scCluBench: Comprehensive Benchmarking of Clustering Algorithms for Single-Cell RNA Sequencing

Ping Xu[1,3], **Zaitian Wang**[1,3], **Zhirui Wang**[2,3], **Pengjiang Li**[1,3], **Jiajia Wang**[1], **Ran Zhang**[1,3], **Pengfei Wang**[1,2,3,*], **Yuanchun Zhou**[1,2,3]

[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]Hangzhou Institute for Advanced Study,University of Chinese Academy of Sciences, Hangzhou, China
[3]University of Chinese Academy of Sciences, Beijing, China
xuping0098@gmail.com, wpf2106@gmail.com, zyc@cnic.cn

Abstract

Cell clustering is crucial for uncovering cellular heterogeneity in single-cell RNA sequencing (scRNA-seq) data by identifying cell types and marker genes. Despite its importance, existing benchmarks for scRNA-seq clustering remain fragmented, lacking standardized protocols and often omitting recent advances in artificial intelligence. To fill these gaps, we present **scCluBench**, a comprehensive **bench**mark of **clu**stering algorithms for **sc**RNA-seq data. scCluBench provides 36 scRNA-seq datasets collected from diverse public sources, covering multiple tissues, which are uniformly processed to ensure consistency for systematic evaluation and downstream analyses. To assess performance, we collect and reproduce a range of scRNA-seq clustering methods, including traditional, deep learning-based, graph-based, and biological foundation models. We comprehensively evaluate each method both quantitatively and qualitatively, using core performance metrics and visualization analyses. Furthermore, we construct representative downstream biological tasks, such as marker gene identification and cell type annotation, to further assess the practical utility. scCluBench then investigates the performance differences and applicability boundaries of various clustering models across diverse analytical tasks, systematically assessing their robustness and scalability in real-world scenarios. Overall, scCluBench offers a standardized and user-friendly benchmark for scRNA-seq clustering, with standardized datasets, unified evaluation protocols, and transparent analyses, facilitating informed method selection and providing valuable insights into model generalizability and application scope.

Code — https://github.com/XPgogogo/scCluBench
Datasets — https://github.com/XPgogogo/scCluBench/Data
Extended version — https://arxiv.org/abs/2512.02471

Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed biological research by enabling the high-resolution exploration of cellular diversity, developmental processes, and tissue organization (Shapiro, Biezuner, and Linnarsson 2013). scRNA-seq clustering, which groups cells based on gene expression profiles, is a cornerstone analysis in scRNA-seq studies and underpins critical tasks such as cell type characterization, atlas construction, and marker gene discovery (Kiselev, Andrews, and Hemberg 2019; Wang et al. 2025a). As scRNA-seq datasets grow in size and complexity, the challenges of achieving robust, reproducible, and biologically meaningful clustering results become increasingly prominent, highlighting the urgent need for advanced computational techniques. However, there is currently no comprehensive and standardized benchmarking framework for scRNA-seq clustering methods, making it difficult to objectively compare model performance, assess robustness and reproducibility across datasets, and select appropriate tools for specific biological contexts (Xu et al. 2025c; Krzak et al. 2019).

Powered by traditional and artificial intelligence methods, we propose **scCluBench**, a comprehensive **bench**marking framework for **s**ingle-**c**ell RNA sequencing **clu**stering. scCluBench systematically compares clustering algorithms under unified conditions, providing standardized solutions in all major stages of scRNA-seq clustering benchmarking, including data resources, evaluation metrics, biological interpretation pipelines, and unified benchmarking workflows.

(1) **Standardization of benchmark resources.** Existing scRNA-seq clustering benchmarks often lack dataset diversity, such as limited species or tissue types, and insufficient coverage of emerging models, particularly recent advances in biological foundation models built upon Transformer architectures. scCluBench present a collection of 36 human and mouse datasets spanning diverse tissues. This standardized resource, encompassing traditional, deep learning–based, graph-based, and foundation models, enables systematic evaluation and fair comparison of single-cell clustering methods.

(2) **Standardization of evaluation protocols.** Assessment of scRNA-seq clustering methods often relies on limited quantitative and qualitative metrics. Thus, we standardize the evaluation process by incorporating diverse quantitative indicators on multiple datasets, along with qualitative assessments such as 2D visualization of cell embeddings. In particular, we offer quantitative analyses of embedding similarity-visualized to systematically evaluate phenomena like representation collapse and provide broader perspectives for model selection and optimization.

(3) **Standardization of biological interpretation.** Downstream analyses such as marker gene identification and cell type annotation are essential for interpreting cluster-

*Corresponding author.

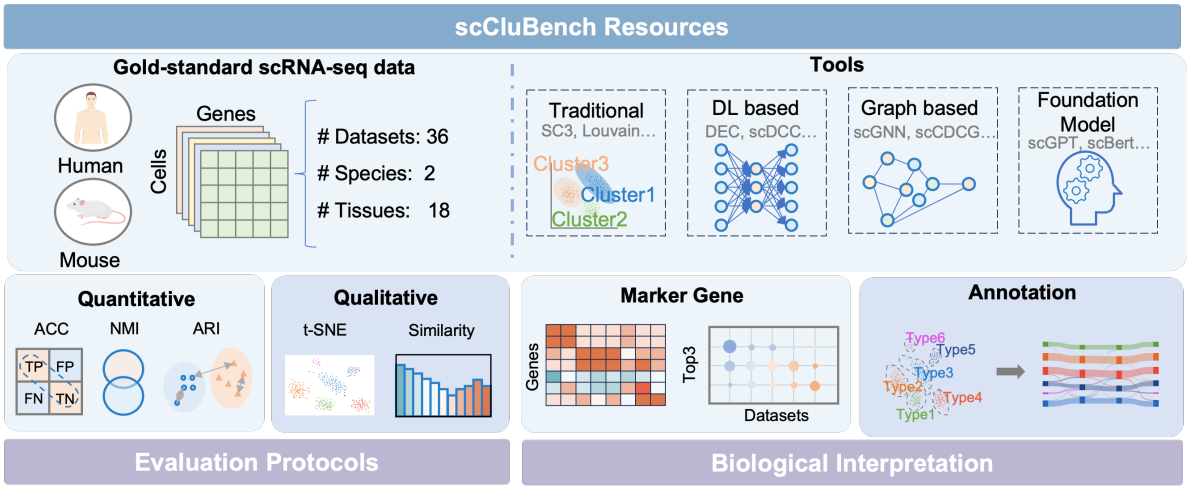


Figure 1: Overview of scCluBench: resources, evaluation protocols, and biological interpretation.

ing results, yet they are often inconsistently addressed. scCluBench deliver standardized, reproducible pipelines for marker gene detection and cell type labeling, complemented by gold-standard references for annotation. This ensures clustering outputs can be validated and interpreted in biological contexts, facilitating applications in single-cell research.

(4) **Unified benchmarking Workflow and Modular Code.** scCluBench provides an integrated and reproducible workflow covering data preprocessing, clustering, and cell type annotation. Standardized input–output formats and modularized implementations ensure ease of use and enable fair and consistent performance comparisons across all models.

By constructing scCluBench, we systematically enabled comparative analyses and identified several key findings:

- We identified three critical components for fair and effective evaluation of scRNA-seq clustering methods: diverse and representative datasets, broad coverage computational methods, and a unified and reproducible analysis pipeline with standardized input/output formats.
- We find that existing scRNA-seq clustering methods suffer from distinct but significant limitations. Traditional methods perform poorly in handling sparse, high-noise data. Deep learning approaches, while effective in dimensionality reduction and denoising, often fail to capture underlying relationships between cells. Graph-based models, although improving structural awareness, suffer from issues such as over-smoothing and embedding collapse. More fundamentally, most methods decouple embedding learning from clustering optimization, resulting in embedding spaces that are not fully conducive to clustering, thereby limiting overall performance.
- We find that current scRNA-seq foundation models are often designed to construct a unified embedding space transferable to multiple downstream tasks, prioritizing general cell representation rather than task-specific optimization. Although such a general-purpose design enhances cross-task transferability, it also diminishes performance in specific tasks, such as clustering.

scCluBench

Benchmark Framework

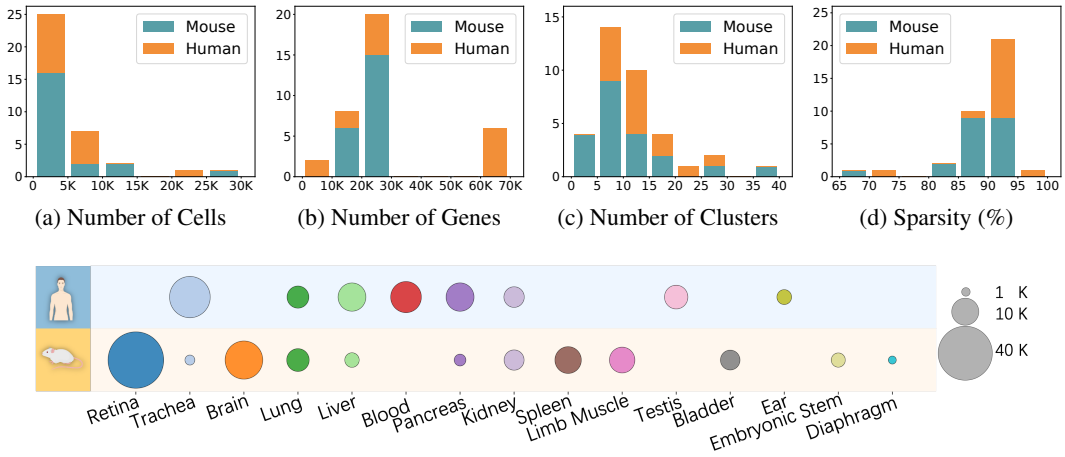
The scCluBench framework, as shown in Fig. 1, offers an extensive benchmark for scRNA-seq clustering. It features a curated collection of 36 diverse datasets derived from human and mouse, spanning 18 tissue types, which serve as comprehensive testbeds for evaluating clustering algorithms. The benchmark encompasses a wide spectrum of scRNA-seq clustering methods, including traditional, deep learning-based, graph-based, and, notably, emerging biological foundation models. This diverse combination of datasets and clustering methods underpins a thorough evaluation, combining quantitative metrics and qualitative analyses, such as 2D cell embedding visualizations for comprehensive insights. Additionally, scCluBench standardizes biological interpretation through reproducible pipelines for marker gene detection and cell type annotation, ensuring clustering results are effectively validated within the context of real biological applications.

Benchmark Datasets

As data quality and diversity are critical to model performance (Wang et al. 2025b,c), scCluBench comprises a diverse collection of 36 single-cell gene expression datasets from human and mouse specimens, covering 18 distinct tissue types, as shown in Fig. 2. Notably, scCluBench includes 2 large-scale datasets with over 20,000 cells and 5 high-dimensional datasets containing more than 60,000 genes. Additionally, 4 datasets contain at least 20 cell types, and 34 datasets exhibit sparsity rates exceeding 80%, with overall sparsity levels ranging from 65.76% to 95.42%.

Benchmark Models

The scCluBench benchmarks a representative collection of state-of-the-art (SOTA) clustering algorithms, spanning four methodological categories: traditional clustering, deep learning-based, graph-based, and biological foundation models, offering a diverse and standardized framework for systematic and fair evaluation of single-cell clustering performance.



(a) Number of Cells (b) Number of Genes (c) Number of Clusters (d) Sparsity (%)

(e) Distribution of Sample Numbers in Human and Mouse Data

Figure 2: Dataset distributions. (a) to (d) show dataset distributions by cell count, gene number, clusters, and sparsity, while (e) displays the distribution of samples between human and mouse datasets.

Methods	Methods	Framework	# of Clusters	Language	Ref.	Journal
Traditional Models	SC3	SingleCellExperiment	Automatic	R	(Kiselev et al. 2017)	Nature Methods
	Louvain	Seurat	Automatic	R	(Stuart et al. 2019)	Cell
	Leiden	Seurat	Automatic	R	(Stuart et al. 2019)	Cell
Deep Learning -based Models	DEC	AE	Automatic	Python	(Xie, Girshick, and Farhadi 2016)	ICML
	scDeepCluster	AE	Hyperparameter	Python	(Tian et al. 2019)	Nature Machine Intelligence
	DESC	Stacked AE	Automatic	Python	(Li et al. 2020)	Nature Communications
	scziDesk	AE	Hyperparameter	Python	(Chen et al. 2020)	NAR Genomics and Bioinformatics
	scDCC	AE	Hyperparameter	Python	(Tian et al. 2021)	Nature Communications
	scNAME	Masked AE	Hyperparameter	Python	(Wan, Chen, and Deng 2022)	Bioinformatics
	scMAE	Masked AE	Hyperparameter	Python	(Fang, Zheng, and Li 2024)	Bioinformatics
Graph-based Models	scGNN	GNN	Hyperparameter	Python	(Wang et al. 2021)	Nature Communications
	scDSC	AE + GNN	Hyperparameter	Python	(Gan et al. 2022)	Briefings in Bioinformatics
	AttentionAE-sc	AE + GNN	Automatic	Python	(Li et al. 2023)	PLOS Computational Biology
	scCDCG	Cut-informed graph embedding	Hyperparameter	Python	(Xu et al. 2024)	DASFAA
Foundation Models	scGPT	Masked Language Model	Hyperparameter (Finetune)	Python	(Cui et al. 2024)	Nature Methods
	GeneFormer	Context-aware BERT	Hyperparameter (Finetune)	Python	(Theodoris et al. 2023)	Nature
	GeneCompass	Knowledge-informed Transformer	Hyperparameter (Finetune)	Python	(Yang et al. 2024)	Cell Research

Table 1: Benchmark Clustering methods (AE: Autoencoder; GNN: Graph Neural Network).

A comprehensive list of all methods with brief descriptions is provided in Tab. 1. All methods follow parameter settings from original publications. When unspecified, we perform controlled tuning to ensure stable and broadly applicable performance. Each dataset-method pair is independently run five times, and results are reported as mean ± standard deviation.

Benchmark Evaluation

Evaluation Protocols. We propose standardized evaluation protocols encompassing three quantitative metrics to assess clustering accuracy and two qualitative approaches to examine the distinguishability of cell representations.

Quantitative Analysis. The primary aim of single-cell clustering is to assign cells of each type with faithful and consistent class labels. Our evaluation of clustering performance focuses on three established metrics from the public domain: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). For these measurements, higher values indicated better performance.

Qualitative Analysis. In machine learning, building accurate decision boundaries relies on distinguishable features. So it is important to understand the quality of cell representations learned by different methods and how the quality affects model performance. To this end, we offer 2D visualizations of cell embeddings using t-SNE to enable a more straightforward elaboration on how cell features are learned and clustered. Additionally, we calculate representation similarities of all learned embeddings and qualitatively analyze the probability distributions. By observing embeddings' concentration on the high-similarity region, we can assess the severity of the over-smoothing problem.

Biological Interpretation. We propose a biological evaluation framework to systematically assess the concordance between predicted clusters and true cell types.

Marker Gene Identification. Differentially expressed genes (DEGs) are genes with significant expression differences across cell populations or experimental conditions. Among these, genes that display strong cluster-specific expression patterns can be further selected as marker genes for cell type annotation. We detected DEGs for each cluster

with "`rank_genes_groups`" (default settings) from the Scanpy package. Using ground-truth cell type labels, we first extract the top 100 DEGs per reference cluster to form a *gold-standard* marker list. The same procedure is then applied to clusters predicted by each model, yielding a comparable list of 100 marker genes per cluster. We further plot the top 3 marker genes using the **tracksplot** diagram to provide straightforward presentations and compare how the expression values of DEGs stand out among other genes.

Cell Type Annotation. Cell type annotation serves as the primary downstream task and main objective of single-cell clustering. In scCluBench, we perform and compare two annotation approaches. **a. Best-mapping annotation**: As a rapid approach, it directly aligns model-predicted cluster labels to ground-truth labels by maximizing one-to-one correspondence using the Hungarian algorithm, disregarding gene expression. **b. Marker-overlap annotation**: For each model-predicted cluster, we compute the proportion of shared genes between its marker list and the marker list of each gold-standard cluster. This overlap is calculated as $\text{score}(p, g) = |\text{DEG}_p \cap \text{DEG}_g|/100$, where p and g are the model-predicted cluster label and gold-standard cluster label. The cluster's cell type is assigned to the gold-standard cell type with the highest overlap score. To elucidate discrepancies between the two annotation methods and quantify their deviations from the gold standard labels, we construct **Sankey diagrams** to trace the relationships between best-mapping annotations, marker-overlap annotations, and gold standard cell types.

Observation and Analysis

Quantitative Analysis

Overall Clustering Performance. Tab. 2 summarizes the clustering accuracy of each method on single-cell clustering tasks, where scCDCG stands out among all methods owing to its cut-informed graph embedding mechanism. By integrating detailed dataset characteristics, we derive the following consistent observations: (1) *Traditional methods* perform well on datasets with fewer than 5,000 cells and moderate gene dimensions. However, as the scale and complexity of data increase, their reliance on low-dimensional distance metrics leads to a marked decline in accuracy and stability. (2) *Among deep learning-based methods*, scMAE demonstrates robustness across varying data scales and sparsity levels. Its self-supervised feature reconstruction effectively mitigates expression bias caused by sparsity. scNAME ranks second, exhibiting robustness but performing slightly inferior on large-scale datasets. (3) *Graph-based methods* show distinct advantages in handling sparse data; however, their performance varies depending on graph construction strategies. Compared to hard graphs based on binary adjacency relationships, soft graphs, exemplified by the continuous edge-weight mechanism in scCDCG, provide a more refined characterization of intercellular similarities and differences, resulting in improved clustering accuracy on complex datasets. Overall, although traditional methods offer simplicity and interpretability for basic tasks, deep learning and graph neural network methods, with superior modeling capacity and robustness, are more suitable for large-scale, high-sparsity single-cell data.

Performance of Biological Foundation Models. To evaluate biological foundation models for single-cell data, we conducted classification and clustering experiments, with results summarized in Tab. 3. Compared with the clustering methods listed in Tab. 2, biological foundation models exhibit a marked performance gap in clustering tasks. Specifically, GeneFormer consistently underperforms in clustering accuracy across most datasets, whereas scGPT shows moderate gains on select datasets. Conversely, these models achieve consistently higher accuracy and F1 score in classification tasks, with stable performance across multiple datasets, indicating superior generalization capabilities. This performance gap reveals a limitation of current foundation models, which prioritize learning generalizable cell representations to enable broad transferability across downstream tasks but often sacrifice task-specific optimization, especially for fundamental tasks such as clustering that require dedicated mechanisms.

Program Exceptions. During our experiments, we have observed some exceptions. Some out-of-memory errors occur when handling large datasets, such as when processing *Macosko mouse retina* (14K+ samples) and *Shekhar mouse retina* (20K+ samples) with AttentionAE-sc and processing *Muris Brain* (13K+ samples) with scziDesk. In a rare case, namely when running scDCC on *QS Limb Muscle*, the program fails due to a NaN-valued loss error.

Qualitative Analysis

Cell Cluster Visualization. To enhance the interpretability of the quantitative results, we perform dimensionality reduction on the embeddings derived by each method and plot the cells on a 2D plane (Fig. 3). We highlight 3 considerations to evaluate the quality of cell clustering through the visualization: (1) the number of clusters, (2) the boundary of each cluster, and (3) the compactness of cells within each cluster. Regarding the number of clusters, we notice that some methods frequently generate mismatched assignments with ground-truth labels. For example, DESC assigns cells to 12 clusters for *Muris Limb Muscle*, while the ground-truth has 6 cell types; it assigns cells to 2 clusters for *Sapiens Ear Utricle*, while the ground-truth has 5 cell types. Such inconsistency leads to inferior clustering accuracies as shown in Tab. 2. As for the boundary, we notice that some graph-based methods yield vague boundaries, such as those by scDSC and scGNN, likely related to their inferior performance compared with scCDCG, which boasts a much clearer boundary and better performance. In terms of compactness, we can see that scDCC represents cells of the same categories with highly similar embeddings and locates them in proximal locations, suggesting that common patterns of each type are well-recognized, in accordance with its decent performance.

Cell Representation Distinguishability. Despite effectively leveraging graph structures, graph-based clustering methods, particularly GNN-based ones, often suffer from over-smoothing and representation collapse (Wang et al. 2024; Ning et al. 2025), where node representations across classes become indistinguishable, due to the inductive bias

	Traditional			Deep Learning-based							Graph-based			
	SC3	louvain	leiden	DEC	DESC	scDeepCluster	scMAE	scNAME	scDCC	scziDesk	scGNN	scDSC	AttentionAE-sc	scCDCG
Human Pancreas 1	87.35±1.56	76.56±7.65	71.76±0.21	37.97±1.10	76.20±1.02	64.50±3.78	87.99±3.41	75.94±12.61	69.51±3.25	68.13±1.47	56.65±0.42	58.57±5.15	81.13±1.88	**92.15±0.84**
Human Pancreas 2	83.87±1.49	**90.66±0.69**	90.55±1.60	45.05±1.79	67.13±3.10	54.83±1.66	79.12±3.07	67.07±5.72	69.95±3.65	54.10±1.15	55.05±3.67	77.76±1.26	84.37±8.11	84.66±4.11
Human Pancreas 3	79.97±4.29	91.40±0.32	**93.31±0.22**	45.47±1.54	91.93±3.99	50.88±3.15	90.11±3.94	71.01±5.37	63.62±7.48	78.78±12.26	67.84±4.53	83.73±0.59	91.20±2.79	88.04±0.34
Human Pancreas4	68.30±2.84	73.52±0.00	72.99±2.47	46.58±5.15	70.51±4.75	55.56±1.21	77.56±7.98	67.00±4.52	57.55±3.13	64.67±9.76	44.90±0.00	76.93±3.77	82.33±4.65	**86.57±0.29**
Mauro Pancreas	83.51±1.77	88.69±0.00	92.08±0.13	65.61±1.10	76.22±2.33	73.55±1.37	**95.62±0.16**	89.73±9.92	86.72±9.05	69.97±17.08	64.84±2.25	70.31±5.59	92.43±2.72	92.65±2.93
68K PBMC	79.44±0.23	65.04±0.09	60.88±6.50	59.36±1.12	55.48±3.54	81.06±4.74	75.84±0.16	78.40±1.67	**82.79±2.96**	63.76±2.01	41.63±2.88	44.60±0.00	62.63±2.31	78.52±0.67
CITE CMBC	73.67±2.99	70.51±18.30	**78.04±3.27**	38.67±5.51	67.07±2.29	65.95±2.27	72.91±5.55	63.53±3.14	72.44±2.94	47.47±12.26	64.56±2.49	30.78±0.59	64.77±8.75	71.45±1.85
Human Kidney	69.67±3.49	67.77±0.00	72.68±6.32	38.82±3.19	60.83±3.39	63.26±4.55	**83.99±3.45**	73.78±0.75	60.64±3.41	80.08±0.57	40.44±0.00	40.09±0.00	63.77±4.54	79.55±0.29
Sonya liver	79.09±4.36	58.04±0.01	69.84±5.17	40.90±1.22	57.17±3.80	69.21±2.47	**80.73±1.86**	79.97±8.80	70.64±4.59	76.58±1.52	33.60±0.00	42.44±0.09	78.60±8.70	75.34±3.67
Sapiens Liver	**85.74±8.08**	79.30±0.00	71.19±0.00	65.76±3.20	42.24±0.00	49.08±1.90	67.51±2.30	63.33±1.70	57.88±4.60	68.19±0.60	73.83±3.00	64.67±4.40	68.76±11.20	73.09±1.40
Sapiens Ear Crista Ampullaris	56.96±10.69	44.84±0.00	43.37±0.94	57.73±2.07	29.83±0.00	53.73±2.50	67.00±0.40	67.35±0.70	80.67±3.70	39.58±0.00	83.12±1.00	76.20±8.10	81.46±7.90	**85.45±4.30**
Sapiens Ear Utricle	59.62±0.00	53.07±0.00	51.21±0.00	59.38±8.40	60.23±0.00	54.83±1.20	73.16±0.60	71.95±1.40	68.51±5.70	73.58±0.60	69.89±2.30	**84.94±6.70**	62.49±9.30	79.58±0.70
Sapiens Lung	51.75±0.00	53.58±1.44	48.38±0.11	60.11±5.00	55.50±0.00	44.07±1.70	63.24±1.10	59.09±2.30	57.40±3.20	71.18±1.70	**74.81±1.60**	61.88±5.80	69.64±9.00	62.06±1.60
Sapiens Testis	42.63±0.00	62.93±13.22	62.86±0.00	45.38±1.50	20.79±0.00	35.17±1.70	53.71±0.80	63.05±9.10	43.89±1.00	74.71±2.80	**79.66±0.20**	69.13±12.00	71.37±15.90	67.18±3.80
Sapiens Trachea	43.46±0.00	56.98±24.56	48.49±0.95	52.30±8.40	50.87±0.00	39.62±0.60	65.78±3.70	70.71±4.00	56.12±3.70	68.97±6.30	66.83±0.00	71.45±6.00	**87.85±4.93**	52.46±2.90
Mouse cerebral cortex	80.53±0.40	65.32±0.47	73.64±0.00	49.87±9.58	58.62±3.53	73.60±2.73	72.99±0.62	**80.70±1.59**	74.27±6.53	76.68±3.65	36.22±0.22	32.80±0.62	71.80±2.36	71.73±0.12
Mouse embryonic stem	88.15±2.34	83.47±1.87	83.44±2.05	66.75±11.59	70.92±8.65	97.47±0.42	80.73±1.93	85.38±0.52	73.86±7.73	88.82±1.61	62.84±5.00	71.84±0.00	78.77±2.90	**98.96±0.06**
Mouse hypothalamus	60.15±3.76	24.40±2.15	18.05±1.98	54.60±5.90	43.93±4.47	59.82±2.81	**89.56±0.20**	88.92±1.61	67.45±4.34	89.54±0.39	38.25±0.00	38.49±0.55	79.13±6.91	85.31±0.34
Mouse Pancreas 1	58.15±2.67	73.84±1.92	73.36±2.41	48.25±1.92	62.14±5.63	43.45±1.51	74.18±7.57	60.49±3.43	51.39±4.84	68.76±3.08	47.93±4.35	49.05±2.96	81.70±3.73	**82.64±0.51**
Mouse Pancreass 2	65.04±3.12	69.92±2.54	68.98±1.87	32.55±0.73	52.39±3.55	49.68±2.80	91.89±0.20	72.62±8.95	57.38±0.34	74.81±12.83	42.54±2.91	75.47±2.83	84.53±7.39	**93.97±0.23**
Shekhar mouse retina	67.94±2.89	80.68±1.76	70.14±2.32	26.17±2.84	86.47±8.72	63.83±4.61	**93.51±0.02**	89.93±0.43	70.21±1.77	51.72±0.25	27.99±0.74	37.10±6.26	OOM	76.04±1.85
Macosko mouse retina	64.88±2.45	70.04±1.98	63.58±2.17	31.60±2.37	84.85±2.20	54.52±2.13	**87.68±1.01**	80.15±5.16	62.74±3.71	72.65±3.59	27.15±0.00	42.96±0.00	OOM	69.78±0.70
Mouse Kidney	92.27±1.87	69.32±2.34	81.07±1.65	24.65±1.83	68.07±0.00	75.60±2.13	**93.45±0.17**	87.49±7.45	80.59±5.39	89.14±9.02	20.70±0.13	19.89±3.22	30.33±16.14	61.47±1.07
Mouse bladder	63.29±2.76	73.81±1.92	**82.52±1.43**	50.12±3.02	76.09±5.14	62.94±3.62	66.21±6.92	64.94±2.70	68.99±0.97	43.09±6.75	52.32±3.35	46.83±1.55	52.27±3.90	75.61±1.23
QS Diaphragm	94.48±1.23	78.97±2.45	78.51±1.87	49.96±7.16	65.06±9.82	71.34±0.35	**98.97±0.11**	98.01±0.29	39.77±7.58	71.38±15.54	50.25±2.99	56.71±2.85	95.83±1.21	98.71±0.25
QS Lung	53.70±2.87	66.11±1.76	52.68±2.34	36.99±1.27	57.24±2.13	47.93±3.00	74.50±1.21	69.77±1.02	51.19±10.54	76.31±4.18	41.86±0.84	49.29±2.23	**76.73±3.92**	70.58±0.87
QS Trachea	80.96±2.13	39.04±3.21	57.26±2.87	47.51±6.05	33.90±3.74	67.41±0.90	82.62±10.50	82.86±5.37	59.14±3.47	**85.88±5.56**	48.07±2.86	49.48±16.76	79.10±10.18	85.42±0.03
QS Limb Muscle	91.65±1.45	73.58±2.76	71.28±2.34	49.05±0.43	52.46±5.65	69.30±8.48	**98.96±0.14**	98.35±0.24	ERR	89.66±7.11	47.28±0.89	50.25±0.58	87.33±6.52	92.94±1.00
Qx Limb Muscle	83.09±2.34	58.79±3.12	76.08±1.98	75.82±0.81	51.15±8.49	79.35±4.08	**99.05±0.86**	98.73±0.45	84.54±3.37	97.25±0.87	56.15±0.36	61.73±0.00	97.47±1.23	96.51±0.60
Qx Bladder	77.40±2.56	46.52±3.45	48.28±2.87	74.56±4.61	52.38±2.56	78.53±0.84	84.41±12.91	91.96±12.92	73.58±3.63	**99.59±0.09**	79.12±1.32	77.32±6.28	94.80±3.81	98.84±0.31
Qx Spleen	55.79±2.87	43.63±3.21	43.81±2.65	48.84±6.22	53.81±8.00	65.28±0.88	96.06±0.18	**96.52±1.66**	65.91±16.76	96.04±1.24	59.46±1.97	75.60±3.65	87.47±13.77	94.48±0.49
Muris Limb Muscle	**98.60±0.87**	97.13±1.23	96.72±1.45	54.79±6.50	39.22±0.00	59.57±3.90	66.13±3.40	61.34±3.10	70.38±4.20	53.31±4.30	48.62±2.30	64.37±4.00	53.35±10.50	94.50±7.10
Muris Brain	54.60±3.21	33.90±2.87	40.84±3.45	55.70±3.20	15.02±0.00	85.36±18.10	71.37±0.00	90.24±0.30	65.02±2.00	OOM	91.40±0.10	**96.02±2.50**	73.41±26.09	95.55±1.10
Muris Kidney	65.29±2.34	44.10±3.12	38.16±2.76	47.46±2.60	49.42±7.54	42.30±5.20	55.52±3.40	47.47±2.30	56.94±5.20	41.97±3.10	46.48±1.50	36.38±2.80	46.32±10.60	**80.65±1.60**
Muris Liver	48.86±2.98	55.86±2.45	45.72±3.21	46.51±5.10	48.90±0.00	42.62±3.20	53.48±0.40	49.72±4.10	45.39±4.30	44.50±3.40	51.58±2.90	55.76±7.50	41.04±8.70	**68.13±1.40**
Muris Lung	42.54±3.12	40.35±2.87	50.45±2.65	50.54±3.80	53.26±0.00	37.98±2.40	51.06±2.20	38.15±1.80	50.10±1.80	37.73±4.30	40.98±4.80	36.85±4.20	64.54±26.60	**65.68±1.70**
AVG	70.34±3.56	64.49±4.21	65.06±2.87	49.48±3.83	57.15±3.28	60.64±3.07	78.24±2.60	74.88±3.69	64.78±4.51	69.96±4.47	53.75±1.77	57.71±3.65	74.08±7.48	81.29±1.45
Model Rank AVG	5	9	7	14	12	10	2	3	8	6	13	11	4	1

Table 2: ACC scores (mean ± std) across 36 datasets; the best score is shown in **bold**, and the second-best is underlined.

Dataset	Clustering Performance						Classification Performance					
	scGPT		GeneFormer		GeneCompass		scGPT		GeneFormer		GeneCompass	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Sapiens Ear Crista Ampullaris	52.28±2.37	45.78±2.70	20.57±0.47	13.82±0.13	41.24±1.91	27.28±1.56	98.14±1.06	95.77±2.84	93.39±0.98	84.64±4.30	94.92±1.04	88.50±2.46
Sapiens Ear Utricle	51.33±0.87	45.34±3.77	31.29±2.00	26.43±2.53	36.56±0.55	26.53±1.71	97.10±3.10	96.72±2.72	82.90±0.88	63.49±5.88	98.06±1.77	96.26±3.36
Sapiens Liver	43.47±4.37	35.17±2.47	24.49±1.10	22.14±2.31	27.77±1.46	20.88±0.75	88.43±3.09	72.72±6.43	80.00±1.44	54.56±4.32	87.78±1.21	70.96±1.40
Sapiens Lung	41.39±1.85	37.56±2.12	12.54±0.43	10.35±0.41	24.75±1.21	13.88±0.65	93.38±1.91	84.39±3.70	83.52±0.37	63.74±3.88	87.44±1.61	76.60±2.12
Sapiens Testis	50.99±3.03	44.32±6.01	18.74±0.64	8.76±0.20	43.45±7.35	17.55±1.41	97.52±1.22	92.11±6.84	96.51±0.58	75.90±4.78	96.90±0.76	84.58±9.14
Sapiens Trachea	39.14±0.73	32.85±2.23	8.66±0.31	4.98±0.07	18.46±2.23	9.76±0.97	97.96±0.39	84.95±1.78	96.37±0.00	77.39±0.56	97.92±0.00	88.93±0.00
Muris Brain	59.54±0.64	39.49±0.28	62.71±0.11	40.64±0.04	54.82±0.02	37.25±0.01	99.84±0.10	98.01±1.58	99.67±0.14	95.88±1.49	100.00±0.00	100.00±0.00
Muris Kidney	61.92±5.31	51.54±6.24	29.87±1.37	23.79±1.15	18.70±0.04	13.32±0.65	96.59±1.52	96.58±2.42	77.25±3.86	74.21±3.78	93.85±3.83	93.04±3.44
Muris Limb Muscle	29.22±0.28	21.88±1.42	23.25±1.29	17.22±0.43	24.47±0.05	19.76±0.09	97.05±1.25	94.89±1.94	90.41±1.00	80.38±2.20	96.63±0.76	94.53±1.30
Muris Liver	32.44±2.46	20.99±1.30	13.84±0.50	9.26±0.29	28.80±2.73	19.42±1.38	95.52±1.22	89.87±3.57	86.71±1.50	59.18±4.27	97.55±0.00	94.91±0.00
Muris Lung	14.28±0.48	13.29±0.71	8.34±0.22	5.95±0.13	12.87±0.46	10.66±0.64	94.82±1.36	89.64±2.92	80.23±4.16	54.68±8.04	94.58±0.00	84.53±0.00

Table 3: Clustering and classification performance(means ± std over 5 runs) of biological foundation models.

of the GNN models that adjacent nodes are highly similar. To investigate the extent of the over-smoothing problem of each method, we calculate pair-wise cosine similarities for all sample embeddings in each dataset and derive their probability distribution. A representative digest of embedding similarity distribution is illustrated in Fig. 4, where red bars indicate the probabilities of embedding pairs with high similarity (up to 1), and blue bars indicate the probabilities of embedding pairs with low similarity (down to 0). We can discover that deep learning methods are free from over-smoothing and representation collapse problems, with a substantial portion of embeddings considerably dissimilar and easily distinguished, as suggested by the blue and white bars. Graph-based methods, namely, AttentionAE-sc, scGAE, scDSC, and scGNN, on the other hand, suffer from severe representation collapse, where almost all embeddings are highly similar and indistinguishable. The only exception is scCDCG, which, thanks to its cut-informed graph embedding mechanism, outperforms other graph-based methods and even some deep learning approaches in embedding distinguishability.

Biological Analysis

Case Study. Taking the results of the scCDCG model on *Mauro Pancreas* as an example (Fig. 5), we performed an extensive biological analysis, including marker gene identification, cell type annotation tasks, and annotation comparison.

Marker Gene Identification. To evaluate marker gene expression across different clusters, we employed tracksplot to visualize the expression patterns of the top three marker genes predicted by the scCDCG model (Fig. 5a). For instance, GCG, TTR, and GC were identified as the marker genes for cluster 1, whereas SCG2, TMEM76B, and TMEM76A were assigned to cluster 8. Tracksplot provides a clear visualization of gene expression across cell clusters, enabling the evalu-

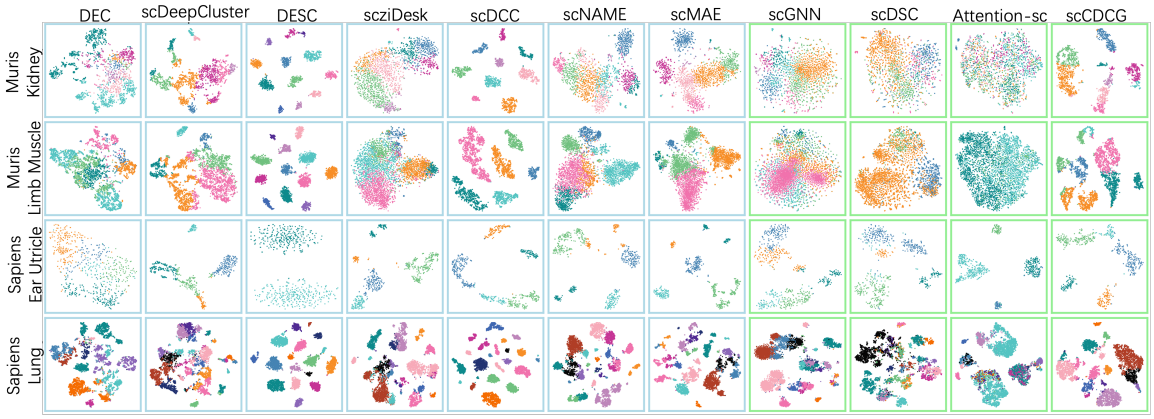


Figure 3: A digest of the visualization of all baselines.

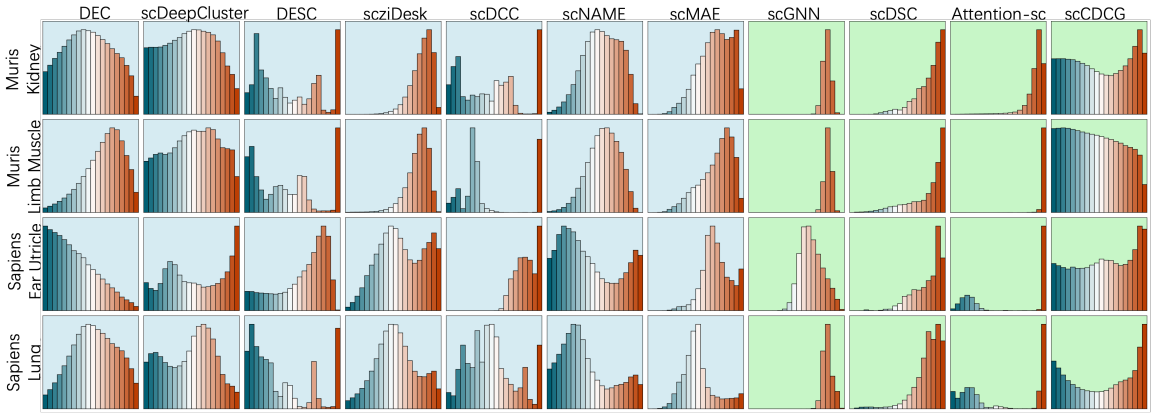


Figure 4: A digest of representation similarity of all baselines.

ation and refinement of cluster label assignments. Notably, clusters 1 and 8 exhibited similar expression profiles among their highly expressed genes, suggesting that these clusters may represent subtypes within a broader cell category, rather than entirely distinct cell populations. The top 3 marker genes may be insufficient to distinguish these two clusters.

Cell Type Annotation. To evaluate the biological interpretability of the clustering results, we performed cell type annotation on scCDCG-predicted clusters using marker-overlap annotation method (detailed in Evaluation Section). As shown in Fig. 5b, clusters 3 and 1 were annotated as "type B pancreatic cell" and "pancreatic A cell", respectively. Notably, "endothelial cell" and "pancreatic epsilon cell" were severely underrepresented (3 and 21 samples, respectively) in the reference dataset. Indeed, the scarcity of reference samples increases the annotation difficulty for rare cell types. Nevertheless, scCDCG successfully annotated the remaining seven major cell types with high accuracy.

Annotation Comparison. Further, we employed the best-mapping annotation method to assign cell types to the clusters predicted by scCDCG. Fig. 5c illustrates the correspondence among the best-mapping annotation, marker-overlap annotation, and the gold standard cell types, highlighting the differences between the two annotation strategies and quantifying their deviations from the ground truth. While the best-mapping method preserves the number of clusters consistent with reference labels, it often introduces ambiguous assignments (e.g., "pancreatic A cell" simultaneously annotated as both "pancreatic A cell" and "pancreatic epsilon cell", or erroneous merging of distinct types). In contrast, the marker-overlap annotation effectively rectifies these errors, and its corrections to the best-mapping annotations are also reflected in the figure. Nevertheless, this method also faces limitations with extremely small populations (e.g., 3 "endothelial cells", 21 "pancreatic epsilon cells"), where insufficient marker gene expression impedes accurate annotation. The results indicate that, compared to the best-mapping method, the marker-overlap annotation approach provides greater biological interpretability by aligning more closely with gene expression patterns and established biological knowledge.

Result Correction. The results of the two annotation methods were compared against the gold standard cell types, and ACC was calculated to reflect the model's performance in cell type identification more accurately. During the evaluation, the ACC obtained by the best-mapping annotation was consistent with the average values reported in Tab. 2 across five experi-

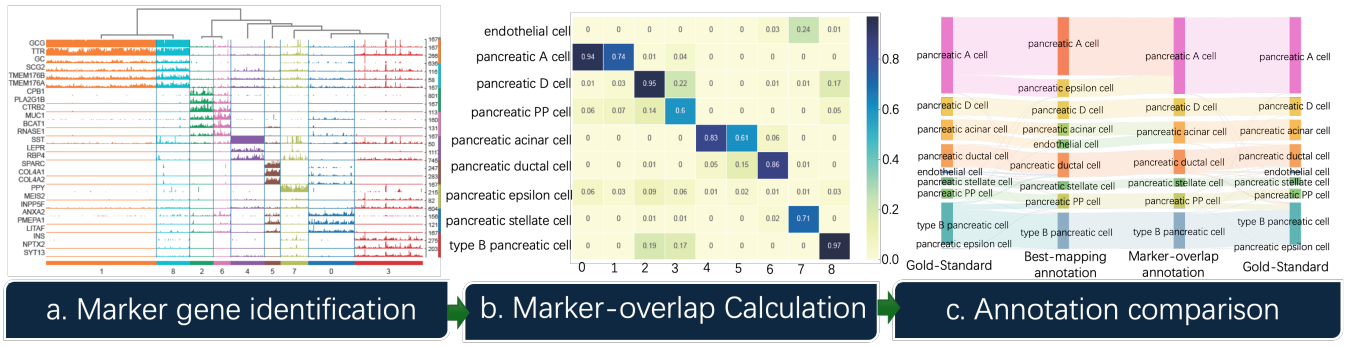


Figure 5: All biological analysis of scCDCG on *Mauro Human Pancreas cells*.

Dataset	DEC		DESC		scDeepCluster		scNAME		scMAE		scDCC		scziDesk		scGNN		scDSC		AttentionAE-sc		scCDCG	
	BM	MO	BM	MO	BM	MO	BM	MO	BM	MO	BM	MO	BM	MO	BM	MO	BM	MO	BM	MO	BM	MO
Mauro Pancreas	65.27	83.74	78.98	95.29	75.87	94.63	22.15	21.21	95.52	95.52	74.65	85.34	85.30	90.29	68.24	77.66	50.71	54.85	81.90	87.23	80.44	93.50
Sonya Liver	50.83	47.23	60.45	88.70	55.77	89.38	21.57	15.60	86.96	95.67	64.93	84.18	65.73	79.41	42.25	63.23	32.77	38.11	72.06	77.57	48.95	73.61
Sapiens Ear Utricle	65.96	65.96	60.23	60.23	53.19	62.36	71.03	88.22	73.81	91.49	70.54	86.74	74.30	90.18	69.56	84.29	90.18	90.18	73.32	93.94	74.30	90.02
Sapiens Liver	62.59	66.45	42.24	64.31	51.72	73.05	61.52	68.40	68.49	70.77	53.35	68.12	67.61	71.24	66.64	65.94	66.08	66.91	85.73	80.34	50.42	67.29
Sapiens Lung	62.54	67.24	55.50	70.32	45.39	54.29	62.34	76.19	61.67	83.71	55.60	76.00	71.26	79.62	68.39	82.40	62.79	72.57	64.84	64.84	63.92	67.17
Sapiens Testis	43.98	77.92	20.8	75.1	38.40	65.44	54.00	79.92	54.84	80.88	44.52	80.20	80.29	79.86	79.81	84.89	66.93	84.39	99.29	100.00	71.67	80.08
Sapiens Trachea	54.59	78.23	50.86	83.18	38.99	65.57	68.78	90.75	64.08	89.89	58.87	85.3	64.27	85.84	56.67	84.04	80.17	80.22	90.41	90.47	51.15	72.96
Muris Brain	50.73	50.73	86.15	80.76	58.96	58.96	84.65	97.86	71.37	97.86	68.82	68.82	-	-	91.54	91.54	92.70	97.86	100.00	100.00	94.61	94.61
Muris Kindey	45.35	38.86	51.79	70.89	36.10	36.43	47.66	51.73	54.38	51.51	48.43	61.36	43.64	43.81	41.11	44.63	41.66	40.51	94.88	94.88	57.62	56.58
Muris Limb Muscle	49.29	62.54	39.22	92.53	55.56	66.02	61.82	84.07	71.18	71.85	67.70	81.32	51.34	76.42	43.66	60.10	67.60	67.68	29.11	76.65	60.03	82.98
Muris Liver	42.11	66.07	50.65	87.58	48.79	71.70	38.75	41.48	52.75	80.59	42.85	74.40	47.71	60.16	56.17	71.49	56.10	68.24	97.48	97.48	53.13	79.61
Muris Lung	52.78	75.58	51.36	79.99	37.47	66.00	47.42	42.54	48.29	73.74	48.13	77.49	40.97	0.54	34.88	54.13	31.82	38.11	50.07	50.07	39.09	55.24
ACC Mean	54.23	65.31	52.09	80.19	50.21	67.76	54.62	65.35	66.94	82.70	60.15	78.5	62.26	72.85	61.78	73.39	63.29	68.53	79.20	85.65	63.43	76.68
Mean gain	11.08		28.1		17.55		10.73		15.76		18.35		10.59		11.61		5.24		6.45		13.25	

Table 4: Accuracy correction performance (BM: Best-mapping; MO: Marker-overlap).

ments, with differences falling within the expected statistical variance. Notably, the marker-overlap annotation corrected specific misclassifications, yielding performance gains across all methods. The revised results are summarized in Tab. 4. Overall, across most datasets, the marker-overlap annotation achieved higher ACC than the best-mapping annotation, indicating that incorporating biological prior knowledge not only improves clustering performance evaluation accuracy but also enhances the biological interpretability of the results.

Related Work

Clustering methods for scRNA-seq have evolved from traditional models grounded in low-dimensional distance metrics to techniques leveraging deep learning and graph-based modeling, and most recently, to biological foundation models built upon Transformer architectures. Early approaches such as SC3, Louvain, and Leiden are limited by low-dimensional assumptions, restricting their ability to capture complex cellular heterogeneity. Deep learning-based methods (e.g., scMAE, scDeepCluster) enhance robustness against data sparsity and noise through unsupervised feature reconstruction, yet often suffer from instability and limited interpretability. Graph-based approaches, exemplified by scSiameseClu (Xu et al. 2025a) and scSGC (Xu et al. 2025b), further improve clustering accuracy by incorporating intercellular relationships, though they remain sensitive to graph construction strategies. Recently, biological foundation models like scGPT and GeneCompass have achieved broad generalization through large-scale pretraining, yet their clustering performance remains limited by non-specific task design.

Although several studies have benchmarked scRNA-seq clustering methods across aspects such as parameter sensitivity, cell number estimation, batch effect correction, and spatial transcriptomics, most evaluations remain limited to specific methodological categories or assessment dimensions (Yuan et al. 2024; Dai et al. 2022; Yu et al. 2022; Tran et al. 2020). For instance, (Krzak et al. 2019) provided an early systematic evaluation of scRNA-seq clustering, but focused solely on R-based algorithms. To date, a comprehensive benchmarking framework spanning the full spectrum of clustering approaches, from traditional models to biological foundation models, has yet to be established. A unified platform integrating diverse clustering algorithms and evaluation metrics is essential for systematic benchmarking and further methodological advancement in single-cell analysis.

Conclusion

We present **scCluBench**, a comprehensive and standardized benchmarking framework for scRNA-seq clustering that integrates diverse datasets, algorithmic paradigms, and multi-dimensional evaluation protocols. The scCluBench systematically compares traditional, deep learning, graph-based, and foundation model, offering detailed insights into their performance trade-offs and applicability boundaries across diverse clustering scenarios, thereby informing future method development and practical tool selection. Looking ahead, scCluBench will be expanded to include larger-scale datasets, integrate multi-modal single-cell data, and refine benchmarking protocols to address emerging biological challenges.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grant No. 92470204 and 62406306), the National Key Research and Development Program of China (Grant No. 2024YFF0729201).

References

Chen, L.; Wang, W.; Zhai, Y.; and Deng, M. 2020. Deep soft K-means clustering with self-training for single-cell RNA sequence data. *NAR genomics and bioinformatics*, 2(2): lqaa039.

Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; and Wang, B. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8): 1470–1480.

Dai, C.; Jiang, Y.; Yin, C.; Su, R.; Zeng, X.; Zou, Q.; Nakai, K.; and Wei, L. 2022. scIMC: a platform for benchmarking comparison and visualization analysis of scRNA-seq data imputation methods. *Nucleic Acids Research*, 50(9): 4877–4899.

Fang, Z.; Zheng, R.; and Li, M. 2024. scMAE: a masked autoencoder for single-cell RNA-seq clustering. *Bioinformatics*, 40(1): btae020.

Gan, Y.; Huang, X.; Zou, G.; Zhou, S.; and Guan, J. 2022. Deep structural clustering for single-cell RNA-seq data jointly through autoencoder and graph neural network. *Briefings in Bioinformatics*, 23(2): bbac018.

Kiselev, V. Y.; Andrews, T. S.; and Hemberg, M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5): 273–282.

Kiselev, V. Y.; Kirschner, K.; Schaub, M. T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K. N.; Reik, W.; Barahona, M.; Green, A. R.; et al. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods*, 14(5): 483–486.

Krzak, M.; Raykov, Y.; Boukouvalas, A.; Cutillo, L.; and Angelini, C. 2019. Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. *Frontiers in genetics*, 10: 1253.

Li, S.; Guo, H.; Zhang, S.; Li, Y.; and Li, M. 2023. Attention-based deep clustering method for scRNA-seq cell type identification. *PLOS Computational Biology*, 19(11): e1011641.

Li, X.; Wang, K.; Lyu, Y.; Pan, H.; Zhang, J.; Stambolian, D.; Susztak, K.; Reilly, M. P.; Hu, G.; and Li, M. 2020. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature communications*, 11(1): 2338.

Ning, Z.; Wang, Z.; Zhang, R.; Xu, P.; Liu, K.; Wang, P.; Ju, W.; Wang, P.; Zhou, Y.; Cambria, E.; et al. 2025. Deep cut-informed graph embedding and clustering. *arXiv preprint arXiv:2503.06635*.

Shapiro, E.; Biezuner, T.; and Linnarsson, S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9): 618–630.

Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W. M.; Hao, Y.; Stoeckius, M.; Smibert, P.; and Satija, R. 2019. Comprehensive integration of single-cell data. *cell*, 177(7): 1888–1902.

Theodoris, C. V.; Xiao, L.; Chopra, A.; Chaffin, M. D.; Al Sayed, Z. R.; Hill, M. C.; Mantineo, H.; Brydon, E. M.; Zeng, Z.; Liu, X. S.; et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965): 616–624.

Tian, T.; Wan, J.; Song, Q.; and Wei, Z. 2019. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4): 191–198.

Tian, T.; Zhang, J.; Lin, X.; Wei, Z.; and Hakonarson, H. 2021. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. *Nature communications*, 12(1): 1873.

Tran, H. T. N.; Ang, K. S.; Chevrier, M.; Zhang, X.; Lee, N. Y. S.; Goh, M.; and Chen, J. 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology*, 21: 1–32.

Wan, H.; Chen, L.; and Deng, M. 2022. scNAME: neighborhood contrastive clustering with ancillary mask estimation for scRNA-seq data. *Bioinformatics*, 38(6): 1575–1583.

Wang, J.; Ma, A.; Chang, Y.; Gong, J.; Jiang, Y.; Qi, R.; Wang, C.; Fu, H.; Ma, Q.; and Xu, D. 2021. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nature communications*, 12(1): 1882.

Wang, P.; Liu, W.; Wang, J.; Liu, Y.; Li, P.; Xu, P.; Cui, W.; Zhang, R.; Long, Q.; Hu, Z.; et al. 2025a. scCompass: An Integrated Multi-Species scRNA-seq Database for AI-Ready. *Advanced Science*, 2500870.

Wang, P.; Wu, D.; Chen, C.; Liu, K.; Fu, Y.; Huang, J.; Zhou, Y.; Zhan, J.; and Hua, X. 2024. Deep adaptive graph clustering via von Mises-Fisher distributions. *ACM Transactions on the Web*, 18(2): 1–21.

Wang, Z.; Wang, P.; Liu, K.; Wang, P.; Fu, Y.; Lu, C.-T.; Aggarwal, C. C.; Pei, J.; and Zhou, Y. 2025b. A comprehensive survey on data augmentation. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, Z.; Zhang, J.; Zhang, X.; Liu, K.; Wang, P.; and Zhou, Y. 2025c. Diversity-oriented data augmentation with large language models. *arXiv preprint arXiv:2502.11671*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Xu, P.; Ning, Z.; Li, P.; Liu, W.; Wang, P.; Cui, J.; Zhou, Y.; and Wang, P. 2025a. scsiameseclu: A siamese clustering framework for interpreting single-cell rna sequencing data. *arXiv preprint arXiv:2505.12626*.

Xu, P.; Ning, Z.; Xiao, M.; Feng, G.; Li, X.; Zhou, Y.; and Wang, P. 2024. scCDCG: Efficient Deep Structural Clustering for Single-Cell RNA-Seq via Deep Cut-Informed Graph Embedding. In *International Conference on Database Systems for Advanced Applications*, 172–187. Springer.

Xu, P.; Wang, P.; Ning, Z.; Xiao, M.; Wu, M.; and Zhou, Y. 2025b. Soft graph clustering for single-cell RNA sequencing data. *BMC bioinformatics*, 26(1): 195.

Xu, P.; Wang, Z.; Wang, Z.; Li, P.; Zhang, R.; Li, G.; Xie, H.; Wang, J.; Zhou, Y.; and Wang, P. 2025c. scUnified:

An AI-Ready Standardized Resource for Single-Cell RNA Sequencing Analysis. *arXiv preprint arXiv:2509.25884*.

Yang, X.; Liu, G.; Feng, G.; Bu, D.; Wang, P.; Jiang, J.; Chen, S.; Yang, Q.; Miao, H.; Zhang, Y.; et al. 2024. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 1–16.

Yu, L.; Cao, Y.; Yang, J. Y.; and Yang, P. 2022. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome biology*, 23(1): 49.

Yuan, Z.; Zhao, F.; Lin, S.; Zhao, Y.; Yao, J.; Cui, Y.; Zhang, X.-Y.; and Zhao, Y. 2024. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4): 712–722.