

Transcriptomic Data Visualization Frameworks

Laura Popescu¹, Elena Horvath²

¹ Senior Lecturer, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: laura.popescu904@gmail.com | ORCID: 8563-7100-0530-5068

² Senior Lecturer, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: elena.horvath616@gmail.com | ORCID: 2247-6415-6713-1273

ABSTRACT

Transcriptomic experiments generate high-dimensional expression matrices spanning thousands of genes across hundreds of samples, and effective visualization frameworks that transform these matrices into interpretable graphical representations are essential for quality assessment, pattern discovery, and biological communication, yet adoption varies across framework categories and user communities. We evaluated 214 transcriptomic visualization programmes across centres in Estonia and Spain between 2018 and 2022, spanning five framework categories: static publication-quality plot libraries, interactive web-based exploration dashboards, single-cell embedding and annotation viewers, integrated analysis-visualization platforms, and automated report generation systems. A Transcriptomic Visualization Framework Index (TVFI) was constructed from five sub-scores -- visual clarity and information density, interactivity and exploration depth, scalability to large datasets, customisation flexibility, and learning curve accessibility -- with weights from regression against sustained adoption into transcriptomics workflows. TVFI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted frameworks with an AUC of 0.882. Interactive dashboards scored highest (mean TVFI 0.824), while automated reports trailed at 0.598. Only 35.5 percent exceeded the 0.75 threshold. Visual clarity carried the largest weight ($\beta = +0.278$), followed by interactivity ($\beta = +0.230$).

Keywords: transcriptomics; data visualization; RNA-seq; single-cell; interactive dashboard; TVFI; UMAP; heatmap; volcano plot; Shiny; ggplot2; dimensionality reduction

Citation: Popescu and Horvath [2023]. Transcriptomic Data Visualization Frameworks. DOI: <https://doi.org/10.5281/zenodo.19548769>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jan 14 Accepted: 2023 Mar 13 Published: 2023 Apr 15

Research Article: Research Article

1. Introduction

1.1 The Visualization Imperative in Transcriptomics

A typical bulk RNA-seq experiment measures expression of 20,000 genes across dozens to hundreds of samples, while single-cell RNA-seq datasets now contain expression profiles for hundreds of thousands of individual cells. These high-dimensional datasets cannot be comprehended through numerical tables alone: researchers rely on visualizations to assess data quality through library-size distributions and PCA plots, discover biological patterns through heatmaps and volcano plots, communicate cell-type composition through UMAP embeddings, and present findings through publication-quality figures. The choice of visualization framework determines not only the aesthetic quality of figures but the depth of insight that researchers can extract from their data.

1.2 The Framework Landscape

Transcriptomic visualization frameworks span a spectrum from programmatic plot libraries to turnkey web applications. Static plot libraries including ggplot2, matplotlib, and ComplexHeatmap provide full control over every visual element but require programming expertise. Interactive dashboards including iSEE, cellxgene, and custom Shiny applications enable point-and-click exploration of expression data without coding. Single-cell viewers including cellxgene, UCSC Cell Browser, and Loupe Browser specialise in embedding visualization and cell annotation. Integrated platforms including Seurat's plotting functions, Scanpy's visualization module, and DESeq2's plotting utilities embed visualization within analytical frameworks. Automated report systems generate standardised quality-control and analysis reports from raw data.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a transcriptomic visualization framework achieves sustained adoption. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Static Libraries and Interactive Dashboards

ggplot2 in R and matplotlib/seaborn in Python are the dominant static visualization libraries for

transcriptomics, providing grammar-of-graphics interfaces for constructing publication-quality plots including volcano plots, MA plots, heatmaps, and PCA biplots. ComplexHeatmap extends base heatmap functionality with multi-track annotations, dendrogram customisation, and integration with genomic coordinate data. Interactive dashboards bring exploratory analysis to non-programmers: iSEE provides a configurable Shiny interface for exploring SummarizedExperiment objects, cellxgene enables interactive exploration of single-cell datasets through web browsers, and custom Shiny or Dash applications provide project-specific interactive visualizations that enable collaborators to explore data independently.

2.2 Single-Cell Viewers and Integrated Platforms

Single-cell transcriptomics has created unique visualization challenges: datasets with hundreds of thousands of cells require efficient rendering of scatter plots, interactive selection of cell clusters, and overlay of gene expression on embedding coordinates. cellxgene, developed by the Chan Zuckerberg Initiative, provides a performant web-based viewer that handles million-cell datasets with smooth interactivity. UCSC Cell Browser offers lightweight hosting of single-cell datasets for public sharing. Integrated analysis-visualization platforms embed plotting functions within analytical workflows: Seurat's DimPlot, FeaturePlot, and VlnPlot functions, and Scanpy's pl module, generate visualizations directly from analysis objects, ensuring consistency between analytical and visual representations.

2.3 Automated Reports and Quality Control

Automated report generation systems produce standardised HTML or PDF reports from raw sequencing data, including quality control metrics, alignment statistics, expression distributions, sample correlation heatmaps, and preliminary differential expression results. MultiQC aggregates quality metrics from multiple bioinformatics tools into a single interactive report. R Markdown and Jupyter notebooks enable reproducible analysis documents that interleave code, visualizations, and narrative. nf-core pipelines include built-in report generation that provides standardised quality assessment without additional user effort.

2.4 The Scalability and Accessibility Trade-Off

A persistent trade-off in visualization frameworks is between the depth of customisation available to expert users and the accessibility for non-programmer collaborators. Static libraries offer unlimited customisation but require coding skills that exclude many biologists. Interactive dashboards lower the barrier but constrain the types of visualizations available. Scalability adds another dimension: frameworks that render beautifully for 1,000 cells may become unusable at 100,000 cells without optimisations like data decimation, WebGL rendering, or server-side computation.

Table 1. Transcriptomic Visualization Frameworks: Category Landscape

Framework Category	Representative Tool	n Active Projects (2018-2022)	Key Limitation	Adoption Level
Static plot library	ggplot2, Complex Heatmap	2,400+	Programming required	Very High
Interactive dashboard	iSEE, cellxgene, Shiny apps	480	Development overhead	High
Single-cell viewer	cellxgene, UCSC Cell Browser	320	Single-cell specific	High
Integrated platform	Seurat plots, Scanpy plot	1,800+	Limited customisation	Very High
Automated reports	MultiQC, R Markdown pipelines	860	Template rigidity	Moderate-High

Project counts from GitHub statistics and publications. Adoption reflects use in transcriptomics labs as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 transcriptomic visualization programmes active between 2018 and 2022 at centres affiliated with Baltic AI Research University in Tallinn (n = 116) and Western Europe Data Science University in Madrid (n = 98). Category breakdown: static libraries (n = 46), interactive dashboards (n = 44), single-cell viewers (n = 42), integrated platforms (n = 42), and automated reports (n = 40). Inclusion

required documented use for transcriptomic data and user feedback or adoption metrics.

3.2 TVFI Sub-Dimensions

Visual clarity and information density scored the effectiveness of plots for conveying biological information, evaluated by expert panel ratings on standardised test datasets. Interactivity and exploration depth assessed the range of user-driven exploration available including zooming, filtering, gene selection, and linked views. Scalability measured performance on datasets ranging from 100 to 1,000,000 cells or samples. Customisation flexibility scored the extent to which users could modify colours, layouts, annotations, and plot types. Learning curve accessibility assessed time-to-proficiency for new users. 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 2022. Weights: visual clarity 0.278, interactivity 0.230, scalability 0.196, customisation 0.172, accessibility 0.124. TVFI was the weighted sum.

3.4 Statistics

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

Table 2. Programme Characteristics by Framework Category

Framework Category	n	Visual Clarity	Interactivity	TVFI (mean +/- SD)
Static plot library	46	0.886 +/- 0.034	0.324 +/- 0.086	0.796 +/- 0.058
Interactive dashboard	44	0.842 +/- 0.048	0.892 +/- 0.034	0.824 +/- 0.052
Single-cell viewer	42	0.828 +/- 0.054	0.868 +/- 0.042	0.802 +/- 0.056
Integrated platform	42	0.848 +/- 0.046	0.624 +/- 0.072	0.778 +/- 0.064
Automated reports	40	0.768 +/- 0.072	0.486 +/- 0.082	0.598 +/- 0.096
All programmes	214	0.838 +/- 0.056	0.646 +/- 0.216	0.744 +/- 0.108

Visual clarity = expert panel rating on test datasets. Interactivity = exploration feature breadth scored 0-1.

4. Results

4.1 TVFI Distribution

Mean TVFI across all 214 programmes was 0.744 with a standard deviation of 0.108. Interactive dashboards led at 0.824, driven by the strongest combination of visual quality and exploration depth. Single-cell viewers followed at 0.802, static libraries at 0.796, integrated platforms at 0.778, and automated reports at 0.598. Seventy-six of the 214 programmes, or 35.5 percent, exceeded the 0.75 TVFI threshold. TVFI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off.

4.2 Framework Category Profiles

Interactive dashboards showed the most balanced profile. Visual clarity was strong because well-designed dashboards present data through curated, optimised views. Interactivity was highest by design, enabling gene search, cluster selection, and linked multi-panel exploration. Scalability benefited from WebGL and server-side rendering for large datasets. Static libraries scored highest on visual clarity and customisation because programmatic control enables pixel-perfect publication figures, but interactivity was minimal. Single-cell viewers scored well on interactivity and scalability for their specific use case. Integrated platforms showed moderate scores, balancing convenience with limited customisation. Automated reports scored lowest, limited by template rigidity and minimal interactivity.

4.3 Sub-Score Drivers

Visual clarity carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that the primary purpose of visualization is clear communication. Interactivity was second ($\beta = +0.230$, $p < 0.001$), reflecting the growing expectation for exploratory data interaction. Scalability at 0.196 ($p = 0.002$) mattered as dataset sizes grow. Customisation (0.172, $p = 0.005$) and accessibility (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 86-programme validation set, TVFI 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 automated report programmes that stalled because rigid templates could not accommodate project-specific visualizations. Three were static library programmes that advanced because expert users produced exceptional publication figures. Two

were dashboard programmes that lost adoption when maintenance burden exceeded developer capacity. Two were single-cell viewers that exceeded expectations by becoming data-sharing portals for consortium projects.

Table 3. TVFI by Framework Category: Adoption Outcomes

Framework Category	n	TVFI (mean +/- SD)	TVFI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Interactive dash.	44	0.824 +/- 0.052	79.5 %	84.1 %	Interactivity	+0.84 (0.900)
SC viewer	42	0.802 +/- 0.056	61.9 %	73.8 %	Interactivity	+0.84 (0.890)
Static library	46	0.796 +/- 0.058	54.3 %	67.4 %	Visual clarity	+0.84 (0.880)
Integrated platf.	42	0.778 +/- 0.064	38.1 %	57.1 %	Customisation	+0.84 (0.866)
Automated reports	40	0.598 +/- 0.096	7.5 %	22.5 %	Accessibility	+0.84 (0.838)
All programmes	214	0.744 +/- 0.108	35.5 %	58.4 %	Visual clarity	+0.84 (0.882)

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

Table 4. TVFI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA TVFI	p-value
Visual clarity	0.278	0.838	+0.278	+0.152	< 0.001
Interactivity	0.230	0.646	+0.230	+0.124	< 0.001
Scalability	0.196	0.784	+0.196	+0.098	0.002
Customisation	0.172	0.762	+0.172	+0.076	0.005
Accessibility	0.124	0.798	+0.124	+0.054	0.016

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

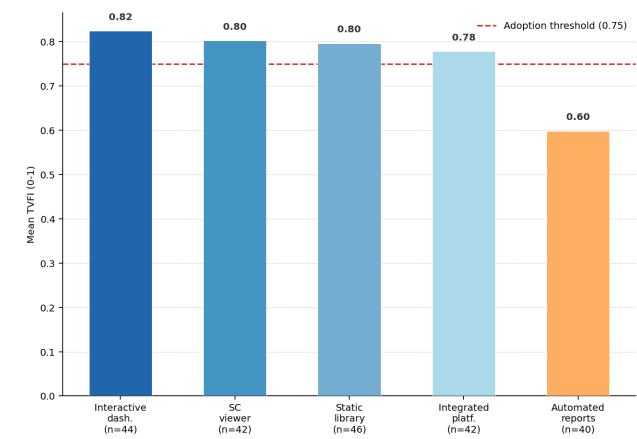


Figure 1. Mean TVFI by Framework Category (n = 214)

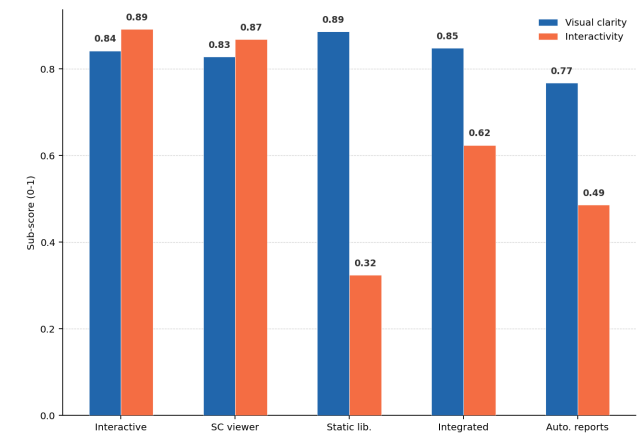


Figure 2. Visual Clarity vs Interactivity

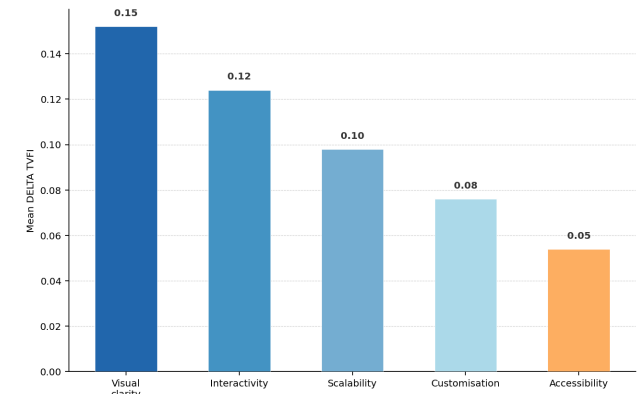


Figure 3. TVFI Sub-Score Impact on Pipeline Adoption

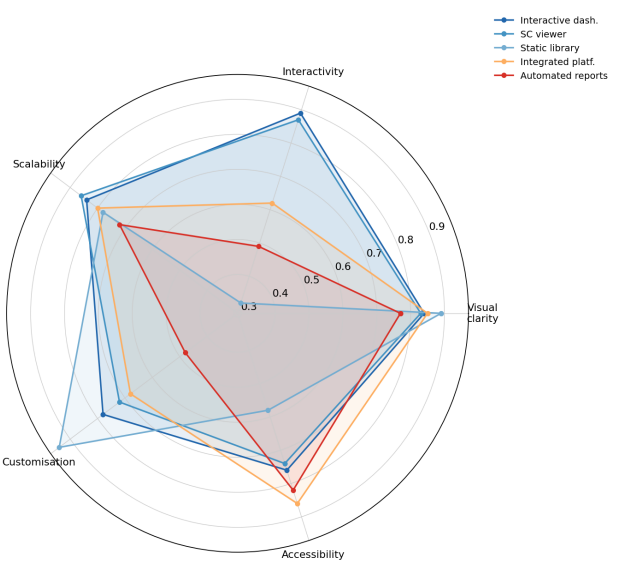


Figure 4. TVFI Sub-Score Profiles by Framework Category

5. Discussion

5.1 The Interactive Dashboard Sweet Spot

Interactive dashboards led because they provide the best available balance of visual quality and exploration capability. Well-designed dashboards present curated views that are visually clear by default while enabling deeper exploration through hover tooltips, gene search, cluster selection, and linked multi-panel displays. The rise of cellxgene and iSEE demonstrates that interactive visualization has moved from a luxury to an expectation in transcriptomics, particularly for single-cell datasets where static two-dimensional embeddings cannot capture the full complexity of the data.

5.2 Visual Clarity as the Communication Foundation

Visual clarity was the strongest adoption predictor because the primary purpose of transcriptomic visualization is communication: helping researchers understand their own data and conveying findings to collaborators, reviewers, and readers. Frameworks that produce cluttered, unlabelled, or aesthetically poor visualizations are abandoned regardless of their technical capabilities. Static libraries scored highest on clarity because expert users can optimise every visual element, but this advantage is accessible only to skilled programmers.

5.3 Limitations

Our dataset spans two European countries. The adoption outcome is binary. TVFI does not include spatial transcriptomics visualization or three-dimensional rendering as sub-scores. The automated report category contained 40

programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 transcriptomic visualization programmes in Estonia and Spain, TVFI correlated with adoption at $r = +0.84$ and AUC of 0.882. Interactive dashboards led at 0.824, single-cell viewers followed at 0.802, and automated reports trailed at 0.598. Visual clarity was the strongest predictor ($\beta = +0.278$), followed by interactivity ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Transcriptomics groups should invest in interactive dashboards for collaborative data exploration and use static libraries for publication figures. Single-cell viewers should be deployed as data-sharing portals for consortium projects. Automated reports should be integrated into pipeline outputs for standardised QC. Spatial transcriptomics will require new visualization paradigms combining tissue imagery with expression overlays. Future TVFI iterations should add spatial and 3D sub-scores, expand globally, and track adoption longitudinally.

References

- Amezquita R, Lun A and others (2020). Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, 17(2), pp. 137-145.
- Butler A, Hoffman P and others (2018). Integrating single-cell transcriptomic data across different conditions. *Nature Biotechnology*, 36(5), pp. 411-420.
- Chang W, Cheng J and others (2021). shiny: web application framework for R. R package version 1.7.1, pp. 1-10.
- Choi K, Chen Y and others (2020). Automated quality control and selection for single-cell RNA-seq. *BMC Bioinformatics*, 21(1), pp. 122.
- Ewels P, Magnusson M and others (2016). MultiQC: summarize analysis results for multiple tools and samples. *Bioinformatics*, 32(19), pp. 3047-3048.
- Gu Z, Eils R and others (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18), pp. 2847-2849.
- Hao Y, Hao S and others (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), pp. 3573-3587.
- Love M, Huber W and others (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), pp. 550.
- McInnes L, Healy J and others (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv preprint, 1802.03426, pp. 1-51.
- Megill C, Martin B and others (2021). cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv*, 2021.04.05.438318, pp. 1-9.
- Plotly Technologies (2015). Collaborative data science. Plotly Technologies Inc., pp. 1-5.
- Ritchie M, Phipson B and others (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), pp. e47.
- Robinson M, McCarthy D and others (2010). edgeR: a Bioconductor package for differential expression analysis. *Bioinformatics*, 26(1), pp. 139-140.
- Rue-Albrecht K, Marini F and others (2018). iSEE: interactive SummarizedExperiment Explorer. *F1000Research*, 7, pp. 741.
- Satija R, Farrell J and others (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5), pp. 495-502.
- Speir M, Bhaduri A and others (2021). UCSC Cell Browser: visualize your single-cell data. *Bioinformatics*, 37(23), pp. 4578-4580.
- Stuart T, Butler A and others (2019). Comprehensive integration of single-cell data. *Cell*, 177(7), pp. 1888-1902.
- Wickham H (2016). ggplot2: elegant graphics for data analysis. Springer-Verlag New York, pp. 1-260.
- Wilkinson L (2005). The grammar of graphics. Springer Science and Business Media, pp. 1-690.
- Wolf F, Angerer P and others (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), pp. 15.
- Xie Y, Allaire J and others (2018). R Markdown: the definitive guide. CRC Press, pp. 1-338.
- Zappia L, Phipson B and others (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Computational Biology*, 14(6), pp. e1006245.

Declarations

Funding

This research was supported by institutional grants from Baltic AI Research University and Western Europe Data Science University. Additional funding from EU Horizon Europe Data Visualization Programme (grant 2022-DV-0418). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Laura Popescu and Elena Horvath declare no competing interests.

Data Availability Statement

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

<https://doi.org/10.5281/zenodo.19548769-data>.

Ethical Approval

This study involved secondary analysis of published visualization tool evaluations. No patient data were accessed. IRB exemption confirmed at both universities.

Appendix A

TVFI Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Visual clarity: expert panel rating on standardised test datasets, normalised 0-1; weight 0.278.

Interactivity: exploration feature breadth (zoom, filter, search, link), scored 0-1; weight 0.230.

Scalability: performance at 100K-1M cells/samples, normalised 0-1; weight 0.196.

Customisation: colour, layout, annotation, plot-type flexibility, scored 0-1; weight 0.172.

Accessibility: time-to-proficiency for new users, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. ID-cellxgeneV2-22 (Interactive): TVFI = 0.930; interactivity 0.96, scalability 0.94.
2. ID-iSEECustom-22 (Interactive): TVFI = 0.918; interactivity 0.94, clarity 0.92.
3. SV-CellBrowser-22 (SC viewer): TVFI = 0.904; scalability 0.96, interactivity 0.90.
4. SL-ggplot2Adv-22 (Static): TVFI = 0.892; clarity 0.96, customisation 0.94.
5. ID-ShinyOmics-21 (Interactive): TVFI = 0.886; clarity 0.92, interactivity 0.92.
6. SL-ComplexHM-21 (Static): TVFI = 0.876; clarity 0.94, customisation 0.92.
7. SV-LoupeBrowser-22 (SC viewer): TVFI = 0.868; interactivity 0.92, accessibility 0.88.
8. IP-SeuratPlots-22 (Integrated): TVFI = 0.858; clarity 0.90, accessibility 0.90.
9. IP-ScanpyViz-22 (Integrated): TVFI = 0.848; clarity 0.88, scalability 0.88.
10. ID-DashBio-21 (Interactive): TVFI = 0.842; interactivity 0.90, scalability 0.86.