

生信方法 | 2026-08 月度文献报告

碱基日读 · 作者 碱基 · 数据来源 PubMed / bioRxiv 官方接口 · 本月收录 46 篇，正文详述其中 30 篇

本月概览

2026年8月，生信方法方向的30篇文献可粗略划分为四个线程。第一，单细胞转录组和3D基因组工具以高密度态势堆叠，transformer架构的渗透率高到不再被作为独立卖点提及，而是悄然成为方法默认选项。第二，跨组学模态整合在基础研究与临床应用两端同步推进：从scATAC-seq标注迁移到纵向临床多模态合成，从BCR序列与转录组轨迹融合到组织切片上的代谢-身份联合光学读出。第三，宏基因组分箱的竞争重心从算法精度转向规模化可重复部署，两项独立工作在工程层面并行推进。第四，也是最值得关注的：一篇方法论反思性预印本以形式化数学框架指出，当前模型评估体系存在系统性天花板，而这个天花板由实验可重复性固定，与被评估的模型无关。这批文献整体质量参差，亮点集中在少数方法论清晰、基准测试规范的工作，大多数工具论文停留在与现有方法的公开数据集比较，独立验证稀缺。

值得细看的研究

Assay concordance 揭示预测模型的评估天花板

这篇bioRxiv预印本 (DOI: 10.64898/2026.08.24.746774) 是本月最值得认真读的一篇，且它不做任何新算法。作者的核心论点是一个数学事实：当多个实验测定结果被当作“可互换真值”平均评分时，模型性能的上界由这些测定之间的一致性固定，与模型本身无关，且该天花板通过一个单一的支撑函数等式随评估指标变化。作者横跨2,822个MaveDB score set、217个ProteinGym assay、两个药物筛选数据集和1,150个CRISPR细胞系测量了这一天花板。关键数字：同一靶标的两次测定之间的一致性为0.56-0.68；对541个被两种蛋白酶分别测定的结构域，估计测定可靠性为0.897，意味着70-90%的天花板来自“不可约的生物学”而非实验噪声。结果：已发表的预测器在相关系数基准上只实现了可达天花板的63%，在用户真正关心的top-1%选择任务上仅实现了18%。

这个发现的意义不在于某一个模型好不好，而在于整个评估体系是否合理。如果两次测同一件事的内在一致性只有0.6，那么模型跑到0.5可能已接近可达极限，而非“还差得远”。该文尚未经过同行评审，但其核心逻辑在数学上是封闭的，应当优先验证而非忽视。

PROFET：轨迹推断的系统性改进与耐药亚群识别

PROFET (PMID: 42664975) 解决从稀疏时序scRNA-seq数据推断连续非线性细胞动态轨迹的问题，技术路线是粒子梯度流算法与无仿真力匹配的结合。基准测试覆盖小鼠与人类体外数据集及一个体内蝾螈再生数据集，对比十种现有轨迹推断方法，PROFET实现了2.6-12.5倍更低的预测误差。生物学应用部分更有实质内容：对palbociclib处理的MCF7细胞系新生成scRNA-seq数据及三个已发表乳腺癌患者数据集，PROFET重建了治疗响应轨迹，识别出一个具有大幅表型漂移的耐药亚群，富集表面标志物UNC5B、TLR3、PCDH19、PROCR、SLITRK6和SEMA6B。证据强度中等：算法基准设计规范，生物学发现依赖已发表数据集，独立队列验证缺失。该文通过透明同行评审流程，可信度略高于未评审的预印本。

scPILOT：跨生物学情境的扰动响应迁移

scPILOT (PMID: 42667121) 针对扰动响应预测中"情境泛化"难题：在某一细胞类型中学到的扰动效应，能否迁移到新的细胞类型、患者或物种。框架采用查询条件化设计，通过判别器辅助训练学习生成式潜在表示，并用潜在最优传输进行跨情境响应迁移。在跨细胞类型、患者和物种三个保留集基准上，context-averaged R^2 / MMD 2 分别为0.945/0.137、0.598/0.025和0.853/0.287。

跨患者 R^2 仅0.598应当被单独突出，而不该被跨细胞类型0.945的成绩所遮掩。患者间个体差异是通向临床应用的真实壁垒，而这个数字说明模型在此明确受限。跨物种0.853的表现反而更好，值得追问是否因为所用测试物种的生物学差异较小。总体而言，scPILOT是把扰动响应预测从训练分布内插值向跨分布外推推进了一步，但距离临床个体化预测仍有结构性差距。

REDCAT：代谢状态与细胞身份的同步光学读出

REDCAT (PMID: 42660991) 将拉曼散射显微（读取蛋白、脂质、核代谢物和氧化还原代谢）与高通量免疫荧光（提供细胞身份标记）整合在一个全光学平台，在人类组织切片上实现亚细胞分辨率的代谢-身份联合图谱，无需破坏细胞。在淋巴结中揭示了细胞类型特异的代谢分化；在淋巴瘤中捕获了肿瘤转化过程中的代谢重编程和过渡状态；在肝脏中解析了区域化依赖的代谢梯度。

这个工作的价值不在于发现新的生物学假说，而在于提供了一种此前不存在的测量能力：同时知道"这是什么细胞"和"这个细胞当前的代谢状态是什么"，两者在同一切片上联合读出。主要弱点是摘要未提供定量性能数据，通量与成本的边界不清晰，从原理验证扩展到系统性人群研究的可行性未被讨论。

PULSE：纵向多模态临床数据的自监督框架

PULSE (PMID: 42649447) 的核心出发点是患者数据与单细胞数据的根本区别：患者数据是纵向的、具有历史配对模态的，而非跨细胞的横截面快照。PULSE用自监督方式显式编码历史配对模态来重建后续未配对测量的完整图谱。在英国生物银行数据上，PULSE从常规血液检测生成代谢组学（251个生物标志物）和蛋白质组学图谱，准确率超过所有对照方法。加入视网膜图像、电子健康记录和血液标志物后，基于生成蛋白质组学图谱对六种常见疾病的预测AUC为0.72-0.83，与使用真实蛋白质组数据的结果相当。

证据强度偏高：在大规模真实临床队列上验证，覆盖多疾病，有明确数字支撑。主要局限：其一，AUC 0.72-0.83的绝对值在临床分诊场景是否足够，原文未讨论；其二，"生成的"蛋白质组在真正需要精确预测的极端病例上的表现未被检验；其三，英国生物银行的人群代表性问题客观存在。

Hi-Cformer 与 HiDT：单细胞3D基因组工具并行推进

同月发表的Hi-Cformer (PMID: 42664339) 和HiDT (PMID: 42665445) 分别攻克单细胞Hi-C分析的两个不同层面。Hi-Cformer用专门设计的多尺度注意力机制同时建模不同距离上的染色质接触块，从极稀疏的单细胞接触图谱中学习低维表示，在细胞类型分离和TAD样边界插补上优于现有方法，并支持跨数据集的细胞类型注释。HiDT基于图神经网络加入了注意力增强的边特征层，整合深度特异性归一化，在低至数百万接触数的数据集上维持差异TAD检测的准确性，并被用于肿瘤进展的TAD重组分析和单细胞Hi-C转录异质性识别。

两篇同月发表本身就是一个信号：单细胞3D基因组学已进入工具快速迭代阶段，不同层面的分析需求（表示学习对差异分析）正在被独立工具分别满足。

nf-core/mag v5 与 QuickBin：宏基因组分箱的工程转向

nf-core/mag v5 (PMID: 42640826) 历时七年、经过四个主要维护团队，此次更新增加了长读长分箱与bin精化支持、五个新分箱工具、病毒和真核生物分类扩展，以及新工具和最新数据库支持的bin质量评估。QuickBin (PMID: 42642622) 是CPU原生、无标记基因的分箱算法，将GC-覆盖空间索引与早退式级联相似性检测结合，紧凑神经网络仅用于模糊合并。在297个真实宏基因组的基准测试中，QuickBin完成所有运行，恢复的高质量MAG ($\geq 95\%$ 完整性， $\leq 1\%$ 污染) 数量多于频繁运行失败的资源密集型替代方法。

两者的共同信号：宏基因组分箱领域的竞争重心已从“哪个算法在基准上最优”转向“哪个流程在真实规模数据上可靠完成运行”。可重复性和部署成本正在成为评价维度，与纯精度指标同等重要。

趋势判断

- Transformer已完成在单细胞方法中的工程常规化。 ** scProtoTransformer (PMID: 42664347)、Hi-Cformer (PMID: 42664339)、CARA (PMID: 42665446) 均以transformer为核心，但没有一篇把它当作主要创新点大写特写——它就是工具箱选项之一。方法竞争已从“用不用transformer”转移到“如何设计transformer的归纳偏置”：Hi-Cformer针对多尺度接触图谱设计专用注意力机制；scProtoTransformer引入通路原型tokenizer将基因表达投射到生物学可解释空间。这种局部化创新比整体架构替换更值得关注，因为它说明领域对模型失败模式的理解在加深。

- 跨模态对齐是主旋律，但动机分化导致失败模式不同。 ** CARA (PMID: 42665446) 的跨模态迁移是为弥补scATAC标注稀缺；ClonoTrace (PMID: 42664803) 整合BCR序列是为利用免疫受体的发育信息；PULSE (PMID: 42649447) 的多模态融合是为从廉价测量合成昂贵测量；REDCAT (PMID: 42660991) 的光学整合是为获得原本不可能同时获得的测量通道。把这四类工作放在同一个“多模态”标签下，会掩盖它们截然不同的验证要求和失败条件。

- 评估框架本身正在被形式化质疑，但尚未形成共识。 ** Assay concordance预印本 (DOI: 10.64898/2026.08.24.746774) 提供了迄今最系统的数学处理，PROFET (PMID: 42664975) 的2.6-12.5倍改进声明也隐含对基准设计的质疑。这不是新话题，但本月出现了更严格的形式化处理，且数据规模足够大。如果该天花板框架被主流期刊采纳，方法论文的基准测试写法将被迫改变。

- 被高估的方向：单细胞扰动响应的跨患者预测。 ** 这一方向正在吸引大量资源，scPILOT (PMID: 42667121) 是典型代表。但跨患者 R^2 仅0.598的成绩是一个清晰的警示：在最具临床相关性的场景（跨患者个体异质性）上，当前最优方法的表现仍然距离实用有距离。预测框架被给予的期望——成为AI虚拟细胞研究和药物发现的基础组件——远超当前证据所支持的范围。

这批研究的共同弱点

- 基准测试的内循环问题。 ** 本月大多数算法论文在已发表公开数据集上做基准，这些数据集被领域反复使用，模型开发阶段难以完全排除数据泄露。PROFET (PMID: 42664975) 设计了系统性保留集，scPILOT (PMID: 42667121) 有跨情境保留基准，是少数例外；cIsoFun (PMID: 42664103)、scProtoTransformer (PMID: 42664347) 等工作的基准数据集独立性未被充分说明。

- 样本量与人群代表性局限。 ** POI多组学靶标研究 (PMID: 42663799) 整合了deCODE (N=35,559) 和英国生物银行 (N=54,219) 的蛋白质组数据，人群几乎为欧洲裔，分子对接结论（如cycloheximide结

合EPHA4, -7.8 kcal/mol , $\text{RMSD} < 2.0 \text{ \AA}$) 停留在计算层面, 缺乏实验验证。HOSCA (PMID: 42640827) 整合了超过388万单细胞转录组图谱、704个样本、60个数据集, 跨越16个解剖区域和15种疾病, 但60个来源数据集的批次效应校正质量边界未被定量评估。

- 预印本结论有待同行评审核验。 **

本月清单中6篇为bioRxiv预印本, 其中包含最锐利的方法论反思文章 (assay concordance, DOI:

10.64898/2026.08.24.746774) 、DNA甲基化熵框架 (DOI:

10.64898/2026.08.25.746714) 和随机基因表达PGF推断 (DOI: 10.64898/2026.08.24.746673) 。统计框架类文章的理论证明完备性需要独立核验, 在此之前结论应当保持保留态度。

- 外推性声明超前于实证证据。 ** PULSE (PMID:

42649447) 声称"生成的"蛋白质组图谱与真实蛋白质组数据在六种疾病预测上效果相当 (AUC

0.72-0.83) , 但对真正需要精准预测的高风险边缘案例, 生成图谱的误差分布未被检验。Data-centric

framework综述 (PMID: 42665646) 是规范性倡议而非实证研究, 大量使用"提议"和"可能", 缺乏原创数据支撑, 作为观点文章被当作方法依据引用时应特别注意。

- 测量噪声在整批文献中被系统性低估。 ** Assay concordance预印本的核心发现——70-90%的天花板来自不可约生物学而非噪声——反向说明当前大多数方法论文在讨论模型上限时, 没有将测定本身的变异纳入考量。这是一个系统性缺失, 不针对任何单篇文章, 而是整批文献共同的盲区。lncRNA亚细胞定位分析 (PMID: 42645529) 是少数例外, 该文以两种零模型估计了预期比例, 并报告了在独立实验室、独立协议和独立定量工具下的重现相关性 (Spearman $\rho = 0.71$ 、 0.80 、 0.56) , 设计值得学习。

下个月值得关注

- Assay concordance天花板估计的独立验证。 ** 如果 (DOI: 10.64898/2026.08.24.746774) 的框架被认真对待, 接下来应当出现对其估计量的独立实现, 以及MaveDB和ProteinGym社区的正式回应。值得盯的信号是: 顶级方法期刊的投稿要求是否开始讨论是否应报告assay concordance。

- 单细胞扰动预测的跨患者瓶颈分析。 ** scPILOT (PMID: 42667121) 跨患者 $R^2 0.598$ 是一个清晰的失败位点, 预期接下来会有两类工作出现: 专门针对患者间异质性建模的新方法, 以及对现有方法在患者层级失败模式的诊断性分析。第二类工作往往比第一类更有价值, 因为它能指导社区在哪里集中资源。

- 宏基因组分箱流程的头对头比较。 ** QuickBin (PMID: 42642622) 与nf-core/mag v5 (PMID: 42640826) 代表不同定位, 两套流程在相同真实数据集上的独立对比研究尚缺失。尤其是长读长数据的分箱, nf-core/mag v5本月刚加入支持, 其与现有短读长流程在混合测序数据上的性能比较值得关注。

- lncRNA亚细胞定位的动态维度。 ** lncRNA等显性异构体分析 (PMID: 42645529) 基于静态ENCODE分馏数据, 报告了21/98基因 (21.4%) 存在亚细胞定位被基因层面分析掩盖的情况 (对照零模型预期3.1%) 。自然的延伸问题是: 在动态生物学条件下 (细胞周期、应激响应、分化) 这种异构体特异性定位差异是否改变? 该结论目前有两个独立实验室的重现支持, 但功能性因果关系证据仍缺失。

- REDCAT通量扩展与计算流程开发。 ** 拉曼-免疫荧光整合平台 (PMID: 42660991) 的技术原理已在三个组织类型中演示, 但摘要未量化通量和成本, 从原理验证扩展到系统性人群研究需要配套的计算分析流程。接下来值得关注是否有专门处理这类多模态原始数据的自动化分析工具出现。

附录：2026-08 全部收录文献（46 篇）

#	标题	期刊	日期	PMID
1	Predicting Single-Cell Perturbation Responses Across Biological Contexts With a Deep Generative Model Integrating Optimal Transport.	Advanced science (Weinheim, Baden-Wurtte	2026-Aug-2	42667121
2	Engineering epicardium-integrated human iPSC-derived heart tissue for modelling Hutchinson-Gilford progeria syndrome.	Biofabrication	2026-Aug-2	42486147
3	Identifying novel druggable targets and repurposable drugs for premature ovarian insufficiency by integrated multiomics and causal inference analysis.	GeroScience	2026-Aug-2	42663799
4	Cell Type-specific Isoform Function Prediction by Multiplex Heterogeneous Network.	IEEE transactions on computational biolo	2026-Aug-2	42664103
5	Hi-Cformer enables multiscale chromatin contact map modeling for single-cell Hi-C data analysis.	Science advances	2026-Aug-2	42664339
6	scProtoTransformer: Scalable reference mapping across molecules, cells, and donors.	Science advances	2026-Aug-2	42664347
7	Multimodal computational framework resolves B cell maturation in autoimmunity and ageing.	Journal of autoimmunity	2026-Aug-2	42664803
8	PROFET predicts continuous gene expression dynamics from scRNA-seq data to elucidate heterogeneity of cancer treatment responses.	Cell systems	2026-Aug-2	42664975
9	Identification of differential topologically associating domains from low sequencing depth and pseudo-bulk chromatin contact maps.	Genome research	2026-Aug-2	42665445
10	Robust annotation and discovery of novel cell types in single-cell ATAC-seq data through cross-modal reference alignment.	Genome research	2026-Aug-2	42665446
11	Data-centric feedback loops for next-generation immunotherapy development.	Nature biomedical engineering	2026-Aug-2	42665646
12	Novel immune-related susceptibility loci associated with pediatric steroid-sensitive nephrotic syndrome identified by a transethnic genome-wide association study.	Nephrology, dialysis, transplantation :	2026-Aug-2	41728924
13	All-optical multimodal mapping of single-cell-type-specific metabolic activities via REDCAT.	Nature methods	2026-Aug-2	42660991
14	Label-free quantification of early virus-induced cellular responses by digital holographic tomography.	mSphere	2026-Aug-2	42549912
15	Systematic Review and Transcriptomic Meta-analysis of Environmental Enrichment Reveal Core Molecular Programs of Brain Plasticity.	Molecular neurobiology	2026-Aug-2	42642673
16	Four-Dimensional Electrochemiluminescence Microscopy Enables Depth-Resolved Imaging of Single, Living Spheroids.	Angewandte Chemie (International ed. in	2026-Aug-2	42644292
17	Subcellular localization of human long non-coding RNAs is an isoform-resolved property masked by gene-level analysis.	Molecular genetics and genomics : MGG	2026-Aug-2	42645529
18	Longitudinal alignments and syntheses of multimodal clinical data for personalized medicine with the PULSE framework.	Nature computational science	2026-Aug-2	42649447
19	The immunophenotype and proviral landscape of HIV-infected CD4 T cells during antiretroviral therapy.	Proceedings of the National Academy of S	2026-Aug-2	42611989
20	Engineering Tissue-Relevant Collagen Bundle Architectures via Macromolecular Crowding.	Advanced healthcare materials	2026-Aug-2	42639745

#	标题	期刊	日期	PMID
21	Community-driven updates for comprehensive long-read metagenomics and enhanced binning in nf-core/mag v5.	Bioinformatics (Oxford, England)	2026-Aug-2	42640826
22	HOSCA: Human Ocular Single-Cell Atlas for Decoding Ocular Biology and Disease Complexity.	Genomics, proteomics & bioinformatic	2026-Aug-2	42640827
23	A Scalable High-Density Microwell Assay for Single-Cell Clonal Expansion Profiling.	SLAS technology	2026-Aug-2	42641793
24	Deployable high-fidelity metagenome binning at scale with QuickBin.	Communications biology	2026-Aug-2	42642622
25	Computational Pipeline Reveals Nature's Untapped Reservoir of Halogenating Enzymes.	ACS omega	2026-Aug-2	42662190
26	The molecular similarity landscape of preclinical cancer models to patient tumors.	JCI insight	2026-Aug-2	42378033
27	A whole-blood cryopreservation method streamlines clinical sample collection for multimodal single-cell immune profiling in sepsis.	JCI insight	2026-Aug-2	42446933
28	Rapid assessment of intracellular lipid using a dissolvable fluorescent sensing layer for single-cell, population-level monitoring.	The Analyst	2026-Aug-2	42483767
29	Selective photoacoustic handling of microparticles via annular laser beams.	Microsystems & nanoengineering	2026-Aug-2	42634074
30	Desert-derived Ensifer sp. SA403 enhances potato salt tolerance by reshaping rhizosphere microbiome functions and host responses.	Microbiological research	2026-Aug-2	42636661
31	Identifying memory gene expression from single-sample scRNA-seq data using power-law signatures.	Cell systems	2026-Aug-2	42636813
32	Target enriched metagenomics in cerebrospinal fluid in central nervous system infections.	The Journal of infection	2026-Aug-2	42636981
33	Perspective on Scanning Electrochemical Probe Microscopy in Single-Cell Imaging.	Chemical & biomedical imaging	2026-Aug-2	42657285
34	Bacteria tracking and life cycle state classification using graph neural networks and pretrained vision transformers.	Medical image analysis	2026-Aug-2	42664662
35	A Streamlined Workflow for Purkinje Cell Labeling and High-Resolution Analyses of Dendrites and Spines in Mice.	eNeuro	2026-Aug	42557105
36	An entropy-based diagnostic framework for characterizing methylation state dynamics during preimplantation development	biorxiv preprint	2026-08-26	10.64898/2026.08.25.746714
37	Unsupervised machine-learning identifies latent pyrenoid states linked to mitotic remodeling defects and CO2-dependent growth	biorxiv preprint	2026-08-26	10.64898/2026.08.25.746929
38	Peptidoglycan remodeling prevents antibiotic resistance during oxidative stress	biorxiv preprint	2026-08-26	10.64898/2026.08.19.745765
39	Likelihood-Based Inference and Model Selection for Stochastic Gene Expression in Probability-Generating-Function Space	biorxiv preprint	2026-08-26	10.64898/2026.08.24.746673
40	BfBio: a graph-based tool for the prediction of Angiogenic Stalk Cell genes using a Personalized PageRank algorithm	biorxiv preprint	2026-08-26	10.64898/2026.08.23.746493
41	Assay concordance sets exact ceilings on what one biological score can predict	biorxiv preprint	2026-08-26	10.64898/2026.08.24.746774

#	标题	期刊	日期	PMID
42	Biological Features Underlying Factor-Specific Predictability of Cross-Species Regulatory Occupancy	biorxiv preprint	2026-08-26	10.64898/2026.04.14.718438
43	The distribution of particle-associated Prochlorococcus across the global oceans	biorxiv preprint	2026-08-26	10.64898/2026.08.24.746807
44	Cell Cycle Phases, Spindle Dynamics and Kinesin-5 Motor LocalizationCharacterized by Deep Learning, Dual Segmentation and Decision-Tree Pipeline	biorxiv preprint	2026-08-26	10.64898/2026.08.24.746832
45	Uni-TINT: Unifying T and Innate Lymphoid Cell Taxonomy with a high resolution, pan-disease and pan-tissue single-cell transcriptomic atlas	biorxiv preprint	2026-08-26	10.64898/2026.07.20.739448
46	CDKN3, PKP2, ADM, MS4A1, FAM83A, and DKK1 Define a Posttranslational Modification (PTM)-Associated Prognostic Model for Lung Adenocarcinoma.	Human mutation	2026	42638932

本报告由自动化流水线生成：文献来自 PubMed / bioRxiv

官方接口，正文中的每一处数字与结论都经程序比对原文摘要，比对不通过的内容不会出现在这里。本报告不构成医疗建议。