

精读导读 · Bunne, Roohani et al. 2024

《How to build the virtual cell with artificial intelligence: Priorities and opportunities》

Cell 187 卷第 25 期 7045–7063 页, 2024 年

42 位作者, 第一作者 Charlotte Bunne (EPFL)、Yusuf Roohani、Yanay Rosen; 通讯作者包括 Theofanis Karaletsos (CZI)、Aviv Regev (Genentech)、Emma Lundberg、Jure Leskovec、Stephen R. Quake

免费预印本 (作者自行公开): <https://arxiv.org/abs/2409.11654> 网页版可直接读: <https://arxiv.org/html/2409.11654v1> Cell 正式版: [https://www.cell.com/cell/fulltext/S0092-8674\(24\)01332-1](https://www.cell.com/cell/fulltext/S0092-8674(24)01332-1)

关于这份导读

这不是论文的翻译或复述, 也不能替代原文。它是配着原文一起用的教学材料: 告诉你这篇是什么性质的文献、该用什么方式读、核心概念是什么、该警惕什么、和教程哪几章对应。

原文请从上面的 arXiv 链接自取 (作者自己公开的预印本, 免费)。

配套教程: 《虚拟细胞: 从入门到精通》第 15、17、18、19 章。

一、先搞清楚：这是一篇什么性质的文献

这一节最重要。如果你用读研究论文的方式读这篇，你会读错。

它不是研究论文。它没有做实验，没有提出新方法，没有报告结果。它是一篇 **Perspective**（观点/纲领文章）——它在定义一个议程。

理解这一点，先看作者列表：**42 位作者**。来自 EPFL、斯坦福、CZI、Genentech、Google、微软、Arc 研究所、Broad、EMBL、哈佛、哥伦比亚、多伦多……名单里有 Aviv Regev、Stephen Quake、Jure Leskovec、Fabian Theis、Eric Horvitz、Jay Shendure、Marinka Zitnik、Christina Theodoris、王波、Theofanis Karaletsos。

这基本上把 AI 虚拟细胞领域的顶级玩家一网打尽了。

所以这篇文献的性质是：**一份集体共识文件**。它不是"某个实验室的观点"，而是"这个领域的主要力量商量好了要往哪走"。

这决定了它的读法：

	研究论文	这篇（纲领文章）
它在说	我发现了什么	我们应该去做什么
你该问	结果可信吗？	这个议程合理吗？我的机会在哪？
重点在	方法与验证	概念框架与优先级
读完你该	会复现/质疑它	知道领域往哪走、自己往哪站

一句话：读 Karr 2012 是为了学会"怎么做"，读这篇是为了知道"要往哪做"、以及"哪里还是空白"。

二、核心主张

如果只能记住一句话：

不再由人手把机制写成方程，而是让大规模神经网络直接从海量实验数据中"学"出细胞的模型——他们称之为 AI 虚拟细胞（AIVC）：一个多尺度、多模态、基于大型神经网络的模型，能表征并模拟分子、细胞、组织在不同状态下的行为。

注意这句话与 Karr 2012 的对立有多彻底。这是两种世界观的正面碰撞：

	Karr 2012（机理）	Bunne 2024（AI）
规则从哪来	人手写方程	网络从数据里学
你必须知道	每个反应和参数	尽可能多的数据
强项	可解释、有因果、能问"为什么"	可扩展、能抓住没方程的规律
死穴	参数、扩展性	可解释性、泛化、数据依赖

读这篇时，请始终把 Karr 那篇放在脑子里做对照。这两篇合起来，就是这个领域的两极。你如果先读了 Karr 的导读，你已经知道答案：正因为手写方程的路在复杂真核细胞面前几乎走不通（参数问题、扩展性问题），才有了这篇的存在。这篇是对那个瓶颈的回应。

三、两个必须吃透的核心概念

这篇文章贡献了两个概念工具。其他内容你都可以忘，这两个不能忘——它们是这个领域后续所有讨论的通用词汇。

3.1 通用表征 (Universal Representations, UR)

要解决的问题：生物数据是碎片化的。单细胞测序是一种数据，蛋白质图像是另一种，空间转录组是第三种，扰动筛选是第四种。它们格式不同、尺度不同、模态不同。传统做法是每种数据配一套专门的分析方法，彼此不通。

思路：把所有这些不同类型、不同尺度的生物数据，统一编码到一个**共享的嵌入空间**里——也就是说，无论输入是什么模态，都变成同一个空间里的向量。而且关键是，这个表征要能**泛化到训练数据之外的新状态**。

为什么你该在意：这正是教程第 1 章说"线性代数在两条路线里都是地基"的地方。第 16 章你用 scanpy 做 PCA、把细胞压到几十维——那就是最简单的"表征"。scVI 用神经网络学表征，是进阶版。UR 是这个思路推到极致的愿景：一个能容纳一切生物数据的空间。

动手对照：读到这个概念时，回去看你跑过的那张 UMAP 图。图上每个点是一个细胞，它的位置就是它的"表征"。UR 的愿景，是让这个空间不只装得下 scRNA-seq，还装得下图像、空间信息、扰动响应——而且还能推广到没见过的细胞。

3.2 虚拟仪器 (Virtual Instruments, VI)

要解决的问题：有了表征之后，拿它干什么？

思路：在这个表征空间里做**计算机里的实验**——虚拟地扰动一个细胞、预测它的反应、观察结果。做那些在湿实验室里成本过高、或者根本做不到的实验。而且强调这种体外 (in silico) 实验要是**可解释的**。

为什么"仪器"这个比喻很妙：显微镜是一种仪器，流式细胞仪是一种仪器——它们让你**观察**细胞。虚拟仪器让你在计算机里**观察和操纵**细胞。这个类比在暗示一个野心：AI 虚拟细胞不是一个"模型"，而是一整套**新的实验设备**。

连接：这正是教程第 18 章"扰动预测"的理论包装。GEARS、CPA 这些方法，都是虚拟仪器的具体实现尝试。而虚拟细胞挑战赛（第 18 章），就是在检验这些"仪器"到底准不准。

四、怎么读这类纲领文献

4.1 读法建议

这篇不长，但别硬读。建议路径：

1. 先读摘要和引言，抓住 AIVC 的愿景是什么。
2. 重点啃 UR 和 VI 这两个概念的定义部分。这是全文的钢筋。
3. 浏览它对能力、数据、评测的讨论——不必逐字，抓住"他们认为需要什么"。
4. 特别注意它自己承认的挑战那部分。纲领文章里最诚实、也最有价值的，往往是它列出的困难，而不是它描绘的愿景。那些困难就是没人解决的问题，就是你的机会。
5. 看参考文献。这篇的参考文献是一份高质量的领域书单——GEARS、CPA、scVI、各种基础模型，都在里面。顺着它的引用往下挖，比自己盲目搜索高效得多。

4.2 别做的事

- 别把它当技术手册。它不会教你怎么训练一个模型。它是路线图，不是施工图。
 - 别指望它给你确定的答案。它提的是优先级和机会，很多地方是开放的。
 - 别被作者阵容吓住而放弃独立判断。见下一节。
-

五、批判性阅读：读纲领文献最需要警惕什么

这一节是这份导读的核心价值。

5.1 愿景与现实之间的差距

这篇描绘的是一个愿景，不是一个已经做成的东西。读的时候必须时刻分清哪些是"已经做到了"，哪些是"我们希望做到"。

对照现实：教程第 17 章讲过，多项严肃的基准研究发现，现有的细胞基础模型在某些任务上并不一定超过简单的、针对特定任务设计的传统方法；在扰动特异性效应、罕见细胞类型、分布外泛化上常常表现挣扎。

所以正确的姿势是：读这篇了解愿景 → 再去读一篇基准/批评论文看现实 → 差距就是研究机会。只读纲领不读批评，你会被 hype 带走；只读批评不读纲领，你会看不到方向。

5.2 "42 位作者的共识"也可能是"42 位利益相关者的共识"

这话说得直白一点：这个名单里有 CZI（正在建 GPU 集群和虚拟细胞平台）、有 Genentech、Google、微软、Arc 研究所。他们既是这个议程的制定者，也是这个议程的受益者。

这不是在指控任何人不诚实。一个领域的领军者推动他们相信的方向，这完全正当，而且科学就是这么进步的。但作为读者，你要清楚：

一份集体共识文件的权威性，来自它的作者阵容；而它的盲区，也同样来自它的作者阵容。

问自己：如果一个持怀疑态度的、不在这个名单里的研究者来写这份路线图，会写得一样吗？哪里不一样？这个问题的答案，往往就藏着最有意思的研究方向。

5.3 数据的因果性问题

这是我认为整个 AI 路线最深的软肋，也值得你在读的时候特别留意：

现有的大部分数据是观察性的快照——你测了一堆细胞在某个时刻的状态。但你想要的是因果：改变 X 会导致 Y 吗？

没有因果数据，模型学到的就只是相关。这是教程第 22 章列的第一大挑战。扰动数据（CRISPR 敲除、加药）正是为了解决这个，但它昂贵、覆盖有限、而且组合空间是天文数字（两万个基因的两两组合有两亿种）。

读的时候问：他们打算怎么解决这个？他们的答案够有说服力吗？

5.4 评测的问题

什么才算"一个好的虚拟细胞"?

这个领域至今没有像 CASP 之于蛋白质结构那样公认的评测标准。没有好的评测，就无法衡量进步，也无法比较方法。这也是教程第 22 章列的挑战之一。

留意这篇怎么谈评测——然后去看虚拟细胞挑战赛（下一节）如何把这件事往前推了一步。

六、与教程的对应关系

教程章节	对应这篇的什么
第 15 章	为什么要数据驱动；AIVC 愿景；UR 与 VI 的引入
第 16 章	单细胞数据与表征（UR 的技术基础）
第 17 章	细胞基础模型；以及对它们的批判性评估
第 18 章	扰动预测（VI 的核心实现）
第 19 章	CZI 平台、Arc 研究所、基准生态
第 20 章	两种文化的融合——读完这篇再回看这章
第 22 章	开放问题（这篇自己承认的挑战）

建议：读这篇之前，先跑通教程第 16 章的 scanpy 流程。你亲手把细胞压成向量、画出 UMAP 之后，"通用表征"这个概念对你就不再是抽象名词。

七、读完自测

1. 这篇是研究论文还是纲领文章？这个区别如何改变你的读法？
2. 用你自己的话解释"通用表征"和"虚拟仪器"，并各举一个你在教程里动手做过的、与之对应的具体例子。
3. 这篇的主张与 Karr 2012 在世界观上如何对立？为什么这个对立是**历史性的**（提示：Karr 那条路遇到了什么瓶颈）？
4. 这篇描绘的愿景中，哪些已经实现了，哪些还没有？你如何验证你的判断？
5. 42 位顶级作者的共识，在什么意义上增加了这篇的分量？在什么意义上可能构成盲区？
6. 为什么"因果性"是 AI 虚拟细胞最深的软肋？扰动数据能完全解决它吗？
7. 如果让你挑一个这篇提出但还没人解决好的问题去做，你会挑哪个？为什么？

第 7 题没有标准答案——但它是这七题里最重要的一题。 它是你从"读者"转向"研究者"的那一步。

八、下一步

8.1 立刻接着读的三样东西

读完这篇，**不要停在愿景里**。按顺序做这三件事，你才能看到全貌：

1. **Roohani et al. 2025, Cell** —— **虚拟细胞挑战赛** ("迈向虚拟细胞的图灵测试")。看这个愿景如何被转化成一个**可衡量的具体任务**：给定起始细胞状态和一个基因扰动，预测扰动后的基因表达。特别注意它的设计智慧——用一个与常见训练数据分布不同的细胞系，**强迫模型泛化而非记忆**。
2. 一篇**基准 / 批评**论文。去搜近期评测细胞基础模型的论文（教程附录 A 有提示）。**这是解毒剂**。读完你会对"现在到底做到了什么"有清醒认识。
3. **CZI 虚拟细胞平台** (virtualcellmodels.cziscience.com)。别只读，去用一次。下载一个模型跑跑看。亲手用一次，胜过读十篇综述。

8.2 你的机会在哪

这篇最实际的用处，是当作一张**机会地图**来读。带着这个问题读它：

哪些地方，既是这篇承认的空白，又是我这个懂生物的人有优势的地方？

我的判断（供你参考，不必同意）：

- **评测与基准**：这个领域缺"什么才算好"的标准。而判断"一个预测在生物学上是否合理"，恰恰需要真懂生物的人。**这是懂生物的人的主场**。
- **数据的因果设计**：什么样的扰动实验最有信息量？这需要生物学判断，不是算法问题。
- **机理与学习的融合**（教程第 20 章）：能同时说两种语言的人极少。你读完 Karr + Bunne 两篇，已经站在这个稀缺位置上了。
- **可解释性**：模型预测准了，但它"为什么"这么预测？把黑箱的输出翻译成生物学洞见，需要生物学家。

最后一句

这篇文章的价值不在于它说对了什么——纲领文章的预测常常会被现实修改。它的价值在于：**它让整个领域第一次有了共同的词汇（UR、VI）和共同的议程**。

而对你个人来说，它最大的用处是让你看清一件事：**这个领域还年轻到，一个懂生物、肯补计算的人，仍然来得及站进去**。那 42 位作者列出的困难，没有一个是已经解决的。