-Multimodal	RNA + Protein + Imaging	K562, liver tissue of mice	CRISPRi	10000000	—	—	—	[73]
化学扰动	10×scRNA + perturbation	RNA	T3 murine methylcholanthrene (MCA)-induced sarcoma line	Drug treatment	14618	—	1 dose for each ICI	Yes	[100]
		RNA	MEF	Drug treatment	96680	Every 12 h in 18D	—	—	[101]
	Drop-seq	RNA	PC9、M14	Drug treatment	255116	PC9: D0, D3 PC9: D0, D1, D3 M14: D0, D2, D11, D19 PC9: D0, D11	—	—	[78]
	MIX-seq	RNA	cancer cell lines	Drug treatment	134682	6 h, 24 h	1 dose for each drug	—	[64]
	sci-Plex	RNA	A549、MCF7、K562	Drug treatment	885454	—	Multiple	—	[10]
	96-plex scRNA-seq	RNA	Neural stem cells	Drug treatment	20382	—	Multiple	Yes	[65]
	Microwell-based scRNA	RNA	GBM cell	Drug treatment	165748	—	1 dose for each drug	—	[66]
	Mosaic	RNA	50 cell lines	Drug treatment	100648790	—	Multiple	—	[68]
	TraCe-seq	RNA	PC9	Drug treatment	42277	—	—	—	[67]
	JUMP Cell Painting	Imaging	U2OS	Drug treatment	1600000000	—	—	—	[13]

188种化合物实施4个剂量处理, 总计4608种条件, 为药物机制分析提供了重要资源. MIX-seq^[64]实现了在单细胞分辨率下对上百种癌细胞系在不同时间点或剂量下的化学扰动响应的解析.

在临床样本中, Zhao等人^[66]在胶质母细胞瘤(glioblastoma multiforme, GBM)组织切片中实现了免疫检查点抑制剂和HDAC抑制剂的扰动研究, 揭示了细胞

类型特异性的响应模式. Aissa等人^[78]利用Drop-seq技术, 在非小细胞肺癌(non-small cell lung cancer, NSCLC)临床前模型中, 分析了酪氨酸激酶抑制剂(如厄洛替尼)处理后的单细胞转录动态, 揭示了治疗过程中耐药性形成及联合治疗响应的潜在机制. 此外, 虽然LINCS(The Library of Integrated Network-Based Cellular Signatures) Connectivity Map (CMap)^[103]主要以Bulk

RNA测序数据为主,但也包含部分单细胞RNA测序数据,为药物重定位、机制解析和靶点预测等提供了重要参考资源。

除小分子药物外,其他化学干预(如免疫检查点抑制剂、干扰素、转录因子等)的单细胞化学扰动研究也取得了重要进展。Gubin等人^[100]利用10× Genomics平台,分析了T3小鼠甲基胆蒎(T3 methylcholanthrene, T3 MCA)诱导肉瘤模型在接受免疫检查点治疗(immune checkpoint cancer therapy, ICT)后的免疫微环境变化。Gehring等人^[65]实现了单细胞水平上对同一细胞系在四种干扰素及四种剂量组合(共96种扰动条件)下的转录响应高效解析。Schiebinger等人^[101]基于315000个单细胞RNA测序图谱,系统重建了细胞重编程过程中的转录动态,研究了转录因子Obox6和细胞因子GDF9在小鼠胚胎成纤维细胞(mouse embryonic fibroblasts, MEF)重编程为诱导多能干细胞(induced pluripotent stem cells, iPSC)过程中的功能轨迹。此外,病毒^[104]和细菌^[105,106]等生物因素扰动也提供了百万级别的多患者跨器官图谱。

在规模和复杂性不断提升的趋势下, Tahoe-100M^[68]是迄今最大规模的化学扰动数据集,覆盖50个癌细胞系、1100多种药物与多剂量条件,总计逾1亿单细胞表达谱。其代表性细胞系涵盖13个器官来源及多种驱动突变,所用药物中近七成为已上市药物。JUMP Cell Painting^[13]则聚焦于形态组学数据整合,覆盖11万种化合物、1.3万遗传扰动,包含超16亿个细胞的图像与表达数据,为深度学习与药物发现提供了系统资源。

2.3.3 单细胞扰动数据集

随着单细胞扰动数据的积累,已有多项研究尝试整合大规模单细胞扰动数据集并开展系统性分析。scPerturb^[23](<http://projects.sanderlab.org/scperturb/>)整合了44个公开可用的数据集,涵盖了多种扰动技术和组学类型。此外,scPerturb引入了E-distance指标,用于量化扰动强度并比较不同实验变量对数据结构的影响,为大规模单细胞扰动数据的标准化与整合分析提供了重要参考。在此基础上,PerturbBase^[107](<http://www.perturbbase.cn/>)进一步扩展了数据整合的广度与深度整合了来自46项公开研究的122个scPerturbation数据集,其中包括115个单模态数据集(如单细胞转录组)和7个多模态数据集(如转录组与表面蛋白共测),支持多模态信息与组合扰动的查询、可视化与分析,已成为单细胞扰动领域的重要基础设施。此外,配套的分析工具也在

不断完善。PertPy(<https://pertpy.readthedocs.io/en/latest/installation.html>)是一个针对单细胞扰动数据开发的、文档完善的Python软件包,内置了多个标准化的h5ad格式扰动数据集,提供了多种建模工具(如扰动效应分析、轨迹推断等),并配备了详细的部署与使用教程,为单细胞扰动研究者提供了便捷高效的分析框架。总体而言,这些数据库与工具的建立,显著推动了单细胞扰动研究的数据整合、标准化与可重复性分析。

3 单细胞扰动建模方法

解析和预测单细胞层面的扰动效应是单细胞扰动建模的核心任务。该过程不仅有助于理解扰动如何驱动细胞状态转变,还为基因功能注释、疾病机制解析和药物开发提供理论基础。研究者们提出了多种方法以从不同建模视角理解和预测细胞对扰动的响应。根据模型构建的理论基础与技术框架,这些方法大致划分为:浅层模型(shallow models)、图神经网络模型(graph neural networks based Models, GNN-based Models),以及深度生成模型(deep generative models),后者是当前的主流,涵盖了变分自编码器(variational autoencoder, VAE)、最优传输(optimal transport Models, OT)、扩散模型(diffusion models)以及受大语言模型启发而迅速发展的通用基础模型(foundation models)等多种前沿方法,广泛应用于扰动效应解析、扰动响应预测、扰动组合预测及扰动生成与设计等四大核心任务(表3)。

在单细胞扰动建模中,不同任务通常依赖于多样化的建模方法。浅层模型主要包括传统的统计推断方法和经典机器学习算法,如线性回归、逻辑回归和随机森林等。此类方法参数少、计算效率高、结构透明,具备良好的可解释性,尤其适用于数据规模较小、噪声水平较低的分析场景。在单细胞扰动研究的早期阶段或特定基准分析中,浅层模型仍发挥着重要作用。图神经网络模型是一类专为处理图结构数据而设计的深度学习模型,能够有效建模节点(如基因、蛋白质)之间的关系及其在网络中的传播效应。在单细胞扰动建模中,图神经网络模型通过构建基因调控网络或基因共表达网络,引入结构先验以模拟扰动在细胞系统内的动态传导过程,广泛应用于扰动响应预测、基因互作识别以及因果推断等任务。深度生成模型是当前单细胞扰动建模的主流方法,其核心思想是学习高维输入数据的潜在概率分布,以生成与目标数据分布一致的

表 3 扰动建模方法总结

Table 3 Summary of perturbation modeling methods

方法	核心框架	核心创新/主要原始应用	扰动响应预测			扰动效应解析			扰动组合预测	扰动生成与设计	参考文献
			分布外扰动组学响应预测	扰动响应表型预测	扰动组学响应重建	基因和基因模块层面解析	细胞层面解析	因果推断解析			
ALFOI	Active learning framework	利用主动学习高效识别出最优扰动					√	√		√	[38]
ASGARD	Predictive pipeline	利用差异表达分析与顺序统计, 识别可逆转持续表达异常基因的候选药物				√					[16]
Augur	Random forest classifier	通过分类性能计算亚群的敏感性评分, 识别高响应性细胞类型					√				[18]
Beyondcell	Therapeutic clustering	构建“治疗群”, 用于个体化治疗评估		√			√				[108]
Biolord	VQVAE	利用VQVAE通过部分监督学习解耦各类扰动属性, 进行跨条件响应预测和缺失属性推断	√		√		√				[24]
CaDRReS-Sc	Matrix factorization techniques for Recommender systems	引入矩阵分解与推荐系统思想, 支持组合药物的个体化优化设计		√			√		√		[109]
CellBox	Interpretable ODE-model + deeplearning optimization	构建非线性常微分方程模型, 构建扰动-蛋白-表型动态网络结构, 通过梯度下降与自动微分联合优化, 支持大规模扰动模拟与组合扰动策略设计	√	√					√		[110]
CellDrift	Generalized linear model-based functional data analysis method	基于广义线性模型结合函数数据分析方法解析扰动响应的时间动态模式	√				√				[111]
CellOracle	GRN modeling	构建细胞类型特异的基因调控网络, 并通过马尔可夫过程模拟扰动对基因表达及细胞命运的影响, 实现细胞状态变化轨迹的可视化与响应重建	√					√			[20]
CellOT	Optimal transport	利用最优传输理论构建状态映射函数, 重建扰动条件下的细胞状态分布	√				√				[25]
CellRank2	Markov chains	基于马尔可夫链模型, 结合稳态概率推断扰动对细胞命运的调控趋势					√	√			[17]
chemCPA	AutoEncoder	基于变分自编码器进行扰动表达谱重建和扰动响应预测	√	√	√		√				[112]
CINEMA-OT	Causal inference +optimal transport	引入因果推断理论, 结合独立成分分析与无分布检验方法区分扰动驱动因子与潜在混杂因子, 建立扰动与结果之间的因果关系估计个体扰动效应			√			√	√		[26]
CondOT	Optimal transport+ context	采用多任务学习策略扩展了OT框, 泛化至未观测的组合扰动条件响应预测	√						√		[113]
contrastiveVI	Contrastive variationalinference	基于对比变分自编码器处理治疗反应的异质性	√				√				[114]
CPA	AutoEncoder	基于变分自编码器进行扰动表达谱重建和扰动条件区分		√	√		√				[33]
cycleCDR	AutoEncoder+ cycle consistency	通过变分自编码器和循环一致性损失提升对未见扰动响应预测的泛化能力	√		√		√				[115]

(续表3)

方法	核心框架	核心创新/主要原始应用	扰动响应预测			扰动效应解析			扰动组合预测	扰动生成与设计	参考文献
			分布外扰动组学响应预测	扰动响应表型预测	扰动组学响应重建	基因和基因模块层面解析	细胞层面解析	因果推断解析			
DESeq2	Negative binomial distribution modeling	基于负二项分布构建模型进行差异基因分析				√					[116]
DREEP	Functional enrichment analysis	基于大量公开数据从肿瘤的单细胞转录谱中确定药物的优先级		√			√		√		[35]
DrugCell	Dual-branch deep learning	构建了可见神经网络, 将神经元映射至通路与功能模块, 进行药物组合预测, 实现对药物反应的机制性解读		√				√	√		[22]
DRUG-NEM	Nested effect model using drug-subpopulation effect matrix, maximize combined effect probabilities with the fewest compounds	分析CyTOF单药扰动数据, 以实现个体化药物组合					√		√		[117]
D-SPIN	Probabilistic Markovrandom field model	构建基因表达程序, 采用马尔可夫随机场模型来推断这些表达程序之间及其与外源扰动之间的调控关系	√			√	√	√	√		[118]
edgeR	Negative binomial distribution modeling	基于负二项分布构建模型识别差异表达基因				√					[119]
expiMap	Interpretable conditional variational autoencoder	将条件VAE与可编程的通路模块结合, 在重建扰动响应的同时识别关键调控路径			√		√				[120]
GEARS	Graph neural networks +knowledge graph	引入知识图谱与图神经网络, 预测扰动响应和基因互作关系	√	√			√	√	√		[21]
GeneCompass	Foundation model	知识嵌入的跨物种基础模型, 迁移下游任务	√	√			√	√	√		[121]
Geneformer	Foundation model	通用基础模型, 下游任务实现小样本迁移学习	√		√						[27]
GenKI	Variational graph autoencoder	采用变分图自编码器实现虚拟基因KO实验	√					√			[122]
GexMolGen	First-align-then-generate" strategy with encoder and decoder	跨模态生成, 使用“先对齐后生成”策略从转录组预测扰动								√	[39]
GraphVCI	Variational Bayesiancausal inference +deep graph representationlearning	结合了深度图表示学习和变分贝叶斯因果推断, 学习未观测到的隐层变量, 并进行因果预测	√		√			√			[123]
Gromov-Wasserstein OT	Gromov-Wasserstein optimal transport	基于Gromov-Wasserstein最优传输实现不同模态间的扰动信息对齐与重建预测			√			√			[124]
GrouNdGAN	GRN-generative model	将因果GAN与GRN推断结合, 基于因果结构生成扰动后表达谱	√		√			√			[125]
ICAT	Two-test algorithm (sparse feature weighting with semi-supervised Louvain community detection)	两步式细胞状态识别方法识别扰动诱导的状态变化					√				[126]
IMPA	A generative style-transfer model	生成风格迁移模型预测药物扰动后图像	√	√							[127]

(续表3)

方法	核心框架	核心创新/主要原始应用	扰动响应预测			扰动效应解析			扰动组合预测	扰动生成与设计	参考文献
			分布外扰动组学响应预测	扰动响应表型预测	扰动组学响应重建	基因和基因模块层面解析	细胞层面解析	因果推断解析			
IterPert	Active learning on a budget	主动学习方法优化Perturb-seq实验中扰动的选择								√	[37]
MELD	Graph signal processing+vertex frequency clustering	将细胞视为嵌于转录组流形上的概率样本, 结合图信号处理识别响应相似的细胞簇			√		√	√			[128]
Milo	K-nearest neighbor graphs	利用细胞邻域密度差异评估扰动对丰度的影响					√				[129]
MIMOSCA	Regularized linear model	解析个体扰动的影响以及遗传相互作用对转录本、程序和细胞状态层面的影响				√	√				[6]
MixScape	Mixture discriminant analysis	识别和消除混杂的变异源, 提高了单细胞扰动筛选的信噪比					√				[7]
Monocle2	Dynamic biological processes analysis pipeline	单细胞拟时分析					√	√			[130]
multiCPA	AutoEncoder	基于变分自编码器进行单细胞多组学扰动响应预测	√		√		√		√		[131]
MUSIC	Topic modeling	基于主题模型(topic model)对单细胞基因扰动数据进行分析, 提升扰动效应推断的统计功效, 进行扰动效应的生物学解释				√					[15]
Nicheformer	Foundation model	跨物种空间转录组基础模型			√						[132]
Norman et al.	Linear regression model	线性组合模型解释遗传扰动组合的表型	√	√					√		[87]
OntoVAE	VAE	本体引导型VAE, 用于预测扰动响应	√		√			√			[133]
PAGA	Partition-based graph abstraction	基于对流形分区连通性的估计, 提供了可解释的、类似于图的流形图谱, 用于细胞聚类, 轨迹推断					√	√			[134]
PDGrapher	Causally-inspired graph neural network	利用因果图神经网络解决反向扰动设计问题						√	√	√	[135]
Perturbation-response score	Constrained quadratic optimization	扰动反应评分描述了扰动剂量与其异质性影响之间的关系					√				[136]
PerturbNet	Deep generative model+conditional normalizing flow	基于VAE预训练, 通过条件归一化流在扰动特征空间与单细胞表达分布之间构建可逆映射, 并能进行虚拟扰动模拟与扰动优化设计	√		√					√	[41]
PopAlign	Alignment of subpopulations represented as Gaussian probability distributions in oNMF latent space	识别并比对配对单细胞群体中的细胞状态, 并量化比对群体之间细胞状态丰度和基因表达的变化					√				[137]
PRESCIENT	Generative modeling framework based on diffusion maps	生成建模与时间分辨轨迹学习相结合, 建模分化路径、命运发展与基因扰动效应	√		√			√		√	[42]
PRnet	Perturbation-conditioned generative model	基于扰动条件生成模型预测扰动后响应, 并进行虚拟药物筛选	√			√	√				[34]

(续表3)

方法	核心框架	核心创新/主要原始应用	扰动响应预测			扰动效应解析			扰动组合预测	扰动生成与设计	参考文献
			分布外扰动组学响应预测	扰动响应表型预测	扰动组学响应重建	基因和基因模块层面解析	细胞层面解析	因果推断解析			
RENGE	Regression with propagation(on graph)	基于图模型模拟扰动在调控网络中的时序传播路径	√					√			[138]
scBERT	Foundation model	基于BERT架构利用双向注意力机制与自监督预训练进行细胞类型注释和迁移至下游任务	√		√		√				[139]
scDEAL	Domain-adaptive Neural Network	利用领域自适应神经网络, 实现从bulk药物标签向单细胞层面的迁移学习, 并通过集成梯度识别药物反应的关键基因特征		√			√				[32]
scDist	Mixed-effects model	基于线性混合效应模型, 在控制个体和技术变异的背景下精确评估转录组差异					√				[140]
scDR	Drug response scoring	整合数据中的药物反应基因和基因表达来计算每个细胞的药物反应评分		√							[141]
scDrug	Bioinformatic workflow	从预处理到药物响应预测的一体化流程		√			√				[142]
scFoundation	Foundation model	基于人类单细胞转录组谱的预训练基础模型, 迁移至一系列单细胞下游分析任务	√	√	√		√	√	√		[143]
scGen	Variational autoencoder+vector arithmetic in the latent space	基于变分自编码器重建扰动表达谱, 并实现扰动响应预测	√		√						[19]
scGPT	Foundation model	基于人类单细胞转录组谱的生物学基础模型, 迁移至各种下游应用	√		√			√	√		[28]
SCING	Gradient boosting and mutual information	一种基于梯度提升的基因调控网络推断方法, 专为大规模单细胞转录组数据设计				√		√			[144]
scMAGeCK	Robust rank aggregation	包含两个分析模块: scMAGeCK-RRA模块采用基于秩的假设检验策略, 评估扰动组基因是否偏离均匀分布并判定显著性; scMAGeCK-LR模块采用L2范数正则化线性回归方法筛选出显著受扰动基因				√		√			[14]
scPerturb	E-distance	借助E-distance衡量扰动间或扰动与对照之间的表型相似性					√				[23]
scRank	Target-perturbed gene regulatory network	采用靶标扰动的基因调控网络, 模拟药物扰动对药物反应细胞群体进行排序		√			√	√			[145]
scVIDR	Linear and log-linear interpolation	剂量依赖响应预测, 通过为每个细胞赋予“伪剂量”实现剂量梯度建模	√	√							[146]
STAMP	Multitask	将基因扰动预测分为子任务: 识别扰动后的差异表达基因、确定基因的表达变化方向及估计基因表达变化的幅度	√						√		[147]
State	Foundation model	由两大核心模块组成: STATE Transition(ST)和STATE Embedding(SE)大规模预训练, 用于预测扰动响应	√	√	√	√	√	√			[148]
sVAE+	Variational autoencoder with sparsity	采用贝叶斯方法学习稀疏机制偏移, 实现解耦指定扰动和细胞的表征和因果推理	√		√		√	√			[149]

(续表3)

方法	核心框架	核心创新/主要原始应用	扰动响应预测			扰动效应解析			扰动组合预测	扰动生成与设计	参考文献
			分布外扰动组学响应预测	扰动响应表型预测	扰动组学响应重建	基因和基因模块层面解析	细胞层面解析	因果推断解析			
TCPR	Few-shot machine-learning	通过移除特征在任务中心方向上的投影,缓解了少样本学习中的样本选择偏差问题,从而提升分类性能	√	√							[150]
trVAE	CVAE+maximum mean discrepancy to handle multiple perturbations	结合条件VAE和最大均值差异正则项,提升跨批次、跨扰动条件下的泛化预测能力	√		√		√				[151]
UNAGI	VAE-GAN	结合VAE-GAN架构与疾病轨迹分析,支持时间序列数据下的扰动响应模拟与药物筛选	√		√		√	√			[152]
VEGA	VAE enhanced by gene annotations	将潜在变量与已知功能模块对齐,实现对调控因子和通路状态的直接解释	√		√			√			[153]
Waddington-OT	Optimal transport	将细胞状态随时间演化建模为分布迁移过程,通过OT估算细胞命运变化			√		√	√			[101]
xTrimoGene	Scalable transformer with asymmetric encoder-decoder	采用稀疏自注意力机制与非对称编码器-解码器结构进行预训练的通用基础模型,并迁移到下游任务	√		√		√				[154]

新样本。常见的生成模型包括变分自编码器、最优传输模型、扩散模型、通用基础模型等。变分自编码器是一种概率生成模型,由编码器与解码器组成。在扰动建模中,变分自编码器可将细胞状态嵌入潜在空间,通过对扰动条件建模生成扰动后的表达谱,广泛用于扰动响应预测、组合干预模拟与跨条件泛化建模。最优传输模型是一种数学框架,用于寻找两个概率分布之间的最小代价映射。最优传输模型可用于估计扰动前后细胞状态分布的映射路径,从而建模扰动引起的状态转变、轨迹演化与细胞命运变化,具有良好的跨样本对齐与因果推断能力。扩散模型是一类基于概率扩散过程的生成模型,通过逐步引入随机噪声并反向学习恢复数据,实现对复杂分布的建模。在扰动建模中,扩散模型用于构建细胞状态的潜能景观,模拟扰动条件下的动态演化路径,支持轨迹生成与状态预测等任务。通用基础模型是指在大规模单细胞数据上进行自监督预训练的深度学习模型,通常基于Transformer架构,具备跨任务迁移与泛化能力。该类模型通过在多个细胞类型、扰动条件下学习上下文信息,并迁移至多种下游任务(如扰动响应预测、表型分类、药物敏感性评估等),具有良好的可扩展性与通用性。不同模型各具优势。浅层模型适用于小数据场景和解释性任务,图神经网络强调结构建模与调控机制解释,深度生成模型适

合模拟高维扰动响应,通用基础模型则在大规模数据背景下展示出强大的统一建模与跨任务迁移能力。模型的选择将更加依赖任务类型、数据特性以及对泛化性与可解释性的需求权衡(建模方法的发展及对对比见图3和图4)。

3.1 扰动效应解析

扰动效应解析旨在识别扰动对多层次细胞状态的显著影响,涵盖关键基因与基因模块的识别、通路富集分析及调控机制推断,以及细胞状态、类型及命运的整体影响。该任务中,浅层模型因结构简洁、计算高效、结果易解释,在早期的单细胞扰动数据分析或在数据量有限、噪声较低的情境下广泛应用。

3.1.1 基因层面扰动分析

基因层面分析旨在识别在扰动条件下表达显著变化的基因,包括直接靶点与以及关键转录调控因子。最常用的方法为基于假设检验的差异表达分析,如Wilcoxon秩和检验和t检验。DESeq2^[116]与edgeR^[119]等方法通过引入负二项分布构建广义线性模型,考虑基因特异离散度及样本间测序深度差异。DESeq2^[116]引入经验贝叶斯策略对离散度估计进行收缩,以提升小样本下的稳健性;edgeR^[119]则采用TMM(trimmed mean of M-values)标准化与拟似然检验框架控制假阳性率。然

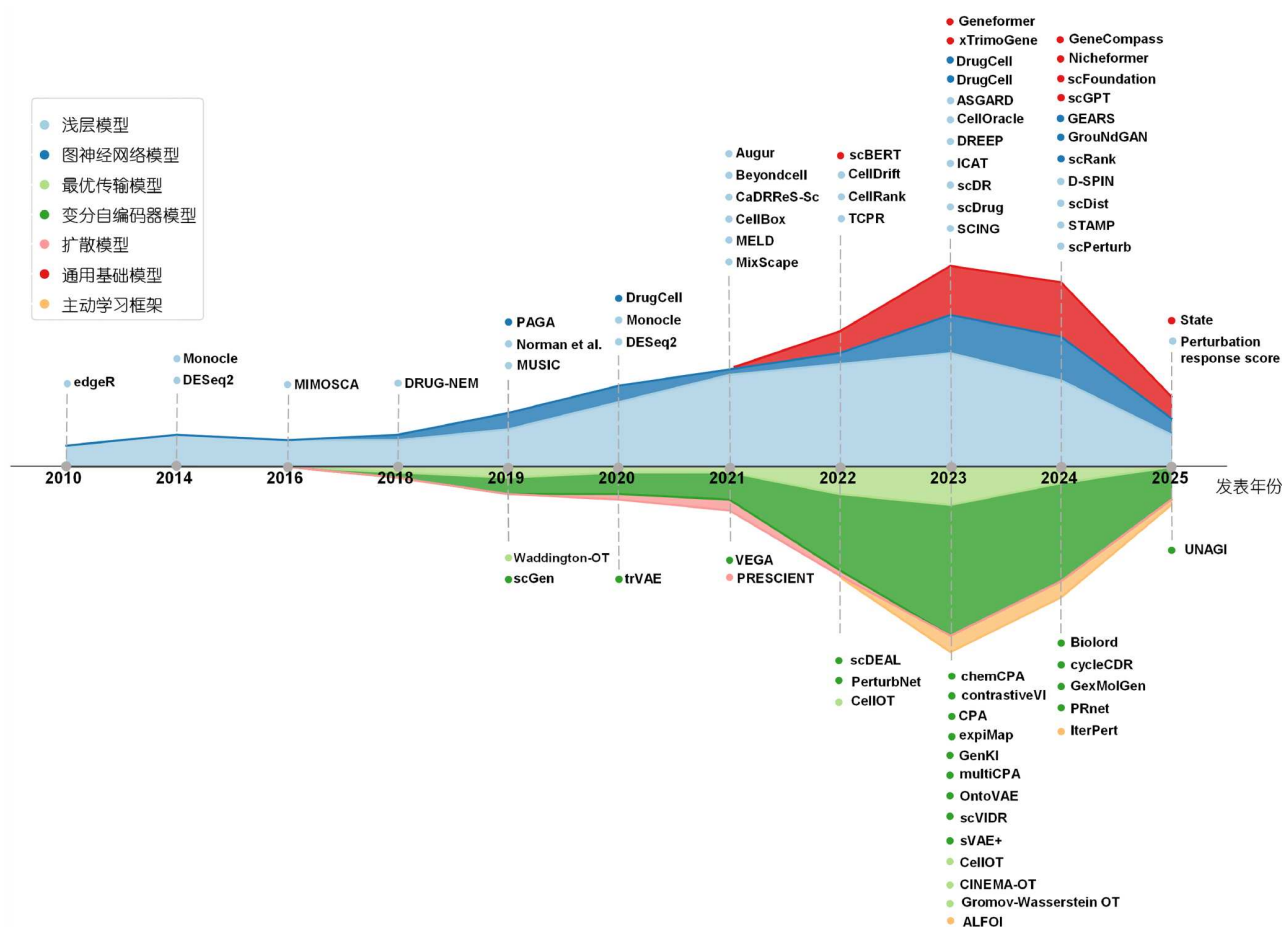


图 3 单细胞扰动建模方法发展时间线

Figure 3 The timeline of the development of single-cell perturbation modeling methods

而单细胞测序数据具有高噪声、高离散度、非正态性以及零膨胀(zero-inflated)等特性, DESeq2^[116]与 edgeR^[119]直接应用于单细胞水平数据时, 需注意其对数据稀疏性和零膨胀建模能力有限, 通常需借助伪细胞合并或正则化手段提升鲁棒性. 这类差异表达分析方法广泛应用于基因层面扰动分析, 更详细的基准评测可以参阅最近的文献^[155,156].

面向单细胞扰动数据, MIMOSCA^[6]和scMAGeCK^[14]引入正则化线性建模框架检测差异表达基因. MIMOSCA^[6]将标准化后的表达矩阵作为因变量, 扰动标签及混杂因子(如细胞周期状态、测序深度等)作为自变量, 并通过ElasticNet对回归系数正则化; 在组合扰动场景下, 还可通过交互项线性建模识别基因互作关系并进行功能注释. scMAGeCK^[14]提供两种互补分析模块: 基于秩检验的RRA(robust rank aggregation)模块

和基于线性模型的LR(linear regression)模块. RRA模块基于秩和检验评估基因表达在扰动组中的秩次偏离, 并采用置换检验评估显著性; LR模块则构建L2正则化回归模型, 同时纳入扰动标签与技术混杂因子, 以全基因范围同时估计扰动效应并通过置换法确定统计显著性. MIMOSCA在交互作用解析和特征稀疏建模方面更为灵活, 而scMAGeCK计算效率更高, 适用于高通量筛选任务, 具备较强的扩展性与实用性.

3.1.2 基因模块层面扰动分析

基因模块层面扰动分析旨在识别扰动条件下协同表达变化的基因集合, 揭示潜在通路结构与调控网络, 并支持下游的机制解析和药物适应证关联. 相较于单基因层面分析, 模块级建模可整合多基因信息, 更适用于复杂扰动背景下的系统性解析.

MUSIC^[15]基于主题模型(topic model)解析单细胞

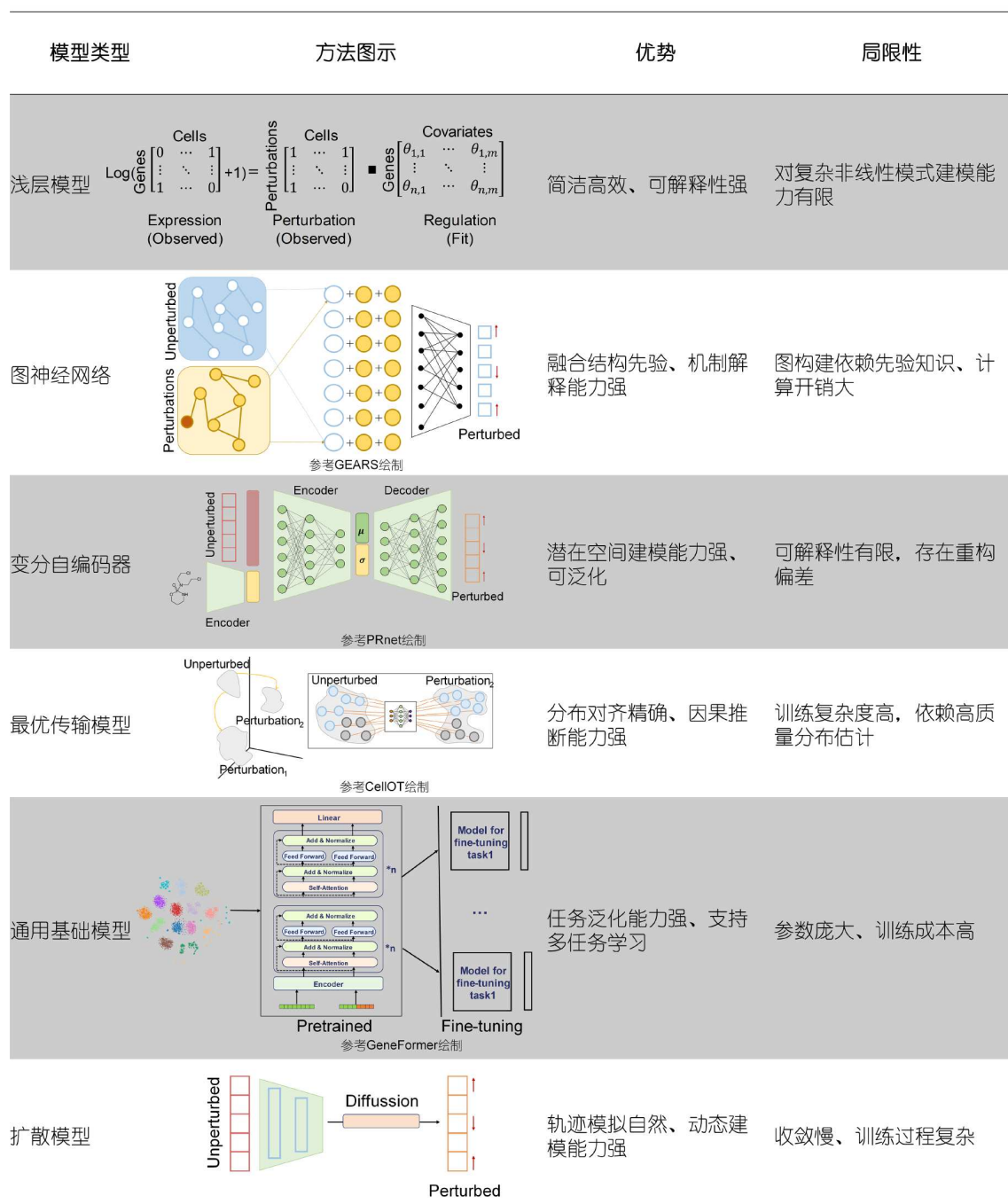


图4 单细胞扰动建模方法对比图

Figure 4 The comparison of single-cell perturbation modeling methods

扰动数据。MUSIC首先通过SAVER^[157]进行数据插补缓解稀疏性，结合细胞筛选剔除标签错误或扰动无效细胞；随后将细胞转录组视为“文本”，基因表达视为“词汇”，通过Gibbs采样推断细胞-主题和主题-基因分布。在主题空间中比较扰动组与对照组的分布差异，可定

量评估扰动效应，并进行功能注释与扰动排序。

D-SPIN^[118]面向大规模数据，基于马尔可夫随机场(Markov random field, MRF)构建调控网络，将细胞状态表示为多个交互的“基因表达程序”，推断这些表达程序之间以及外部扰动之间的调控关系。随后构建概率

模型量化调控程度, 可用于揭示通路结构、功能模块及调控逻辑, 并用于解析异质细胞群的药物响应机制, 如免疫治疗组合如何重塑表达程序、诱导新细胞状态等。

SCING^[144]是一种结合梯度提升(gradient boosting)和互信息(mutual information)的基因调控网络(gene regulatory network, GRN)推断方法, 适用于scRNA-seq、snRNA-seq和空间转录组学数据。其核心策略包括: 通过袋装法(bagging)整合多个子网络构建共识GRN; 利用边的重要性评分与条件互信息进行方向化与剪枝, 增强网络的准确性与生物学可靠性。构建的GRN可进一步划分为功能模块, 支持细胞聚类、表型关联分析和通路注释。

此外, 基于模块层面的扰动特征可用于适应症关联。ASGARD^[16]利用差异表达分析与顺序统计, 基于L1000药物响应数据集, 识别可逆转持续表达异常基因的候选药物, 支持药物重定位。通过综合细胞类型比例、表达逆转重要性及药物调控失调基因的覆盖程度, ASGARD^[16]为多细胞簇设定药物评分, 实现更精确的扰动策略评估。

3.1.3 细胞层面的扰动分析

细胞层面的扰动分析旨在刻画扰动对细胞状态、类型及命运的整体影响, 揭示其如何驱动细胞表型重塑及不同细胞类型的异质性响应。该类分析不仅有助于识别扰动引起的特定细胞亚群变化, 还可推断细胞状态转变路径, 揭示潜在的发育或疾病相关轨迹。

一种常见策略是基于低维表示, 如主成分分析(principal component analysis, PCA)、均匀流形近似和投影(uniform manifold approximation and projection, UMAP)、 t 分布随机邻域嵌入(t -distributed stochastic neighbor embedding, t -SNE)与聚类算法(如Louvain、Leiden)对扰动前后的细胞状态进行可视化与亚群识别, 分析扰动引发的细胞状态重塑。Scanpy^[158]和Seurat^[159]等分析框架提供了从数据预处理、聚类到差异表达分析的完整流程, 是此类分析的基础工具。

为深入解析扰动引起的动态变化, 轨迹推断方法如Monocle^[130]、PAGA^[134]等被广泛应用, 以拟时序重建和图模型建模揭示扰动对细胞分化方向与速度的影响。CellRank2^[17]在此基础上引入马尔可夫链模型, 结合稳态概率推断扰动对细胞命运的调控趋势, 提升了轨迹推断的生物学解释力。

针对不同细胞亚群的响应异质性, Augur^[18]通过分

类性能计算亚群的敏感性评分, 识别高响应性细胞类型。Beyondcell^[108]则从药物反应角度构建“治疗群”(therapeutic clusters), 解析肿瘤细胞异质性并支持个体化治疗策略; scRank^[145]和scDEAL^[32]等方法基于细胞类型的药物敏感性评估异质性。此外, scGen^[19]、trVAE^[151]等利用深度生成模型的一类方法在隐层空间建模扰动引起的表型变化, 揭示细胞对扰动的特异性响应。

部分方法则通过评分机制或距离度量评估扰动效应。Perturbation-response Score^[136]可用于量化扰动强度与表型影响的关系, 适用于剂量梯度条件下的评估; scPerturb^[23]借助E-distance衡量扰动间或扰动与对照之间的表型相似性, 适用于跨平台响应一致性分析。

此外, 一些方法聚焦于扰动引起的细胞丰度变化与分布偏移。Milo^[129]基于K近邻图构建细胞邻域, 比较扰动组与对照组的局部细胞比例以评估丰度差异; MELD^[128]将细胞视为嵌于转录组流形上的概率样本, 结合图信号处理识别响应相似的细胞簇。Mixscape^[7]借助混合判别分析区分表达相同gRNA的细胞中“受扰动”与“逃逸扰动”群体, 提升了ECCITE-seq等多组学数据的分析信噪比。此外, ICAT^[126]提出一种两步式细胞状态识别方法, 先在对照组中进行基因特征选择与加权聚类, 再在扰动组中进行半监督聚类, 能敏感地识别扰动诱导的状态变化。scDist^[140]基于线性混合效应模型, 在控制个体和技术变异的背景下精确评估转录组差异, 适用于小规模患者队列。PopAlign^[137]则通过概率模型对细胞亚群状态进行建模与对齐, 能够在大规模免疫扰动实验中追踪细胞状态的演化路径, 并识别亚群的丰度与表达变化。

3.2 扰动响应预测

在单细胞扰动建模中, 扰动响应预测是关键研究方向, 是核心任务之一, 其目标是预测细胞在受到扰动(如药物处理、基因编辑等)后的动态变化。根据具体建模任务的不同, 该任务可分为以下三类: (1) 分布外扰动组学响应预测, 即在未见的扰动条件或组合下预测细胞的组学响应; (2) 扰动响应表型预测, 即预测扰动诱导下的细胞状态、分类、药物敏感性标签等表型特征; (3) 扰动组学响应重建, 即在给定扰动条件下模拟或生成完整的扰动后组学表达谱。下文将依照上述分类, 并结合方法类型系统性地介绍代表性研究进展。

3.2.1 分布外扰动组学响应预测

分布外扰动响应预测的核心在于对训练集中未出现过的扰动类型或组合(如新药物、新剂量、新基因敲除/激活或其组合)进行建模,提升模型的泛化能力与真实场景适应性。

浅层模型方法主要包括线性模型和回归模型等。CellBox^[110]采用非线性常微分方程(ordinary differential equation, ODE)模拟扰动-蛋白-表型三者间的动态网络关系,从数据中直接学习相互作用参数,并通过梯度下降与自动微分技术优化网络结构,适合构建大规模时序扰动响应模型。CellDrift^[111]基于广义线性模型结合函数数据分析方法,引入细胞类型与时间的交互项,识别扰动响应的主要动态趋势,并借助模糊C均值聚类与主成分函数分析揭示细胞类型特异的时间动态变化模式。CellOracle^[120]利用scRNA-seq与scATAC-seq数据构建细胞类型特异的GRN,并通过马尔可夫过程模拟TFs扰动对细胞状态轨迹的影响。其核心优势在于可模拟TFs敲除或过表达后细胞命运的转变方向,从而提供扰动下的可解释性机制图谱。

基于图神经网络的扰动建模典型方法,如GEARS^[121]将图神经网络与知识图谱结合,在建模共表达网络的同时,嵌入基因本体(gene ontology, GO)信息,在单基因或基因组合扰动条件下预测表达响应。D-SPIN^[118]基于马尔可夫随机场建模基因表达程序间的扰动传播,适用于大规模数据下的药物及药物组合响应与预测。GRouNdGAN^[125]则创新性地将对因果对抗生成网络(generative adversarial networks, GAN)与GRN推断结合,基于因果结构生成逼真的扰动后scRNA-seq数据,保留细胞谱系、伪时序等动态信息,可用于模拟扰动效应或评估GRN推断算法性能。GraphVCI^[123]融合深度图表示学习和变分贝叶斯因果推断,学习未观测到的隐层变量,并基于该模型生成因果预测。RE-NGE^[138]面向时间序列scCRISPR数据建模,通过模拟调控网络中的扰动传播动态,区分早期与晚期响应,实现直接与间接调控效应的解析。总体而言, GNN类方法强调图结构建模与调控机制推断,广泛适用于扰动响应预测、GRN重建及机制模拟等任务。

基于变分自编码器及其变体的深度生成模型是当前扰动响应预测的主流方法。代表性方法如scGen^[119],通过在潜在空间中进行向量运算,假设扰动引起的细胞状态变化在低维嵌入空间中呈线性结构,从而有效捕捉高维单细胞表达数据中的扰动响应,具有良好的

跨细胞类型、物种和实验条件的泛化能力。其后续方法CPA^[133]系列,包括chemCPA^[112]和MultiCPA^[131],在scGen^[119]基础上进一步引入小分子化学结构信息和多模态数据,通过编码器解码器结构预测扰动响应,在潜在空间中同时建模扰动因素与细胞背景,从而增强组合扰动的预测能力和对未见扰动条件的泛化性。IMPA^[127]将预测任务扩展至高内涵成像数据,预测未见扰动的形态学和群体水平变化。Biolord^[24]利用VQVAE(vector quantized variational autoencoder)强调表征可解耦性,通过部分监督学习解耦各类扰动属性,使模型具备跨条件预测能力和缺失属性推断能力。cycleCDR^[115]强调扰动前后状态映射的可逆性,通过循环一致性损失提升对未见扰动的泛化能力。针对跨域迁移和复杂生物条件建模, trVAE^[151]和contrastiveVI^[114]分别通过对抗训练与对比学习策略,提升了在复杂生物条件下的迁移能力与稳定性。sVAE+^[149]、OntoVAE^[133]和VEGA^[153]等则通过引入稀疏先验、生物本体结构或已知通路模块,使潜在变量与生物功能之间建立直接对应关系,显著提升模型的可解释性与机制推理能力。在虚拟扰动模拟方面, GenKI^[122]结合变分图自编码器(variational graph autoencoder, VGAE)从无先验知识的scRNA-seq数据中推断GRN,并进行虚拟基因敲除。PRnet^[34]针对小分子药物响应建模,提出包含扰动适配器、扰动编码器与解码器的结构,借助SMILES化学式表示和未扰动表达谱,实现对新化合物扰动效应的预测,并应用于药物筛选。此外, scVIDR^[146]专注于剂量依赖响应预测,通过为每个细胞赋予“伪剂量”实现剂量梯度建模。UNAGI^[152]基于VAE-GAN^[160]架构结合时间序列分析,构建病理轨迹图谱,并通过潜在空间中的扰动操作,进行药物筛选与治疗方案推演。PerturbNet^[41]则基于条件归一化流(conditional normalizing flow),在扰动特征空间与单细胞表达分布之间构建可逆映射,支持在多种扰动类型(如小分子、CRISPR激活、点突变)下的状态预测,并能进行虚拟扰动模拟与扰动优化设计。综上, VAE及其扩展方法通过潜在空间建模、属性解耦、生物结构融合与时间动态建模等策略,构建了系统化、模块化的单细胞扰动建模框架,具备强大的泛化性、可解释性与可扩展性。

在扰动响应预测中,除基于VAE类方法外,近年来基于最优传输与扩散模型的生成策略也逐渐成为扰动响应预测的新兴方向。代表性方法CellOT^[25]基于神经

网络参数化的最优传输理论, 构建凸网络形式的双对偶势函数(dual potentials), 学习对照组与扰动组间的分布映射, 能有效捕捉细胞群体的异质性响应, 在多物种与多疾病模型中展现出出色的泛化能力. CINEMA-OT^[26]则引入因果推断理论结合最优传输, 通过独立成分分析与无分布检验方法区分扰动驱动因子与潜在混杂因子, 其核心思想是在干预组和对照组细胞中找出“潜在对应”, 建立扰动与结果之间的因果关系, 进而估计个体扰动效应(individual treatment effect, ITE). CONDOT^[113]扩展了OT框架的条件建模能力, 采用多任务学习策略, 在多个标注上下文下学习一个统一的上下文感知映射器, 使模型能够泛化至未观测的组合扰动条件(如新组合药物或基因敲除). PRESCIENT^[42]则基于扩散过程建模细胞状态的潜能景观(potential landscape), 支持在时间维度上模拟细胞命运的动态演化, 支持从任意初始状态出发的高维轨迹生成. 总体而言, 基于OT和扩散过程的生成模型通过构建扰动引起的分布映射与时间演化路径, 展现出在处理样本偏移、推断因果机制及模拟未见扰动方面的独特优势, 正在成为扰动建模领域的重要技术分支.

近年来, 受到大语言模型成功经验的启发, 基于Transformer架构的通用基础模型在单细胞扰动建模中迅速发展, 成为生成式模型的重要方向. 这类方法将单细胞转录组数据类比为自然语言, 将细胞视作“句子”, 基因视作“词汇”, 通过Transformer架构在大规模scRNA-seq数据上进行自监督预训练, 从而获得具备上下文感知能力的通用表示, 用于多种扰动响应预测等下游任务. 代表方法, 如Geneformer^[27], 在近三千万个人类单细胞样本上进行预训练, 利用掩码语言模型学习基因调控上下文, 实现对剂量敏感基因及其靶标的精确预测, 并在多种下游任务中实现小样本迁移学习. scBERT^[139]是首个引入BERT(bidirectional encoder representations from transformers)架构的模型, 利用双向注意力机制与自监督预训练, 显著提升了细胞类型注释的准确性和鲁棒性, 并具备良好的基因层级可解释性. scFoundation^[143]构建了亿级参数规模的Transformer模型, 并引入read-depth-aware预训练任务以增强对测序深度差异的建模能力, 支持无需微调地跨任务泛化至扰动预测与药物响应分类等任务. scGPT^[28]采用统一的生成式预训练框架, 支持细胞与基因的联合编码, 广泛适用于细胞类型注释、遗传扰动预测和多组学整合等场景, 并展示出数据规模与模型性能正相关

的可扩展性. 在模型架构优化方面, xTrimoGene^[154]采用稀疏自注意力机制与非对称编码器-解码器结构, 显著降低训练复杂度, 支持在最大规模数据集上高效预训练, 并在组合扰动预测与药物筛选中表现突出. GeneCompass^[121]则引入跨物种知识引导, 在逾1亿细胞的跨物种数据集上构建模型, 并结合启动子序列、TF调控网络、基因家族等先验信息, 提高了调控网络泛化能力, 支持从小鼠推理至人类的关键通路药物靶点. 此外, STAMP^[147]将遗传扰动预测任务拆解为差异表达基因识别、调控方向判断和表达变化幅度估计三个子任务, 提升基础模型在小样本条件下的表达预测能力与基因网络解读性能. STATE^[148]是首个专为扰动响应预测设计的多尺度基础模型, 由两大核心模块组成: STATE Transition(ST)和STATE Embedding(SE). 其中SE模型由1.67亿个细胞的数据训练, ST模型由1亿个扰动细胞数据训练, 能够预测干细胞、癌细胞和免疫细胞在药物、细胞因子或遗传干预下的响应情况. 该类通用基础模型依托大规模预训练, 结合生物学先验与任务定制建模策略, 展现出卓越的迁移能力、可解释性与预测精度.

3.2.2 扰动响应表型预测

扰动响应表型预测旨在根据扰动条件和细胞初始状态, 预测扰动诱导的表型变化, 如细胞类型、状态转变、药物响应等级或敏感性标签. 该任务通常被建模为监督学习问题, 模型输入为扰动变量(如药物或基因)及细胞表达谱或其嵌入特征, 输出为表型分类或连续响应分值.

在浅层模型方面, Beyondcell^[108]通过计算药物特征集的富集分值识别药物响应异质性, 定义“治疗簇”(therapeutic clusters, TCs). DREEP^[35]则结合富集分析与大规模药物筛选数据库, 预测上千种药物的细胞敏感性并动态评估耐药演化. CaDRReS-Sc^[109]引入矩阵分解与推荐系统思想, 支持在肿瘤内异质性背景下进行组合药物的个性化优化设计. scDrug^[142]构建了一体化生物信息学工作流, 覆盖从scRNA-seq数据预处理、细胞亚群识别到药物响应预测的完整流程. Norman等人^[87]提出的线性组合模型可用于解释遗传扰动组合的表型效应, 在提高预测准确性的同时保留较好的可解释性.

为增强模型的生物机制解释能力, DrugCell^[122]构建了可见神经网络(visible neural network, VNN), 将神经元映射至通路与功能模块, 实现对药物反应的机制性

解读. scDEAL^[32]利用领域自适应神经网络(domain-adaptive neural network, DaNN), 实现从bulk药物标签向单细胞层面的迁移学习, 并通过集成梯度识别药物反应的关键基因特征. scRank^[145]则基于扰动靶点调控网络(target-perturbed GRN, tpGRN), 通过网络扰动模拟与流形对齐方法, 实现从未经扰动的样本中推断潜在药物响应细胞类型. 此外, scDR^[141]则基于药物响应相关基因集信息, 实现了对未见扰动或细胞类型的泛化预测. 面向数据稀缺情境, TCPR^[150]等小样本学习方法通过去除特征向量中与任务质心方向一致的维度, 有效缓解样本选择偏倚问题, 提升模型的稳定性与泛化能力.

此外, 结合生成模型并利用潜在空间表示进行表型预测的方法也逐渐兴起. 例如, scVIDR^[146]不仅能够预测不同剂量下的基因表达谱, 还引入“伪剂量”变量捕捉细胞对化学扰动的潜在敏感性, 在生成与分类任务间提供平衡解决方案. 类似地, 基于VAE框架的CPA^[33]、chemCPA^[112]等VAE框架方法通过扰动向量嵌入, 支持单扰动、组合扰动及跨物种药物响应预测. 更进一步, scFoundation^[143]、GeneCompass^[121]和scGPT^[28]等通用基础模型在预训练后, 已将细胞类型注释与药物敏感性预测等任务纳入下游任务, 在预训练后具备良好的迁移能力.

综上所述, 扰动响应表型预测方法已逐步形成涵盖机制可解释性、迁移学习、生成建模与基础模型预训练等多层次的技术体系, 在个体化药物筛选、耐药机制解析与复杂疾病治疗策略设计中展现出广泛的应用潜力.

3.2.3 扰动组学响应重建

扰动组学响应的重建任务旨在在给定扰动条件(如药物、基因干预或其组合)下, 重建完整的扰动后组学表达谱, 为虚拟实验、数据补全、机制推断及分布外泛化提供建模基础^[120,161].

当前主流方法多基于深度生成模型, 尤其以变分自编码器及其扩展为核心架构. 典型方法如scGen^[19]、CPA^[33]、chemCPA^[112]与Biolord^[24]均通过在潜在空间中嵌入扰动向量, 实现高维转录组表达谱的精确重建, 具备良好的跨扰动类型、细胞类型及物种的泛化能力. 为提升模型的生物学可解释性与机制一致性, 部分方法进一步引入结构建模与先验知识引导策略. expi-Map^[120]将条件VAE架构与可编程的通路模块(gene programs, GPs)相结合, 在重建扰动响应的同时识别关

键调控路径, 增强对通路层面表型变异的解析能力. 通用基础模型, 如scFoundation^[143]、GeneCompass^[121]与scGPT^[28]等在大规模scRNA-seq数据上进行预训练时, 通常将表达谱重建作为核心任务之一, 以获取上下文表示, 从而支持后续多样化的下游扰动建模任务. Nicheformer^[132]更进一步将预训练扩展至空间转录组场景, 在空间组学数据上训练Transformer模型, 实现空间上下文的建模与扰动表达谱的迁移生成, 推动空间分辨率下的扰动响应预测能力.

基于最优传输的建模策略也展现出强大能力. Cel-IOT^[25]、CONDOT^[113]与CINEMA-OT^[26]等模型通过最优传输映射学习对照组与扰动组之间的分布变换路径, 在建模细胞状态跃迁与扰动诱导变化方面表现优越. 针对多模态干预数据, Gromov-Wasserstein OT^[124]可实现不同模态(如图像与转录组)间的扰动信息对齐与重建预测. Waddington-OT^[101]可从时间分辨scRNA-seq数据中推演命运转换路径, 并预测关键调控因子. 其他方法, MELD^[128]则基于图信号处理技术构建扰动响应的连续表示谱, 捕捉扰动驱动下细胞亚群结构与基因特征的微妙变动, 适用于复杂背景下的扰动解析.

3.3 扰动组合预测

扰动组合预测聚焦于建模多个扰动因素(如药物-药物、基因-基因)之间的协同或拮抗效应, 旨在揭示复杂调控网络中的功能性相互作用关系, 是联合用药策略优化、基因调控网络解析及组合扰动设计的关键环节.

在基因组合扰动建模方面, Norman等人^[87]提出线性回归框架, 将遗传扰动组合的表达效应表示为单独扰动的加权和, 并通过残差项捕捉非线性交互. GEARS^[21]在此基础上引入知识图谱与图神经网络, 支持多类型基因互作关系(包括协同与抑制、异性功能变化、冗余与上位互作)的泛化预测, 具备良好的跨条件迁移能力. 通用基础模型如scFoundation^[143]、GeneCompass^[121]与scGPT^[28]等将基因互作预测作为下游任务, 通过预训练嵌入提升任务表示能力, 提升模型在扰动组合预测任务准确性.

在药物组合建模方面, CaDRReS-Sc^[109]基于推荐系统中的矩阵分解技术, 融合患者内异质细胞亚群的药物反应模式, 提出细胞死亡叠加策略以预测组合疗效提升幅度. CellBox^[110]基于ODE学习扰动-蛋白-表型的动态网络结构, 支持大规模组合扰动设计.

DrugCell^[22]利用可见神经网络(visible neural network, VNN), 将药物结构与突变信息映射至功能模块, 能够精确预测药物响应并辅助协同药物组合筛选, 其预测结果已获实验验证. DREEP^[35]通过对不同细胞亚群的敏感性建模, 精准识别潜在耐药或敏感亚群, 并提出个性化联合用药建议. MultiCPA^[131]采用多模态生成模型, 结合RNA与蛋白等多组学特征, 通过潜在空间向量叠加预测未见单药或组合扰动下的表达谱响应. CINEMA-OT^[26]结合最优传输与因果推断, 构建细胞-基因层面的协同效应矩阵, 实现组合扰动的非线性交互建模. DRUG-NEM^[117]针对CyTOF单细胞数据开发嵌套效应建模策略, 旨在最小化所需药物数量的同时实现对关键信号通路的最大激活, 用于个体化药物组合推荐.

3.4 扰动生成与设计

扰动生成与设计旨在给定目标表型、期望响应或疾病状态的前提下, 通过计算方法从高维扰动空间中高效筛选、生成或优化能够诱导期望表型转变的扰动条件, 以提升实验设计效率并实现更具生物学意义的干预效果. 针对组合扰动数量庞大、实验资源有限等挑战, 近年来涌现出融合因果建模、主动学习(active learning)与跨模态生成的策略, 显著提升探索效率与干预针对性.

在遗传扰动实验设计方面, 一些方法通过主动学习优化实验扰动选择. 主动学习是一种迭代的机器学习策略, 旨在通过让模型主动选择最具信息价值的数据进行标注或实验, 从而在有限数据或预算条件下最大化模型性能. 主动学习通过在巨大的扰动空间迭代选择诱导目标表型转变的关键扰动, 大幅提升了单细胞扰动实验的效率和策略性. 主动学习通过在巨大的扰动空间迭代选择诱导目标表型转变的关键扰动, 大幅提升了单细胞扰动实验的效率和策略性. IterPert^[37]提出一种迭代式主动学习策略, 用于优化Perturb-seq实验中的扰动选择. 该方法结合当前模型的预测结果与信息增益指标, 在每轮实验中优选最具信息价值的扰动组合, 显著减少实验轮次与筛选成本, 特别适用于“预算受限主动学习”情境. ALFOI^[38]则在因果图框架下建立贝叶斯更新机制与封闭形式的获取函数, 优化干预后表达分布的均值与目标状态之间的差异, 方法在合成数据和Perturb-CITE-seq单细胞转录组数据上进行了验证, 成功识别出可诱导特定细胞状态转变的最

优干预组合. PDGrapher^[135]基于因果图神经网络模型, 能够高效直接预测逆转疾病表型的扰动组合, 显著提升了表型驱动药物发现的准确性与速度. 此外, PerturbNet^[41]和PRESCIENT^[42]也可用于指导扰动实验的策略优化, 通过模拟扰动效应和预测状态转移轨迹辅助干预条件的筛选与验证.

在分子生成方面, GexMolGen^[39]提出基于跨模态生成的扰动设计范式, 利用期望与基线表达状态之间的差异构建目标转录特征, 通过“先对齐后生成”策略将表达签名映射至分子嵌入空间, 并结合图结构解码器生成满足预期转录响应的候选小分子. 该方法同时融合了基于scaffold的结构生成策略和单细胞基础模型(如scGPT^[28]、Geneformer^[27]、scBERT^[139]等)作为表达特征编码器, 显著提升生成分子的结构有效性与跨模态对齐的鲁棒性, 突破传统方法对训练域的一致性依赖. 类似方法如DLEPS^[162]通过从小分子结构预测差异表达特征, Gex2SGen^[163]直接以基因表达生成小分子, 但两者均未针对单细胞扰动数据进行训练.

4 总结与展望

单细胞扰动建模已成为基因功能解析、药物机制研究与精准治疗设计的关键技术手段. 近年来, 随着高通量单细胞扰动测序技术的迅速发展, 包括Perturb-seq^[6]、CROP-seq^[44]、sci-Plex^[10]与Mosaic^[68]等方法, 研究者能够在细胞分辨率下系统获取大规模扰动实验数据. 如Tahoe-100M^[68]基于Mosaic平台生成了超过1亿个单细胞化学扰动数据, sci-fi-RNA-seq^[71]和Perturb-seq^[6,102]也分别实现了千万级至百万级的遗传扰动筛选. 这些高通量平台的广泛应用, 使得基因调控网络、细胞状态转变及药物响应研究达到了前所未有的分辨率与深度.

伴随数据规模提升, 单细胞扰动建模方法也迅速发展. 从早期的统计模型到广泛应用的深度学习方法, 如变分自编码器、图神经网络、扩散模型, 以及基于Transformer架构的通用基础模型, 均已在扰动响应解析与预测、表型分类、组学生成、扰动组合与扰动设计等任务中展现出显著性能. 这些方法各具优势, 为扰动效应的模拟与预测提供了多元化、可扩展的工具体系.

尽管取得诸多进展, 当前单细胞扰动建模仍面临的一系列关键挑战, 存在包括数据标准化与评测体系尚不健全、数据扰动类型与细胞多样性不足、多模态

融合能力仍较薄弱等问题。首先是数据标准化与评测体系尚不健全,现有单细胞扰动数据在实验设计、平台技术等方面存在较大差异,缺乏统一的数据标准和评估框架。尽管如scPerturb^[23]和Perturbbase^[107]等数据库已整合了大量单细胞扰动样本,但标准化的基准数据集、建模任务定义与系统性的性能评估体系仍不完善。此外,已有数据的扰动类型与细胞多样性仍显不足。尽管现有数据在细胞数量上具备优势,但在扰动类型和细胞背景上仍不充分。例如, Tahoe-100M^[68]虽覆盖约50种细胞系,但仅包含1100种化合物;而经典Perturb-seq^[6,102]多集中于单一细胞系(如K562)中上千基因的扰动。这在一定程度上限制了模型的泛化能力,尤其是跨细胞类型、跨扰动类型的迁移建模能力。与此同时,现有数据普遍缺乏多模态(RNA、蛋白、表观)、纵向(多时间点、多剂量)及空间定位信息,难以支撑对扰动机制的全方位解析。JUMP Cell Painting^[13]提供了一种高内涵成像与大规模扰动整合的思路,为低成本大规模扰动筛选提供新方向。单细胞扰动技术将朝着多样化和高复杂性方向发展,例如NIS-seq^[63]等为代表的创新技术,有望在原位条件下以单细胞分辨率更全面地揭示细胞对扰动的响应。单细胞扰动技术的发展,也为扰动建模提出了挑战,多模态融合是算法面临的一项挑战。目前多数模型仍主要集中于转录组数据,面对空间组学,单细胞多组学(如ECCITE-seq^[7,59]、Perturb-ATAC^[11])等多模态数据,缺乏统一、高效、可扩展的融合建模策略,难以充分捕捉扰动下细胞状态的多维表征。

同时目前模型的实用性也存在挑战,限制了其从方法开发向研究实用甚至临床转化的推进。当前模型的泛化性与可解释性仍待提升。例如,尽管深度生成模型在部分任务上表现出色,如细胞类型注释,但在真实扰动预测场景下,其准确性甚至不如浅层线性模型稳定^[164]。图神经网络模型虽然具备较强的结构建模能力,但高度依赖预先构建的图结构,受限于先验知识的质量,难以适应不同扰动条件或新的生物系统,限制了其泛化能力。以上大多数模型基于一两个细胞系训练,难以泛化至不同剂量、时间点、细胞亚群或未见扰动条件。同时,深度模型的黑箱特性也限制了其在生物发现和临床应用中的信任度。在面对单细胞数据中天然存在的高噪声和潜在混杂因素时,如何进行因果推断,提供机制解释,提升模型的任务适应性、跨条件泛化能力是后续研究的核心方向。同时,组合扰动空间庞

大,全面实验探索代价高昂,实验设计受限于资源“预算瓶颈”。传统全量实验方法效率低下,无法满足大规模组合扰动探索需求。虽然部分模型已具备预测潜力,但缺乏系统的湿实验支持,阻碍了其在药物发现和功能基因筛选中的转化落地。在资源受限条件下,亟需结合主动学习、因果建模与实验优化策略,发展高效且智能的实验设计框架,实现数据-模型-实验的闭环联动。

面向未来,单细胞扰动建模的发展可从以下几个方向展开。首先从数据层面,可以构建系统性的扰动图谱。当前如The Human Cell Atlas^[165,166]及Arc Virtual Cell Atlas等国际协作计划,正推动跨组织、跨疾病、跨时空尺度的系统性扰动数据资源建设。这些图谱将为模型的泛化学习、任务无关预训练与跨物种迁移建模提供坚实基础。此外,在算法层面,受大语言模型启发多个团队探索发展通用型扰动基础模型,scGPT^[28]、Geneformer^[27]、scFoundation^[143]等单细胞基础模型在多研究数据整合、跨任务迁移方面初现成效。未来需进一步提升其对未知扰动、跨细胞类型与跨模态任务的适应能力,拓展其在组合预测、机制解析与药物筛选中的实用性与可靠性。同时结合多模态融合和因果机制建模也是备受关注的方向,随着空间转录组、蛋白质组、染色质可及性等新兴技术的成熟,构建融合基因组-转录组-表观组-空间信息的多模态模型,将有望全面揭示扰动下细胞命运转变的关键调控轴线。同时,结合因果表示学习与生物知识图谱,有望提升模型的可解释性与推理能力,支持基于机制的扰动策略设计。最后构建虚拟细胞与模拟实验平台,是推动数据-模型-实验闭环发展的关键突破点。随着生成模型的发展,构建“虚拟细胞”成为备受关注的方向,支持在体外模拟不同扰动条件下的响应过程,为离体药物筛选、个体化治疗预测及合成生物学提供关键支撑。随着美国食品药品监督管理局(Food and Drug Administration, FDA)宣布将计划逐步取消单克隆抗体及其他药物的动物实验要求,使用“新方法论”(包括AI计算模型、人类细胞系、类器官以及器官芯片系统等)替代动物实验是重要发展方向。

综上所述,单细胞扰动建模正加速迈向数据驱动与机制解释相结合、任务特化与通用模型并重,以及静态预测向动态干预转化的演进阶段。未来研究应进一步聚焦于跨模态数据的高效整合、因果推断与模型可解释性提升,增强模型在多任务、多条件下的泛化

与迁移能力。同时,亟需推动数据资源、计算建模与实验技术的深度融合,构建协同发展的跨学科研究框架,

为基础生物学机制解析、临床精准决策与新药发现提供强有力的技术支撑与理论保障。

参考文献

- 1 Han X, Zhou Z, Fei L, et al. Construction of a human cell landscape at single-cell level. *Nature*, 2020, 581: 303–309
- 2 Wen L, Tang F. Recent advances in single-cell sequencing technologies. *Precision Clin Med*, 2022, 5: pbac002
- 3 Hannon G J. RNA interference. *Nature*, 2002, 418: 244–251
- 4 Ballou L M, Lin R Z. Rapamycin and mTOR kinase inhibitors. *J Chem Biol*, 2008, 1: 27–36
- 5 Zhao Y, Adjei A A. The clinical development of MEK inhibitors. *Nat Rev Clin Oncol*, 2014, 11: 385–400
- 6 Dixit A, Parnas O, Li B, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 2016, 167: 1853–1866.e17
- 7 Papalexi E, Mimitou E P, Butler A W, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet*, 2021, 53: 322–331
- 8 Adamson B, Norman T M, Jost M, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 2016, 167: 1867–1882.e21
- 9 Jaitin D A, Weiner A, Yofe I, et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, 2016, 167: 1883–1896.e15
- 10 Srivatsan S R, McFaline-Figueroa J L, Ramani V, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 2020, 367: 45–51
- 11 Rubin A J, Parker K R, Satpathy A T, et al. Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, 2019, 176: 361–376.e17
- 12 Frangieh C J, Melms J C, Thakore P I, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat Genet*, 2021, 53: 332–341
- 13 Chandrasekaran S N, Ackerman J, Alix E, et al. JUMP cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023, doi: 10.1101/2023.03.23.534023
- 14 Yang L, Zhu Y, Yu H, et al. scMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biol*, 2020, 21: 19
- 15 Duan B, Zhou C, Zhu C, et al. Model-based understanding of single-cell CRISPR screening. *Nat Commun*, 2019, 10: 2233
- 16 He B, Xiao Y, Liang H, et al. ASGARD is a single-cell guided pipeline to aid repurposing of drugs. *Nat Commun*, 2023, 14: 993
- 17 Weiler P, Lange M, Klein M, et al. CellRank 2: unified fate mapping in multiview single-cell data. *Nat Methods*, 2024, 21: 1196–1205
- 18 Skinnider M A, Squair J W, Kathe C, et al. Cell type prioritization in single-cell data. *Nat Biotechnol*, 2021, 39: 30–34
- 19 Lotfollahi M, Wolf F A, Theis F J. scGen predicts single-cell perturbation responses. *Nat Methods*, 2019, 16: 715–721
- 20 Kamimoto K, Stringa B, Hoffmann C M, et al. Dissecting cell identity via network inference and *in silico* gene perturbation. *Nature*, 2023, 614: 742–751
- 21 Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat Biotechnol*, 2024, 42: 927–935
- 22 Kuenzi B M, Park J, Fong S H, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 2020, 38: 672–684.e6
- 23 Peidli S, Green T D, Shen C, et al. scPerturb: harmonized single-cell perturbation data. *Nat Methods*, 2024, 21: 531–540
- 24 Piran Z, Cohen N, Hoshen Y, et al. Disentanglement of single-cell data with biolord. *Nat Biotechnol*, 2024, 42: 1678–1683
- 25 Bunne C, Stark S G, Gut G, et al. Learning single-cell perturbation responses using neural optimal transport. *Nat Methods*, 2023, 20: 1759–1768
- 26 Dong M, Wang B, Wei J, et al. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nat Methods*, 2023, 20: 1769–1779
- 27 Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. *Nature*, 2023, 618: 616–624
- 28 Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods*, 2024, 21: 1470–1480
- 29 Ji Y, Lotfollahi M, Wolf F A, et al. Machine learning for perturbational single-cell omics. *Cell Syst*, 2021, 12: 522–537
- 30 Gavriilidis G I, Vasileiou V, Orfanou A, et al. A mini-review on perturbation modelling across single-cell omic modalities. *Comput Struct Biotechnol J*, 2024, 23: 1886–1896
- 31 Rood J E, Hupalowska A, Regev A. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*,

- 2024, 187: 4520–4545
- 32 Chen J, Wang X, Ma A, et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat Commun*, 2022, 13: 6494
- 33 Lotfollahi M, Klimovskaia Susmelj A, De Donno C, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol Syst Biol*, 2023, 19: e11517
- 34 Qi X, Zhao L, Tian C, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nat Commun*, 2024, 15: 9256
- 35 Pellecchia S, Viscido G, Franchini M, et al. Predicting drug response from single-cell expression profiles of tumours. *BMC Med*, 2023, 21: 476
- 36 Preuer K, Lewis R P I, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 2018, 34: 1538–1546
- 37 Huang K, Lopez R, Hütter J C, et al. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. In: *Proceedings of the International Conference on Research in Computational Molecular Biology*. Cham: Springer Nature Switzerland, 2024. 17–37
- 38 Zhang J, Cammarata L, Squires C, et al. Active learning for optimal intervention design in causal models. *Nat Mach Intell*, 2023, 5: 1066–1075
- 39 Cheng J, Pan X, Fang Y, et al. GexMolGen: cross-modal generation of hit-like molecules via large language model encoding of gene expression signatures. *Brief BioInf*, 2024, 25: bbae525
- 40 Méndez-Lucio O, Baillif B, Clevert D A, et al. *De novo* generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat Commun*, 2020, 11: 10
- 41 Yu H, Welch J D. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *bioRxiv*, 2022, doi: 10.1101/500854
- 42 Yeo G H T, Saksena S D, Gifford D K. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nat Commun*, 2021, 12: 3222
- 43 Xie S, Duan J, Li B, et al. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol Cell*, 2017, 66: 285–299. e5
- 44 Datlinger P, Rendeiro A F, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods*, 2017, 14: 297–301
- 45 Roth T L, Li P J, Blaesche F, et al. Pooled knock in targeting for genome engineering of cellular immunotherapies. *Cell*, 2020, 181: 728–744. e21
- 46 Schraivogel D, Gschwind A R, Milbank J H, et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat Methods*, 2020, 17: 629–635
- 47 Song Q, Ni K, Liu M, et al. Direct-seq: programmed gRNA scaffold for streamlined scRNA-seq in CRISPR screen. *Genome Biol*, 2020, 21: 136
- 48 Replogle J M, Norman T M, Xu A, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol*, 2020, 38: 954–961
- 49 Datlinger P, Rendeiro A F, Boenke T, et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods*, 2021, 18: 635–642
- 50 Huang A C, Hsieh T-H S, Zhu J, et al. X-Atlas/Orion: genome-wide Perturb-seq datasets via a scalable fix-cryopreserve platform for training dose-dependent biological foundation models. *bioRxiv*, 2025, doi: 10.1101/2025.06.11.659105
- 51 Parekh U, Wu Y, Zhao D, et al. Mapping cellular reprogramming via pooled overexpression screens with paired fitness and single-cell RNA-sequencing readout. *Cell Syst*, 2018, 7: 548–555. e8
- 52 Pierce S E, Granja J M, Greenleaf W J. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat Commun*, 2021, 12: 2969
- 53 Liscovitch-Brauer N, Montalbano A, Deng J, et al. Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat Biotechnol*, 2021, 39: 1270–1277
- 54 Wroblewska A, Dhainaut M, Ben-Zvi B, et al. Protein barcodes enable high-dimensional single-cell CRISPR screens. *Cell*, 2018, 175: 1141–1155. e16
- 55 Wang C, Lu T, Emanuel G, et al. Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization. *Proc Natl Acad Sci USA*, 2019, 116: 10842–10851
- 56 Wheeler E C, Vu A Q, Einstein J M, et al. Pooled CRISPR screens with imaging on microrraft arrays reveals stress granule-regulatory factors. *Nat Methods*, 2020, 17: 636–642
- 57 Yenkin A L, Bramley J C, Kremitzki C L, et al. Pooled image-base screening of mitochondria with microrraft isolation distinguishes pathogenic mitofusins 2 mutations. *Commun Biol*, 2022, 5: 1128
- 58 Joung J, Ma S, Tay T, et al. A transcription factor atlas of directed differentiation. *Cell*, 2023, 186: 209–229. e26
- 59 Mimitou E P, Cheng A, Montalbano A, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods*, 2019, 16: 409–412

- 60 Legut M, Gajic Z, Guarino M, et al. A genome-scale screen for synthetic drivers of T cell proliferation. *Nature*, 2022, 603: 728–735
- 61 Binan L, Jiang A, Danquah S A, et al. Simultaneous CRISPR screening and spatial transcriptomics reveal intracellular, intercellular, and functional transcriptional circuits. *Cell*, 2025, 188: 2141–2158.e18
- 62 Dhainaut M, Rose S A, Akturk G, et al. Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell*, 2022, 185: 1223–1239.e20
- 63 Fandrey C I, Jentsch M, Konopka P, et al. NIS-seq enables cell-type-agnostic optical perturbation screening. *Nat Biotechnol*, 2025, 43: 1337–1347
- 64 McFarland J M, Paoletta B R, Warren A, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat Commun*, 2020, 11: 4296
- 65 Gehring J, Hwee Park J, Chen S, et al. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nat Biotechnol*, 2020, 38: 35–38
- 66 Zhao W, Dovas A, Spinazzi E F, et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. *Genome Med*, 2021, 13: 82
- 67 Chang M T, Shanahan F, Nguyen T T T, et al. Identifying transcriptional programs underlying cancer drug response with TraCe-seq. *Nat Biotechnol*, 2022, 40: 86–93
- 68 Zhang J, Ubas A A, de Borja R, et al. Tahoe-100M: a giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025, doi: 10.1101/2025.02.20.639398
- 69 Przybyla L, Gilbert L A. A new era in functional genomics screens. *Nat Rev Genet*, 2022, 23: 89–103
- 70 Cheng J, Lin G, Wang T, et al. Massively parallel CRISPR-based genetic perturbation screening at single-cell resolution. *Adv Sci*, 2023, 10: e2204484
- 71 Rosenberg A B, Roco C M, Muscat R A, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 2018, 360: 176–182
- 72 Stoeckius M, Zheng S, Houck-Loomis B, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol*, 2018, 19: 224
- 73 Saunders R A, Allen W E, Pan X, et al. Perturb-Multimodal: a platform for pooled genetic screens with imaging and sequencing in intact mammalian tissue. *Cell*, 2025, 188: 4790–4809.e22
- 74 Ma S, Zhang B, LaFave L M, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, 2020, 183: 1103–1116.e20
- 75 Zhu C, Yu M, Huang H, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol*, 2019, 26: 1063–1070
- 76 Feldman D, Funk L, Le A, et al. Pooled genetic perturbation screens with image-based phenotypes. *Nat Protoc*, 2022, 17: 476–512
- 77 Feldman D, Singh A, Schmid-Burgk J L, et al. Optical pooled screens in human cells. *Cell*, 2019, 179: 787–799.e17
- 78 Aissa A F, Islam A B M M K, Ariss M M, et al. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat Commun*, 2021, 12: 1628
- 79 Wagle M V, Vervoort S J, Kelly M J, et al. Antigen-driven EGR2 expression is required for exhausted CD8⁺ T cell stability and maintenance. *Nat Commun*, 2021, 12: 2782
- 80 Parekh U, McDonald D, Dailamy A, et al. Charting oncogenicity of genes and variants across lineages via multiplexed screens in teratomas. *iScience*, 2021, 24: 103149
- 81 Schmidt R, Steinhart Z, Layeghi M, et al. CRISPR activation and interference screens decode stimulation responses in primary human T cells. *Science*, 2022, 375: eabj4008
- 82 McDonald D, Wu Y, Dailamy A, et al. Defining the teratoma as a model for multi-lineage human development. *Cell*, 2020, 183: 1402–1419.e18
- 83 Lalli M A, Avey D, Dougherty J D, et al. High-throughput single-cell functional elucidation of neurodevelopmental disease-associated genes reveals convergent mechanisms altering neuronal differentiation. *Genome Res*, 2020, 30: 1317–1331
- 84 LaFleur M W, Nguyen T H, Cox M A, et al. PTPN2 regulates the generation of exhausted CD8⁺ T cell subpopulations and restrains tumor immunity. *Nat Immunol*, 2019, 20: 1335–1347
- 85 Dong M B, Wang G, Chow R D, et al. Systematic immunotherapy target discovery using genome-scale *in vivo* CRISPR screens in CD8 T cells. *Cell*, 2019, 178: 1189–1204.e23
- 86 Son J, Ding H, Farb T B, et al. BACH2 inhibition reverses β cell failure in type 2 diabetes models. *J Clin Invest*, 2021, 131: e153876
- 87 Norman T M, Horlbeck M A, Replogle J M, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 2019, 365: 786–793
- 88 Hein M Y, Weissman J S. Functional single-cell genomics of human cytomegalovirus infection. *Nat Biotechnol*, 2022, 40: 391–401

- 89 Jin X, Simmons S K, Guo A, et al. *In vivo* Perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, 2020, 370: eaaz6063
- 90 Ursu O, Neal J T, Shea E, et al. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat Biotechnol*, 2022, 40: 896–905
- 91 Gasperini M, Hill A J, McFaline-Figueroa J L, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 2019, 176: 377–390.e19
- 92 Tian R, Gachechiladze M A, Ludwig C H, et al. CRISPR interference-based platform for multimodal genetic screens in human iPSC-derived neurons. *Neuron*, 2019, 104: 239–255.e12
- 93 Shifrut E, Carnevale J, Tobin V, et al. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, 2018, 175: 1958–1971.e15
- 94 Tian R, Abarientos A, Hong J, et al. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nat Neurosci*, 2021, 24: 1020–1034
- 95 Hill A J, McFaline-Figueroa J L, Starita L M, et al. On the design of CRISPR-based single-cell molecular screens. *Nat Methods*, 2018, 15: 271–274
- 96 Xie S, Armendariz D, Zhou P, et al. Global analysis of enhancer targets reveals convergent enhancer-driven regulatory modules. *Cell Rep*, 2019, 29: 2570–2578.e5
- 97 Genga R M J, Kernfeld E M, Parsi K M, et al. Single-cell RNA-sequencing-based CRISPRi screening resolves molecular drivers of early human endoderm development. *Cell Rep*, 2019, 27: 708–718.e10
- 98 Wessels H H, Méndez-Mancilla A, Hao Y, et al. Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq. *Nat Methods*, 2023, 20: 86–94
- 99 Mimitou E P, Lareau C A, Chen K Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol*, 2021, 39: 1246–1258
- 100 Gubin M M, Esaulova E, Ward J P, et al. High-dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful immune-checkpoint cancer therapy. *Cell*, 2018, 175: 1443
- 101 Schiebinger G, Shu J, Tabaka M, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 2019, 176: 1517
- 102 Replogle J M, Saunders R A, Pogson A N, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 2022, 185: 2559–2575.e28
- 103 Lamb J, Crawford E D, Peck D, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 2006, 313: 1929–1935
- 104 Ren X, Wen W, Fan X, et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 2021, 184: 5838
- 105 Blattman S B, Jiang W, Oikonomou P, et al. Prokaryotic single-cell RNA sequencing by *in situ* combinatorial indexing. *Nat Microbiol*, 2020, 5: 1192–1201
- 106 Kuchina A, Brettner L M, Paleologu L, et al. Microbial single-cell RNA sequencing by split-pool barcoding. *Science*, 2021, 371: eaba5257
- 107 Wei Z, Si D, Duan B, et al. PerturbBase: a comprehensive database for single-cell perturbation data analysis and visualization. *Nucleic Acids Res*, 2025, 53: D1099–D1111
- 108 Fustero-Torre C, Jiménez-Santos M J, García-Martín S, et al. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. *Genome Med*, 2021, 13: 187
- 109 Suphavitai C, Chia S, Sharma A, et al. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Med*, 2021, 13: 189
- 110 Yuan B, Shen C, Luna A, et al. CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell Syst*, 2021, 12: 128–140.e4
- 111 Jin K, Schnell D, Li G, et al. CellDrift: inferring perturbation responses in temporally sampled single-cell data. *Brief BioInf*, 2022, 23: bbac324
- 112 Hetzel L, Boehm S, Kilbertus N, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Adv Neur Inform Proc Systems*, 2022, 35: 26711–26722
- 113 Bunne C, Krause A, Cuturi M. Supervised training of conditional monge maps. *Adv Neur Inform Proc Systems*, 2022, 35: 6859–6872
- 114 Weinberger E, Lin C, Lee S I. Isolating salient variations of interest in single-cell data with contrastiveVI. *Nat Methods*, 2023, 20: 1336–1345
- 115 Huang W, Liu H. Predicting single-cell cellular responses to perturbations using cycle consistency learning. *Bioinformatics*, 2024, 40: i462–i470
- 116 Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 2014, 15: 550
- 117 Anchang B, Davis K L, Fienberg H G, et al. DRUG-NEM: optimizing drug combinations using single-cell perturbation response to account for intratumoral heterogeneity. *Proc Natl Acad Sci USA*, 2018, 115: E4294
- 118 Jiang J, Chen S, Tsou T, et al. D-SPIN constructs gene regulatory network models from multiplexed scRNA-seq data revealing organizing

- principles of cellular perturbation response. *bioRxiv*, 2024, doi: 10.1101/537364
- 119 Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010, 26: 139–140
 - 120 Lotfollahi M, Rybakov S, Hrovatin K, et al. Biologically informed deep learning to query gene programs in single-cell atlases. *Nat Cell Biol*, 2023, 25: 337–350
 - 121 Yang X, Liu G, Feng G, et al. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res*, 2024, 34: 830–845
 - 122 Yang Y, Li G, Zhong Y, et al. Gene knockout inference with variational graph autoencoder learning single-cell gene regulatory networks. *Nucleic Acids Res*, 2023, 51: 6578–6592
 - 123 Wu Y, Barton R A, Wang Z, et al. Predicting cellular responses with variational causal inference and refined relational information. In: Proceedings of the 11th International Conference on Learning Representations, ICLR 2023, <https://openreview.net/forum?id=ICYasJBIZNs>
 - 124 Ryu J, Lopez R, Bunne C, et al. Cross-modality matching and prediction of perturbation responses with labeled Gromov-Wasserstein optimal transport. In: Proceedings of the ICML 2024 Workshop on Structured Probabilistic Inference and Generative Modeling, 2024, <https://openreview.net/forum?id=ZXq6M04MX6>
 - 125 Zinati Y, Takiddeen A, Emad A. GRouNdGAN: gRN-guided simulation of single-cell RNA-seq data using causal generative adversarial networks. *Nat Commun*, 2024, 15: 4055
 - 126 Hawkins D Y, Zuch D T, Huth J, et al. ICAT: a novel algorithm to robustly identify cell states following perturbations in single-cell transcriptomes. *Bioinformatics*, 2023, 39: btad278
 - 127 Palma A, Theis F J, Lotfollahi M. Predicting cell morphological responses to perturbations using generative modeling. *Nat Commun*, 2025, 16: 505
 - 128 Burkhardt D B, Stanley Iii J S, Tong A, et al. Quantifying the effect of experimental perturbations at single-cell resolution. *Nat Biotechnol*, 2021, 39: 619–629
 - 129 Dann E, Henderson N C, Teichmann S A, et al. Differential abundance testing on single-cell data using k -nearest neighbor graphs. *Nat Biotechnol*, 2022, 40: 245–253
 - 130 Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*, 2014, 32: 381–386
 - 131 Inecik K, Uhlmann A, Lotfollahi M, et al. Multicpa: multimodal compositional perturbation autoencoder. *bioRxiv*, 2022, doi: 10.1101/2022.07.08.499049
 - 132 Schaar A C, Tejada-Lapuerta A, Palla G, et al. Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv*, 2024, doi: 10.1101/2024.04.15.589472
 - 133 Doncevic D, Herrmann C, Martelli P L. Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations. *Bioinformatics*, 2023, 39: btad387
 - 134 Wolf F A, Hamey F K, Plass M, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol*, 2019, 20: 59
 - 135 Gonzalez G, Herath I, Veselkov K, et al. Combinatorial prediction of therapeutic perturbations using causally-inspired neural networks. *bioRxiv*, 2024, doi: 10.1101/2024.01.03.573985
 - 136 Song B, Liu D, Dai W, et al. Decoding heterogeneous single-cell perturbation responses. *Nat Cell Biol*, 2025, 27: 493–504
 - 137 Chen S, Rivaud P, Park J H, et al. Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. *Proc Natl Acad Sci USA*, 2020, 117: 28784–28794
 - 138 Ishikawa M, Sugino S, Masuda Y, et al. RENGE infers gene regulatory networks using time-series single-cell RNA-seq data with CRISPR perturbations. *Commun Biol*, 2023, 6: 1290
 - 139 Yang F, Wang W, Wang F, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell*, 2022, 4: 852–866
 - 140 Nicol P B, Paulson D, Qian G, et al. Robust identification of perturbed cell types in single-cell RNA-seq data. *Nat Commun*, 2024, 15: 7610
 - 141 Lei W, Yuan M, Long M, et al. scDR: predicting drug response at single-cell resolution. *Genes*, 2023, 14: 268
 - 142 Hsieh C Y, Wen J H, Lin S M, et al. scDrug: from single-cell RNA-seq to drug response prediction. *Comput Struct Biotechnol J*, 2023, 21: 150–157
 - 143 Hao M, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics. *Nat Methods*, 2024, 21: 1481–1491
 - 144 Littman R, Cheng M, Wang N, et al. SCING: inference of robust, interpretable gene regulatory networks from single cell and spatial transcriptomics. *iScience*, 2023, 26: 107124
 - 145 Li C, Shao X, Zhang S, et al. scRank infers drug-responsive cell types from untreated scRNA-seq data using a target-perturbed gene regulatory

- network. *Cell Rep Med*, 2024, 5: 101568
- 146 Kana O, Nault R, Filipovic D, et al. Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. *Patterns*, 2023, 4: 100817
- 147 Gao Y, Wei Z, Dong K, et al. Toward subtask-decomposition-based learning and benchmarking for predicting genetic perturbation outcomes and beyond. *Nat Comput Sci*, 2024, 4: 773–785
- 148 Adduri A, Gautam D, Bevilacqua B, et al. Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv*, 2025, doi: 10.1101/2025.06.26.661135
- 149 Lopez R, Tagasovska N, Ra S, et al. Learning causal representations of single cells via sparse mechanism shift modeling. In: *Proceedings of the Conference on Causal Learning and Reasoning*, PMLR, 2023. 662–691
- 150 Xu J, Luo X, Pan X, et al. Alleviating the sample selection bias in few-shot learning by removing projection to the centroid. *Adv Neur Inform Proc Systems*, 2022, 35: 21073–21086
- 151 Lotfollahi M, Naghipourfar M, Theis F J, et al. Conditional out-of-distribution generation for unpaired data using transfer VAE. *Bioinformatics*, 2020, 36: i610–i617
- 152 Zheng Y, Schupp J C, Adams T, et al. A deep generative model for deciphering cellular dynamics and *in silico* drug discovery in complex diseases. *Nat Biomed Eng*, 2025, doi: 10.1038/s41551-025-01423-7
- 153 Seninge L, Anastopoulos I, Ding H, et al. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat Commun*, 2021, 12: 5684
- 154 Gong J, Hao M, Cheng X, et al. xTrimoGene: an efficient and scalable representation learner for single-cell RNA-seq data. *Adv Neur Inform Proc Systems*, 2023, 36: 69391–69403
- 155 Nguyen H C T, Baik B, Yoon S, et al. Benchmarking integration of single-cell differential expression. *Nat Commun*, 2023, 14: 1570
- 156 Squair J W, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. *Nat Commun*, 2021, 12: 5692
- 157 Huang M, Wang J, Torre E, et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods*, 2018, 15: 539–542
- 158 Wolf F A, Angerer P, Theis F J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*, 2018, 19: 15
- 159 Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell*, 2019, 177: 1888–1902.e21
- 160 Razghandi M, Zhou H, Erol-Kantarci M, et al. Variational autoencoder generative adversarial network for synthetic data generation in smart home. In: *Proceedings of the ICC 2022-IEEE International Conference on Communications*, IEEE, 2022. 4781–4786
- 161 Wang J, Wang K, Yu Y, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat Med*, 2025, 31: 609–617
- 162 Zhu J, Wang J, Wang X, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nat Biotechnol*, 2021, 39: 1444–1452
- 163 Das D, Chakrabarty B, Srinivasan R, et al. Gex2SGen: designing drug-like molecules from desired gene expression signatures. *J Chem Inf Model*, 2023, 63: 1882–1893
- 164 Ahlmann-Eltze C, Huber W, Anders S. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear methods. *bioRxiv*, 2024, doi: 10.1101/613342
- 165 Regev A, Teichmann S A, Lander E S, et al. The Human Cell Atlas. *eLife*, 2017, 6: e27041
- 166 Abhulimen E S. The Human Cell Atlas: promises, recent developments, and bridging the African single-cell data gap. *Afr J Lab Med*, 2024, 13: 2583

Summary for “单细胞扰动建模研究进展”

Advances in single-cell perturbation modeling

Xiaoning Qi^{1,2}, Lianhe Zhao¹, Runsheng Chen³ & Yi Zhao^{1*}

¹ Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

² University of Chinese Academy of Sciences, Beijing 100049, China

³ Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

* Corresponding author, E-mail: biozy@ict.ac.cn

Single-cell perturbation technologies have emerged as powerful tools for dissecting cellular responses to diverse perturbations at single-cell resolution. By integrating perturbations with high-throughput single-cell sequencing techniques, these methods enable the comprehensive profiling of transcriptomic, epigenomic, proteomic, and morphological responses across thousands to millions of individual cells. Such advancements have significantly broadened the scope of biological research, facilitating the detailed characterization of gene function, the investigation of disease mechanisms, and the acceleration of drug discovery.

Recently, the surge of large-scale single-cell perturbation datasets has driven a paradigm shift in biomedical research, from traditional hypothesis-driven frameworks to data-driven discovery. Perturbation modeling aims to reveal how perturbations influence cell states and behaviors, encompassing key tasks including: (1) perturbation effect interpretation: quantifying perturbation-induced changes at gene, gene program and cell levels; (2) perturbation response prediction: predicting cellular omics and phenotype outcomes under unobserved perturbations or novel cell type; (3) perturbation interaction prediction: identifying synergistic or antagonistic interactions between multiple perturbations; and (4) perturbation generation and design: computationally optimizing or synthesizing perturbations to achieve desired cellular states. Recent advances in artificial intelligence have significantly promoted the development of single-cell perturbation modeling. These data-driven approaches enable more accurate and scalable modeling of complex cellular systems, leading to deeper insights into the dynamic landscape of perturbation responses.

In this review, we provide a comprehensive overview of the recent progress in single-cell perturbation modeling. We systematically summarize key modeling tasks and representative studies. In terms of single-cell perturbation technologies, we curate and categorize 32 distinct single-cell perturbation technologies, encompassing both genetic (e.g., CRISPR-based perturbations) and chemical (e.g., drug-induced perturbations) perturbation technologies. These technologies have generated a rich and growing collection of 54 datasets. Some of these datasets include multi-omics and multi-modal data, providing a valuable resource for model training, evaluation, and biological discovery. In terms of perturbation modeling, we systematically categorize computational models according to the core modeling tasks and underlying algorithms. These models range from traditional statistical inference and early machine learning models to more advanced deep learning-based models. Deep learning-based models include graph neural networks (GNNs) for modeling cellular and regulatory relationships, and deep generative models, such as variational autoencoders (VAEs), optimal transport (OT) models, diffusion models, and foundation models, that enable perturbation response prediction and data generation. We further discuss the emerging trend of developing large-scale, foundation models inspired by large language models (LLMs), which aim to capture universal patterns of perturbation effects across diverse biological contexts.

Finally, we critically evaluate the current limitations and challenges in single-cell perturbation modeling. Despite significant advances, several critical challenges remain. These include the integration of heterogeneous data modalities, the lack of standardized benchmarks, and the limited interpretability of complex models. We also highlight several promising directions for future research, such as the construction of comprehensive perturbation atlases, the development of generalizable perturbation foundation models, integration of multi-modal perturbation modeling, and the exploration of virtual cells. Single-cell perturbation, perturbation modeling, single-cell omics, perturbation response, artificial intelligence.

single-cell perturbation, perturbation modeling, single-cell omics, perturbation response, artificial intelligence

doi: [10.1360/CSB-2025-0682](https://doi.org/10.1360/CSB-2025-0682)