

scUnified: An AI-Ready Standardized Resource for Single-Cell RNA Sequencing Analysis

Ping Xu^{1,3}, Zaitian Wang^{1,3}, Zhirui Wang^{2,3}, Pengjiang Li^{1,3}, Ran Zhang^{1,3}, Gaoyang Li^{1,3},
Hanyu Xie⁴, Jiajia Wang^{1,3}, Yuanchun Zhou^{1,2,3}, Pengfei Wang^{1,2,3,*}

¹Computer Network Information Center, Chinese Academy of Sciences, Beijing, China

²Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China

³University of Chinese Academy of Sciences, Beijing, China

⁴Department of Computer Science, Columbia University, New York, USA

xuping0098@gmail.com, wpf2106@gmail.com

Abstract—Single-cell RNA sequencing (scRNA-seq) technology enables systematic delineation of cellular states and interactions, providing crucial insights into cellular heterogeneity. Building on this potential, numerous computational methods have been developed for tasks such as cell clustering, cell type annotation, and marker gene identification. To fully assess and compare these methods, standardized, analysis-ready datasets are essential. However, such datasets remain scarce, and variations in data formats, preprocessing workflows, and annotation strategies hinder reproducibility and complicate systematic evaluation of existing methods. To address these challenges, we present scUnified, an AI-ready standardized resource for single-cell RNA sequencing data that consolidates 13 high-quality datasets spanning two species (human and mouse) and nine tissue types. All datasets undergo standardized quality control and preprocessing and are stored in a uniform format to enable direct application in diverse computational analyses without additional data cleaning. We further demonstrate the utility of scUnified through experimental analyses of representative biological tasks, providing a reproducible foundation for the standardized evaluation of computational methods on a unified dataset.

Index Terms—AI-Ready, Dataset, scRNA-seq data, Standardized processing, Multi-Task analysis

I. INTRODUCTION

With the rapid advancement of single-cell RNA sequencing (scRNA-seq) technologies, it is now possible to characterize complex cellular populations and their functional states at unprecedented resolution [1]. Although scRNA-seq data are inherently high-dimensional, sparse, and subject to substantial technical noise, they capture rich biological information that provides a robust foundation for studying cellular heterogeneity and elucidating disease mechanisms [2]–[5]. Such data enable a wide array of computational analyses, including cell clustering, cell type annotation, trajectory inference, gene regulatory network reconstruction, and so on [6], [7].

Recent years have witnessed the development of diverse computational strategies for scRNA-seq data, spanning traditional statistical modeling, machine learning, deep learn-

ing, and foundation models informed by biological priors to cope with the complexity and noise inherent in single-cell datasets [8]–[13]. Representative methods include graph-based clustering algorithms such as Louvain and Leiden [14], probabilistic generative models like scVI [15], deep clustering frameworks such as scCDCG [5], and large-scale foundation models including scGPT [16] and GeneCompass [17].

Despite the rapid progress and methodological diversity, these computational models face significant obstacles in rigorous evaluation and comparison [18]–[21]. Specifically, the lack of standardized, high-quality datasets limits reproducibility and hinders fair benchmarking. Three main challenges can be identified. (i) Unrigorous cluster number setting: In single-cell clustering benchmarks, the number of annotated cell types is often directly used as the number of clusters. This practice is not always biologically justified and may introduce biases in performance evaluation. (ii) Inconsistent data standards leading to unfair evaluation: Significant variations exist across datasets in terms of format, preprocessing workflows, and annotation quality. These inconsistencies not only hinder model training across studies but also limit the ability to perform fair and reproducible comparisons of multiple methods on the same dataset. (iii) Limited availability of multi-task datasets: Few existing single-cell datasets can simultaneously support diverse downstream biological analyses, such as clustering, cell-type annotation, and marker gene identification. This limitation restricts the scope of systematic benchmarking and constrains the advancement of AI-driven single-cell research.

To cope with the aforementioned issues, we present **scUnified**, an **AI-ready standardized resource for single-cell RNA sequencing analysis**. scUnified integrates 13 high-quality publicly available datasets covering two species and nine tissue types, with consistent quality control, preprocessing, and multi-level annotations provided in the `.h5ad` format to ensure compatibility with widely used single-cell analysis frameworks. By providing analysis-ready data, scUnified eliminates the need for additional data cleaning or format conversion, offering a standardized and reliable resource that facilitates reproducible evaluation of computational methods across diverse models and tasks. Our principal contributions are summarized as follows:

* Corresponding author.

The resource is publicly available at the link: <https://github.com/XPgogogo/scUnity-AI>.

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

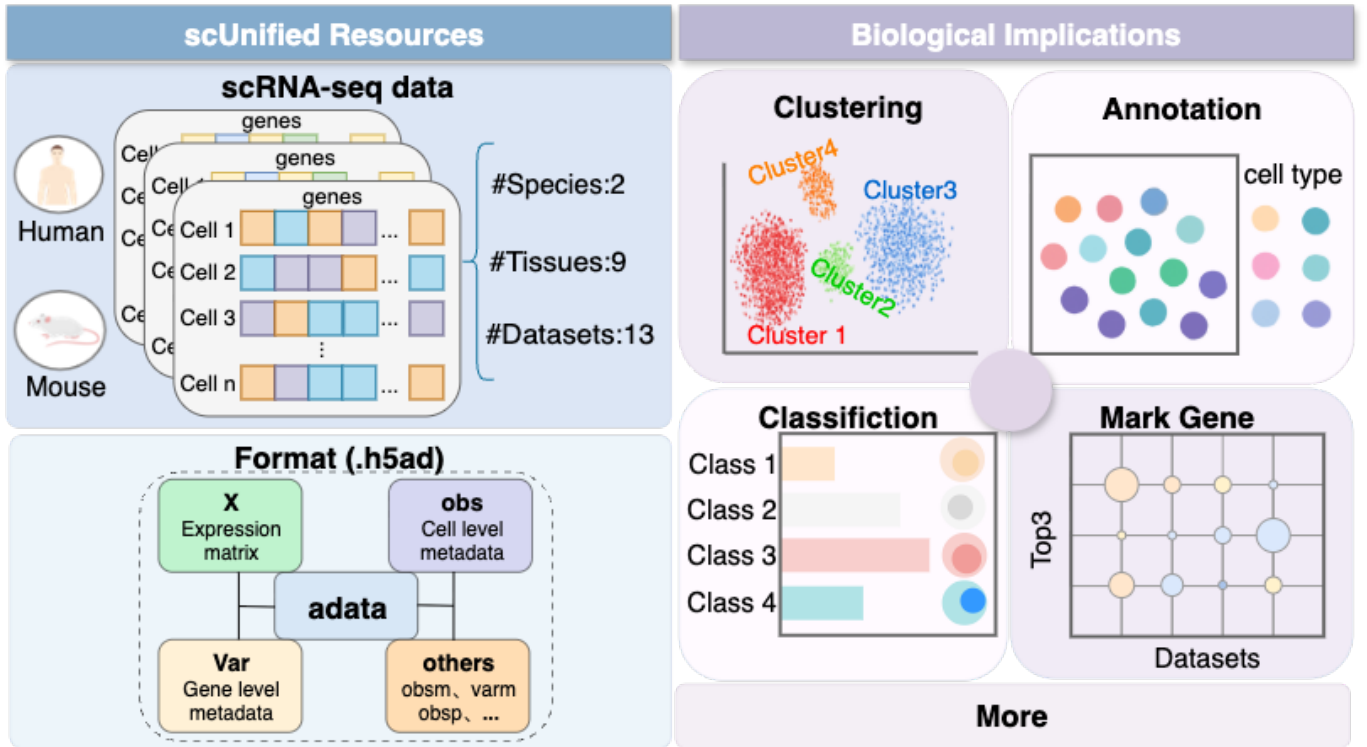


Fig. 1: Overview of scUnified: standardized single-cell RNA sequencing datasets across species and tissues, supporting AI-driven biological research and discovery.

- comprehensive collection and systematic curation of high-quality scRNA-seq datasets with uniform quality control and preprocessing;
- provision of standardized, analysis-ready data format to support a broad spectrum of biological tasks, including clustering, cell type annotation, marker gene identification, and beyond;
- validation of dataset utility through representative biological case studies, establishing a reproducible foundation for method development, fair model comparison, and AI-driven discovery in single-cell research.

II. SCUNIFIED

We present **scUnified**, an **AI-ready standardized resource for single-cell RNA sequencing analysis**, offering uniformly processed datasets that facilitate fair evaluation of diverse computational models on a consistent and reproducible dataset collection. As depicted in Fig. 1, scUnified provides a standardized workflow that integrates and curates 13 publicly available single-cell RNA sequencing datasets spanning two species and nine tissue types. All datasets undergo uniform quality control, preprocessing, and multi-level annotation, stored in the .h5ad format to ensure seamless compatibility with widely adopted single-cell analysis frameworks. This standardized resource enables direct application to a wide range of computational tasks, including core analyses such as cell clustering, cell type annotation, cell type classification, and marker gene identification, and supports fair evaluation of multiple models

on the same dataset or a single model across multiple datasets, without the need for additional curation or format conversion. By unifying diverse, high-quality scRNA-seq datasets into a consistent and accessible framework, scUnified provides a reproducible foundation for AI-driven single-cell research, reducing technical barriers, facilitating method development and training, and supporting rigorous systematic evaluation across biological contexts.

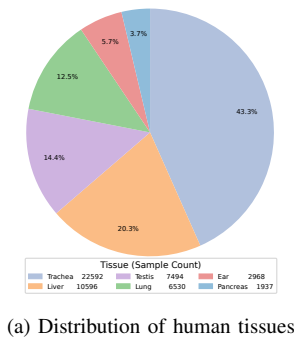
III. DATASET

A. Dataset Coverage

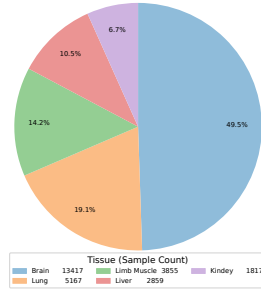
scUnified comprises a curated collection of 13 single-cell RNA-seq datasets from human and mouse, spanning 9 distinct tissue types. Comprehensive dataset characteristics, covering species, tissue origin, cell counts, gene dimensionality, annotated clusters, sparsity, and sequencing protocols, are summarized in Fig.2, 3 and Tab. I. Specifically, cell counts range from 611 to 22,592, including three large-scale datasets with more than 10,000 cells. Gene dimensionality varies from 4,999 to 61,759, with six high-dimensional datasets exceeding 60,000 genes. The number of annotated cell types per dataset ranges from 2 to 39, with four datasets containing at least 20 clusters, reflecting substantial cellular heterogeneity. Sparsity is generally high, with most datasets (12/13) exceeding 80% and overall values ranging from 73.02% to 95.42%. The datasets were generated using diverse experimental protocols, including Smart-seq2, 10X Genomics, and CEL-seq2, capturing a wide spectrum of technical platforms and biological contexts. Taken

Species	Dataset Name	#Cell	#Gene	#Cluster	Organ	Seq. Method	Sparsity (%)	Ref.
Human	Mauro Pancreas	2,122	19,046	9	Pancreas	CEL-seq2	73.02	[22]
	Sonya Liver	8,444	4,999	11	Liver	10X Genomics	90.77	[23]
	Sapiens Liver	2,152	61,759	15	Liver	Smart-seq2	95.42	[24]
	Sapiens Ear Crista Ampullaris	2,357	61,759	7	Ear	Smart-seq2	93.59	[24]
	Sapiens Ear Utricle	611	61,759	5	Ear	Smart-seq2	93.75	[24]
	Sapiens Lung	6,530	61,759	25	Lung	Smart-seq2	93.88	[24]
	Sapiens Testis	7,494	61,759	8	Testis	Smart-seq2	93.91	[24]
	Sapiens Trachea	22,592	61,759	20	Trachea	Smart-seq2	94.73	[24]
Mouse	Muris Limb Muscle	3,855	21,609	6	Limb Muscle	Smart-seq2	91.38	[24]
	Muris Brain	13,417	21,609	2	Brain	Smart-seq2	91.83	[24]
	Muris Kidney	1,817	21,609	9	Kidney	Smart-seq2	92.25	[24]
	Muris Liver	2,859	21,609	11	Liver	Smart-seq2	88.20	[24]
	Muris Lung	5,167	21,609	25	Lung	Smart-seq2	89.90	[24]

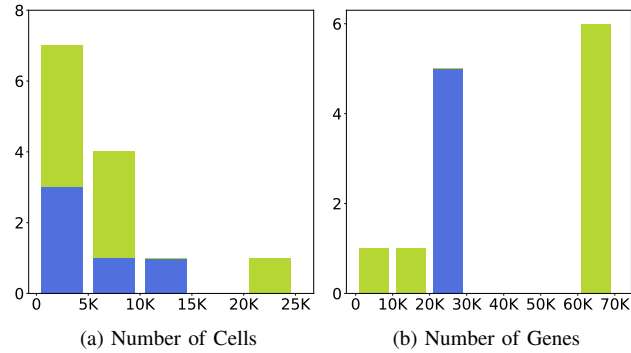
TABLE I: Details of 13 selected single-cell gene expression datasets. Large datasets (with > 20000 cells) and high-dimensional datasets (with > 60000 genes) are highlighted with bold fonts.



(a) Distribution of human tissues

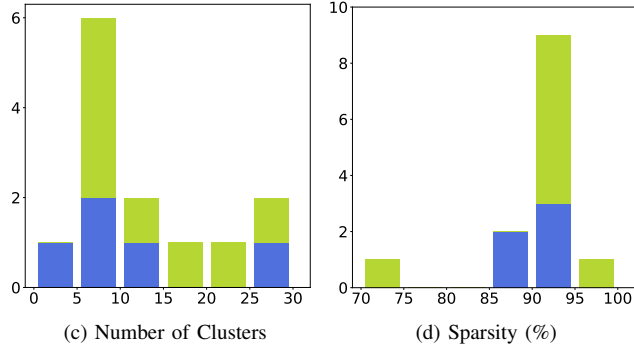


(b) Distribution of mouse tissues



(a) Number of Cells

(b) Number of Genes



(c) Number of Clusters

(d) Sparsity (%)

Fig. 2: Data distribution of sample numbers in human and mouse data

Fig. 3: Dataset distributions by cell count, gene number, clusters, and sparsity.

together, these characteristics underscore the diversity and comprehensiveness of scUnified, establishing it as a standardized, AI-ready resource for cell clustering, cell type annotation, marker gene identification, cross-dataset benchmarking, and other downstream single-cell analysis tasks.

B. Dataset Format

All scRNA-seq datasets are distributed in the .h5ad format, ensuring compatibility with widely used single-cell analysis frameworks such as Scanpy while supporting efficient data storage and manipulation. Datasets can be readily accessed

using `scanpy.read_h5ad("path/to/file.h5ad")`, which returns an AnnData object. Within this object, multiple layers of information are organized in a standardized manner, and the unified structure comprises:

- **Gene names:** stored in `data.var["feature_name"]`, representing standardized gene identifiers.
- **Cell type annotations:** recorded in `data.obs["cell_type"]`, providing ground-truth labels for each cell.
- **Expression matrix:**

stored in `adata.X` (cells-by-genes), which can be converted to a Pandas DataFrame if needed.

The standardized data structure ensures consistent retrieval of both expression profiles and cell type annotations, thereby supporting reproducible benchmarking across heterogeneous datasets. The dataset can be directly inspected via `print(adata)`, which displays the available keys across major components, including `.obs`, `.var`, `.uns`, and `.obsm`. These attributes respectively encode cell-level and gene-level metadata, raw and normalized expression matrices, as well as derived features such as PCA or UMAP embeddings. By integrating multiple layers of information within a unified framework, this representation offers transparent, versatile, and reproducible access for downstream analyses.

C. Data Preprocessing

In this study, we implemented a rigorous and standardized preprocessing pipeline to ensure consistency and comparability across single-cell RNA sequencing datasets. Raw data were loaded from `.h5ad` files into `AnnData` objects, which contain gene expression matrices along with corresponding metadata and cell annotations. The preprocessing workflow systematically examined whether datasets had undergone normalization, log-transformation ($\log_{1p}$), and scaling. For datasets lacking normalization, library size normalization was applied to mitigate variability associated with sequencing depth, followed by the computation of cell-specific size factors to standardize expression levels across cells. Untransformed data were subjected to $\log_{1p}$ transformation to reduce skewness in expression distributions. Finally, z-score scaling was applied to center and scale gene expression values, yielding features with uniform variance. This standardized pipeline ensures data quality and stability, while providing a standardized input foundation that supports reproducible and comparable downstream single-cell analyses.

Overall, `scUnified` provides a high-quality, AI-ready resource that integrates diverse large-scale datasets from multiple experimental platforms. Its standardized preprocessing and consistent annotation make it well-suited for AI model development, fair method evaluation, and systematic single-cell analyses, thereby establishing a robust foundation for advancing computational single-cell research.

IV. EXPERIMENT

A. Experiment Setup

1) *Baseline Methods*: To illustrate the versatility of `scUnified`, we assessed its ability to support multiple analytical paradigms by applying three representative clustering algorithms and a biological foundation model designed for classification. Overall, these methods span a diverse methodological spectrum, from traditional community detection to deep learning and foundation models, thereby underscoring the broad applicability of `scUnified` as a unified resource across different computational paradigms. The purpose of this evaluation is not to conduct an exhaustive performance comparison among methods, but rather to showcase that the

AI-ready datasets in `scUnified` make it possible to apply different modeling approaches in a consistent and reproducible manner. Specifically, we include the following methods.

- **Leiden**. A graph-based clustering method implemented in *Seurat* (R) that improves upon Louvain to produce stable and well-resolved partitions [14]. The number of clusters is determined automatically.
- **scMAE**. A masked autoencoder framework for scRNA-seq implemented in Python that reconstructs perturbed expression profiles to learn robust latent cell representations, enabling flexible clustering resolution through adjustable hyperparameters [25].
- **scCDCG**. A deep graph clustering framework implemented in Python that integrates graph construction, self-supervised representation learning, and autoencoder-based feature extraction to capture higher-order structural signals, offering adjustable clustering resolution through hyperparameter tuning [5].
- **GeneCompass**. A knowledge-informed cross-species foundation model implemented in Python that leverages biological priors to characterize gene regulation and cell-state transitions, supporting fine-tuning for downstream classification tasks [17].

2) *Implementation Details*: All methods were run with the parameter settings recommended in their original publications. When such settings were not specified, minimal adjustments were applied to ensure stable execution. Each method–dataset pair was evaluated over five independent runs, and the mean result was reported.

3) *Evaluation Metrics*: To comprehensively evaluate the performance of clustering and classification methods, we adopted two distinct sets of evaluation metrics. For clustering assessment, three widely adopted indices were considered: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) [26]. Specifically, ACC quantifies the proportion of correctly assigned cells by aligning predicted clusters with ground-truth labels, NMI measures the amount of shared information between predicted and true partitions, ARI evaluates assignment similarity while correcting for random chance. For classification assessment, three standard metrics were adopted: Accuracy (ACC), Precision (PRE), and Recall (REC). Here, ACC measures the overall correctness of label prediction, PRE reflects the fraction of predicted positives that are true positives, and REC indicates the proportion of true positives successfully recovered.

B. Performance

Table II presents the comparative performance of representative clustering and classification approaches evaluated on the 13 standardized datasets in `scUnified`.

For clustering, Leiden demonstrates robust and consistent performance on relatively small or less complex datasets, e.g., *Mauro Pancreas*. However, its accuracy declines markedly when applied to datasets with higher dimensionality or greater cellular diversity, such as *Sapiens Lung* and *Muris Kidney*. In contrast, deep learning-based methods, `scMAE` and `scCDCG`,

Dataset	Metrics	Clustering			Metrics	Classification	
		Leiden	scMAE	scCDCG		Genecompass	
Mauro Pancreas	ACC	92.08±0.13	95.62±0.16	92.65±2.93	ACC	98.35±0.17	
	NMI	89.96±0.12	88.49±0.44	86.81±0.98	PRE	97.29±0.27	
	ARI	93.57±0.12	92.38±0.35	91.37±1.21	REC	98.26±0.13	
Sonya Liver	ACC	69.84±5.17	80.73±1.86	75.34±3.67	ACC	98.58±0.14	
	NMI	70.70±0.05	85.59±1.44	71.34±2.59	PRE	98.39±0.31	
	ARI	54.86±0.53	88.92±1.78	81.26±2.69	REC	97.73±0.50	
Sapiens Liver	ACC	71.19±0.00	67.51±2.30	73.09±1.40	ACC	87.78±1.21	
	NMI	70.15±0.00	78.25±0.60	62.54±3.20	PRE	71.68±1.17	
	ARI	48.05±0.00	63.82±4.10	41.11±4.40	REC	72.52±3.03	
Sapiens Ear Crista Ampullaris	ACC	43.37±0.94	67.00±0.40	85.45±4.30	ACC	94.92±1.04	
	NMI	73.42±0.63	74.31±0.20	69.54±2.90	PRE	93.42±3.84	
	ARI	59.38±0.31	57.35±0.10	66.16±4.80	REC	85.34±2.26	
Sapiens Ear Utricle	ACC	51.21±0.00	73.16±0.60	79.58±0.70	ACC	98.06±1.77	
	NMI	71.20±0.00	78.28±1.20	66.82±5.40	PRE	98.46±2.15	
	ARI	62.76±0.00	64.09±0.70	60.16±3.30	REC	94.80±4.18	
Sapiens Lung	ACC	48.38±0.11	63.24±1.10	62.06±1.60	ACC	87.44±1.61	
	NMI	70.16±0.17	79.65±0.60	66.94±1.90	PRE	77.78±2.91	
	ARI	44.19±0.37	58.40±1.60	60.15±1.50	REC	77.14±2.41	
Sapiens Testis	ACC	62.86±0.00	53.71±0.80	67.18±3.80	ACC	97.33±0.38	
	NMI	44.26±0.00	57.09±0.40	57.42±3.30	PRE	94.02±2.08	
	ARI	51.52±0.00	43.27±0.50	55.38±7.20	REC	88.03±0.84	
Sapiens Trachea	ACC	48.49±0.95	65.78±3.70	52.46±2.90	ACC	98.21±0.09	
	NMI	63.81±1.62	77.12±1.50	63.25±1.00	PRE	91.17±1.32	
	ARI	41.80±2.34	53.69±4.80	42.92±2.40	REC	91.06±0.65	
Muris Limb Muscle	ACC	96.72±1.12	66.13±3.40	94.50±7.10	ACC	96.63±0.76	
	NMI	0.24±0.11	59.44±3.80	56.54±7.60	PRE	94.66±1.26	
	ARI	8.54±0.66	51.54±3.30	53.37±8.50	REC	94.73±1.35	
Muris Brain	ACC	40.84±0.01	71.37±0.00	95.55±1.10	ACC	100.00±0.00	
	NMI	26.52±0.01	1.33±0.00	22.48±8.30	PRE	100.00±0.00	
	ARI	1.46±0.01	2.22±0.00	35.56±7.80	REC	100.00±0.00	
Muris Kidney	ACC	38.16±1.22	55.52±3.40	80.65±1.60	ACC	93.85±3.83	
	NMI	21.38±0.58	54.37±1.90	55.82±1.30	PRE	93.96±3.67	
	ARI	18.43±1.44	35.79±1.40	42.88±2.10	REC	92.61±3.53	
Muris Liver	ACC	45.72±0.14	53.48±0.40	68.13±1.40	ACC	94.76±0.49	
	NMI	50.59±0.10	65.39±1.00	62.06±2.40	PRE	86.47±0.12	
	ARI	38.63±0.09	47.55±0.60	46.96±3.70	REC	86.58±1.88	
Muris Lung	ACC	50.45±2.96	51.06±2.20	65.68±1.70	ACC	93.62±1.37	
	NMI	64.32±0.46	64.49±0.90	49.53±3.80	PRE	85.46±3.91	
	ARI	31.22±5.55	35.69±2.40	26.46±3.80	REC	83.61±2.00	

TABLE II: Performance comparison of clustering and classification methods across datasets.

exhibit stronger adaptability under these challenging conditions, achieving superior ARI and ACC scores. This suggests that models based on representation learning are better suited to capture subtle cellular heterogeneity and complex non-linear structures in single-cell data. For classification, GeneCompass attains consistently high accuracy across all datasets, often exceeding 95% and reaching 100% on *Muris Brain*. The high precision and recall further confirm the robustness of GeneCompass, highlighting the effectiveness of foundation models in transferring knowledge from well-annotated references to diverse tissues and species.

Overall, these results highlight the unique value of scUnified as a comprehensive and standardized single-cell resource. This unified resource ensures that methodological advances can be assessed systematically and applied broadly, maximizing the impact and comparability of single-cell analyses.

C. Case Study

To further illustrate the versatility and practical utility of scUnified, we present several representative case studies that demonstrate its support for diverse analytical tasks. Using the *Muris Limb Muscle* dataset as a primary example, we applied

the scCDCG model to learn cell representations, revealing clear separation of cell populations in both cosine similarity heatmaps and two-dimensional projections (Fig. 4a-b). By integrating prior biological knowledge, we identified highly expressed marker genes informative for cell-type annotation. As shown in Fig. 4c-e, the resulting cell-type assignments exhibit strong concordance with known labels, demonstrated through heatmaps, dot plots of top differentially expressed genes, and Sankey diagrams.

1) *Cell Representations and Cluster Visualization*: Accurate clustering depends on the ability to learn discriminative cell representations. To evaluate this, we first measured the similarity structure of the learned embeddings and visualized it as a heatmap, which highlights strong intra-cluster coherence. We then employed t-SNE to project the embeddings into a two-dimensional space, offering an intuitive representation of the underlying feature distribution and cluster organization. As illustrated in Fig. 4b and 4a, clusters are clearly separated, and cells within each cluster display strong internal consistency. These results demonstrate that scCDCG effectively preserves transcriptional heterogeneity across cells and yields highly discriminative representations for downstream analysis.

2) *Marker Gene Identification*: To examine whether the datasets provided by scUnified support downstream biological interpretation, we performed differential expression gene (DEG) analysis for each cluster using the `rank_genes_groups` function in Scanpy with default parameters. For each cluster, the top 100 genes with pronounced cluster-specific expression were selected as candidate marker genes for subsequent cell-type annotation. As shown in Fig. 4c, the top three representative marker genes of each cluster clearly delineate the expression signatures of distinct cell populations. For example, *Chodl*, *Des*, and *Cd82* were identified as marker genes for cluster 1, whereas *Gpx3*, *Cd63*, and *Sdc4* were identified for cluster 3.

3) *Cell Type Annotation*: Building upon the identified marker genes, we next evaluated the biological interpretability of the predicted clusters through a marker-overlap annotation strategy. Specifically, we first performed differential expression gene (DEG) analysis on the reference clusters provided in the scUnified dataset, and retained the top 100 cluster-specific genes to construct a *gold-standard* marker set. For each cluster predicted by scCDCG, we quantified its concordance with reference clusters by computing an overlap score as $\text{overlap}(p, g) = \frac{|\text{DEG}_p \cap \text{DEG}_g|}{100}$, where p and g denote the predicted and reference clusters, respectively. Each predicted cluster was then assigned to the reference cluster with which it shared the highest overlap score, thereby determining its putative cell-type identity.

As illustrated in Fig. 4d, this approach successfully annotated scCDCG-predicted clusters 2 and 4 as ‘endothelial cell’ and ‘mesenchymal stem cell’, respectively, highlighting the biological coherence of the learned representations and the reliability of scUnified for cell-type annotation tasks.

4) *Annotation Comparison*: Cell type annotation is a central step in scRNA-seq data analysis, and different strategies

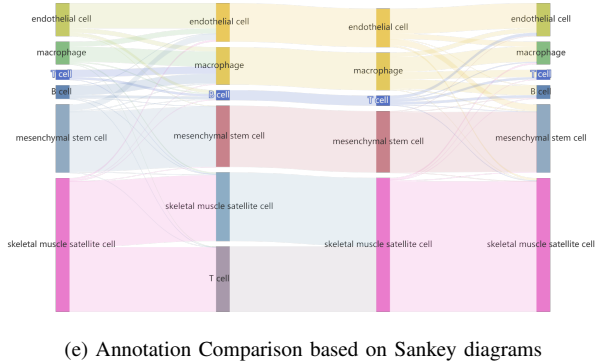
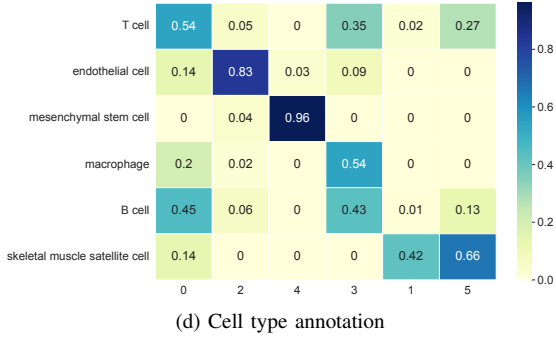
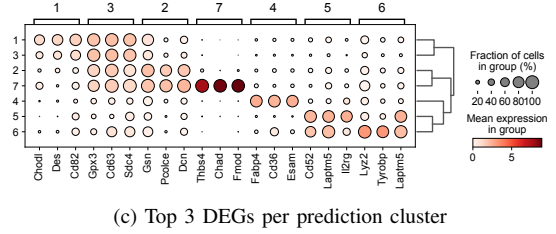
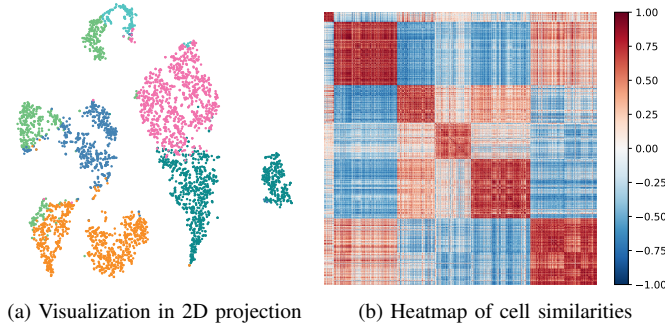


Fig. 4: Case study of scCDCG on *Muris Limb Muscle*, demonstrating integrated representation learning, two-dimensional visualization, and biologically guided cell-type annotation. (e) presents four columns from left to right: Gold-standard labels, results of the Best-mapping annotation, results of the Marker-overlap annotation, and the Gold-standard labels.

may yield varying levels of biological interpretability. In addition to the *marker-overlap annotation* strategy introduced above, we implemented an alternative approach, termed *best-mapping annotation*. This method applies the Hungarian algorithm to establish an optimal one-to-one correspondence be-

tween predicted and reference clusters, thereby achieving rapid label alignment independent of gene expression information.

To systematically compare these strategies and assess their deviation from the gold-standard annotations, we visualized the results using *Sankey diagrams*. The direction and width of the flows effectively capture the mapping patterns and degrees of divergence across clusters. As illustrated in Fig. 4e, this comparative analysis demonstrates that biologically informed strategies such as marker-overlap annotation yield more coherent and interpretable results than purely alignment-based approaches, highlighting the importance of integrating biological knowledge into cell-type annotation.

5) *Validation Across Datasets*: To further validate the quality and utility of the datasets provided by scUnified, we conducted additional analyses on the *Sapiens Ear Utricle* and *Muris Limb Muscle* datasets. For each dataset, representative methods including Leiden, scMAE, and scCDCG were systematically applied.

The results, summarized in Fig. 5 for the *Muris Limb Muscle* dataset and Fig. 6 for the *Sapiens Ear Utricle* dataset, consistently demonstrate that the standardized and high-quality data in scUnified enable reliable, reproducible analyses and support diverse analytical paradigms. This unified and versatile resource thus provides a solid foundation for a wide range of downstream single-cell tasks, from representation learning to cell type annotation, facilitating both methodological evaluation and biologically driven discovery.

V. CONCLUSION

We present **scUnified**, an AI-ready standardized resource for single-cell RNA sequencing analysis that consolidates diverse biological datasets into a unified, analysis-ready framework. By providing uniform preprocessing, standardized formatting, and multi-level annotations, scUnified addresses key challenges in reproducibility, comparability, and cross-dataset evaluation. Importantly, scUnified offers high-quality, standardized datasets that support a broad range of downstream analyses, including but not limited to clustering, classification, marker gene identification, and cell-type annotation, serving as a foundational resource to facilitate AI-driven model development and computational analyses in single-cell biology. Looking forward, we aim to expand scUnified to include more species, additional tissue types, and complementary omics data, further strengthening its utility for single-cell research and AI-driven model development.

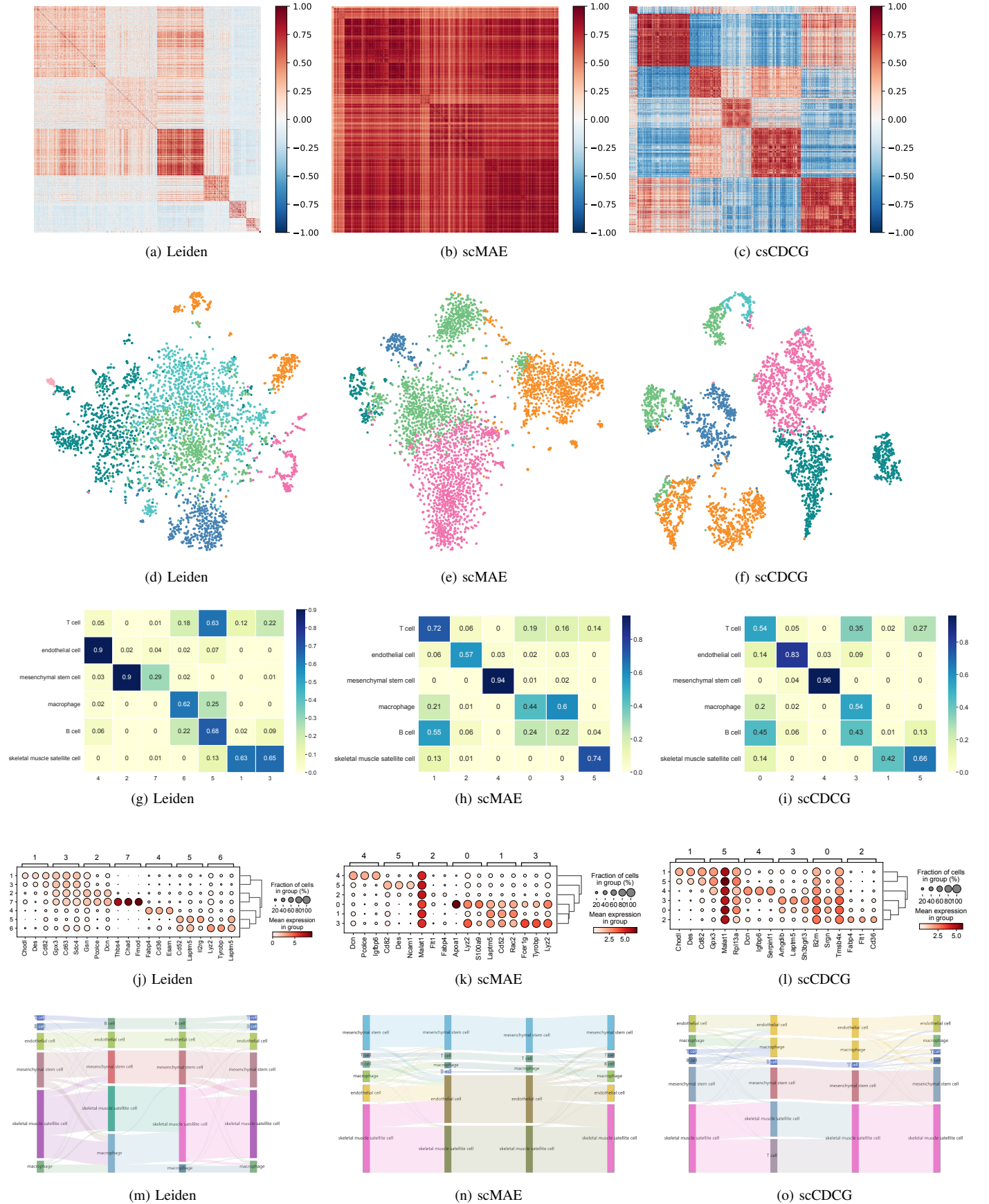


Fig. 5: Case study on the *Muris Limb Muscle*, summarizing the results of Leiden, scMAE, and scCDCG models, including representation learning, two-dimensional visualization, and marker gene-based cell type annotation. (m)-(o) presents four columns from left to right: Gold-standard labels, results of the Best-mapping annotation, results of the Marker-overlap annotation, and the Gold-standard labels.

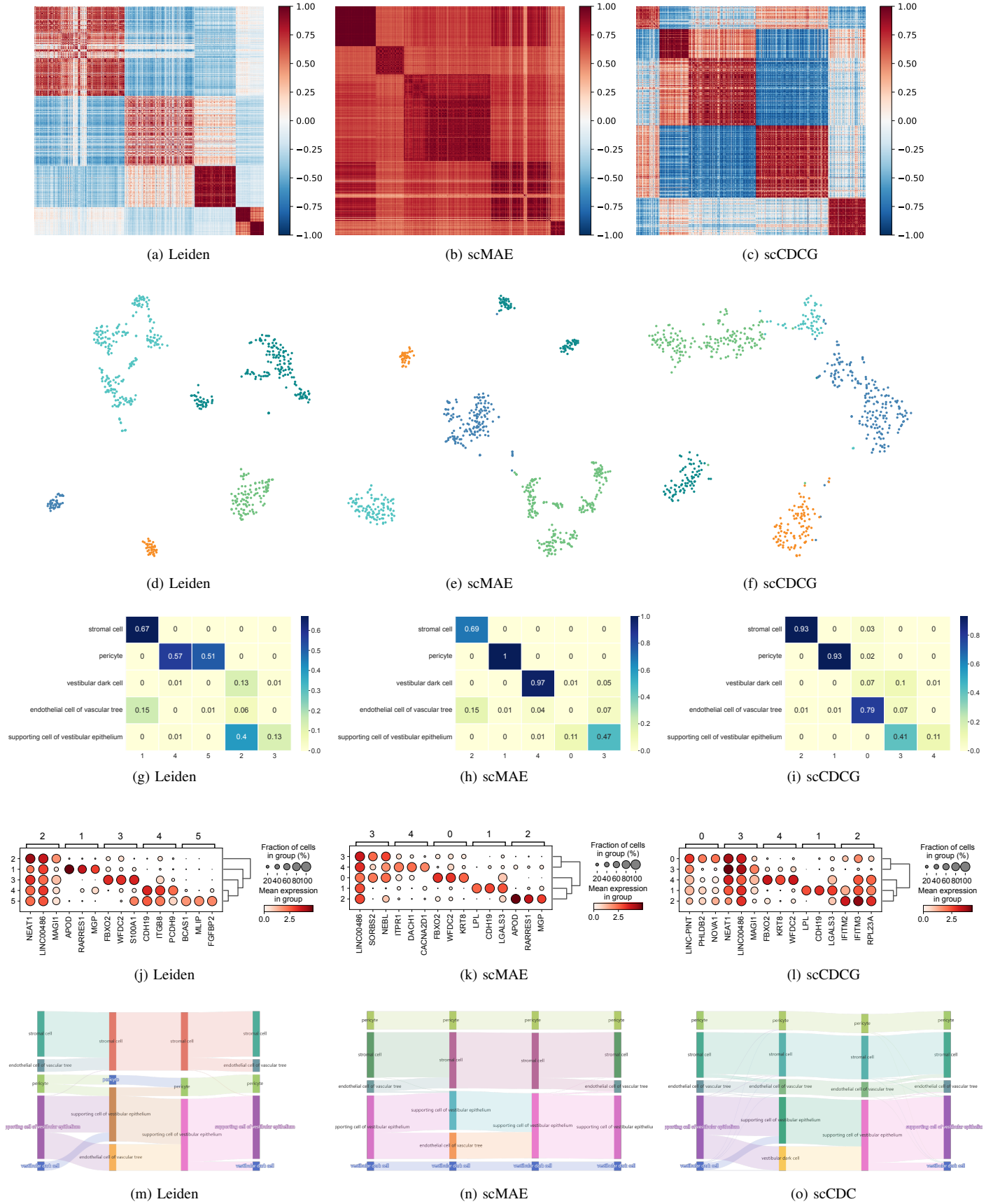


Fig. 6: Case study on the *Sapiens Ear Utricle*, summarizing the results of Leiden, scMAE, and scCDCG models, including representation learning, two-dimensional visualization, and marker gene-based cell type annotation.(m)-(o) presents four columns from left to right: Gold-standard labels, results of the Best-mapping annotation, results of the Marker-overlap annotation, and the Gold-standard labels.

REFERENCES

- [1] E. Shapiro, T. Biezuner, and S. Linnarsson, "Single-cell sequencing-based technologies will revolutionize whole-organism science," *Nature Reviews Genetics*, vol. 14, no. 9, pp. 618–630, 2013.
- [2] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell rna-seq data," *Nature Reviews Genetics*, vol. 20, no. 5, pp. 273–282, 2019.
- [3] R. Petegrosso, Z. Li, and R. Kuang, "Machine learning and statistical methods for clustering single-cell rna-sequencing data," *Briefings in bioinformatics*, vol. 21, no. 4, pp. 1209–1223, 2020.
- [4] V. Menon, "Clustering single cells: a review of approaches on high-and low-depth single-cell rna-seq data," *Briefings in functional genomics*, vol. 17, no. 4, pp. 240–245, 2018.
- [5] P. Xu, Z. Ning, M. Xiao, G. Feng, X. Li, Y. Zhou, and P. Wang, "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*. Springer, 2024, pp. 172–187.
- [6] P. Wang, W. Liu, J. Wang, Y. Liu, P. Li, P. Xu, W. Cui, R. Zhang, Q. Long, Z. Hu *et al.*, "sccompass: An integrated multi-species scrna-seq database for ai-ready," *Advanced Science*, p. 2500870, 2025.
- [7] E. Mereu, A. Lafzi, C. Moutinho, C. Ziegenhain, D. J. McCarthy, A. Álvarez-Varela, E. Battle, N. Sagar, D. Gruen, J. K. Lau *et al.*, "Benchmarking single-cell rna-sequencing protocols for cell atlas projects," *Nature biotechnology*, vol. 38, no. 6, pp. 747–755, 2020.
- [8] S. Zhang, X. Li, J. Lin, Q. Lin, and K.-C. Wong, "Review of single-cell rna-seq data clustering for cell-type identification and characterization," *Rna*, vol. 29, no. 5, pp. 517–530, 2023.
- [9] M. Krzak, Y. Raykov, A. Boukouvalas, L. Cutillo, and C. Angelini, "Benchmark and parameter sensitivity analysis of single-cell rna sequencing clustering methods," *Frontiers in genetics*, vol. 10, p. 1253, 2019.
- [10] P. Xu, Z. Ning, P. Li, W. Liu, P. Wang, J. Cui, Y. Zhou, and P. Wang, "scsiameseclu: A siamese clustering framework for interpreting single-cell rna sequencing data," *arXiv preprint arXiv:2505.12626*, 2025.
- [11] P. Xu, P. Wang, Z. Ning, M. Xiao, M. Wu, and Y. Zhou, "Soft graph clustering for single-cell rna sequencing data," 2025.
- [12] R. Qi, A. Ma, Q. Ma, and Q. Zou, "Clustering and classification methods for single-cell rna-sequencing data," *Briefings in bioinformatics*, vol. 21, no. 4, pp. 1196–1208, 2020.
- [13] Y. Zhai, L. Chen, and M. Deng, "Realistic cell type annotation and discovery for single-cell rna-seq data," in *IJCAI*, 2023, pp. 4967–4974.
- [14] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *cell*, vol. 177, no. 7, pp. 1888–1902, 2019.
- [15] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature methods*, vol. 15, no. 12, pp. 1053–1058, 2018.
- [16] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang, "scgpt: toward building a foundation model for single-cell multi-omics using generative ai," *Nature Methods*, vol. 21, no. 8, pp. 1470–1480, 2024.
- [17] X. Yang, G. Liu, G. Feng, D. Bu, P. Wang, J. Jiang, S. Chen, Q. Yang, H. Miao, Y. Zhang *et al.*, "Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model," *Cell Research*, pp. 1–16, 2024.
- [18] T. G. Brooks, N. F. Lahens, A. Mrčela, and G. R. Grant, "Challenges and best practices in omics benchmarking," *Nature Reviews Genetics*, vol. 25, no. 5, pp. 326–339, 2024.
- [19] L. Yu, Y. Cao, J. Y. Yang, and P. Yang, "Benchmarking clustering algorithms on estimating the number of cell types from single-cell rna-sequencing data," *Genome biology*, vol. 23, no. 1, p. 49, 2022.
- [20] C. Dai, Y. Jiang, C. Yin, R. Su, X. Zeng, Q. Zou, K. Nakai, and L. Wei, "scimc: a platform for benchmarking comparison and visualization analysis of scrna-seq data imputation methods," *Nucleic Acids Research*, vol. 50, no. 9, pp. 4877–4899, 2022.
- [21] J. Wang, Q. Zou, and C. Lin, "A comparison of deep learning-based pre-processing and clustering approaches for single-cell rna sequencing data," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab345, 2022.
- [22] M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. Van Gurp, M. A. Engelse, F. Carlotti, E. J. De Koning *et al.*, "A single-cell transcriptome atlas of the human pancreas," *Cell systems*, vol. 3, no. 4, pp. 385–394, 2016.
- [23] S. A. MacParland, J. C. Liu, X.-Z. Ma, B. T. Innes, A. M. Bartczak, B. K. Gage, J. Manuel, N. Khuu, J. Echeverri, I. Linares *et al.*, "Single cell rna sequencing of human liver reveals distinct intrahepatic macrophage populations," *Nature communications*, vol. 9, no. 1, p. 4383, 2018.
- [24] "A single-cell transcriptomic atlas characterizes ageing tissues in the mouse," *Nature*, vol. 583, no. 7817, pp. 590–595, 2020.
- [25] Z. Fang, R. Zheng, and M. Li, "scmae: a masked autoencoder for single-cell rna-seq clustering," *Bioinformatics*, vol. 40, no. 1, p. btae020, 2024.
- [26] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1, 2014.