研究开发出细胞身份鉴定新型AI引擎

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/34167.html

研究开发出细胞身份鉴定新型AI引擎
。随着单细胞和空间组学技术的快速发展，公开可共享数据量已突破亿级大关。然而，技术平台产生的差异、复杂疾病状态、跨物种研究带来的批次效应和离群细胞等，对数据解读构成挑战。面对动辄百万规模的离群细胞，传统的“先聚类、后注释”分析方法难以快速、精准且可解释地将这些“身份不明”的细胞映射到参考细胞图谱上，进而制约单细胞数据在跨大规模人群队列研究、多模态信息整合以及物种间保守性探索等领域的应用潜力。因此，亟需高效实现细胞的数字化表征、整合与解析。

针对上述问题，中国科学院北京基因组研究所（国家生物信息中心）研究员蒋岚团队联合新加坡国立大学教授刘钿渤、加拿大麦吉尔大学教授
李岳，研发了一款高效、泛化且可解释的有监督细胞表征和解析模型——CellMemory。该模型受全局工作空间理论启发，对传统Transformer架构进行改造，即植入低维记忆空间并通过Cross-Attention机制将高维基因特征压缩、竞争及广播。研究显示，该模型可提高计算效率3至5倍，并显著增强模型泛化能力，无需预训练即可实现单细胞数据跨平台与物种整合。同时，记忆空间可为CellMemory带来分层式“可读窗口”。其中，L1 (Gene
Level)为面对特定细胞，研究可知单个基因对目标细胞表征的贡献分数；L2（Gene Program Level）为模型在记忆空间中，自动归纳协调的共表达/共调控模式。多层可解释性为理解模型决策逻辑与探索表型关联细胞状态提供了可靠解决方案，即“高准确性 + 强可解释性”。

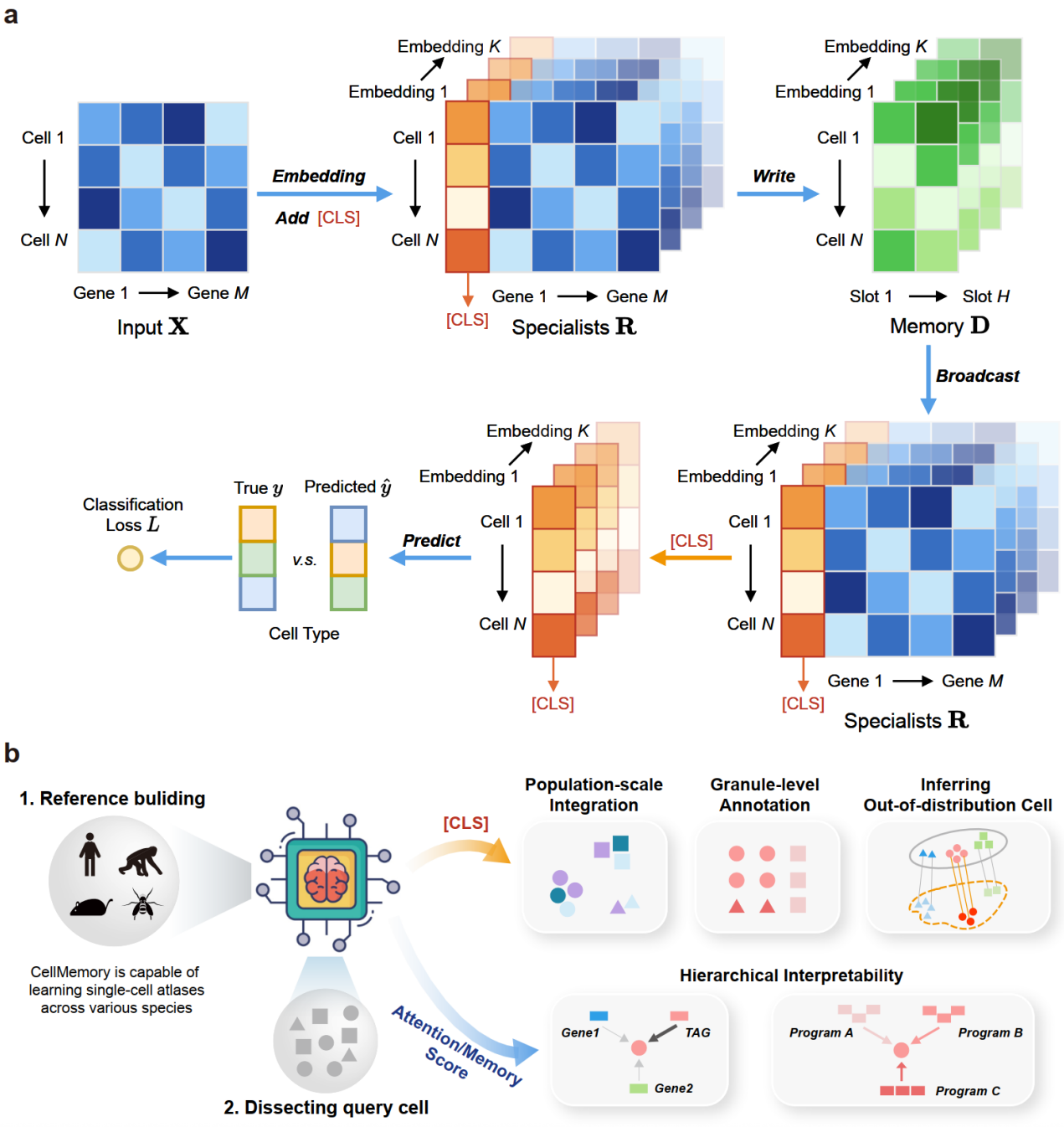
进一步，研究人员将CellMemory与3个单细胞基础大模型、16个任务专用模型在1500万细胞上进行比较。基准评测结果显示，CellMemory在人群尺度的单细胞数据整合、超高分辨率细胞状态注释等任务中均取得了State-of-the-Art级别的表现。同时，面对59张共含400万细胞、338个细胞亚群的MERFISH小鼠脑空间组学切片，与基于传统transformer架构预训练的单细胞基础大模型相比，CellMemory在95%的空间切片上展现领先的注释表现，准确率较传统机器学习方法提升30%，证明了CellMemory较好的泛化能力。

当前，将疾病细胞与健康细胞比对存在挑战。得益于准确与可解释的细胞表征，研究人员利用CellMemory在多个癌症队列单细胞图谱中解析疾病复杂性。例如，在肺腺癌队列中，该模型基于参考图谱定位到$MSLN^{+}$
$CAPN8^{+}$
的肺泡2型过渡态细胞，并观测到其显著的拷贝数变异，提示肺腺癌或利用肺泡2型细胞可塑性获得侵袭能力。同时，在混合表型急性白血病、髓母细胞瘤等数据中，该模型基于健康参考图谱，揭示了不同患者潜在的异质性起源，为耐药和预后研究提供了高分辨率数据解析基础，展示出CellMemory在离群细胞推断场景中较好的表征能力。

从“序列搜索”到“亚群搜索”，参考映射正在重塑单细胞数据分析的技术范式。得益于较好的泛化能力与高效的计算效率，CellMemory有望成为覆盖病理、时空及物种等多维度细胞参考图谱建设与临床精准诊疗的关键引擎。

近日，相关研究成果以*CellMemory: hierarchical interpretation of out-of-distribution cells using bottlenecked transformer*为题，发表在《基因组生物学》（*Genome Biology*）上。研究工作得到科学技术部、中国科学院等的支持。

论文链接



CellMemory模型架构与应用场景

研究团队单位：北京基因组研究所