scLLM-DSC: LLM-Knowledge Enhanced Cross-Modal Deep Structural Clustering for Single-Cell RNA Sequencing

Ping Xu[1,2], **Pengjiang Li**[1,2], **Tian Du**[1,2], **Zaitian Wang**[1,2], **Jiawei Gu**[4], **Zhiyuan Ning**[5], **Ziyue Qiao**[4], **Pengfei Wang**[1,2,3] and **Yuanchun Zhou**[1,2,3*]
[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China
[4]School of Computing and Information Technology, Great Bay University, Dongguan, China
[5]School of Engineering, Westlake University, Hangzhou, China
xuping0098@gmail.com, zyc@cnic.cn

Abstract

Clustering is fundamental to scRNA-seq analysis, serving as a cornerstone for identifying cell populations and resolving tissue heterogeneity. However, existing methods focus on mining numerical statistical patterns, suffering from *semantic agnosticism* by neglecting the intrinsic biological functions encoded by genes. While Large Language Models (LLMs) offer promising semantic capabilities, their direct adaptation to cell clustering is hindered by the structural mismatch between generative pre-training objectives and discriminative downstream tasks. To bridge this gap, we propose **scLLM-DSC**, a novel **LLM**-Knowledge Enhanced Cross-Modal **D**eep **S**tructural **C**lustering framework. Diverging from data-driven paradigms, scLLM-DSC establishes a semantically-grounded representation by synergizing two views: a *Knowledge-Driven Semantic View* derived from NCBI gene priors and contextualized Cell2Sentence embeddings, and a *Structure-Aware Topological View* extracted via a graph-guided encoder. Crucially, we introduce a cross-modal contrastive alignment mechanism to enforce consistency between biological semantics and transcriptomic features within a unified latent space. Extensive benchmarks demonstrate that scLLM-DSC significantly outperforms eleven state-of-the-art baselines in clustering accuracy.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized the characterization of tissue heterogeneity by providing high-resolution cellular atlases [Svensson *et al.*, 2018]. As a pivotal analytical step, cell clustering aims to define functional identities based on expression similarity, serving as the cornerstone for downstream biological discovery [Kiselev *et al.*, 2019].

In retrospect, the evolution of clustering algorithms has been a pursuit of optimal numerical representations. From early statistical linear reductions (e.g., PCA, t-SNE [Maaten and Hinton, 2008]) to deep autoencoder-based non-linear embeddings (e.g., scDeepCluster [Tian *et al.*, 2019], scNAME [Wan *et al.*, 2022]), and further to structure-constrained learning (e.g., scGNN [Wang *et al.*, 2021] and scCDCG [Xu *et al.*, 2024]), existing methods have achieved maturity in mining statistical distributions and topological structures. However, these methods suffer from a fundamental limitation: *semantic agnosticism* [Cui *et al.*, 2024; Theodoris *et al.*, 2023]. By reducing genes to abstract numerical indices, existing frameworks prioritize expression patterns over functional logic. Consequently, while adept at identifying distributional shifts, they lack the semantic grounding required for biological reasoning and interpretability [Xu *et al.*, 2025d].

To address this limitation, the field is witnessing a paradigm shift toward Foundation Models (FMs) [Zhang *et al.*, 2025]. Inspired by LLMs in NLP, these approaches analogize genes as "tokens" and cells as "sentences," pre-training on massive atlases to learn universal representations. Pioneering works like scBERT [Yang *et al.*, 2022] and Geneformer [Theodoris *et al.*, 2023] adopted the BERT paradigm, capturing genome-wide context via masked modeling. Subsequently, generative models such as scGPT [Cui *et al.*, 2024] and scFoundation [Hao *et al.*, 2024] scaled up parameters to reconstruct complex expression patterns across species. More recently, researchers have sought to enhance interpretability by integrating biological priors, such as gene regulatory networks in GeneCompass [Yang *et al.*, 2024] or leveraging textual embeddings in GenePT [Chen and Zou, 2024].

Despite their potential, directly adapting these general-purpose models to the specific downstream task of cell clustering presents structural contradictions [Kedzierska *et al.*, 2025; Li *et al.*, 2025]: (1) **Objective Mismatch between Pre-training and Clustering:** Most models (e.g., scGPT) rely on "generative" pre-training focused on reconstructing global statistics rather than distinguishing local subpopulation boundaries. When applied to clustering (a discriminative task), this lack of boundary constraints often leads to "over-smoothing" representations, failing to separate subtle transitional states. (2) **Numerical representations lack biological semantic support:** While treating genes as tokens, existing methods es-

*Corresponding authors.

sentially capture numerical co-occurrence rather than explicit functional attributes (e.g., transcription factor regulation). Prioritizing probabilistic associations over causal logic makes them prone to hallucinations, thereby generating biologically implausible patterns driven by noise. (3) **Missing Cross-Modal Alignment:** There is a lack of effective mechanisms to align "external biological knowledge" with "internal transcriptomic features" in the latent space. This modal disconnection results in clustering outputs that remain mathematical labels rather than biologically attributable entities [Dip *et al.*, 2025].

To address these impediments, we propose **scLLM-DSC**, an **LLM**-Knowledge Enhanced Cross-Modal **D**eep **S**tructural **C**lustering framework for **s**ingle-**c**ell RNA Sequencing. Diverging from the purely generative paradigm of general foundation models, scLLM-DSC is formulated as a task-specific *knowledge-integrated* framework anchored in three core pillars: **Knowledge-Driven Priors, Structure-Aware Topology, and Cross-Modal Semantic Alignment**. The core logic is to utilize LLMs not as data generators but as semantic mappers, explicitly integrating biological semantics from knowledge bases (e.g., NCBI) into the clustering process and enforcing alignment with numerical features via cross-modal contrastive learning. scLLM-DSC comprises three coupled modules: (1) **Dual-Path Cell Semantic Encoding (Knowledge-driven):** It incorporates external priors by encoding NCBI gene summaries and using a Cell2Sentence strategy to capture dynamic regulation contexts. (2) **Structure-Aware Cell Feature Encoding(Structure-aware):** Leveraging the scCDCG backbone, it extracts high-order topological information to ensure the fidelity of the native data manifold. (3) **Cross-Modal Alignment and Fusion (Semantic-alignment):** Utilizing bidirectional InfoNCE loss and variance regularization, this module bridges the cognitive gap between "data-driven" structural signals and "knowledge-driven" semantic signals. Extensive benchmarks demonstrate that scLLM-DSC significantly outperforms 11 mainstream baselines in both accuracy and efficiency, while uniquely endowing clustering results with explicit biological attribution.

Our contributions are summarized as follows:

- **A Knowledge-Data Dual-Driven Paradigm:** We transcend the limitations of traditional numerical indexing by systematically incorporating NCBI functional priors. This establishes a novel paradigm that bridges the gap between 'structure-aware' deep learning and semantic-aware' biological reasoning.
- **Task-Specific Synergistic Optimization:** To overcome the objective mismatch of FMs for clustering, we devise a specialized optimization objective. By integrating semantic alignment with discriminative structural constraints, we ensure the model focuses on defining precise cell subpopulation boundaries rather than global reconstruction.
- **Unified Interpretable Representation:** We leverage cross-modal alignment to construct a unified latent space that preserves both topological fidelity and biological meaning, enabling interpretable mechanism discovery.

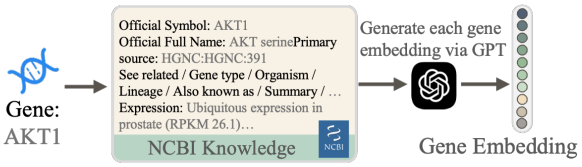


Figure 1: Details of Generating Gene Embedding Generation.

2 Related Work

Conventional scRNA-seq Clustering. Early pipelines like Seurat [Butler *et al.*, 2018] and Scanpy [Wolf *et al.*, 2018] rely on linear reduction and community detection. To capture non-linear manifolds and data sparsity, deep generative models such as scVI [Lopez *et al.*, 2018] and DCA [Eraslan *et al.*, 2019] utilize ZINB-integrated autoencoders for explicit count modeling. Subsequent frameworks, including scDeepCluster [Tian *et al.*, 2019] and scDCC [Tian *et al.*, 2021], introduced joint optimization of reconstruction and clustering-constrained objectives. To incorporate intercellular topological information, initial frameworks like scGNN [Wang *et al.*, 2021] and scDSC [Gan *et al.*, 2022] utilize graph convolutional networks for neighborhood aggregation, while scSiameseClu [Xu *et al.*, 2025a] employs contrastive learning to bolster representation robustness. To enhance scalability and circumvent GCN-induced over-smoothing, structural models based on deep graph-cut objectives have emerged. scCDCG [Xu *et al.*, 2024] and its extension scSGC [Xu *et al.*, 2025b] utilize deep spectral clustering to efficiently capture global topological constraints. Despite these advancements, these numerical-driven methods often lack explicit biological semantics, limiting their interpretability in complex tissues.

Biological Foundation Models. The focus has recently shifted toward Transformer-based foundation models that learn universal biological representations from large-scale cell atlases. scBERT [Yang *et al.*, 2022] and Geneformer [Theodoris *et al.*, 2023] pioneered this paradigm by employing masked gene modeling to capture complex genomic contexts. To enhance generalization, scGPT [Cui *et al.*, 2024] and scFoundation [Hao *et al.*, 2024] further scaled parameter capacities for multi-species generative embeddings. Beyond sequence-only modeling, knowledge-enhanced frameworks such as GeneCompass [Yang *et al.*, 2024] and GenePT [Chen and Zou, 2024] integrate gene regulatory networks and textual semantics, respectively, to infuse prior biological constraints. Recently, scMamba [Yuan *et al.*, 2025] was introduced to achieve linear computational complexity for ultra-long genomic sequences. Despite their powerful representation, these foundation models are not inherently optimized for partitioning tasks, suffering from a objective mismatch between general pre-training and downstream structural clustering.

3 Methodology

3.1 Problem Formulation

The core challenge addressed in this study is **Knowledge-Enhanced Structural Clustering**. Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ be the single-cell gene expression matrix, where N denotes the number of cells and D is the number of genes. Unlike traditional

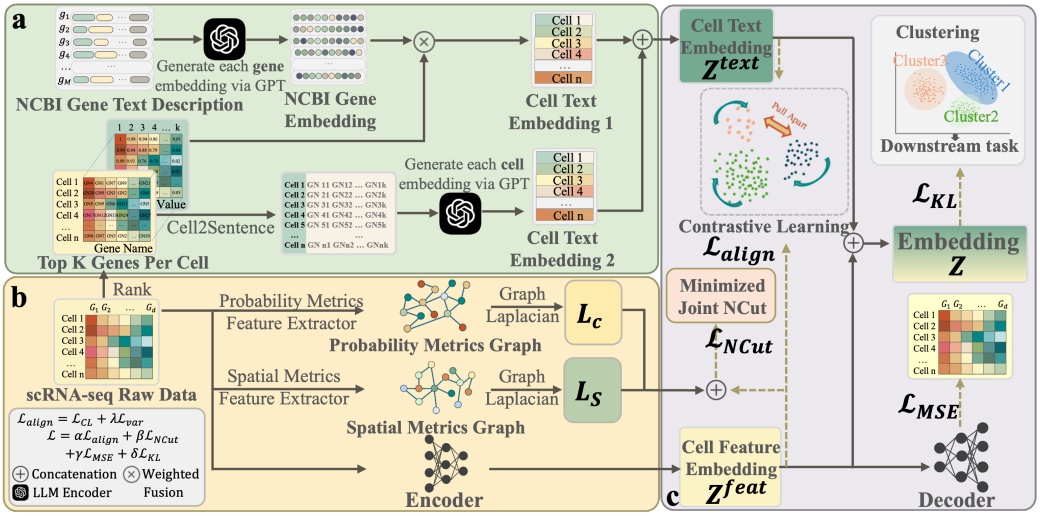


Figure 2: Framework Overview of the Proposed scLLM-DSC. (a) Dual-path cell semantic encoding (Knowledge-driven); (b) Structure-aware cell feature encoding(Structure-aware); (c) Cross-modal alignment and fusion (Semantic-alignment).

data-driven methods, we incorporate an external gene semantic knowledge base, denoted as $\mathbf{G}$, derived from LLMs. Our objective is to learn a mapping function $f : (\mathbf{X}, \mathbf{G}) \rightarrow \mathbf{Z}$ that projects cells into a unified latent space $\mathbf{Z}$ by aligning transcriptomic structural features with biological semantic information. Ultimately, we aim to partition $\mathbf{Z}$ into O cell populations, yielding the clustering assignment $\mathbf{C} = \{c_1, c_2, \ldots, c_N\}$, where $c_i \in \{1, \ldots, O\}$ denotes the cluster label of the i-th cell.

3.2 Framework Overview

The scLLM-DSC framework, illustrated in Fig. 2, establishes a novel paradigm for LLM-Knowledge enhanced cross-modal deep structural clustering by synergizing large-scale biological priors with data-driven representation learning. To integrate biological semantic knowledge, we first retrieve structured biological annotations from NCBI and encode them into gene-level textual representations using a frozen large language model. Subsequently, we construct a unified cell embedding through three modules: (1) **Semantic View**: creating complementary cell-level textual embeddings by aggregating gene semantics and serializing cell profiles; (2) **Structural View**: capturing the intrinsic data manifold via a graph-cut–guided deep structural encoder; and (3) **Cross-Modal Contrastive Fusion**: performing alignment between LLM-encoded biological semantics (Text) and expression-derived structural features (Feature), yielding a biologically grounded latent space optimized for the downstream clustering task.

3.3 Gene Semantic Encoding

To explicitly incorporate biological priors, we leverage a pre-trained LLM as a universal knowledge encoder, as illustrated in Fig. 1. Specifically, we systematically collected comprehensive metadata covering genome-wide genes from the NCBI database [Li *et al.*, 2024]. For the j-th gene ($j \in \{1, \ldots, M\}$), this heterogeneous information (e.g., symbol, functional summary) is serialized into a structured prompt $\mathcal{T}_j$. We then extract the dense vector representation via the LLM encoder:

$$\mathbf{g}_j = f_{\text{LLM}}(\mathcal{T}_j). \tag{1}$$

Consequently, we construct a global gene semantic matrix

$$\mathbf{G} = [\,\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_M\,]^\top \in \mathbb{R}^{M \times d_1}, \tag{2}$$

which projects discrete gene entities into a continuous, machine-interpretable semantic manifold.

3.4 Dual-Path Cell Semantic Encoding

To bridge the gap between gene-level knowledge and cell-level identity, we design a dual-path encoding strategy. Crucially, both paths operate on the most informative genes to filter noise and ensure computational efficiency. For each cell, we strictly select the top-K genes ranked by expression intensity (experimentally set to $K = 2048$). Let $\widetilde{\mathbf{X}} \in \mathbb{R}^{N \times K}$ denote this filtered and sorted expression matrix, where each row contains the expression values of the top-K genes for a cell.

Abundance-Weighted Semantic Aggregation. This branch explicitly utilizes the expression magnitudes in $\widetilde{\mathbf{X}}$ as importance weights. Let $\widetilde{\mathbf{G}} \in \mathbb{R}^{K \times d_1}$ be the subset of LLM-derived gene embeddings corresponding to the genes in $\widetilde{\mathbf{X}}$. The first cell-level embedding $\mathbf{Z}^{(1)}$ is obtained by aggregating these gene semantics weighted by their expression values:

$$\mathbf{Z}^{(1)} = \widetilde{\mathbf{X}}\,\widetilde{\mathbf{G}} \in \mathbb{R}^{N \times d_1}, \tag{3}$$

which projects the cell's gene expression abundance directly into the semantic manifold defined by $\mathbf{G}$.

Sequence-Based Contextual Modeling. To capture non-linear dependencies and gene co-occurrence patterns, we adopt a serialization approach inspired by Cell2Sentence [Levine *et al.*, 2024]. We adopt a serialization approach where the row vector $\tilde{\mathbf{x}}_i$ from $\widetilde{\mathbf{X}}$ is treated as a sentence. Specifically, the names of the top-K genes (ordered by their rank in $\widetilde{\mathbf{X}}$)

are tokenized into a sequence $\mathcal{S}_i$. These sequences are then encoded by the shared LLM encoder f_{LLM}:

$$\mathbf{Z}_i^{(2)} = f_{\text{LLM}}(\mathcal{S}_i), \quad \mathbf{Z}^{(2)} \in \mathbb{R}^{N \times d_1}. \tag{4}$$

Unlike $\mathbf{Z}^{(1)}$, this representation leverages the LLM's attention mechanism to perceive the underlying regulatory logic of gene expressions preserved in the sorted gene list.

Dual-Path Semantic Fusion. We integrate the abundance-based and context-based representations to yield the final unified cell semantic embedding $\mathbf{Z}^{\text{text}}$:

$$\mathbf{Z}^{\text{text}} = \omega \mathbf{Z}^{(1)} + (1-\omega)\mathbf{Z}^{(2)}, \tag{5}$$

where $\omega \in [0,1]$ is a balancing coefficient. This fused representation $\mathbf{Z}^{\text{text}} \in \mathbb{R}^{N \times d_1}$ serves as the semantic anchor for the subsequent structure-aware alignment.

3.5 Structure-Aware Cell Feature Encoding

To capture the intrinsic topological manifold of the cell population, we utilize the deep graph clustering framework scCDCG [Xu *et al.*, 2024] as our structural backbone. Given input $\mathbf{X}$, the backbone functions as a non-linear mapping encoder $f_{\text{struc}}(\cdot)$ to extract the cell feature embedding:

$$\mathbf{Z}^{\text{feat}} = f_{\text{struc}}(\mathbf{X}) \in \mathbb{R}^{N \times d_2}. \tag{6}$$

This encoding process is governed by three complementary unsupervised objectives:

- **Topology Preservation ($\mathcal{L}_{NCut}$):** A differentiable Normalized Cut loss that captures high-order structural information via long-range dependencies, effectively preserving the global manifold while circumventing the over-smoothing limitations of traditional GCNs.
- **Expression Fidelity ($\mathcal{L}_{MSE}$):** An autoencoder-based reconstruction loss that guarantees the latent embeddings retain essential gene expression information.
- **Cluster Refinement ($\mathcal{L}_{KL}$):** A KL-divergence objective that aligns soft assignments with a target distribution optimized via Optimal Transport (OT). This mechanism enforces a global balance constraint to prevent cluster collapsing and stabilize the partitioning manifold.

The resulting $\mathbf{Z}^{\text{feat}}$ encapsulates the data-driven structural priors, serving as the topological anchor for the subsequent cross-modal alignment.

3.6 Cross-Modal Alignment and Fusion

To bridge the heterogeneity between biological semantics and topological features, we align the semantic embedding $\mathbf{Z}^{\text{text}}$ and the structural embedding $\mathbf{Z}^{\text{feat}}$ into a unified latent space via contrastive learning.

Projection and Similarity. We first map the modality-specific embeddings to a shared d-dimensional space using two separate non-linear projection heads (two-layer MLPs with ReLU), denoted as $f_\phi(\cdot)$ and $g_\psi(\cdot)$:

$$\hat{\mathbf{Z}}^{\text{text}} = f_\phi(\mathbf{Z}^{\text{text}}), \quad \hat{\mathbf{Z}}^{\text{feat}} = g_\psi(\mathbf{Z}^{\text{feat}}), \tag{7}$$

where $\hat{\mathbf{Z}} \in \mathbb{R}^{N \times d}$. We then compute the cross-modal cosine similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$. By normalizing the projected embeddings to the unit hypersphere (i.e., $\|\hat{\mathbf{z}}\| = 1$), the similarity is efficiently calculated via scaled dot product:

$$\mathbf{S} = (\hat{\mathbf{Z}}^{\text{text}}(\hat{\mathbf{Z}}^{\text{feat}})^\top)/\tau, \tag{8}$$

where τ is a temperature parameter.

Bidirectional Alignment Objective. To enforce semantic consistency, we minimize the symmetric InfoNCE loss. For a mini-batch of N cells, considering the paired representations (i,i) as positives and all other pairs as negatives, the loss is formulated as:

$$\mathcal{L}_{CL} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\underbrace{\log\frac{\exp(S_{ii})}{\sum_{k=1}^{N}\exp(S_{ik})}}_{\text{Text}\,\to\,\text{Feature}} + \underbrace{\log\frac{\exp(S_{ii})}{\sum_{k=1}^{N}\exp(S_{ki})}}_{\text{Feature}\,\to\,\text{Text}}\right). \tag{9}$$

The first term aligns the textual view to the structural view, while the second aligns the structural view to the textual view.

Variance Regularization. To prevent the projection heads from mapping diverse cells to a trivial point (i.e., dimensional collapse), we impose a hinge loss on the standard deviation of each feature dimension:

$$\mathcal{L}_{var} = \sum_{\mathbf{V}\in\{\hat{\mathbf{Z}}^{\text{text}},\hat{\mathbf{Z}}^{\text{feat}}\}}\frac{1}{d}\sum_{k=1}^{d}\max\left(0, 1-\sqrt{\text{Var}(\mathbf{V}_{:,k})+\epsilon}\right), \tag{10}$$

where $\mathbf{V}_{:,k} \in \mathbb{R}^N$ denotes the column vector corresponding to the k-th dimension across the batch, and $\epsilon = 10^{-4}$ is a constant for numerical stability.

The overall loss for the cross-modal alignment module is formulated as:

$$\mathcal{L}_{align} = \mathcal{L}_{CL} + \lambda\mathcal{L}_{var}, \tag{11}$$

where λ is a hyperparameter.

3.7 Optimization and Inference

Joint Training. The framework is optimized end-to-end via a weighted composite objective:

$$\mathcal{L} = \alpha\mathcal{L}_{align} + \beta\mathcal{L}_{NCut} + \gamma\mathcal{L}_{MSE} + \delta\mathcal{L}_{KL}, \tag{12}$$

where α, β, γ, and δ are hyperparameters balancing cross-modal semantic alignment, global topology preservation, reconstruction fidelity, and cluster refinement, respectively.

Inference. Upon convergence, we obtain the unified embeddings by averaging the aligned semantic and structural views:

$$\mathbf{Z}^{\text{cluster}} = \frac{1}{2}\left(\hat{\mathbf{Z}}^{\text{text}} + \hat{\mathbf{Z}}^{\text{feat}}\right). \tag{13}$$

The final clustering assignments $\mathbf{C}$ are then generated by applying K-Means (or Leiden) directly to $\mathbf{Z}^{\text{cluster}}$.

4 Experiments

4.1 Experimental Setup

Dataset. We utilized a subset of datasets provided by scCluBench [Xu *et al.*, 2025c], consisting of 6 real scRNA-seq datasets from human and mouse tissues (Table 1). All data were used as provided, without further preprocessing.

(a) scDeepCluster (b) scMAE (c) scSiameseClu (d) scCDCG (e) scLLM-DSC

Figure 3: Visualization of scLLM-DSC and four typical baselines on *Mauro Pancreas* in 2D t-SNE projection.

Species	Dataset Name	#Cell	#Gene	#Cluster	Sparsity (%)
Human	Mauro Pancreas	2,122	19,046	9	73.02
	Sonya Liver	8,444	4,999	11	90.77
	Sapiens Liver	2,152	61,759	15	95.42
Mouse	Muris Brain	13,417	21,609	2	91.83
	Muris Liver	2,859	21,609	11	88.20
	Muris Limb Muscle	3,855	21,609	6	91.38

Table 1: Details of single-cell RNA-seq Datasets.

Baseline Methods. We compare scLLM-DSC against 11 baselines across three categories: (1) Deep Non-linear Embedding Models: These models utilize deep autoencoders to learn latent representations, including scDeepCluster [Tian *et al.*, 2019], scziDesk [Chen *et al.*, 2020], and masked autoencoder variants scNAME [Wan *et al.*, 2022] and scMAE [Fang *et al.*, 2024]. (2) Deep Structural Clustering Models: These methods explicitly incorporate cellular topological information via graph-based architectures, comprising scGNN [Wang *et al.*, 2021], scDSC [Gan *et al.*, 2022], the contrastive scSiameseClu [Xu *et al.*, 2025a], and the cut-informed scCDCG [Xu *et al.*, 2024]. (3) Foundation Models: These large-scale pre-trained frameworks leverage expansive atlases for downstream transfer, represented by scGPT [Cui *et al.*, 2024], GeneFormer [Theodoris *et al.*, 2023], and the knowledge-informed GeneCompass [Yang *et al.*, 2024]. All methods were implemented using their publicly available Python packages, with hyperparameters tuned for optimal performance.

Implementation Details. Implemented in Python 3.12 (PyTorch $\geq$2.9.1, CUDA 13.0), scLLM-DSC follows a two-stage protocol: autoencoder pre-training (joint NCut/MSE) followed by contrastive clustering. We generate gene embeddings from NCBI metadata using OpenAI's *text-embedding-3-small*. Cell embeddings are extracted from NCBI metadata via OpenAI's *text-embedding-3-small*. Cell embeddings aggregate the top-2048 highly expressed genes ($k = 2048$) weighted by expression v_i: $\mathbf{e}_{\text{cell}} = \sum_{i=1}^{k} v_i \cdot \mathbf{e}_{\text{gene}_i}$. The architecture features a [256, 16] autoencoder producing 16-dimensional latent vectors. A 128-dimensional projection head aligns modalities via bi-directional InfoNCE loss ($\tau = 0.1$) with variance regularization ($\lambda = 10^{-2}$). Training spans 200 epochs/stage using Adam, with hyperparameters optimized via Optuna. Clusters are initialized by KMeans and refined via Sinkhorn ($\lambda = 5$, 1k iters). We report the mean of 5 independent runs on an NVIDIA A800-80GB GPU. All datasets and code for scLLM-DSC are available at the link: https://github.com/XPgogogo/scLLM-DSC.

Evaluation Metrics. We assess the clustering quality by measuring the congruence between predicted partitions and biological ground truth using three benchmark indices: Clustering Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). While ACC and NMI range from 0 to 1, ARI can be negative if clustering performs worse than random assignment. For all metrics, higher values indicate superior clustering fidelity.

4.2 Overall Performance

Qualitative Analysis. (1)***Comparison with Deep Clustering Baselines.*** Tab. 2 shows that scLLM-DSC achieves **SOTA performance**, outperforming all baselines in both average metrics (ACC: 88.80%, NMI: 85.35%, ARI: 83.04%) and overall ranking. The superiority stems from its ability to synergize LLM-derived semantic anchors with global structural manifolds. Specifically, the performance margin over *deep non-linear embedding models* underscores the necessity of integrating both textual semantics and topological structures. Furthermore, compared to *deep structural clustering models* (e.g., scCDCG), scLLM-DSC yields higher clustering fidelity, validating that LLM-based priors provide more discriminative guidance for structural alignment than raw numerical features. This robustness is further exemplified on the *Muris Brain* dataset (#Cell=13,417, #Cluster=2). While many baselines (e.g., scziDesk) achieve misleadingly high ACC by predicting majority classes, their near-zero NMI and ARI indicate a collapse into trivial partitions. Such failure occurs because raw numerical features often lack sufficient margin to define clear boundaries in high-volume but low-cluster scenarios, causing structural models to suffer from oscillation or over-smoothing. scLLM-DSC overcomes this by injecting LLM-based semantic anchors, which provide the contrastive signals required to maintain cluster separation and prevent manifold collapse. (2) ***Comparison with Biological Foundation Models.*** Evaluation against large-scale foundation models (Tab. 3) reveals that scLLM-DSC provides more robust partitioning across most datasets. Unlike general-purpose models, scLLM-DSC is specialized for clustering tasks, effectively capturing discrete cell-type boundaries. While Geneformer and GeneCompass show localized peaks on *Sonya Liver*, likely due to pre-training data overlap, scLLM-DSC demonstrates superior generalization elsewhere. This strongly underscores that for effectively resolving highly complex cellular heterogeneity, dedicated structural objectives specifically tailored for clustering are indispensable compared to universal embeddings.

Quantitative Analysis. As depicted in Fig. 3, we project the

Dataset	Metrics	Deep Non-linear Embedding Models				Deep Structural Clustering Models				
		scDeepCluster	scMAE	scNAME	scziDesk	scGNN	scDSC	scSiameseClu	scCDCG	**OURS**
Mauro Pancreas	ACC	73.55$_{\pm 1.37}$	95.62$_{\pm 0.16}$	89.73$_{\pm 9.92}$	69.97$_{\pm 17.08}$	79.32$_{\pm 3.96}$	79.42$_{\pm 1.34}$	94.95$_{\pm 0.15}$	92.65$_{\pm 2.93}$	**96.94**$_{\pm 1.00}$
	NMI	78.89$_{\pm 0.98}$	88.49$_{\pm 0.44}$	85.31$_{\pm 5.11}$	64.94$_{\pm 14.21}$	78.76$_{\pm 5.60}$	75.89$_{\pm 0.61}$	86.19$_{\pm 0.34}$	86.81$_{\pm 0.98}$	**92.43**$_{\pm 1.49}$
	ARI	64.90$_{\pm 1.07}$	92.38$_{\pm 0.35}$	84.51$_{\pm 13.42}$	50.31$_{\pm 29.92}$	78.81$_{\pm 3.17}$	75.42$_{\pm 1.22}$	94.59$_{\pm 0.35}$	91.37$_{\pm 1.21}$	**96.08**$_{\pm 0.67}$
Sonya liver	ACC	69.21$_{\pm 2.47}$	80.73$_{\pm 1.86}$	79.97$_{\pm 8.80}$	76.58$_{\pm 1.52}$	72.33$_{\pm 3.40}$	70.90$_{\pm 2.23}$	88.33$_{\pm 1.74}$	75.34$_{\pm 3.67}$	**90.47**$_{\pm 6.44}$
	NMI	79.69$_{\pm 0.68}$	85.59$_{\pm 1.44}$	85.04$_{\pm 3.65}$	83.34$_{\pm 0.70}$	73.56$_{\pm 2.70}$	71.63$_{\pm 2.75}$	88.82$_{\pm 1.24}$	79.34$_{\pm 2.59}$	**91.92**$_{\pm 9.22}$
	ARI	69.71$_{\pm 1.18}$	88.92$_{\pm 1.78}$	83.96$_{\pm 13.63}$	83.87$_{\pm 1.17}$	76.01$_{\pm 2.60}$	75.34$_{\pm 3.56}$	91.90$_{\pm 0.48}$	81.26$_{\pm 2.69}$	**92.01**$_{\pm 5.58}$
Sapiens Liver	ACC	49.08$_{\pm 1.90}$	67.51$_{\pm 2.30}$	63.33$_{\pm 1.70}$	68.19$_{\pm 0.60}$	73.83$_{\pm 3.00}$	64.67$_{\pm 4.40}$	62.34$_{\pm 0.84}$	73.09$_{\pm 1.40}$	**75.19**$_{\pm 2.82}$
	NMI	60.98$_{\pm 2.20}$	78.25$_{\pm 0.60}$	74.83$_{\pm 0.70}$	75.10$_{\pm 1.20}$	77.21$_{\pm 2.00}$	68.21$_{\pm 5.20}$	68.96$_{\pm 1.60}$	62.54$_{\pm 3.20}$	**79.09**$_{\pm 5.97}$
	ARI	34.52$_{\pm 2.00}$	63.82$_{\pm 4.10}$	58.94$_{\pm 2.50}$	66.12$_{\pm 1.90}$	72.36$_{\pm 1.70}$	56.94$_{\pm 7.90}$	51.10$_{\pm 1.29}$	41.11$_{\pm 4.40}$	**78.03**$_{\pm 0.97}$
Muris Limb Muscle	ACC	59.57$_{\pm 3.90}$	66.13$_{\pm 3.40}$	61.34$_{\pm 3.10}$	53.31$_{\pm 4.30}$	48.62$_{\pm 2.30}$	64.37$_{\pm 4.00}$	81.32$_{\pm 6.88}$	94.50$_{\pm 7.10}$	**95.54**$_{\pm 3.86}$
	NMI	34.55$_{\pm 4.90}$	59.44$_{\pm 3.80}$	63.51$_{\pm 1.30}$	36.66$_{\pm 6.40}$	22.10$_{\pm 4.70}$	33.30$_{\pm 4.50}$	80.04$_{\pm 4.44}$	56.54$_{\pm 7.60}$	**91.69**$_{\pm 8.60}$
	ARI	35.95$_{\pm 8.10}$	51.54$_{\pm 3.30}$	54.70$_{\pm 2.30}$	32.59$_{\pm 7.20}$	21.90$_{\pm 3.20}$	34.28$_{\pm 6.40}$	75.49$_{\pm 8.34}$	53.37$_{\pm 8.50}$	**82.91**$_{\pm 5.60}$
Muris Brain	ACC	85.36$_{\pm 18.10}$	71.37$_{\pm 0.00}$	90.24$_{\pm 0.30}$	OOM	91.40$_{\pm 0.10}$	96.02$_{\pm 2.50}$	99.17$_{\pm 5.19}$	95.55$_{\pm 1.10}$	**99.42**$_{\pm 0.51}$
	NMI	59.48$_{\pm 44.40}$	1.33$_{\pm 0.00}$	0.01$_{\pm 0.00}$	OOM	0.31$_{\pm 0.10}$	0.07$_{\pm 0.00}$	33.01$_{\pm 1.79}$	22.48$_{\pm 8.30}$	**68.58**$_{\pm 1.25}$
	ARI	60.11$_{\pm 46.80}$	−2.22$_{\pm 0.00}$	0.34$_{\pm 0.10}$	OOM	2.77$_{\pm 0.50}$	−0.37$_{\pm 0.80}$	30.87$_{\pm 1.22}$	35.56$_{\pm 7.80}$	**79.36**$_{\pm 1.96}$
Muris Liver	ACC	42.62$_{\pm 3.20}$	53.48$_{\pm 0.40}$	49.72$_{\pm 4.10}$	44.50$_{\pm 3.40}$	51.58$_{\pm 2.90}$	55.76$_{\pm 7.50}$	66.81$_{\pm 4.50}$	68.13$_{\pm 1.40}$	**75.24**$_{\pm 6.58}$
	NMI	45.07$_{\pm 2.20}$	65.39$_{\pm 1.00}$	61.21$_{\pm 1.70}$	43.17$_{\pm 3.40}$	55.15$_{\pm 1.10}$	35.07$_{\pm 4.70}$	71.46$_{\pm 1.51}$	62.06$_{\pm 2.40}$	**88.38**$_{\pm 6.54}$
	ARI	27.92$_{\pm 3.90}$	47.55$_{\pm 0.60}$	38.55$_{\pm 3.40}$	27.27$_{\pm 4.60}$	45.58$_{\pm 3.20}$	43.74$_{\pm 12.30}$	49.56$_{\pm 3.16}$	46.96$_{\pm 3.70}$	**69.85**$_{\pm 1.99}$
AVG	ACC	63.23$_{\pm 5.16}$	72.47$_{\pm 1.35}$	72.39$_{\pm 4.65}$	62.51$_{\pm 5.38}$	69.51$_{\pm 2.61}$	71.85$_{\pm 3.66}$	82.15$_{\pm 3.22}$	83.21$_{\pm 2.93}$	**88.80**$_{\pm 3.54}$
	NMI	59.78$_{\pm 9.23}$	63.08$_{\pm 1.21}$	61.65$_{\pm 2.08}$	60.64$_{\pm 5.18}$	51.18$_{\pm 2.70}$	46.03$_{\pm 2.96}$	71.41$_{\pm 1.82}$	61.63$_{\pm 4.18}$	**85.35**$_{\pm 5.51}$
	ARI	48.85$_{\pm 10.50}$	57.00$_{\pm 1.69}$	53.50$_{\pm 5.89}$	50.03$_{\pm 7.45}$	49.44$_{\pm 2.40}$	47.55$_{\pm 5.36}$	65.59$_{\pm 2.47}$	58.27$_{\pm 4.72}$	**83.04**$_{\pm 2.80}$
RANK	ALL	6	4	5	6	9	8	2	3	1

Table 2: Clustering performance compared with deep embedding and structural clustering models (mean$_{\pm \text{std}}$). Bold and underline denote best/runner-up. (OOM: Out-of-Memory; Rank: avg. of ACC/NMI/ARI; scDeepCluster and sczi Desk tied for 6th).

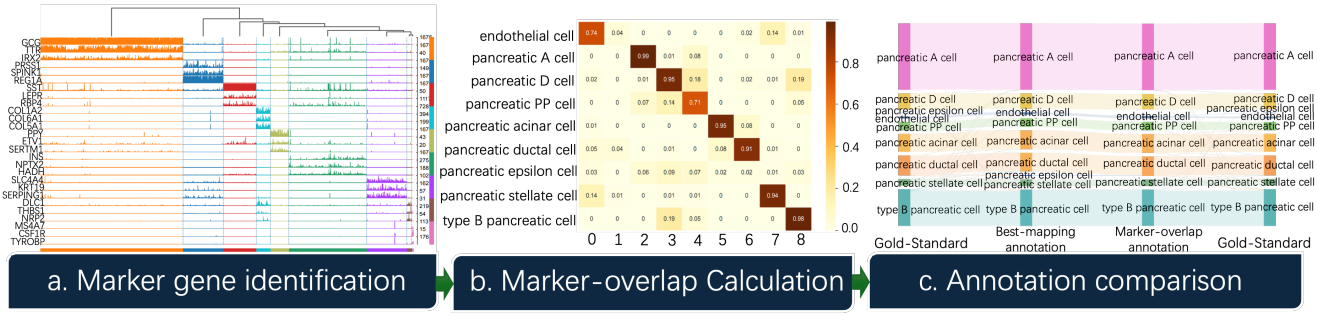


Figure 4: All Biological Analysis of scLLM-DSCon *Mauro Pancreas* Dataset.

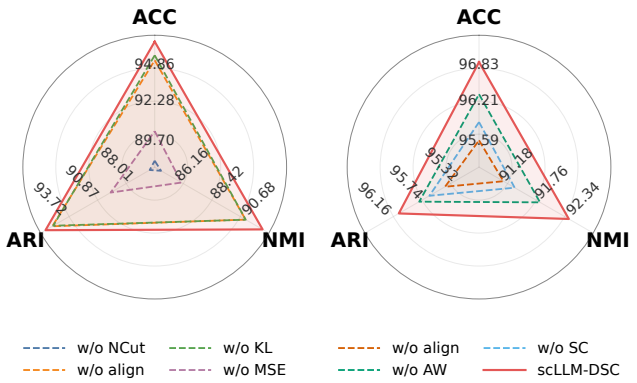
Dataset	Metric	scGPT	Geneformer	GeneCompass	OURS
Mauro Pancreas	ACC	56.61$_{\pm 1.35}$	17.14$_{\pm 0.67}$	25.19$_{\pm 0.95}$	**96.94**$_{\pm 1.00}$
	NMI	53.91$_{\pm 1.26}$	0.87$_{\pm 0.15}$	7.00$_{\pm 0.11}$	**92.43**$_{\pm 1.49}$
	ARI	37.15$_{\pm 1.49}$	0.02$_{\pm 0.04}$	3.00$_{\pm 0.19}$	**96.08**$_{\pm 0.67}$
Sonya Liver	ACC	57.75$_{\pm 8.35}$	**100.00**$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	90.47$_{\pm 6.44}$
	NMI	76.47$_{\pm 3.40}$	**100.00**$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	91.92$_{\pm 9.22}$
	ARI	53.74$_{\pm 9.45}$	**100.00**$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	92.01$_{\pm 5.58}$
Sapiens Liver	ACC	43.47$_{\pm 3.91}$	24.49$_{\pm 0.98}$	27.77$_{\pm 1.31}$	**75.19**$_{\pm 2.82}$
	NMI	59.75$_{\pm 1.25}$	23.78$_{\pm 0.78}$	24.23$_{\pm 0.78}$	**79.09**$_{\pm 5.97}$
	ARI	36.11$_{\pm 6.46}$	10.75$_{\pm 0.64}$	13.48$_{\pm 2.19}$	**78.03**$_{\pm 0.97}$
Muris Limb Muscle	ACC	29.22$_{\pm 0.25}$	23.25$_{\pm 1.15}$	24.47$_{\pm 0.04}$	**95.54**$_{\pm 3.86}$
	NMI	9.44$_{\pm 0.06}$	1.05$_{\pm 0.20}$	4.16$_{\pm 0.01}$	**91.69**$_{\pm 8.60}$
	ARI	5.64$_{\pm 0.31}$	0.77$_{\pm 0.30}$	2.40$_{\pm 0.00}$	**82.91**$_{\pm 5.60}$
Muris Brain	ACC	59.54$_{\pm 0.57}$	62.71$_{\pm 0.10}$	54.82$_{\pm 0.02}$	**99.42**$_{\pm 0.51}$
	NMI	0.04$_{\pm 0.02}$	0.02$_{\pm 0.00}$	0.00$_{\pm 0.00}$	**68.58**$_{\pm 1.25}$
	ARI	0.20$_{\pm 0.05}$	0.17$_{\pm 0.00}$	−0.01$_{\pm 0.00}$	**79.36**$_{\pm 1.96}$
Muris Liver	ACC	32.44$_{\pm 2.20}$	13.84$_{\pm 0.45}$	28.80$_{\pm 2.44}$	**75.24**$_{\pm 6.58}$
	NMI	29.40$_{\pm 1.25}$	2.01$_{\pm 0.15}$	19.70$_{\pm 0.79}$	**83.88**$_{\pm 6.54}$
	ARI	13.61$_{\pm 0.44}$	−0.03$_{\pm 0.17}$	13.48$_{\pm 2.16}$	**69.85**$_{\pm 1.99}$

Table 3: Clustering performance comparison with biological foundation models (mean$_{\pm \text{std}}$). The **bold** values represent the best results.

latent representations of scLLM-DSC and four representative baselines into a 2D space. The visualization results demonstrate that scLLM-DSC yields substantial advantages in both boundary clarity and intra-cluster compactness. Regarding the cluster boundaries, baselines such as scDeepCluster and scCDCG exhibit discernible "bridge" structures, resulting in marginal overlaps between distinct cell populations. Conversely, scLLM-DSC effectively eliminates such structural ambiguity; by leveraging LLM-based semantic priors to provide robust discriminative signals, it defines sharp and clean biological boundaries. Notably, from the perspective of intra-cluster compactness, scLLM-DSC facilitates a higher degree of aggregation within the manifold space, forming a more dense and robust cluster structure than the baselines. This dual enhancement in boundary separation and compactness validates the efficacy of cross-modal alignment in constraining the latent space to capture fine-grained cellular heterogeneity.

4.3 Biological Analysis

We conduct an in-depth analysis of the biological relevance of our clustering results following the standard evaluation pipeline of scCluBench [Xu *et al.*, 2025c]. Taking the *Mauro Pancreas* dataset as a case study, we first identify the top 100 marker genes for each predicted cluster via SCANPY, with their expression specificity visualized in the tracksplot in Fig. 4(a). Subsequently, Fig. 4(b) illustrates the marker-overlap calculation process, which establishes a statistical foundation for biological identity assignments by calculating



(a) Framework Component Study (b) Semantic Component Study

Figure 5: Ablation Study on *Mauro Pancreas* Dataset.

VARIANTS	ACC	NMI	ARI
Jasper-Token-Compression-600M	$96.27_{\pm 1.12}$	$91.63_{\pm 1.42}$	$95.70_{\pm 0.83}$
granite-embedding-small-english-r2	$96.36_{\pm 0.90}$	$91.53_{\pm 1.51}$	$95.71_{\pm 0.69}$
Qwen3-Embedding-0.6B	$96.02_{\pm 1.26}$	$91.42_{\pm 1.48}$	$95.57_{\pm 0.77}$
all-MiniLM-L6-v2	$96.64_{\pm 0.98}$	$91.89_{\pm 1.62}$	$95.85_{\pm 0.77}$
bge-m3	$96.60_{\pm 1.13}$	$91.99_{\pm 1.39}$	$95.85_{\pm 0.74}$
OURS (text-embedding-3-small)	$\mathbf{96.94}_{\pm 1.00}$	$\mathbf{92.43}_{\pm 1.49}$	$\mathbf{96.08}_{\pm 0.67}$

Table 4: Robustness to LLM Variants on *Mauro Pancreas*.

the intersection (overlap score) between predicted markers and gold-standard reference sets. Furthermore, in the Sankey diagram of Fig. 4(c), we compare two annotation paradigms: the purely numerical *Best-Mapping* (BM) strategy, which operates independently of external biological priors, and the biology-guided *Marker-Overlap* (MO) strategy. Analytical results demonstrate that while the BM strategy tends to "enforce mappings" for all nine potential cell types, the MO strategy exhibits superior biological rigor. Concretely, for cluster 1 in Fig. 4(b) mislabeled as " *epsilon cell*" by BM, the MO strategy deliberately labels the cluster as "*pancreatic ductal cell*" according to its highest overlap score, as shown in Fig. 4(c), ensuring higher fidelity cellular identification.

4.4 Ablation Study

Framework Component Study. To evaluate the contribution of each module, we compare scLLM-DSC with four variants on the *Mauro Pancreas* dataset by systematically removing $\mathcal{L}_{align}$ (*w/o align*), $\mathcal{L}_{NCut}$ (*w/o NCut*), $\mathcal{L}_{MSE}$ (*w/o MSE*), and $\mathcal{L}_{KL}$ (w/o KL). Results visualized in Fig. 5a reveal that scLLM-DSC consistently outperforms all ablation variants across all metrics, confirming that each component is vital for optimal performance. Notably, the significant performance degradation in *w/o NCut* and *w/o align* underscores the indispensable roles of both high-order structural manifolds and LLM-derived biological semantics. The performance gain demonstrates that scLLM-DSC effectively synergizes high-order topological structures with explicit biological semantics, leading to more robust and interpretable cell clustering.

Semantic Component Study. We evaluate the contribution of each semantic path by comparing scLLM-DSC with three variants: *w/o AW* (removing Abundance-Weighted Semantic Aggregation), *w/o SC* (removing Sequence-based Contextual Encoding), and *w/o align* (discarding both). Results in Fig. 5b show that while either path alone improves performance, their combination consistently yields the highest accuracy. This confirms that expression-weighted magnitudes and gene-order regulatory logic provide complementary biological insights into complex cellular states.

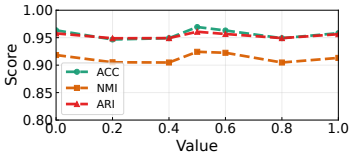


Figure 6: Parameter Sensitivity of ω on *Mauro Pancreas dataset*.

4.5 Robustness to LLM Variants

To assess the model's dependence on a specific language model, we investigate its robustness by substituting the default *text-embedding-3-small* with various LLM backbones of differing architectures and scales. As reported in Tab. 4, scLLM-DSC maintains remarkably stable performance across all evaluation metrics, with negligible fluctuations despite significant variations in the underlying embedding models. This stability indicates that scLLM-DSC is highly compatible with diverse LLM architectures, as the specific backbone choice is not the primary determinant of performance. Combined with Fig. 5, these results underscore that integrating LLM-based semantics remains indispensable for superior accuracy, demonstrating the broad applicability of our framework.

4.6 Parameter Sensitivity

We evaluate the impact of the balancing coefficient ω (Eq. 5) on clustering performance. As illustrated in Fig. 6, metrics including ACC, NMI, and ARI fluctuate with varying fusion ratios of abundance and context semantics. Optimal results across all metrics are achieved at $\omega = 0.5$, confirming that the synergy of dual-path representations is essential for capturing comprehensive cell semantics. The performance decline at $\omega \in \{0, 1\}$ further proves that relying on a single semantic source is insufficient for maintaining peak clustering precision.

5 Conclusion

scLLM-DSC presents a novel paradigm that integrates LLM-derived biological knowledge with deep structural clustering to address the persistent omission of semantic context in traditional scRNA-seq analysis. By synergizing functional gene priors with graph-guided topological encoding, our framework establishes a unified latent space that aligns numerical transcriptomic features with high-level biological meaning. Benchmarks confirm that this task-specific cross-modal alignment not only achieves state-of-the-art accuracy but also provides an interpretable foundation for mechanism-aware discovery. Looking ahead, we plan to scale this framework to atlas-level datasets and incorporate spatial transcriptomics to further decode complex tissue architectures.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grant No. 92470204, 62406306, and 62406056) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515140114).

Contribution Statement

Ping Xu and Pengjiang Li contributed equally to this work. Specifically, they led the project, provided theoretical support, and were responsible for overall model design, code implementation, experimental design, and paper writing. Tian Du, Zaitian Wang, Jiawei Gu, and Zhiyuan Ning provided guidance in solving complex problems and assisted with paper writing. Ziyue Qiao, Pengfei Wang, and Yuanchun Zhou provided valuable feedback on manuscript drafts. All authors reviewed and approved the final version.

References

[Butler *et al.*, 2018] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[Chen and Zou, 2024] Yiqun Chen and James Zou. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10, 2024.

[Chen *et al.*, 2020] Liang Chen, Weinan Wang, Yuyao Zhai, and Minghua Deng. Deep soft k-means clustering with self-training for single-cell rna sequence data. *NAR genomics and bioinformatics*, 2(2):lqaa039, 2020.

[Cui *et al.*, 2024] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

[Dip *et al.*, 2025] Sajib Acharjee Dip, Adrika Zafor, Bikash Kumar Paul, Uddip Acharjee Shuvo, Muhit Islam Emon, Xuan Wang, and Liqing Zhang. Llm4cell: A survey of large language and agentic models for single-cell biology. *arXiv preprint arXiv:2510.07793*, 2025.

[Eraslan *et al.*, 2019] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):390, 2019.

[Fang *et al.*, 2024] Zhaoyu Fang, Ruiqing Zheng, and Min Li. scmae: a masked autoencoder for single-cell rna-seq clustering. *Bioinformatics*, 40(1):btae020, 2024.

[Gan *et al.*, 2022] Yanglan Gan, Xingyu Huang, Guobing Zou, Shuigeng Zhou, and Jihong Guan. Deep structural clustering for single-cell rna-seq data jointly through autoencoder and graph neural network. *Briefings in Bioinformatics*, 23(2):bbac018, 2022.

[Hao *et al.*, 2024] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

[Kedzierska *et al.*, 2025] Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

[Kiselev *et al.*, 2019] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[Levine *et al.*, 2024] Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: teaching large language models the language of biology. *BioRxiv*, pages 2023–09, 2024.

[Li *et al.*, 2024] Cong Li, Qingqing Long, Yuanchun Zhou, and Meng Xiao. screader: Prompting large language models to interpret scrna-seq data. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 665–672. IEEE, 2024.

[Li *et al.*, 2025] Longfan Li, Wangzilu Lu, Yuxin Ji, Zhiwen Gu, Huajie Huang, Yuhang Zhang, Jian Zhao, and Yongfu Li. Sceval: An open-source platform for standardized evaluation and optimization of standard cell libraries in next-generation process nodes. In *2025 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4. IEEE, 2025.

[Lopez *et al.*, 2018] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Svensson *et al.*, 2018] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[Theodoris *et al.*, 2023] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

[Tian *et al.*, 2019] Tian Tian, Ji Wan, Qi Song, and Zhi Wei. Clustering single-cell rna-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191–198, 2019.

[Tian *et al.*, 2021] Tian Tian, Jie Zhang, Xiang Lin, Zhi Wei, and Hakon Hakonarson. Model-based deep embedding for constrained clustering analysis of single cell rna-seq data. *Nature communications*, 12(1):1873, 2021.

[Wan *et al.*, 2022] Hui Wan, Liang Chen, and Minghua Deng. scname: neighborhood contrastive clustering with ancil-

lary mask estimation for scrna-seq data. *Bioinformatics*, 38(6):1575–1583, 2022.

[Wang *et al.*, 2021] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1882, 2021.

[Wolf *et al.*, 2018] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[Xu *et al.*, 2024] Ping Xu, Zhiyuan Ning, Meng Xiao, Guihai Feng, Xin Li, Yuanchun Zhou, and Pengfei 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*, pages 172–187. Springer, 2024.

[Xu *et al.*, 2025a] Ping Xu, Zhiyuan Ning, Pengjiang Li, Wenhao Liu, Pengyang Wang, Jiaxu Cui, Yuanchun Zhou, and Pengfei Wang. scsiameseclu: A siamese clustering framework for interpreting single-cell rna sequencing data. *arXiv preprint arXiv:2505.12626*, 2025.

[Xu *et al.*, 2025b] Ping Xu, Pengfei Wang, Zhiyuan Ning, Meng Xiao, Min Wu, and Yuanchun Zhou. Soft graph clustering for single-cell rna sequencing data. *BMC bioinformatics*, 26(1):195, 2025.

[Xu *et al.*, 2025c] Ping Xu, Zaitian Wang, Zhirui Wang, Pengjiang Li, Jiajia Wang, Ran Zhang, Pengfei Wang, and Yuanchun Zhou. scclubench: Comprehensive benchmarking of clustering algorithms for single-cell rna sequencing. *arXiv preprint arXiv:2512.02471*, 2025.

[Xu *et al.*, 2025d] Ping Xu, Zaitian Wang, Zhirui Wang, Pengjiang Li, Ran Zhang, Gaoyang Li, Hanyu Xie, Jiajia Wang, Yuanchun Zhou, and Pengfei Wang. scunified: An ai-ready standardized resource for single-cell rna sequencing analysis. *arXiv preprint arXiv:2509.25884*, 2025.

[Yang *et al.*, 2022] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[Yang *et al.*, 2024] Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, et al. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, pages 1–16, 2024.

[Yuan *et al.*, 2025] Zhen Yuan, Shaoqing Jiao, Yihang Xiao, and Jiajie Peng. scmamba: A scalable foundation model for single-cell multi-omics integration beyond highly variable feature selection. *arXiv preprint arXiv:2506.20697*, 2025.

[Zhang *et al.*, 2025] Fan Zhang, Hao Chen, Zhihong Zhu, Ziheng Zhang, Zhenxi Lin, Ziyue Qiao, Yefeng Zheng, and Xian Wu. A survey on foundation language models for single-cell biology. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 528–549, 2025.