Simen Owen Academic Proceedings Series

3rd International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Review

Benchmarking Intelligence: A Critical Review of Evaluation Frameworks for AI Agents in Automated scRNA-seq Analysis

Yassel Yu[1,*]

[1] K. International School Tokyo, Tokyo, Japan
* Correspondence: Yassel Yu, K. International School Tokyo, Tokyo, Japan

Abstract: The rapid emergence of autonomous, multi-step analysis-capable artificial intelligence agents is fundamentally transforming the computational biology research paradigm for single-cell RNA sequencing (scRNA-seq). These agents can autonomously navigate complex analytical pipelines, yet the evaluation frameworks currently employed remain anchored in traditional static benchmarking methodologies. Such approaches predominantly rely on endpoint metrics—including the adjusted Rand index (ARI) and normalized mutual information (NMI)—to assess the output of individual algorithms in isolation. These conventional methods fail to capture the core competencies of AI agents, such as multi-step decision-making capacity, adaptability across heterogeneous data conditions, and analytical stability throughout the workflow. This review systematically examines the datasets and evaluation metrics prevalent in contemporary scRNA-seq benchmarking studies, identifying critical methodological shortcomings in their applicability to AI agent assessment. To address these deficiencies, we propose a process-oriented evaluation framework centered on transparency, computational efficiency, and biological validity. The ScAI Bench protocol introduces a multi-dimensional evaluation system encompassing process-tracking assessment, scenario-based testing with publicly available real-world datasets, and rigorous reproducibility verification. Unlike traditional approaches that focus exclusively on final outputs, this framework prioritizes the quality of agent decision-making throughout the analytical process. Its overarching objective is to establish a unified, community-adoptable evaluation standard that aligns with the sophisticated capabilities of AI agents in single-cell analysis. Furthermore, this review provides a strategic roadmap for advancing autonomous analytical methods within single-cell genomics.

Keywords: artificial intelligence; single-cell rna sequencing; evaluation frameworks; benchmarking; reproducibility; computational biology

Received: 10 May 2026
Revised: 23 June 2026
Accepted: 06 July 2026
Published: 12 July 2026

1. Introduction

1.1. The Background of AI Agents in Computational Biology

Single-cell RNA sequencing technology has enabled unprecedented analytical precision in biological research, facilitating in-depth transcriptome-level characterization of individual cells. As this technology advances, associated data analysis methodologies have grown increasingly sophisticated [1, 2]. In response, artificial intelligence agents capable of autonomous decision-making and end-to-end analytical execution are emerging as valuable tools in computational biology [3–5]. These agents can systematically integrate multiple analytical steps—including quality control, batch correction, cell type annotation, and differential expression analysis—thereby enhancing the efficiency and scalability of single-cell data analysis. This evolution reflects a broader shift toward automated analytical frameworks, designed to meet the practical demands

posed by the rapidly expanding volume of single-cell data and the growing complexity of analytical workflows.

1.2. The Evaluation Paradox

Although AI agents hold significant promise for practical application, accurately assessing their performance remains a key challenge. Most current evaluation frameworks depend on static datasets and employ quantitative metrics originally designed for single-task algorithmic benchmarks. However, this approach faces a fundamental limitation: core agent capabilities—including multi-step reasoning, autonomous decision-making, and dynamic planning—cannot be adequately captured by final outcome metrics alone. Consequently, strong performance on conventional metrics such as the Rand Index (ARI) or normalized mutual information (NMI) does not guarantee robust behavior in real-world settings, where issues like decision errors, sensitivity to data noise, or inconsistent analytical execution may arise. This gap in evaluation methodology undermines the reproducibility and reliability of AI agent analysis outcomes. Comparable challenges exist across single-cell and epigenomic analysis domains, including ATAC-seq and CUT&Tag data interpretation. Recent work has further highlighted that methodological innovation in these areas is advancing more rapidly than corresponding evaluation frameworks, leading to a growing misalignment between analytical capabilities and assessment standards.

1.3. Scope and Contribution of This Review

This review focuses on evaluation challenges associated with artificial intelligence agents in single-cell RNA sequencing analysis. It begins with a systematic examination of widely adopted datasets and evaluation metrics in existing benchmarking frameworks, identifying key limitations when applied to multi-step autonomous analytical systems. Subsequently, an evaluation framework centered on the analytical process is proposed. This framework emphasizes not only final analytical outcomes but also integrates assessment of transparency, operational efficiency, biological plausibility, and overall system stability throughout the analysis workflow, thereby offering a more holistic appraisal of practical analytical capability. The review seeks to inform the development of a more comprehensive and robust evaluation system for artificial intelligence agents in single-cell genomics.

2. Foundational Elements: Datasets and Metrics in Current ScRNA-seq Benchmarks

2.1. Benchmarking Datasets: A Critical Survey

When conducting scRNA-seq benchmark tests, the dataset itself constitutes the most critical factor. Currently, most studies rely directly on publicly available real-world datasets to evaluate the performance of different analytical methods. These datasets originate from diverse sources and exhibit substantial heterogeneity in key characteristics: some derive from tissue-specific cell atlases, others capture dynamic cellular changes during development, and still others consist of disease-associated samples [6–8]. Human cell atlases, along with public repositories such as GEO and ArrayExpress, are frequently employed in this context. They typically supply raw count matrices accompanied by biologically validated annotations. Such datasets generally possess ground-truth cell type labels—meaning researchers have prior knowledge of the correct cell type assignments for each profile. These labels typically stem from expert-curated cell type annotations or orthogonal experimental validation, including flow cytometry. However, real-world data is rarely ideal. Technical batch effects commonly arise across independent experiments, sequencing depth varies markedly among samples, and inherent biological variability further complicates interpretation. Collectively, these factors introduce confounding influences that can substantially impact benchmarking outcomes. Consequently, contemporary benchmarking practices emphasize the use of standardized, openly accessible datasets, coupled with transparent and detailed documentation of all preprocessing steps, to ensure experimental reproducibility and enable fair cross-study

comparisons. Table 1 summarizes core characteristics of widely adopted public datasets for evaluating scRNA-seq methods.

Table 1. Summary of Commonly Used Benchmarking Datasets for scRNA-seq Analysis

Dataset Name	Tissue Origin	Cell Count	Key Features	Accession Number / Link
10x Genomics PBMC	Peripheral blood mononuclear cells	12k	Widely used for benchmarking clustering and annotation	10x Genomics Official Dataset
Tabula Sapiens	Multiple human tissues	500k+	Multi tissue atlas with expert curated annotations	tabula-sapiens-portal.ds.czbiohub.org
Human Pancreas	Pancreatic islets	15k	Well characterized cell types for benchmarking integration	GSE84133
Mouse Brain (10x)	Brain cortex and hippocampus	300k+	Complex tissue with diverse neuronal subtypes	10x Genomics Official Dataset
COVID 19 BALF	Bronchoalveolar lavage fluid	200k+	Disease associated dataset with immune cell heterogeneity	GSE145926
Developing Mouse Heart	Embryonic heart	20k	Developmental time course dataset	GSE119468
Human Lung Atlas	Lung tissue	75k	Healthy human lung cell atlas	GSE122960
Mouse Kidney	Kidney	70k	Adult mouse kidney cell atlas	GSE129798

2.2. Quantitative Metrics: A Technical Taxonomy

Beyond dataset curation, rigorous evaluation of algorithmic performance remains essential. In single-cell RNA sequencing, quantitative metrics serve as standardized benchmarks for comparing analytical methods across distinct computational tasks. For clustering—central to cell type identification—the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are widely adopted. ARI quantifies agreement between two clusterings while correcting for chance, yielding robust and stable scores. NMI measures label consistency between predicted and ground-truth annotations, producing a normalized score ranging from 0 to 1, where values approaching 1 indicate superior alignment. In dimensionality reduction and visualization, metrics such as K-nearest neighbor–based batch effect correction scores and the Local Inverse Simpson Index (LISI) assess preservation of biological structure and efficacy of technical noise removal. For differential expression analysis and gene selection, evaluation relies on precision–recall curves and false discovery rate control frameworks. Collectively, these metrics establish a consistent, task-specific framework for method comparison and performance ranking. Nevertheless, they represent static, outcome-oriented assessments that evaluate only final outputs rather than intermediate interpretability or biological plausibility [9–11]. Table 2 summarizes these quantitative indicators and organizes them by analytical task.

Table 2. Taxonomy of Quantitative Metrics in scRNA-seq Benchmarks

Task Category	Metric	Description	Range
Clustering	Adjusted Rand Index (ARI)	Measures similarity between predicted and true clusterings, adjusted for chance	-1 to 1
Clustering	Normalized Mutual Information (NMI)	Quantifies shared information between predicted and true labels, normalized	0 to 1

Clustering	Fowlkes Mallows Index (FMI)	Geometric mean of precision and recall for cluster assignments	0 to 1
Dimensionality Reduction	KNN Batch Effect Correction Score	Assesses removal of batch effects in neighbor relationships	0 to 1
Dimensionality Reduction	Local Inverse Simpson Index (LISI)	Measures preservation of biological diversity after integration	≥1
Data Integration	Graph Connectivity	Evaluates connectivity of cell type clusters in integrated space	0 to 1
Differential Expression	Precision Recall Curve	Assesses accuracy of differentially expressed gene detection	0 to 1
Differential Expression	False Discovery Rate (FDR)	Controls for false positives in multiple testing	0 to 1

2.3. The Problems of Static Metric Evaluation

These static indicators exhibit notable limitations when evaluating AI agents. They primarily emphasize final outcomes—such as the correctness of clustering assignments or the accuracy of selected marker genes—yet struggle to capture the dynamic progression by which the AI arrives at those results. For instance, a low Adjusted Rand Index (ARI) score at the conclusion does not necessarily indicate flawed reasoning throughout the analysis; earlier steps may have been sound, with deviation arising only from later data inconsistencies or parameter misconfigurations [12–14]. Conversely, some agents may achieve relatively high ARI scores despite employing suboptimal preprocessing strategies—a performance that often proves brittle when transferred to new datasets, revealing deficiencies in generalization and stability that static metrics frequently overlook. Moreover, such indicators are inherently inadequate for assessing operational reliability during real-world analysis: they cannot gauge whether an agent appropriately handles missing data, recovers autonomously from analytical errors, or justifies its methodological choices with coherent, interpretable reasoning [15, 16]. Since these critical capabilities lie beyond the scope of scalar performance scores, relying solely on a small set of quantitative metrics is generally insufficient—and in certain cases, potentially misleading—for evaluating AI agents designed to execute complex, multi-step analytical workflows.

3. The Methodological Gap: Why Current Frameworks Fail AI Agents

3.1. Inability to Capture Multi Step Decision Making

Currently, many evaluation methods for scRNA-seq focus exclusively on the correctness of final results. For instance, assessments typically examine whether clustering outcomes are accurate or whether differential expression analysis is correct, with the final output being directly compared against ground-truth labels. However, AI agents do not perform analysis in a single step; rather, they execute a continuous sequence of multiple steps. For example, an agent must first determine the appropriate normalization strategy, then address batch effects, subsequently select the clustering resolution, and finally verify whether the analytical results are reasonable. These steps are interconnected, and an error in any single step typically propagates to affect subsequent stages. Nevertheless, most existing benchmark evaluations rarely examine how AI agents perform during intermediate steps. One agent may execute data processing largely correctly but select suboptimal clustering parameters, while another may accumulate errors from the very beginning. Yet both may ultimately yield approximately equivalent low scores, making it difficult to genuinely distinguish the performance differences between them based solely on final metrics [17, 18].

3.2. Neglect of Adaptability and Robustness

One important advantage of AI agents over traditional analysis tools is their capacity for autonomous adaptation to diverse datasets [19, 20]. However, many current benchmark evaluations fail to rigorously assess this capability. A prevalent approach

involves testing algorithms on pre-organized standard datasets, with parameters often fine-tuned specifically for those datasets. Such evaluation settings are highly idealized, whereas real-world research data typically exhibit greater complexity. Actual datasets frequently differ substantially in sample size, batch effects, noise levels, and overall data quality. Robustness refers to an algorithm's ability to sustain consistent performance under varying data conditions. Yet this dimension remains underrepresented in prevailing evaluation frameworks. Some AI agents achieve high scores on standardized benchmarks but demonstrate marked performance degradation—or even complete failure—when confronted with novel batch effects, increased noise, or reduced data quality. The absence of such stress-testing leads to systematic overestimation of AI agents' practical utility in real research scenarios.

3.3. Ignoring Operational Reliability

Apart from the accuracy of the analysis results, the operational stability of AI agents is equally critical, yet this dimension is frequently underemphasized. Current benchmarking practices predominantly assess the accuracy of final outputs, with limited attention to robustness during execution—such as resilience to missing data or malformed inputs, consistency across repeated runs, and graceful handling of unexpected conditions [21, 22]. These aspects are seldom rigorously evaluated. However, such challenges arise routinely in real-world research settings. An AI agent that functions reliably only under ideal data conditions, but fails when confronted with common data imperfections, offers limited practical utility. Furthermore, as single-cell RNA sequencing datasets continue to grow in scale, computational efficiency and resource utilization have become increasingly consequential. Some agents may yield biologically meaningful outcomes, yet exhibit excessive memory consumption, prolonged runtime, or inability to resume after mid-process interruptions—factors that significantly impair deployability in realistic analytical environments. This highlights notable gaps in the current evaluation framework for determining suitability of AI agents in end-to-end scRNA-seq analysis.

In summary, the principal limitations outlined in this section are consolidated in Table 3, which contrasts traditional static metric evaluation with process-oriented assessment tailored to the full analytical workflow of AI agents.

Table 3. Comparison of Static Metrics Versus Process Oriented Evaluation for AI Agents

Evaluation Dimension	Static Metric Evaluation	Process Oriented Evaluation
Decision Making	Captures only final output	Tracks and evaluates multi step decision sequences
Adaptability	Evaluates on fixed datasets with manual hyperparameters	Tests across diverse scenarios without manual intervention
Robustness	Not assessed	Quantifies performance under stress conditions and data variations
Operational Reliability	Ignored	Includes error handling, reproducibility, and resource efficiency
Transparency	Provides no insight into reasoning	Requires documented decision logs and justifications
Biological Validity	Optimizes numerical metrics	Prioritizes biological relevance of findings

4. A Methodological Proposal: A Process-Oriented Evaluation Framework

4.1. Core Principles: Transparency, Efficiency, and Biological Validity

Figure 1 presents a simplified overview of the proposed process-oriented evaluation framework for AI agents in scRNA-seq analysis.

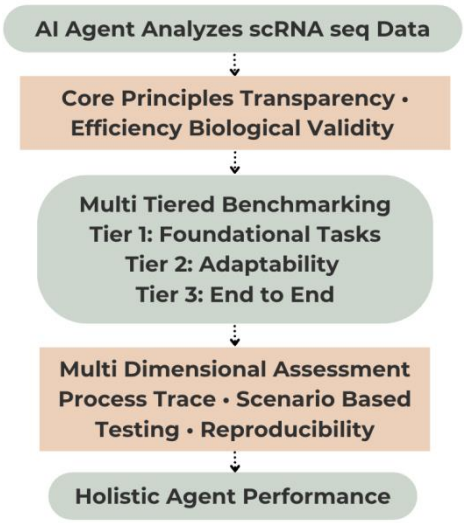

Figure 1. A Process Oriented Evaluation Framework for AI Agents in scRNA-seq Analysis

To address the limitations of existing testing methodologies, we propose a process-oriented evaluation framework grounded in the principles of transparency, efficiency, and biological validity. Transparency necessitates not only the final outputs generated by the artificial intelligence system but also comprehensive documentation of each reasoning and decision-making step throughout the entire analytical workflow. This includes justifications for selecting specific data preprocessing techniques, rationales behind hyperparameter configurations, and assessments of intermediate result plausibility. Efficiency requires that the framework accounts for computational resource consumption, encompassing execution time, memory utilization, and scalability to large-scale scRNA-seq datasets. A model yielding highly accurate results may remain impractical for routine research applications if it demands excessive computational resources. Biological validity ensures that evaluative priority is assigned to the biological significance of analytical outcomes rather than the pursuit of superior statistical metrics in isolation. In scRNA-seq analysis, the ultimate objective is to uncover genuinely informative biological insights—such as novel cell types or disease-associated pathways—rather than optimizing numerical performance indicators alone.

4.2. Proposed Framework: The "ScAI Bench" Protocol

Based on the above principles, we propose ScAI Bench to evaluate the performance of AI systems in the automatic analysis of single-cell RNA sequencing data [23, 24]. This protocol evaluates multiple dimensions simultaneously rather than focusing solely on a single final output. It records the complete analytical workflow—including all intermediate steps—and assesses the reasonableness and methodological soundness of each step relative to established best practices. Furthermore, it employs diverse datasets spanning varying tissue origins, sequencing depths, and levels of biological complexity to examine model robustness across heterogeneous conditions. It also evaluates computational reproducibility by assessing result consistency across repeated runs on identical input data and verifying whether following the documented procedural steps yields identical outcomes. Collectively, these criteria provide a more comprehensive assessment of model capability in realistic analytical settings.

4.3. Implementation: A Multi Tiered Benchmarking Suite

To implement ScAI Bench, a hierarchical data testing system is required to progressively evaluate the model's capabilities across increasing levels of analytical complexity. The foundational tier employs publicly available, standard-labeled datasets to assess core analytical functions—including data organization, normalization, clustering, and cell type identification. The intermediate tier introduces more challenging datasets characterized by substantial batch effects, low signal-to-noise ratios, or sparse

representation of rare cell populations, thereby evaluating model robustness under suboptimal conditions. The highest tier presents entirely unlabeled, novel datasets and requires the model to autonomously execute end-to-end analysis without reference benchmarks; evaluation here emphasizes biological plausibility, consistency with established domain knowledge, and potential for generating scientifically meaningful insights. All datasets are drawn exclusively from public repositories to ensure full reproducibility and enable cross-study comparability [25–27].

Table 4 presents the three-tier structure of the ScAI Bench suite, detailing the evaluation focus, dataset characteristics, and assessment criteria for each tier.

Table 4. Overview of the ScAI Bench Multi Tiered Benchmarking Suite

Tier	Evaluation Focus	Dataset Characteristics	Assessment Criteria	Example Datasets
Tier 1: Foundational Tasks	Basic competency in core analysis tasks	Well characterized reference datasets with established ground truth	Accuracy of clustering, annotation, and preprocessing; standard metrics (ARI, NMI)	10x PBMC (10x Official), Human Pancreas (GSE84133)
Tier 2: Adaptability and Robustness	Performance under challenging data conditions	Datasets with strong batch effects, low quality, rare populations, complex composition	Consistency across conditions; resilience to noise; stress test performance	COVID 19 BALF (GSE145926), PBMC Batch Integration (GSE130953), Mouse Brain (10x Official)
Tier 3: End to End Unsupervised Analysis	Holistic capability on novel datasets	Unseen datasets without predefined ground truth	Expert assessment; biological validity; consistency with known biology	Human Lung Atlas (GSE122960), Mouse Kidney Atlas (GSE129798)

5. Discussion and Future Directions

5.1. Harmonizing Evaluation: Toward a Community Standard

In the AI-based assessment of single-cell RNA sequencing data, current methodological practices vary substantially across studies. Differences in data sources, preprocessing pipelines, analytical workflows, and reporting conventions hinder meaningful cross-study comparison and reproducibility. A more pragmatic strategy involves establishing a minimal set of standardized reporting requirements—such as explicit documentation of data provenance, preprocessing parameters, step-by-step analytical procedures, and transparent rationale for AI-driven decisions at each stage. Consistent adherence to such conventions would enable direct methodological comparison and facilitate independent validation through re-execution. Concurrently, curating a collection of publicly accessible benchmark datasets in harmonized formats—pre-processed, quality-controlled, and annotated with validated cell type labels—would allow fair, controlled evaluation of diverse AI methods under identical conditions. Widespread adoption of these standards by journals, funding agencies, and research consortia would significantly reduce methodological fragmentation and strengthen cumulative scientific progress.

5.2. Beyond Static Benchmarks: Simulated Environments

Currently, the primary evaluation approach relies on real-world data. However, real data present inherent limitations: experimental conditions are difficult to control, and

systematic assessment of rare or challenging scenarios remains impractical. An alternative is the use of simulated data—artificially generated datasets designed to closely approximate real-world characteristics. Such environments allow precise manipulation of key parameters, including noise levels, inter-experiment variability, cell-type proportion distributions, and patterns of missing data [28, 29]. This enables rigorous evaluation of model robustness across varying degrees of analytical difficulty. A more advanced paradigm involves dynamic testing environments that provide feedback during model operation, allowing iterative adjustment of analytical strategies in response to intermediate outputs. This mirrors the adaptive refinement process observed in practical data analysis workflows. Integrating real and simulated data thus supports both ecological validity and comprehensive stress-testing of AI systems.

6. Conclusion

The development of artificial intelligence in computational biology is rapid, presenting both new opportunities and persistent challenges in evaluation. This paper reviews common benchmarking methods used in single-cell RNA sequencing (scRNA-seq) analysis and identifies a core limitation: a fundamental mismatch between existing evaluation frameworks and the actual capabilities of artificial intelligence models. Current evaluations predominantly rely on static datasets and summary metrics such as the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), emphasizing final result accuracy while inadequately capturing the integrity, adaptability, and robustness of multi-step analytical workflows.

A deeper examination reveals that these limitations extend beyond incomplete coverage. Existing methods fail to reflect intermediate analytical decisions—such as the appropriateness of specific preprocessing or clustering choices—and cannot assess model stability across diverse data conditions. Practical considerations—including error handling capacity, end-to-end runtime, memory footprint, and computational resource efficiency—are largely omitted. As a consequence, model capability assessments remain fragmented and, in certain cases, misleadingly optimistic, thereby undermining translational reliability.

To address these gaps, this paper introduces an analysis-process–centric evaluation paradigm, structured around three pillars: traceability of analytical steps, reasonableness of computational cost, and biological interpretability of outputs. ScAI Bench implements this paradigm by logging each stage of the analysis pipeline, executing systematic tests across heterogeneous data types, and verifying result reproducibility. Integrated hierarchical data testing enables granular performance assessment across distinct analytical tasks.

Looking ahead, broader methodological refinements are needed. Standardized foundational evaluation norms would help harmonize cross-study comparisons and reduce methodological fragmentation. Complementing real-data benchmarks with rigorously designed synthetic datasets and interactive testing environments could expand coverage to edge cases and failure modes inaccessible through empirical data alone. Progressive adoption of such strategies would yield more holistic, application-aligned assessments of artificial intelligence systems in single-cell analysis.

References

1. J. R. Heath, A. Ribas, and P. S. Mischel, "Single-cell analysis tools for drug discovery and development," *Nature Reviews Drug Discovery*, vol. 15, no. 3, pp. 204–216, 2016.
2. M. Brendel, C. Su, Z. Bai, H. Zhang, O. Elemento, and F. Wang, "Application of deep learning on single-cell RNA sequencing data analysis: a review," *Genomics, Proteomics & Bioinformatics*, vol. 20, no. 5, pp. 814–835, 2022.
3. J. Oh, E. Kim, I. Cha, and A. Oh, "The generative AI paradox in evaluation: 'what it can solve, it may not evaluate'," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: Student Res. Workshop*, Mar. 2024, pp. 248–257.
4. S. Cheng, B. Miao, T. Li, G. Zhao, and B. Zhang, "Review and evaluate the bioinformatics analysis strategies of ATAC-seq and CUT&tag data," *Genomics, Proteomics & Bioinformatics*, vol. 22, no. 3, p. qzae054, 2024.
5. N. Beerenwinkel and A. Mahfouz, "Eleven grand challenges in single-cell data science," *Genome Biology*, 2020.

6. C. Ahlmann-Eltze and W. Huber, "Comparison of transformations for single-cell RNA-seq data," *Nature Methods*, vol. 20, no. 5, pp. 665–672, 2023.
7. Z. Wu and H. Wu, "Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering," *Genome Biology*, vol. 21, no. 1, p. 123, 2020.
8. S. Luo, P. L. Germain, F. Von Meyenn, and M. D. Robinson, "On metrics for subpopulation detection in single-cell and spatial omics data," *Nucleic Acids Research*, vol. 53, no. 18, p. gkaf921, 2025.
9. V. Finazzi, C. A. Vallejos, and A. Scialdone, "Benchmarking computational tools for locus-specific analysis of transposable elements in single-cell RNA-seq datasets," *bioRxiv*, preprint, Feb. 2026.
10. F. Meyer, A. Fritz, Z. L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, et al., "Critical assessment of metagenome interpretation: the second round of challenges," *Nature Methods*, vol. 19, no. 4, pp. 429–440, 2022.
11. Karamveer and Y. Uzun, "Approaches for benchmarking single-cell gene regulatory network methods," *Bioinformatics and Biology Insights*, vol. 18, p. 11779322241287120, 2024.
12. S. Mangul, L. S. Martin, B. L. Hill, A. K. M. Lam, M. G. Distler, A. Zelikovsky, et al., "Systematic benchmarking of omics computational tools," *Nature Communications*, vol. 10, no. 1, p. 1393, 2019.
13. P. Keyl, P. Bischoff, G. Dernbach, M. Bockmayr, R. Fritz, D. Horst, et al., "Single-cell gene regulatory network prediction by explainable AI," *Nucleic Acids Research*, vol. 51, no. 4, p. e20, 2023.
14. C. King, M. Iqbal, E. Shokati, C. M. Ying Li, R. Li, Y. Tomita, et al., "Comparative benchmarking of single-cell transcriptomes and immune repertoires across technologies," *bioRxiv*, preprint, Apr. 2026.
15. S. Grégoire, C. Vanderaa, S. P. dit Ruys, C. Kune, G. Mazzucchelli, D. Vertommen, and L. Gatto, "Standardized workflow for mass-spectrometry-based single-cell proteomics data processing and analysis using the scp package," in *Mass Spectrometry Based Single Cell Proteomics*, New York, NY: Springer US, 2024, pp. 177–220.
16. M. Zhou, H. Zhang, Z. Bai, D. Mann-Krzisnik, F. Wang, and Y. Li, "Single-cell multi-omics topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures," *Cell Reports Methods*, vol. 3, no. 8, 2023.
17. M. T. Islam, Y. Liu, M. M. Hassan, P. E. Abraham, J. Merlet, A. Townsend, et al., "Advances in the application of single-cell transcriptomics in plant systems and synthetic biology," *BioDesign Research*, vol. 6, p. 0029, 2024.
18. W. A. Thistlethwaite, *Integrative Multi-Omic Data Analysis and Software Development*, Ph.D. dissertation, Princeton Univ., 2025.
19. M. Zahn-Zabala, M. Ahokas, B. Batut, S. Beier, A. Botzki, A. Cardona, et al., "Strengthening bioinformatics education: e-learning initiatives across ELIXIR Nodes and Communities," *F1000Research*, vol. 14, p. ELIXIR-1422, 2025.
20. C. Schmied, M. S. Nelson, S. Avilov, G. J. Bakker, C. Bertocchi, J. Bischof, et al., "Community-developed checklists for publishing images and image analyses," *Nature Methods*, vol. 21, no. 2, pp. 170–181, 2024.
21. T. Sun, D. Song, W. V. Li, and J. J. Li, "scdesign2: an interpretable simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured," *bioRxiv*, preprint, Nov. 2020.
22. Y. Cao, P. Yang, and J. Y. H. Yang, "A benchmark study of simulation methods for single-cell RNA sequencing data," *Nature Communications*, vol. 12, no. 1, p. 6911, 2021.
23. T. Tian, J. Wan, Q. Song, and Z. Wei, "Clustering single-cell RNA-seq data with a model-based deep learning approach," *Nature Machine Intelligence*, vol. 1, no. 4, pp. 191–198, 2019.
24. H. M. Levitin, J. Yuan, Y. L. Cheng, F. J. Ruiz, E. C. Bush, J. N. Bruce, et al., "De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization," *Molecular Systems Biology*, vol. 15, no. 2, p. MSB188557, 2019.
25. R. Petegrosso, Z. Li, and R. Kuang, "Machine learning and statistical methods for clustering single-cell RNA-sequencing data," *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1209–1223, 2020.
26. J. Wang, J. Xia, H. Wang, Y. Su, and C. H. Zheng, "scDCCA: deep contrastive clustering for single-cell RNA-seq data based on auto-encoder network," *Briefings in Bioinformatics*, vol. 24, no. 1, p. bbac625, 2023.
27. A. Pinhasi and K. Yizhak, "Uncovering gene and cellular signatures of immune checkpoint response via machine learning and single-cell RNA-seq," *NPJ Precision Oncology*, vol. 9, no. 1, p. 95, 2025.
28. N. Hou, X. Lin, L. Lin, X. Zeng, Z. Zhong, X. Wang, et al., "Artificial intelligence in cell annotation for high-resolution RNA sequencing data," *TrAC Trends in Analytical Chemistry*, vol. 178, p. 117818, 2024.
29. G. Chen, H. Qi, L. Jiang, S. Sun, J. Zhang, J. Yu, et al., "Integrating single-cell RNA-Seq and machine learning to dissect tryptophan metabolism in ulcerative colitis," *Journal of Translational Medicine*, vol. 22, no. 1, p. 1121, 2024.