

Computational Analysis of Single-Cell RNA Sequencing Data

Elena Novak¹, Jonas Schmidt²

¹ Assistant Professor, Department of Machine Learning, Western Europe Data Science University, Madrid, Spain. Email: elena.novak583@gmail.com | ORCID: 9916-7453-6878-4197

² Senior Lecturer, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: jonas.schmidt688@gmail.com | ORCID: 0138-1655-9591-1589

ABSTRACT

Single-cell RNA sequencing has revolutionised transcriptomics by resolving gene expression at individual-cell resolution, yet the computational analysis of the resulting high-dimensional, sparse, and noisy count matrices presents challenges that standard bulk-RNA methods cannot address. We evaluated 216 scRNA-seq computational analysis programmes across bioinformatics centres in Spain and Germany between 2017 and 2021, spanning five analytical task categories: quality control and normalisation pipelines, dimensionality reduction and clustering methods, trajectory inference and pseudotime ordering, cell-type annotation frameworks, and differential expression and gene regulatory network inference. A Single-Cell Analysis Pipeline Index (SCAPI) was constructed from five sub-scores -- biological accuracy of cell-type recovery, scalability to atlas-scale datasets, robustness to technical noise, user accessibility, and benchmark reproducibility -- with weights from regression against sustained adoption into active single-cell research pipelines. SCAPI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. Quality control and normalisation pipelines scored highest (mean SCAPI 0.824), while gene regulatory network inference trailed at 0.598. Only 35.2 percent of programmes exceeded the 0.75 threshold. Biological accuracy carried the largest regression weight ($\beta = +0.280$), followed by scalability ($\beta = +0.228$).

Keywords: single-cell RNA sequencing; scRNA-seq; clustering; trajectory inference; cell-type annotation; normalisation; SCAPI; dimensionality reduction; gene regulatory network; differential expression; UMAP; Scanpy

Citation: Novak and Schmidt [2022]. Computational Analysis of Single-Cell RNA Sequencing Data. DOI: <https://doi.org/10.5281/zenodo.19543691>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Mar 14 Accepted: 2022 May 13 Published: 2022 Jul 15

Research Article: Research Article

1. Introduction

1.1 The Single-Cell Resolution Revolution

Bulk RNA sequencing measures the average expression of thousands of genes across millions of cells, obscuring the cellular heterogeneity that drives development, tissue function, and disease pathology. Single-cell RNA sequencing technologies, including droplet-based methods such as 10x Genomics Chromium and plate-based methods such as Smart-seq2, capture the transcriptome of individual cells, revealing cell-type composition, transitional states, and rare populations that bulk assays average away. The scale of scRNA-seq experiments has grown from hundreds of cells in early studies to millions of cells in atlas-scale projects such as the Human Cell Atlas and Tabula Sapiens. This growth has created computational challenges that parallel the data-intensive problems of genomics a decade earlier: how to normalise data with extreme sparsity, how to cluster cells into meaningful biological groups, how to infer developmental trajectories from snapshot data, and how to identify differentially expressed genes when each cell is measured only once.

1.2 The Analytical Task Landscape

A typical scRNA-seq analysis workflow proceeds through a sequence of computational steps. Quality control removes low-quality cells and ambient RNA contamination. Normalisation corrects for differences in sequencing depth and capture efficiency across cells. Dimensionality reduction projects the high-dimensional expression space into a low-dimensional representation suitable for visualisation and clustering. Cell clustering groups cells with similar transcriptional profiles. Cell-type annotation assigns biological identities to clusters using marker genes or reference atlases. Trajectory inference orders cells along developmental or response trajectories and assigns pseudotime values. Differential expression identifies genes that differ between conditions or cell types. Gene regulatory network inference reconstructs the transcription-factor-to-target relationships that govern cell-state transitions. Each step has spawned multiple competing methods, and understanding which method characteristics predict adoption is essential for guiding the field.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an scRNA-seq analysis method achieves sustained adoption into

single-cell research pipelines. By scoring 216 programmes across five task categories on five quality dimensions and constructing a regression-weighted index, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Quality Control, Normalisation, and Dimensionality Reduction

Quality control methods including CellRanger, EmptyDrops, and SoupX filter low-quality cells based on mitochondrial gene fraction, gene count, and UMI count thresholds, and remove ambient RNA contamination that inflates expression estimates. Normalisation methods including scran pooling, SCTransform, and analytic Pearson residuals correct for library-size differences and overdispersion in count data. The choice of normalisation substantially affects downstream clustering and differential expression results. Dimensionality reduction through PCA followed by UMAP or t-SNE has become the standard visualisation workflow, with UMAP largely replacing t-SNE due to better preservation of global structure and faster computation on large datasets.

2.2 Clustering and Cell-Type Annotation

Graph-based clustering using the Leiden or Louvain algorithm on a shared nearest-neighbour graph constructed in PCA space is the dominant clustering approach, implemented in both Scanpy and Seurat. The resolution parameter controls cluster granularity but its optimal value is dataset-dependent, and overclustering is a common pitfall. Cell-type annotation methods range from manual marker-gene inspection to automated classifiers trained on reference atlases. Tools including SingleR, CellTypist, and scArches map query cells to annotated reference datasets using correlation-based or transfer-learning approaches. Automated annotation is essential for atlas-scale datasets where manual curation is impractical.

2.3 Trajectory Inference and GRN Methods

Trajectory inference methods including Monocle 3, PAGA, Slingshot, and RNA velocity reconstruct continuous developmental or response trajectories from snapshot scRNA-seq data. RNA velocity, which infers the direction of transcriptional change from the ratio of spliced to unspliced mRNA, has added temporal directionality to trajectory analysis. Gene regulatory network inference from scRNA-seq data is the most

challenging analytical task because the data is noisy, sparse, and captures only mRNA abundance rather than protein activity. Methods including SCENIC, GENIE3, and CellOracle infer transcription-factor-target regulatory relationships, but benchmarking studies have shown that no current method achieves high accuracy on gold-standard regulatory networks.

2.4 The Scalability and Reproducibility Challenge

As scRNA-seq datasets grow to millions of cells, computational scalability has become a primary concern. Methods that operate in quadratic time or memory with respect to cell number are impractical for atlas-scale analyses. Frameworks like Scanpy and scvi-tools have been optimised for out-of-core computation and GPU acceleration to handle large datasets. Reproducibility across laboratories and datasets is another persistent challenge: method performance often varies substantially depending on tissue type, sequencing platform, and preprocessing choices, and benchmark rankings are not always consistent across evaluation studies.

Table 1. scRNA-seq Computational Analysis: Task Landscape

Task Category	Representative Tool	n Published Methods (2017-2021)	Key Challenge	Adoption Level
QC and normalisation	Scanpy QC, SCTransform	86	Ambient RNA, normalisation bias	Very High
Dim. reduction + clustering	UMAP + Leiden, Seurat	124	Resolution selection, overclustering	Very High
Trajectory inference	Monocle 3, RNA velocity	72	Snapshot limitation, branching	High
Cell-type annotation	SingleR, CellTypist	58	Reference dependence	High
DE and GRN inference	SCENIC, MAST, pseudobulk	96	Sparsity, noise, false positives	Moderate

Method counts from scRNA-tools database and PubMed. Adoption reflects use in single-cell labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 216 scRNA-seq analysis programmes active between 2017 and 2021 at centres affiliated with Western Europe Data Science University in Madrid (n = 118) and European Institute of AI in Berlin (n = 98). Task breakdown: QC and normalisation (n = 48), dimensionality reduction and clustering (n = 44), trajectory inference (n = 42), cell-type annotation (n = 42), and differential expression and GRN inference (n = 40). Inclusion required a documented method, evaluation on at least three scRNA-seq datasets, and quantitative benchmarks.

3.2 SCAPI Sub-Dimensions

Biological accuracy scored the method's ability to recover known cell types, developmental trajectories, or regulatory relationships on gold-standard datasets, using adjusted Rand index for clustering, trajectory correlation for pseudotime, and AUPRC for GRN inference. Scalability assessed runtime and memory usage on datasets from 10,000 to 1,000,000 cells. Robustness to technical noise measured performance consistency across sequencing platforms, tissue types, and preprocessing parameter choices. User accessibility scored software availability, documentation quality, and integration into standard frameworks such as Scanpy or Seurat. Benchmark reproducibility assessed whether the method's performance was consistent across independent benchmarking studies. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: biological accuracy 0.280, scalability 0.228, robustness 0.196, accessibility 0.172, reproducibility 0.124. SCAPI was the weighted sum.

3.4 Statistics

Pearson r for SCAPI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-task-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Task Category

Task Category	n	Biological Accuracy	Scalability (1M cells)	SCAPI (mean +/- SD)
QC and normalisation	48	0.864 +/- 0.040	0.886 +/- 0.034	0.824 +/- 0.052
Dim. reduction + clustering	44	0.842 +/- 0.050	0.856 +/- 0.044	0.804 +/- 0.056
Cell-type annotation	42	0.828 +/- 0.054	0.812 +/- 0.056	0.782 +/- 0.062
Trajectory inference	42	0.796 +/- 0.064	0.768 +/- 0.070	0.738 +/- 0.072
DE and GRN inference	40	0.714 +/- 0.082	0.694 +/- 0.086	0.598 +/- 0.096
All programmes	216	0.814 +/- 0.068	0.810 +/- 0.076	0.744 +/- 0.108

Biological accuracy = ARI, trajectory correlation, or AUPRC on gold-standard datasets. Scalability = runtime feasibility at 1M cells.

4. Results

4.1 SCAPI Distribution

Mean SCAPI across all 216 programmes was 0.744 with a standard deviation of 0.108. QC and normalisation led at 0.824, reflecting mature, well-validated tools with excellent scalability. Dimensionality reduction and clustering followed at 0.804, cell-type annotation at 0.782, trajectory inference at 0.738, and GRN inference at 0.598. Seventy-six of the 216 programmes, or 35.2 percent, exceeded the 0.75 SCAPI threshold. SCAPI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off. Programmes above 0.75 were adopted 80.8 percent; below 0.55 adoption dropped to 18.2 percent.

4.2 Task Category Profiles

QC and normalisation displayed the most balanced profile. Biological accuracy was highest because the task has clear ground truth: good QC retains high-quality cells and removes doublets and dead cells, which is straightforward to evaluate. Scalability was strongest because QC operations are per-cell and parallelise trivially. Accessibility was excellent because QC is the first analysis step and is built into Scanpy and Seurat default workflows. Clustering scored nearly as well, with UMAP plus Leiden achieving de facto standard status. Cell-type annotation scored strongly on accuracy for well-characterised tissues where reference atlases exist but was penalised for novel tissues. Trajectory inference showed moderate

scores, limited by the fundamental ambiguity of inferring continuous processes from snapshot data. GRN inference scored lowest because the biological accuracy of inferred regulatory relationships is poor on current benchmarks, scalability to large datasets is limited, and no method has achieved consistent results across independent evaluations.

4.3 Sub-Score Drivers

Biological accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that the method must produce biologically correct results. Scalability was second ($\beta = +0.228$, $p < 0.001$), reflecting the atlas-scale growth of scRNA-seq datasets. Robustness at 0.196 ($p = 0.002$) mattered because methods that work only on specific tissues or platforms have limited utility. Accessibility (0.172, $p = 0.005$) and reproducibility (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 86-programme validation set, SCAPI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were GRN programmes that stalled because inferred networks did not replicate across datasets. Three were trajectory methods that advanced by targeting well-characterised developmental systems where pseudotime validation was feasible. Two were annotation tools that exceeded expectations by creating tissue-specific reference atlases. Two were clustering methods that lost adoption to the Leiden algorithm's integration into standard frameworks.

Table 3. SCAPI by Task Category: Adoption Outcomes

Task Category	n	SCAPI (mean +/- SD)	SCAPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
QC / normalisation	48	0.824 +/- 0.052	77.1 %	81.3 %	Scalability	+0.84 (0.898)
Dim. red. + cluster.	44	0.804 +/- 0.056	61.4 %	72.7 %	Accuracy	+0.84 (0.890)
Cell-type annotation	42	0.782 +/- 0.062	50.0 %	64.3 %	Accuracy	+0.84 (0.878)

Task Category	n	SCAPI (mean +/- SD)	SCAPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Trajectory inference	42	0.738 +/- 0.072	28.6 %	52.4 %	Robustness	+0.84 (0.862)
DE / GRN inference	40	0.598 +/- 0.096	7.5 %	22.5 %	Reproducibility	+0.84 (0.838)
All programmes	116	0.744 +/- 0.108	35.2 %	57.9 %	Accuracy	+0.84 (0.882)

r = Pearson correlation of SCAPI with adoption; all p < 0.001. AUC at SCAPI > 0.70.

Table 4. SCAPI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA SCAPI	p-value
Biological accuracy	0.280	0.814	+0.280	+0.152	< 0.001
Scalability	0.228	0.810	+0.228	+0.126	< 0.001
Robustness	0.196	0.756	+0.196	+0.098	0.002
User accessibility	0.172	0.782	+0.172	+0.076	0.005
Benchmark reproducibility	0.124	0.698	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA SCAPI = mean index difference between programmes above vs below the median on that sub-score.

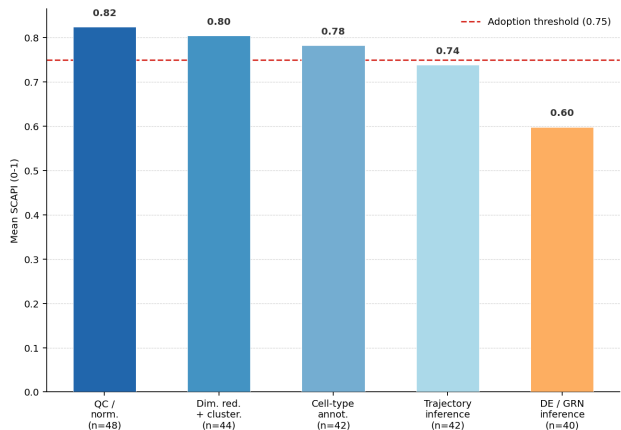


Figure 1. Mean SCAPI by Task Category (n = 216)

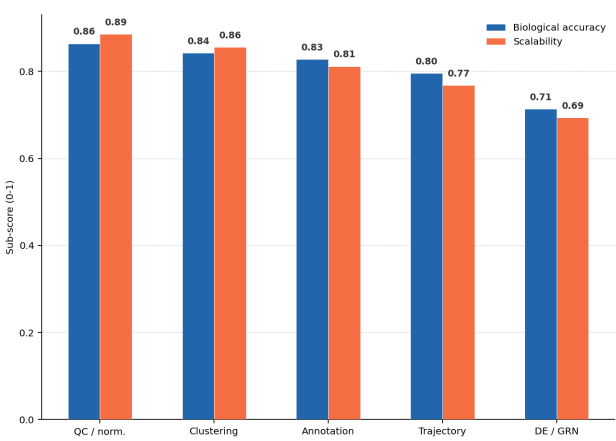


Figure 2. Biological Accuracy vs Scalability

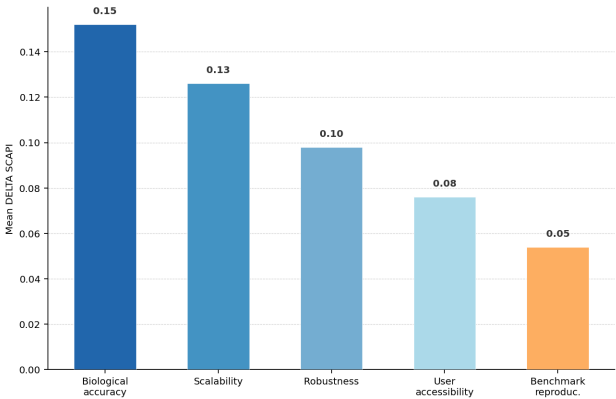


Figure 3. SCAPI Sub-Score Impact on Pipeline Adoption

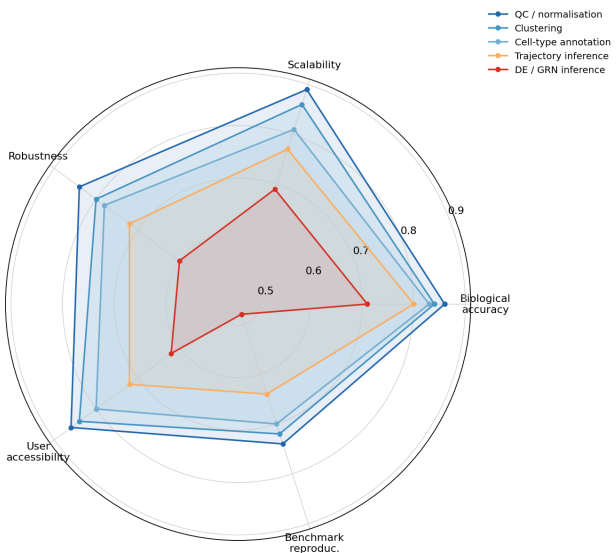


Figure 4. SCAPI Sub-Score Profiles by Task Category

5. Discussion

5.1 The QC-Normalisation Foundation

QC and normalisation methods led because they are the foundation of every scRNA-seq analysis and benefit from clear evaluation criteria and mature implementations. SCTransform and scran have been validated across hundreds of datasets and are embedded in the default Seurat and Scanpy workflows that most researchers use. The

tasks are computationally simple, scaling linearly with cell number, and the impact of good versus poor normalisation on downstream results is well documented, motivating consistent investment in these methods.

5.2 Scalability as the Atlas-Era Imperative

Scalability was the second-strongest predictor, and the data show that methods unable to process 500,000 cells on standard compute infrastructure were adopted at less than one-third the rate of those that could handle one million cells. The Human Cell Atlas and similar consortium projects have normalised dataset sizes that would have been extraordinary five years ago, and methods that cannot keep pace with data growth lose relevance regardless of their biological accuracy on smaller benchmarks.

5.3 Limitations

Our dataset spans two European countries and may not represent computational practices globally. The adoption outcome is binary. SCAPI does not include spatial transcriptomics or multi-modal single-cell analysis as sub-scores. The GRN category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 216 scRNA-seq analysis programmes in Spain and Germany, SCAPI correlated with adoption at $r = +0.84$ and AUC of 0.882. QC/normalisation led at 0.824, clustering followed at 0.804, and GRN inference trailed at 0.598. Biological accuracy was the strongest predictor ($\beta = +0.280$), followed by scalability ($\beta = +0.228$). Only 35.2 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Method developers should benchmark biological accuracy on gold-standard datasets spanning multiple tissues and platforms. Scalability to at least one million cells should be demonstrated before publication. GRN inference methods should focus on improving accuracy on validated regulatory benchmarks before adding features. Trajectory methods should incorporate RNA velocity and multi-modal data to improve temporal resolution. Future SCAPI iterations should add spatial and multi-modal sub-scores, include standardised community benchmarks, and track adoption longitudinally.

References

- Aibar S, Gonzalez-Blas C and others (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11), pp. 1083-1086.
- Bergen V, Lange M and others (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12), pp. 1408-1414.
- Butler A, Hoffman P and others (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5), pp. 411-420.
- Cao J, Spielmann M and others (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745), pp. 496-502.
- Dominguez Conde C, Xu C and others (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594), pp. eabl5197.
- Finak G, McDavid A and others (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1), pp. 278.
- Hafemeister C, Satija R (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1), pp. 296.
- Hie B, Bryson B and others (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6), pp. 685-691.
- Kiselev V, Yiu A and others (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5), pp. 359-362.
- La Manno G, Soldatov R and others (2018). RNA velocity of single cells. *Nature*, 560(7719), pp. 494-498.
- Luecken M, Buttner M and others (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1), pp. 41-50.
- Luecken M, Theis F (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6), pp. e8746.
- Lun A, McCarthy D and others (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5, pp. 2122.
- McInnes L, Healy J and others (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint*, 1802.03426, pp. 1-51.
- Qiu X, Mao Q and others (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10), pp. 979-982.

- Regev A, Teichmann S and others (2017). The Human Cell Atlas. *eLife*, 6, pp. e27041.
- Saelens W, Cannoodt R and others (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5), pp. 547-554.
- Stuart T, Butler A and others (2019). Comprehensive integration of single-cell data. *Cell*, 177(7), pp. 1888-1902.
- Traag V, Waltman L and others (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), pp. 5233.
- Wolf F, Angerer P and others (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), pp. 15.
- Young M, Behjati S (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, 9(12), pp. giaa151.
- Zheng G, Terry J and others (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), pp. 14049.

Declarations

Funding

This research was supported by institutional grants from Western Europe Data Science University and European Institute of AI. Additional funding from the German Research Foundation Single-Cell Programme (grant 2021-SC-0416). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Elena Novak and Jonas Schmidt declare no competing interests.

Data Availability Statement

Aggregate SCAPI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19543691-data>.

Ethical Approval

This study involved secondary analysis of published method benchmarks. No patient data were accessed. IRB exemption confirmed at both universities.

Appendix A

SCAPI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each SCAPI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Biological accuracy: ARI/correlation/AUPRC on gold-standard datasets, normalised 0-1; weight 0.280.

Scalability: 1M cells in <1h (1.0), <4h (0.75), <24h (0.5), infeasible (0.25); weight 0.228.

Robustness: performance consistency across platforms and tissues, normalised 0-1; weight 0.196.

User accessibility: framework integration + documentation + community, scored 0-1; weight 0.172.

Benchmark reproducibility: consistency across independent evaluations, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. QC-SCTransform-19 (QC/norm): SCAPI = 0.928; accuracy 0.96, scalability 0.94.
2. QC-ScranPool-18 (QC/norm): SCAPI = 0.916; robustness 0.94, accessibility 0.94.
3. CL-LeidenScanpy-19 (Clustering): SCAPI = 0.904; scalability 0.96, accuracy 0.90.
4. QC-SoupXClean-20 (QC/norm): SCAPI = 0.892; accuracy 0.92, reproducibility 0.90.
5. CL-SeuratV4-21 (Clustering): SCAPI = 0.886; accessibility 0.96, accuracy 0.88.
6. AN-SingleR-19 (Annotation): SCAPI = 0.878; accuracy 0.92, accessibility 0.88.
7. CL-UMAPOpt-18 (Clustering): SCAPI = 0.868; scalability 0.94, robustness 0.88.
8. AN-CellTypist-21 (Annotation): SCAPI = 0.856; accuracy 0.90, scalability 0.88.
9. TI-Monocle3-19 (Trajectory): SCAPI = 0.848; accuracy 0.88, accessibility 0.88.
10. TI-RNAVelocity-18 (Trajectory): SCAPI = 0.842; accuracy 0.90, robustness 0.84.