Multimodal Self-supervised Single-cell Clustering of scRNA-seq Data

Shentong Mo, Yixuan Chen, Ding Bai
MBZUAI
{shentong.mo,yixuan.chen,ding.bai}@mbzuai.ac.ae

Abstract

Single-cell clustering of scRNA-seq data is a typical and challenging problem that predicts cell subtype clusters given gene expression sequences from single-cell RNA data. Previous models utilized classical clustering (*e.g.*, Principal Component Analysis, K-means) on well-annotated data to classify cells. However, they extremely relied on the expected number of clusters as input. With the success of self-supervised learning, constrastive-sc [1] applied an InfoNCE-based contrastive loss on anchor and augmented expression sequence representations for clustering scRNA-seq data. While achieving promising performance, they can not fully capture interconnections across expressions of long sequences for each cell, as they took a short gene expression sequence in a cell as the anchor. To address the problem, in this work, we propose a novel multimodal self-supervised framework with masked expression modeling on single-cell data, namely **mask-sc**, that can learn compact and discriminative representations by reconstructing masked gene expression for scRNA-seq clustering. Our mask-sc aggregates interconnections across multiple groups of expression sequences via a masked expression encoder applied on expression matrices. Then, a sequence-guided decoder is applied to recover sequence-level features of masked expression matrices. Finally, representations extracted from the gene expression encoder can be used for scRNA-seq clustering. We conduct extensive experiments on 15 real scRNA-seq datasets, where empirical results demonstrate the effectiveness of our proposed mask-sc against previous baselines. Demo, code, and pre-trained models are available at our project website: https://masksc.github.io/.

1 Introduction

In bioinformatics, single-cell RNA sequencing (scRNA-seq) is a platform to discover cellular subpopulations [2] and transcriptional profiles [3] from individual cells. Through analysis of scRNA-seq, cell biologists are able to understand and interpret intrinsic cell identities. This crucial biological perception intelligence of cells attracts many researchers to analyze the scRNA-seq data.

Early scRNA-seq analysis works leveraged traditional machine learning approaches, such as Principal Component Analysis (PCA) [4], K-means [5], and Gaussian Mixture Models [6], to cluster cellular subtypes. Due to the challenge of high dimensional and significantly sparse sequences in the clustering analysis, the following methods [7, 8, 9, 10, 11, 12, 13, 14, 15, 16] tried to explore different frameworks to address the problem. Typically, CIDR [7] proposed a hierarchical clustering with an implicit imputation stage before PCA to alleviate the effect of dropouts. scRNA [9] transferred knowledge from a large and well-annotated reference dataset by non-negative matrix factorization for small disease-specific data. With the advance of deep learning, deep neural networks, such as DCA [12] have been widely used to boost clustering performance. ScDeepCluster [13] adopted a clustering layer on the embedding space learned from DCA to enrich representations. A soft self-training KMeans clustering was introduced in SczиDesk [14] to aggregate similar cells in the

ML-701 Project.

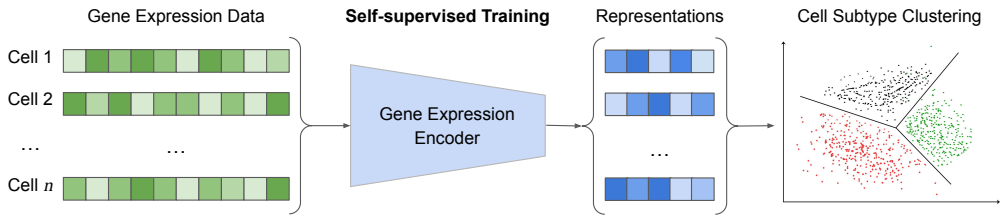

Figure 1: Illustration of single-cell clustering from scRNA-seq representations extracted from a gene expression encoder. The goal of this work is to design a self-supervised training framework to learn such an encoder from gene expression data of multiple cells.

same cluster. However, those methods rely on well-annotated data and take as input the expected number of clusters, which limits their generalization to all circumstances. In contrast, we will solve them in our approach by extracting discriminative and compact representations from expression inputs via a self-supervised encoder, as shown in Figure 1.

Recently, inspired by the success of self-supervised learning [17, 18, 19, 20, 21, 22] in image and text, contrastive-sc [1] applied a contrastive loss on anchor and augmented sample outputs from an encoder to extract representations for clustering scRNA-seq data. While the state-of-the-art baseline achieved promising performance, they can not fully capture interconnections between expressions of different genes for each cell, as they only took a short fragment of expressions in a cell as an input sample during training. For each anchor sample, the contrastive loss closed the distance between embeddings from the anchor and augmented gene expressions, while pushing away embeddings from different gene expressions. Different from them, we group multiple fragments of expressions as an expression matrix, which is fed into a gene expression encoder to capture interconnections across multiple groups by masked expression modeling. In the meanwhile, we apply a sequence-guided decoder to reconstruct sequence-level features of masked expression matrices and propose a novel multimodal self-supervised training framework for single-cell clustering.

The main challenge is that gene expression sequences for each cell are high-dimensional. With only a short fragment of expressions in a cell as input, the state-of-the-art self-supervised approach [1] can not learn discriminative and compact global representations by capturing diverse interconnections across enough genes in this cell. In the meanwhile, previous machine learning approaches [4, 9, 13, 14] are extremely dependent on the well-annotation of the number of cellular subtypes for training. To address the aforementioned challenges, our key idea is to take as input gene expression matrices composed of grouped expression sequences for self-supervised training, which is different from existing clustering and self-supervised methods. During training, we aim to learn compact representations from input gene expression matrices with diverse interconnections across various genes in each cell for discovering potential cellular subtypes.

To this end, we propose a novel multimodal self-supervised training approach based on masked expression modeling for single-cell clustering, termed as masked-sc, that can learn compact representations from diverse interconnection across different genes for each cellular subtype. Specifically, our mask-sc leverages gene expression matrix embeddings as input to the gene expression encoder for masked expression modeling to capture interconnections from grouped gene expression matrices in each cell. Then, a sequence-guided decoder is applied to predict sequence-level features of masked expression tokens, where sequence-level feature targets are extracted from a pre-trained sequence-level encoder. Compared to previous scRNA-seq clustering approaches, our method can extract discriminative representations from interconnections across gene expression matrices.

Empirical experiments on 15 real scRNA-seq datasets comprehensively validate the state-of-the-art performance against the previous scRNA-seq clustering baselines. In addition, qualitative visualizations of clustering results vividly showcase the effectiveness of our mask-sc in learning discriminative representation from Input Matrix Embeddings and Multimodal Encoder-Decoder. Extensive ablation studies also validate the importance of Input Matrix Embeddings for masked expression modeling and Multimodal Encoder-Decoder for sequence-guided reconstruction in learning compact expression representations for single-cell clustering of scRNA-seq data. Meanwhile, qualitative visualizations of learned attention of the gene expression encoder indicate diverse interconnections for different cellular subtypes.

Our main contributions can be summarized as follows:

- We present a novel multimodal self-supervised training framework with masked expression modeling for single-cell clustering, namely mask-sc, that can learn discriminative representations from interconnection across different genes for each cell.
- We introduce input expression matrix embeddings for masked modeling on grouped gene expression matrices and multimodal encoder-decoder for sequence-guided reconstruction.
- Extensive experiments on 15 real scRNA-seq benchmarks comprehensively demonstrate the state-of-the-art superiority of our mask-sc over previous baselines.

2 Related Work

Self-supervised Representation Learning. Self-supervised representation learning has been addressed in many previous works [18, 23, 20, 17, 19, 24] to learn discriminative representations from internal characteristics of data without any label. Such learnable and transferrable features are beneficial to many downstream tasks, such as image classification [18, 23, 20, 25, 26], object detection [17, 19, 24, 27, 28], semantic textual similarity tasks [21, 22, 29, 30, 31], protein structure prediction [32, 33, 34], and transcription factor binding sites prediction [35, 36, 37]. In this work, our main focus is to leverage a self-supervised training framework to learn compact gene expression representations from scRNA-seq data for identifying potential cell clusters, which is more challenging than those tasks listed above.

Masked Representation Learning. Masked representation learning is one type of unsupervised learning, which aims to learn self-supervised representations by reconstructing desired features of masked data given unmasked parts as clues. In the recent years, masked representation learning has achieved promising results in natural language processing [38, 39, 40, 41, 42] and computer vision [43, 44, 45, 46, 47] community. Typically, BERT [38] randomly masked 15% of word tokens in the sentence and recovered them with unmasked words to learn generalizable textual features via a self-attention transformer [48]. A block-wise masking strategy was proposed in BEiT [43] to reconstruct discrete tokens of masked image patches for pre-training transferrable visual representations. To simplify the masked image encoding framework, MAE [44] directly reconstructed missing pixels of 75% masked patches using vision transformers [49] for self-supervised pre-training. More recently, researchers introduced diverse masking pipelines to show the effectiveness of masked modeling in learning meaningful representations from video [50, 51], audio [52], and MRI/CT scans [53]. However, there is little work exploring the mask modeling methodology on single-cell data, especially on scRNA-seq datasets across various cells. To the best of our knowledge, we are the first to adopt masked gene expression modeling on scRNA-seq data for extracting transferrable representations to do scRNA-seq clustering on multiple cellular subpopulations.

scRNA-seq Clustering. scRNA-seq clustering is a challenging problem that predicts cellular subtype clusters from gene expression data of diverse cells. Early methods applied classical Principal Component Analysis (PCA) [4], K-means [5], and Gaussian Mixture Models [6] to cluster cell subpopulations from gene expression data directly. Because of the high dimensionality and sparsity of gene expression sequences, the following work [7, 8, 9, 10, 11, 12, 13, 14, 15, 16] explored diverse pipelines to tackle with this issue. For instance, hierarchical clustering with an implicit imputation stage was introduced in CIDR [7] before PCA to address the dropout problem. In order to classify both pure and transitional cells, SOUP [10] utilized the expression similarity matrix to estimate soft membership for cell-type cluster centers. In recent years, deep neural networks, such as DCA [12] have been widely used for extracting expression representations before clustering. Typically, ScDeepCluster [13] used a clustering layer on the embedding space from DCA to enrich embeddings for boosting the clustering performance. ScziDesk [14] proposed a soft self-training KMeans clustering to aggregate similar cells in the same cluster. However, those scRNA-seq clustering baselines mainly rely on well-annotated data, which limits their generalization and violates the fundamental goal of discovering potential cell subtype clusters. More recently, contrastive-sc [1] introduced the same self-supervised framework as SimCLR [18] to extract embeddings of a short gene expression sequence by an InfoNCE-based contrastive loss. Different from contrastive-sc [1], we develop a novel multimodal self-supervised framework to learn compact and discriminative representations by reconstructing sequence-level features of masked gene expression matrices for

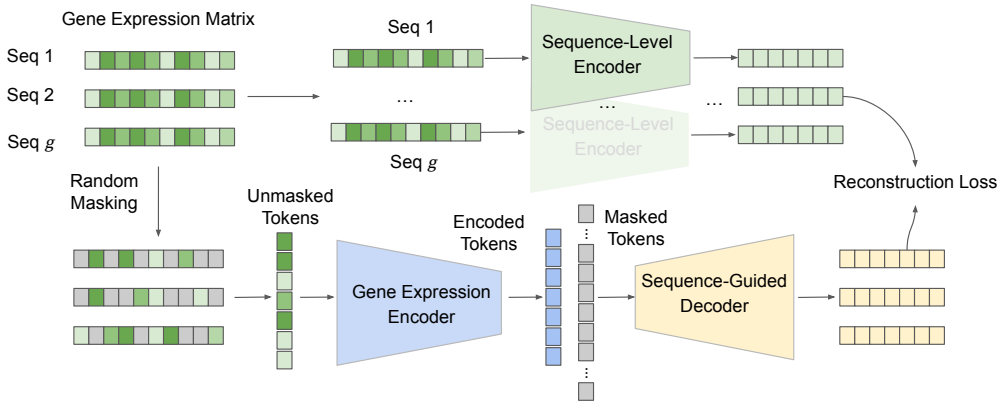

Figure 2: Illustration of the proposed mask-sc for mask self-supervised training on a gene expression matrix with g sequences. The gene expression encoder takes as input unmasked tokens generated from random masking on the expression matrix. With encoded expression tokens, a sequence-guided decoder is utilized to predict sequence-level features of masked expression tokens. Note that all targeted sequence-level features are extracted from a pre-trained sequence-level encoder, and this pre-trained encoder is frozen during training. Finally, a reconstruction loss is minimized between the predicted and targeted sequence-level features, which pushes the gene expression encoder to extract discriminative representations from the gene expression matrix.

scRNA-seq clustering. In addition, masked gene expression modeling is helpful for aggregating interconnections across multiple groups of expression sequences.

3 Method

Given a set of gene expression data from scRNA-seq, our target is to learn a gene expression encoder to extract gene expression embeddings for scRNA-seq clustering. In this work, we propose a novel mask-based self-supervised training framework named mask-sc for extracting compact and discriminative representations from single-cell data, which mainly consists of two modules, Input Matrix Embeddings for masked expression matrix modeling in Section 3.2 and Multimodal Encoder-Decoder for sequence-guided reconstruction in Section 3.3.

3.1 Preliminaries

In this section, we first describe the problem setup and notations and then revisit contrastive-sc, the state-of-the-art self-supervised baseline for scRNA-seq clustering.

Problem Setup and Notations. Given a set of gene expression data with m genes from n cells, our goal is to learn a discriminative global gene expression feature from m genes in each cell. Note that m, n denote the number of genes and cells, respectively. D denotes the dimension of embeddings.

Revisit contrastive-sc. To solve the single-cell representation learning problem, contrastive-sc [1] first extracted a sequence-level embedding $\mathbf{s} \in \mathbb{R}^{s \times D}$ with a length of s from a gene expression encoder, and then concatenated all m/s sequences as the whole feature for m genes. During training, they utilized a self-supervised framework similar to SimCLR [18] with an InfoNCE-based contrastive loss to close the distance between the anchor and augmented embeddings of one sequence in the same cell, which is denoted as:

$$\mathcal{L}_{\text{contrastive-sc}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{s}_i, \hat{\mathbf{s}}_i)\right)}{\sum_{j=1}^{B}\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{s}_i, \hat{\mathbf{s}}_j)\right)} \tag{1}$$

where $\mathbf{s}_i, \hat{\mathbf{s}}_i \in \mathbb{R}^{1\times D}$ denote the anchor and augmented embeddings for ith sample in a min-batch. B is the batch size. $\texttt{sim}(\mathbf{s}_i, \hat{\mathbf{s}}_i) = \mathbf{s}_i^T\hat{\mathbf{s}}_i/(\|\mathbf{s}_i\|\|\hat{\mathbf{s}}_i\|)$ is the cosine similarity, and τ is the temperature parameter. $B^2 - B$ negative sequences are created within a training batch. By optimizing this loss, they successfully extracted discriminative representations of each short sequence in the same cell.

However, this sequence-level contrastive learning framework can not fully capture the interconnections between expressions of different genes for each cell, as the length of each sequence s is much less than the total number of genes m in the same cell, *i.e.*, $s \ll m$. To address this issue, we propose a novel mask-based single-cell self-supervised training framework that can aggregate the inter-connection across multiple groups of gene expressions from a gene expression matrix with grouping g sequences, as shown in Figure 2.

3.2 Input Matrix Embeddings for Masked Expression Modeling

In order to explicitly learn the interconnections from the grouped gene expression matrix for each cell, we introduce learnable matrix patch embeddings that are extracted from raw expression input via a 2D convolutional layer, *i.e.*, $\mathbf{x} \in \mathbb{R}^{G \times D}$, where G denotes the total number of patches for each gene matrix. Assume the patch resolution of each matrix is P, the patch-wise raw input for the gene matrix is formally denoted as $\mathbf{g} \in \mathbb{R}^{G \times P \times P}$. Note that $G = g/P \times s/P$.

With local-patch expression representations $\{\mathbf{x}_i\}_{i=1}^{G}$ for the gene expression matrix, we first apply a self-attention transformer encoder $\phi(\cdot)$ to aggregate the patch-level features from the raw input matrix embeddings and align the features with the global token embedding $\mathbf{c} \in \mathbb{R}^{1 \times D}$ as:

$$\begin{aligned} \hat{\mathbf{c}}, \{\hat{\mathbf{x}}_i\}_{i=1}^{G} &= \{\phi(\mathbf{x}_j, \mathbf{X}, \mathbf{X})\}_{j=1}^{1+G} \\ \mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{G} &= [\mathbf{c}; \{\mathbf{x}_i\}_{i=1}^{G}] \end{aligned} \tag{2}$$

where $[\ ;\ ]$ denotes the concatenation operator. $\hat{\mathbf{c}}, \hat{\mathbf{x}}_i \in \mathbb{R}^{1 \times D}$, and D is the dimension size of embeddings. The self-attention operator is formulated as

$$\phi(\mathbf{x}_j, \mathbf{X}, \mathbf{X}) = \text{Softmax}(\frac{\mathbf{x}_j \mathbf{X}^\top}{\sqrt{D}})\mathbf{X} \tag{3}$$

Then, to capture the interaction across each patch of expression, we exploit a random masking mechanism and a decoder to predict the masked patch tokens given encoded tokens as clues. Specifically, a random mask with M entries as zero is applied along the number dimension of patch-level expression embeddings $\mathbf{x} \in \mathbb{R}^{G \times D}$. Let the number of unmasked patches be U, then we apply the self-attention transformer encoder $\phi(\cdot)$ to aggregate unmasked patch-level embeddings with the global token embeddings as:

$$\begin{aligned} \hat{\mathbf{c}}_u, \{\hat{\mathbf{x}}_i^u\}_{i=1}^{U} &= \{\phi(\mathbf{x}_j^u, \mathbf{X}^u, \mathbf{X}^u)\}_{j=1}^{1+U} \\ \mathbf{X}^u = \{\mathbf{x}_j^u\}_{j=1}^{U} &= [\mathbf{c}; \{\mathbf{x}_i^u\}_{i=1}^{U}] \end{aligned} \tag{4}$$

where $\hat{\mathbf{c}}_u, \hat{\mathbf{x}}_i^u \in \mathbb{R}^{1 \times D}$ denote the global and patch-level tokens encoded from unmasked input expression embeddings. Finally, those encoded expression tokens $\{\hat{\mathbf{x}}_i^u\}_{i=1}^{U}$ are passed into the decoder to generate the targeted features of masked patch tokens. With masked expression modeling, the encoder is optimized to capture the interconnections across different genes in each cell. Furthermore, the pre-trained encoder can learn discriminative representations from input gene expression matrix embeddings for discovering potential cellular subtypes. It is worth noting that, to the best of our knowledge, we are the first to apply masked self-supervised training on expression matrices for scRNA-seq clustering.

3.3 Multimodal Encoder-Decoder for Sequence-guided Reconstruction

With the benefit of masked expression modeling from input matrix embeddings mentioned above, we propose a novel and explicit sequence-guided decoder to reconstruct the sequence-level features of masked tokens. Based on encoded tokens $\{\hat{\mathbf{x}}_i^u\}_{i=1}^{U}$ and learnable masked tokens $\{\mathbf{m}_i\}_{i=1}^{M}$, a light self-attention transformer decoder $\phi(\cdot)$ is applied to generate sequence-level embeddings $\{\hat{\mathbf{s}}_i^m\}_{i=1}^{M}$ as:

$$\begin{aligned} \hat{\mathbf{c}}^d, \{\hat{\mathbf{s}}_i^u\}_{i=1}^{U}, \{\hat{\mathbf{s}}_i^m\}_{i=1}^{M} &= \{\phi(\mathbf{x}_j^d, \mathbf{X}^d, \mathbf{X}^d)\}_{j=1}^{1+U} \\ \mathbf{X}^d = \{\mathbf{x}_j^d\}_{j=1}^{1+U+M} &= [\hat{\mathbf{c}}; \{\hat{\mathbf{x}}_i^u\}_{i=1}^{U}; \{\mathbf{m}_i\}_{i=1}^{M}] \end{aligned} \tag{5}$$

where $\hat{\mathbf{c}}^d, \{\hat{\mathbf{s}}_i^u\}_{i=1}^{U}, \{\hat{\mathbf{s}}_i^m\}_{i=1}^{M}$ denote the global, unmasked, masked prediction from the decoder. To recover the sequence-level features, we apply the pre-trained contrastive-sc [1] as the sequence-level encoder to extract the target features $\{\mathbf{s}_i^m\}_{i=1}^{M}$ for masked patches of gene expression. Note that

Table 1: Comparison results (%) on 10 PBMC cells and worm neuron cells datasets. ↑ denotes that a large value is better.

Method	10 PBMC cells			worm neuron cells		
	ARI (↑)	NMI (↑)	Silhouette (↑)	ARI (↑)	NMI (↑)	Silhouette (↑)
PCA [4]	19.00	31.21	40.24	4.65	17.56	17.69
scRNA [9]	47.15	52.91	23.27	24.58	41.52	10.85
ScDeepCluster [13]	51.54	65.98	48.52	42.40	59.23	23.63
SczíDesk [14]	55.20	68.67	14.46	4.95	22.02	19.78
contrastive-sc [1]	68.63	72.54	58.73	45.03	58.80	27.83
mask-sc (ours)	**80.64**	**79.23**	**69.41**	**57.77**	**71.90**	**48.12**

this sequence-level encoder is frozen for stabilizing training. The overall model with a multimodal encoder-decoder architecture is simply optimized to reconstruct the sequence-level features of masked expression tokens as:

$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^{M} ||\mathbf{s}_i^m - \hat{\mathbf{s}}_i^m||_2^2 \tag{6}$$

Optimizing the loss will promote the model to capture discriminatively sequence-level features by learning expression interconnections across different genes for each cell. During inference, we use the pre-trained self-attention encoder to extract patch-level features $\{\hat{\mathbf{x}}_i\}_{i=1}^{G}$ from input gene expression matrices. Mean pooling is finally applied along the number dimension of the encoded patch-level representations to generate the global feature for each cell.

4 Experiments

4.1 Experimental Setup

Datasets. The 15 real scRNA-seq datasets contain 7 datasets [54] for multiple mouse organs (4 from Smart-seq2 sequencing prefixed with "Quake Smart seq2" and 3 from 10x Genomics sequencing prefixed with "Quake 10x"), 4 public available datasets from human organs (Adam [55], Muraro [56], Romanov [57] and Young [58]), and 4 scDeepCluster datasets from different sequencing platforms (10 PBMC cells [59] from 10x genomic platform, Mouse embryonic stem cells [60] from droplet barcoding, Mouse bladder cell [61] from Microwell-seq, and Worm neuronal cells [62] from sci-RNA-seq. We use the same split in [1] for training and testing, where the number of cells varies from 870 to 9552, and 4-16 cellular subtype clusters are annotated for evaluation.

Evaluation Metrics. Following previous work [1, 13, 14], we apply Adjusted Rand Index (ARI) [63], Normalized Mutual Information (NMI) [64], Silhouette [65] score and Calinski-Harabasz [66] score for evaluation. ARI score calculates the ratio of sample pairs assigned to the correct cluster labels. NMI score measures the agreement of the ground truth and predicted cluster assignments. A larger value of ARI and NMI is better, which means that the predicted cluster matches the ground-truth cluster. Silhouette score measures the compactness of the generated clusters, and a higher score means that the predicted clusters are denser and better separated. Calinski-Harabasz score computes the proportion of the sum between inter-cluster and intra-cluster dispersion for all clusters, which produces an unbounded positive score value. Please see more results about Calinski-Harabasz score in Appendix.

Implementation. Our implementation is based on PyTorch [67] framework. For each gene expression matrix with a total of 4096 expression entries, we apply a patch size of 4×4 to generate 256 patch-level expression embeddings, *i.e.*, $g \times s = 4096, P = 4, G = 256$. The mask ratio is 75% in our experiments, *i.e.*, $U = 64, M = 192$. For the gene expression encoder, the depth of self-attention transformers is 12, and the number of attention heads is 12. For the sequence-guided decoder, the depth of self-attention transformers is 8, and the number of attention heads is 16. The dimension size of embeddings is 768, *i.e.*, $D = 768$. The model is trained with the AdamW [68] optimizer with a learning rate of 1e-4, and hyper-parameters β_1 = 0.9, β_2 = 0.95. The model is trained for 1600

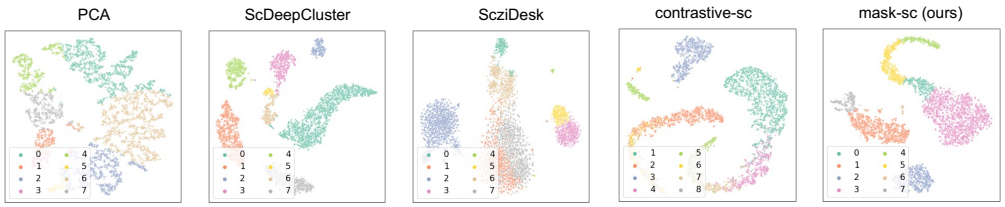

Figure 3: Qualitative comparisons with PCA [4], ScDeepCluster [13], SczíDesk [14], contrastive-sc [1] on 10 PBMC cells dataset. The proposed mask-sc generates much more intra-subtype compact and inter-subtype separable representations for single-cell clustering. Note that each spot denotes the feature of one cell, and each color refers to one cellular subtype, such as "subtype 5" in yellow and "subtype 7" in gray.

epochs with a batch size of 256. After self-supervised training, we use the K-means algorithm [5] in the scikit-learn package for evaluation to cluster gene expression representations.

4.2 Comparison to Prior Work

In this work, we propose a novel and effective self-supervised training framework for scRNA-seq clustering. To demonstrate the effectiveness of the proposed mask-sc, we comprehensively compare it to previous scRNA-seq clustering baselines: 1) PCA [4]: a traditional machine learning approach with raw gene expression sequence as input to extract principal components; 2) scRNA [9]: a baseline based on non-negative matrix factorization by using transferred knowledge from large and well-annotated data for reference; 3) ScDeepCluster [13]: an improved deep count autoencoder by adding a clustering layer to the embedding space; 4) SczíDesk [14]: a soft KMeans clustering method to aggregate similar cells from the same cellular subtype; 5) contrastive-sc [1]: the state-of-the-art self-supervised framework with InfoNCE-based contrastive loss to extract embeddings from sequences only for scRAN-seq clustering.

Table 1 reports the quantitative comparison results on 10 PBMC cells and worm neuron cells datasets. As can be seen, we achieve the best performance in terms of all metrics compared to previous scRAN-seq clustering baselines on 10 PBMC cells benchmark. In particular, the proposed mask-sc significantly outperforms PCA [4], the traditional machine learning approach, by 61.64 ARI, 48.02 NMI, and 29.17 Silhouette. Moreover, we achieve superior performance gains of 33.49 ARI, 26.32 NMI, and 46.14 Silhouette compared to scRNA [9], which indicates the importance of masked self-supervised training for learning discriminative representations from interconnections across different genes in each cell. Meanwhile, our mask-sc outperforms contrastive-sc, the current state-of-the-art self-supervised approach for scRNA-seq clustering, where we achieve the performance gains of 12.01 ARI, 6.69 NMI, and 10.68 Silhouette. These significant improvements demonstrate the superiority of our method in learning compact embeddings from gene expression for clustering.

In addition, significant gains in worm neuron cells benchmark can be observed in Table 1. Compared to SczíDesk [14], the recent baseline based on soft KMeans clustering, we achieve the results gains of 52.82 ARI, 49.88 NMI, and 28.34 Silhouette. Furthermore, when evaluated on this challenging benchmark with more cellular subtypes, the proposed approach still outperforms contrastive-sc [1] by 12.74 ARI, 13.10 NMI, and 20.29 Silhouette. We also achieve highly better results against ScDeepCluster [13], the improved clustering baseline by adding a clustering layer to the embedding space in deep count autoencoder. These results validate the effectiveness of our approach in learning discriminative features with patch-level interconnections from gene expression matrices for each cellular subtype.

To qualitatively evaluate scRNA-seq clustering, we compare the proposed mask-sc with PCA [4], ScDeepCluster [13], SczíDesk [14], and contrastive-sc [1] on the 10 PBMC cells dataset with 8 cellular subtypes in Figure 3. To better evaluate the quality, we visualize the learned representations of each cellular subtype by t-SNE [69]. Note that each spot denotes the feature of one cell, and each color refers to one cellular subtype, such as "subtype 5" in yellow and "subtype 7" in gray. From comparisons, three main observations can be derived: 1) with the raw gene expression sequence as input, PCA [4], the traditional machine learning approach, fails to predict some cellular subtypes, including "subtype 0" and "subtype 4"; 2) the quality of cellular subtypes generated by our method is much better than the strong self-supervised baseline, contrastive-sc [1]. 3) the proposed mask-

Table 2: Ablation studies on Input Matrix Embeddings (IME) and Multimodal Encoder-Decoder (MED).

IME	MED	ARI (↑)	NMI (↑)	Silhouette (↑)
✗	✗	68.63	72.54	58.73
✓	✗	72.98 (**+4.35**)	75.80 (**+3.26**)	60.39 (**+1.66**)
✗	✓	74.15 (**+5.52**)	77.56 (**+5.02**)	62.37 (**+3.64**)
✓	✓	80.64 (**+12.01**)	79.23 (**+6.69**)	69.41 (**+10.68**)

sc achieves competitive even better results on clustering results against ScDeepCluster [13], the improved deep count autoencoder with a clustering layer added to the embedding space. Furthermore, as can be observed in the last column, gene expression representations extracted by the proposed mask-sc are both intra-subtype compact and inter-subtype separable. In contrast to our compact embeddings in the cellular subtype semantic space, there still exists mixtures of multiple cellular subtypes among features learned by SczíDesk [14]. These meaningful visualizations further showcase the superiority of our mask-sc with input matrix embedding and multi-modal encoder-decoder for masked expression modeling in extracting compact gene expression representations for scRNA-seq clustering.

4.3 Experimental Analysis

In this section, we performed ablation studies to validate the benefit of introducing the Input Matrix Embeddings for masked expression modeling and Multimodal Encoder-Decoder for sequence-guided reconstruction. We also conducted extensive experiments to explore the effect of input expression matrix size and mask ratio on scRNA-seq clustering. Furthermore, we visualized learned attention of self-attention layers in the gene expression encoder after self-supervised training.

Input Matrix Embeddings & Multimodal Encoder-Decoder. To demonstrate the effectiveness of the introduced Input Matrix Embeddings for masked expression modeling (IME) and Multimodal Encoder-Decoder for sequence-guided reconstruction, we ablate the necessity of each module and report the quantitative comparison results in Table 2. We can observe that adding IME to the vanilla baseline highly raises the results of scRNA-seq clustering by 4.35 ARI, 3.26 NMI, and 1.66 Silhouette, which validates the benefit of Input Matrix Embeddings in learning discriminative expression representations for discovering accurate cellular subtypes. Meanwhile, introducing only MED in the baseline also increases the clustering performance in terms of all metrics, which indicates the importance of reconstructing sequence-level features for scRNA-seq clustering. More importantly, incorporating IME for masked expression modeling and MED for sequence-guided reconstruction together into the baseline significantly raises the performance by 12.01 ARI, 6.69 NMI, and 10.68 Silhouette. These improving results demonstrate the importance of IME for masked expression modeling and MED for sequence-guided reconstruction in learning discriminative representations with interconnections across different genes for each cell.

Effect of expression matrix size and mask ratio. The expression matrix size and mask ratio used in the proposed masked expression modeling approach affect the extracted expression representations for scRNA-seq clustering. To explore such effects more comprehensively, we ablated the expression matrix size from $\{36 \times 36, 64 \times 64, 80 \times 80, 100 \times 100\}$ and varied the mask ratio from $\{10\%, 25\%, 50\%, 75\%, 90\%\}$. The comparison results of scRNA-seq clustering are shown in Figure 4. When the expression matrix size is 64×64 and the mask ratio is 75%, we achieve the best clustering performance in terms of all metrics. With the increase of expression matrix size from 36×36 to 64×64, the proposed mask-sc consistently raises results, which shows the importance of utilizing input matrix embeddings for masked expression modeling to learning compact representation from interconnections of more genes in each cell. However, increasing the expression matrix size from 64×64 to 80×80 and 100×100 will not continually improve the results of ARI and NMI. In particular, a drastic drop can be observed in the Silhouette score, which means that the generated clusters are not dense and well-separable. This is might be caused by the high sparsity of non-zero values in the gene expression matrix. In this case, masked expression modeling without

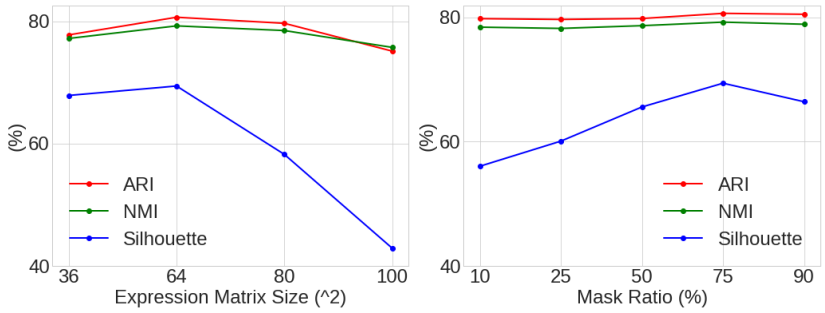

Figure 4: Effect of gene expression matrix size and mask ratio on the final performance of scRNA-seq clustering. The proposed mask-sc achieves the best clustering performance in terms of all metrics when the expression matrix size is 64×64 and the mask ratio is 75%.

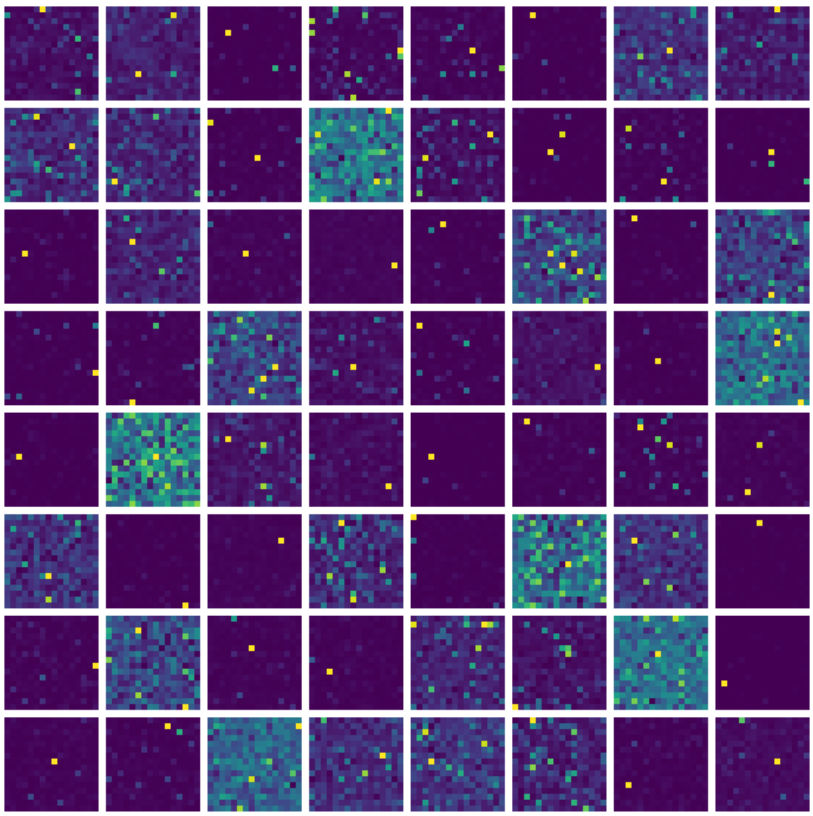

Figure 5: Visualization of input gene expression matrices with attention from the last attention layer of the self-supervised gene expression encoder trained on 10 PBMC cells dataset. We can observe that the pre-trained encoder can learn diverse interconnections across genes for each input expression matrix. Similar cellular subtypes have similar interconnection patterns, such as expression matrices at the locations (Row 2, Column 4), (Row 5, Column 2), and (Row 6, Column 6).

discriminating this sparsity will deteriorate the quality of expression representations pre-trained from many zero entries in the input matrix.

In terms of mask ratio, the performance of the proposed mask-sc climbs with the increase of the mask ratio from 10% to 75%. Compared to the Silhouette score, there are no significant changes in ARI and NMI. This interesting trend could be due to the self-property of these metrics. Higher ARI and NMI indicate that the predicted cluster assignment matches the ground-truth cluster assignment, while a larger value of the Silhouette score refers to denser and better-separated clusters. The

former metrics can not measure the quality of expression embeddings extracted from the pre-trained gene expression encoder, but the latter is a strict metric for measuring the compactness of learned expression representations for scRNA-seq clustering. Meanwhile, when the mask ratio is increased to 90%, the Silhouette score drops significantly but ARI and NMI decrease at an insignificant range. This decreasing trend is also observed in the highly influential masked image modeling approach [44]. For masked expression modeling, we are the first to observe such an effect on input matrix embeddings for extracting discriminative representations from gene expression matrices to conduct clustering on single-cell data.

Learned attention showcases diverse interconnections. Diverse interconnections across different genes are essential for us to learn compact and discriminative representations of scRNA-seq clustering. To better understand how the model captures interconnections across genes during masked expression modeling, we visualized input gene expression matrices with learned attention from the last self-attention layer of the gene expression encoder in Figure 5. Specifically, we computed the mean attention value of the last self-attention layer with 12 heads in the gene expression encoder. As can be seen, the pre-trained encoder can successfully learn diverse interconnections across different genes for each input gene expression matrix. In addition, similar interconnection patterns can be observed in the same cellular subtypes. For instance, the gene expression matrix at the location (Row 2, Column 4) has a similar interconnection pattern as the ones at the location (Row 5, Column 2) and location (Row 6, Column 6). Meanwhile, these three gene expression matrices are indeed from the same cellular subtype. Another similar case can be seen in gene expression matrices at location (Row 5, Column 1), location (Row 3, Column 1), and location (Row 5, Column 5).

Overall, these meaningful visualizations further demonstrate the effectiveness of the proposed mask-sc with input matrix embeddings for masked expression modeling and multimodal encoder-decoder for sequence-guided reconstruction in learning discriminative representations from interconnects across various genes for scRNA-seq clustering. When it comes to future applications, attention-based interconnection patterns potentially provide a discriminative input to replace both raw expression sequence and sequence features for discovering cellular subtypes. The benefit of using attention-based interconnections for scRNA-seq clustering will not be dependent on well-annotated data from human experts. Instead, these interconnections still can be derived from unsupervised pre-training from large-scale gene expression data without high-quality annotations, which is much cheaper than labeling cellular subtypes from high-dimensional gene expression data for each cell.

4.4 Limitation

Although the proposed mask-sc achieves superior results on single-cell clustering of scRNA-seq data, the performance gains of our approach on human organs are not significant. One possible reason is that our model easily overfits the task of masked expression modeling during the training stage, and the solution is to incorporate dropout and momentum encoders together for masked expression modeling. Meanwhile, we notice that our model performs worse on the "Quake Smart seq2 Limb Muscle" dataset with a smaller training size. The future work could be to add more training data or incorporate contrastive learning with masked expression modeling to increase the accuracy and compactness of generated clusters.

5 Conclusion

In this work, we present mask-sc, a novel multimodal self-supervised framework with masked expression modeling on single-cell data, that can learn discriminative gene expression representations by reconstructing sequence-level features of masked expressions for scRNA-seq clustering. Our mask-sc potentially aggregates interconnections across grouped expression sequences through a self-supervised expression encoder with random masked gene expression matrices as input. Then, we leverage a sequence-guided decoder is leveraged to reconstruct sequence-level features of masked expression matrices. After self-supervised training, the pre-trained gene expression encoder is applied to extract representations for scRNA-seq clustering. Extensive experiments on 15 real scRNA-seq datasets demonstrate the state-of-the-art performance of our proposed method, compared to previous baselines. Qualitative visualizations of clustering results and attention showcase the effectiveness of our masks-sc in learning compact representations for intra-subtype cells and diverse interconnections for inter-subtype cells.

References

[1] Madalina Ciortan and Matthieu Defrance. Contrastive self-supervised clustering of scrna-seq data. *BMC bioinformatics*, 22(1):1–27, 2021. 1, 2, 3, 4, 5, 6, 7, 16

[2] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015. 1

[3] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009. 1

[4] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. 1, 2, 3, 6, 7

[5] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297, 1967. 1, 3, 7

[6] Carl Rasmussen. The infinite gaussian mixture model. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 1999. 1, 3

[7] Peijie Lin, Michael Troup, and Joshua W. K. Ho. Cidr: Ultrafast and accurate clustering through imputation for single cell rna-seq data. *Genome Biology*, 18:59, 2017. 1, 3

[8] Yunpei Xu, Hong-Dong Li, Yi Pan, Feng Luo, and Jianxin Wang. Biorank: A similarity assessment method for single cell clustering. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 157–162. IEEE, 2018. 1, 3

[9] Bettina Mieth, James R F Hockley, Nico Görnitz, Marina M-C Vidovic, Klaus-Robert Müller, Alex Gutteridge, and Daniel Ziemek. Using transfer learning from prior reference knowledge to improve the clustering of single-cell rna-seq data. *Scientific reports*, 9(1):20353, 2019. 1, 2, 3, 6, 7

[10] Lingxue Zhu, Jing Lei, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471, 2019. 1, 3

[11] Xiaoshu Zhu, Lilu Guo, Yunpei Xu, Hong-Dong Li, Xingyu Liao, Fang-Xiang Wu, and Xiaoqing Peng. A global similarity learning for clustering of single-cell rna-seq data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 261–266. IEEE, 2019. 1, 3

[12] 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 Communication*, 10:390, 2019. 1, 3

[13] Tian Tian, Wan Ji, Song Qi, and Wei Zhi. Clustering single-cell rna-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191–198, 2019. 1, 2, 3, 6, 7, 8

[14] Chen Liang, Wang Weinan, Zhai Yuyao, and Deng Minghua. Deep soft k-means clustering with self-training for single-cell rna sequence data. *NAR Genom Bioinform*, 2020. 1, 2, 3, 6, 7, 8

[15] Ruiqing Zheng, Zhenlan Liang, Xiangmao Meng, Yu Tian, and Min Li. A robust single cell clustering method based on subspace learning and partial imputation. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 140–145. IEEE, 2020. 1, 3

[16] Florian Schmidt and Bobby Ranjan. Robust clustering and interpretation of scrna-seq data using reference component analysis. *bioRxiv*, 2020. 1, 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2, 3

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020. 2, 3, 4, 16

[19] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3

[20] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[21] Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, 2020. 2, 3

[22] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910, 2021. 2, 3

[23] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[24] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[25] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of International Conference on Machine Learning (ICML)*, 2021. 3

[26] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 3

[27] Shentong Mo, Zhun Sun, and Chao Li. Siamese prototypical contrastive learning. In *Proceedings of British Machine Vision Conference (BMVC)*, 2021. 3

[28] Shentong Mo, Zhun Sun, and Chao Li. Rethinking prototypical contrastive learning through alignment, uniformity and correlation. In *Proceedings of British Machine Vision Conference (BMVC)*, 2022. 3

[29] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 5065–5075, 2021. 3

[30] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022. 3

[31] Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3898–3907, 2022. 3

[32] Amy X. Lu, H. Zhang, Marzyeh Ghassemi, and Alan M. Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *bioRxiv*, 2020. 3

[33] Michael Heinzinger, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics and Bioinformatics*, 4(2), 2022. 3

[34] Yang Li, Guanyu Qiao, Xin Gao, and Guohua Wang. Supervised graph co-contrastive learning for drug–target interaction prediction. *Bioinformatics*, 38(10):2847–2854, 2022. 3

[35] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021. 3

[36] Shentong Mo, Xiao Fu, Chenyang Hong, Yizhen Chen, Yuxuan Zheng, Xiangru Tang, Zhiqiang Shen, Eric P. Xing, and Yanyan Lan. Multi-modal self-supervised pre-training for regulatory genome across cell types. *arXiv preprint arXiv:2110.05231*, 2021. 3

[37] Weizhi An, Yuzhi Guo, Yatao Bian, Hehuan Ma, Jinyu Yang, Chunyuan Li, and Junzhou Huang. Modna: Motif-oriented pre-training for dna language model. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2022. 3

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pref-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[40] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019. 3

[41] Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[42] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*, 2022. 3

[43] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[44] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 3, 10

[45] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[46] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[47] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for BERT pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 3

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, page 5998–6008, 2017. 3

[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 3

[50] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[51] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv:2205.09113*, 2022. 3

[52] Po-Yao Huang, Hu Xu, Juncheng Billy Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *arXiv:2207.06405*, 2022. 3

[53] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, Venkat S. Sethuraman, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. 3

[54] Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018. 6

[55] Mike Adam, Andrew S Potter, and S Steven Potter. Psychrophilic proteases dramatically reduce single-cell rna-seq artifacts: a molecular atlas of kidney development. *Development*, 144(19):3625–3632, 2017. 6

[56] Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon Van Gurp, Marten A Engelse, Francoise Carlotti, Eelco Jp De Koning, et al. A single-cell transcriptome atlas of the human pancreas. *Cell systems*, 3(4):385–394, 2016. 6

[57] Roman A Romanov, Amit Zeisel, Joanne Bakker, Fatima Girach, Arash Hellysaz, Raju Tomer, Alan Alpar, Jan Mulder, Frederic Clotman, Erik Keimpema, et al. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nature neuroscience*, 20(2):176–188, 2017. 6

[58] Matthew D Young, Thomas J Mitchell, Felipe A Vieira Braga, Maxine GB Tran, Benjamin J Stewart, John R Ferdinand, Grace Collord, Rachel A Botting, Dorin-Mirel Popescu, Kevin W Loudon, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science*, 361(6402):594–599, 2018. 6

[59] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017. 6

[60] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015. 6

[61] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018. 6

[62] Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017. 6

[63] Hubert Lawrence and Arabie Phipps. Comparing partitions. *Journal of Classification*, 2:193–218, 1985. 6

[64] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002. 6

[65] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 6

[66] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. 6

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6

[68] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 6

[69] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 7

Table 3: Comparison results of the **MLP-Encoder-Decoder** model on 15 real scRNA-seq datasets. ↑ denotes that large value is better, except for Calinski-Harabasz.

Dataset	ARI (↑)	NMI (↑)	Silhouette (↑)	Calinski-Harabasz
Quake Smart seq2 Trachea	0.82	0.76	0.44	402.23
Quake Smart seq2 Lung	0.67	0.8	0.52	723.15
Quake Smart seq2 Diaphragm	0.97	0.95	0.71	1372.80
Quake Smart seq2 Limb Muscle	0.97	0.95	0.74	1672.58
Quake 10x Bladder	0.73	0.77	0.64	2296.94
Quake 10x Spleen	0.91	0.83	0.67	5378.69
Quake 10x Limb Muscle	0.99	0.98	0.43	1882.33
Adam	0.83	0.84	0.44	1289.16
Muraro	0.92	0.87	0.78	2728.00
Romanov	0.74	0.72	0.50	1441.03
Young	0.75	0.81	0.45	1942.5
10 PBMC cells	0.74	0.77	0.52	4631.53
Mouse ES cells	0.75	0.73	0.59	2705.81
Mouse bladder cells	0.52	0.74	0.43	2872.66
Worm neuron cells	0.37	0.53	0.14	319.13

Appendix

In this appendix, we provide a trivial mask-sc version with an MLP-Encoder-Decoder architecture and more analysis about the state-of-the-art self-supervised baseline, contrastive-sc [1]. We also give the detailed work division and milestones for this work.

A MLP-Encoder-Decoder

In this section, we implemented a trivial mask-sc version with an MLP-Encoder-Decoder network that consists of a multi-layer linear encoder and a single-layer linear decoder. There are three linear layers in the encoder, of which the input sequence's length is 500, the number of prior selected most highly variable genes of each sample, and the output encoded representation is of length 60. When evaluating, the output of this linear encoder is then fed into a clustering algorithm. When training, the output representation is decoded to an expression sequence of which the length is 500 by the single-layer decoder and the loss function is the mean squared error of the decoded sequence and the original gene sequence.

Experimental results on all 15 real scRNA-seq datasets in Table 3 show that the simple structured MLP-Encoder-Decoder performs similarly to the contrastive-sc method on average when measured by the metrics of ARI and NMI scores. However, it cannot make Silhouette and Calinski-Harabasz score better. Since the ARI and NMI scores are supervised and the other two are unsupervised, we conclude that the MLP-Encoder-Decoder is good at the task of classification but relatively bad at the task of clustering without supervised labels.

B More analysis about contrastive-sc [1]

In this part, we reproduced the state-of-the-art baseline named contrastive-sc, which utilized a self-supervised framework similar to SimCLR [18] with an InfoNCE-based contrastive loss to bring close anchor and augmented embeddings of the same sequence gene expression in one cell and to push away embeddings from other sequences.

Table 4 reports the comparison results on 15 real scRNA-seq datasets. As can be seen on "Quake Smart seq2" series datasets, the baseline achieves worse performance on Lung organ than other organs in terms of ARI, NMI, and Silhouette scores. This is might be caused by the limited capacity of the shallow encoder architecture, as more non-zero values exist on the input gene expression sequences on average. A similar trend can be observed on "Quake 10x Spleen" dataset. In the meanwhile, contrastive-sc performs the best result on the "Quake 10x Limb Muscle" dataset. For example, the approach achieves 0.99 ARI and 0.98 NMI, which means that the predicted clustering result is almost the same as the ground-truth clusters.

Table 4: Comparison results of the state-of-the-art approach (**contrastive-sc**) on 15 real scRNA-seq datasets. ↑ denotes that large value is better, except for Calinski-Harabasz.

Dataset	ARI (↑)	NMI (↑)	Silhouette (↑)	Calinski-Harabasz
Quake Smart seq2 Trachea	0.86	0.84	0.58	856.94
Quake Smart seq2 Lung	0.59	0.76	0.53	1512.29
Quake Smart seq2 Diaphragm	0.97	0.94	0.80	4754.66
Quake Smart seq2 Limb Muscle	0.97	0.95	0.79	3992.59
Quake 10x Bladder	0.75	0.78	0.80	7235.09
Quake 10x Spleen	0.57	0.67	0.39	13876.19
Quake 10x Limb Muscle	0.99	0.98	0.61	3196.62
Adam	0.79	0.82	0.49	2131.84
Muraro	0.86	0.82	0.64	4602.89
Romanov	0.73	0.72	0.52	2770.27
Young	0.62	0.73	0.46	3106.40
10 PBMC cells	0.68	0.72	0.60	16724.88
Mouse ES cells	0.69	0.66	0.57	3917.39
Mouse bladder cells	0.42	0.63	0.39	4315.95
Worm neuron cells	0.16	0.30	0.29	2623.48

When it comes to 4 public human organ datasets, contrastive-sc does not achieve very high scores on both ARI and NMI. The averaged overall score is lower than average scores for "Quake Smart seq2" series datasets from mouse organs. This is because the total number of cells in the training set is less than those in "Quake Smart seq2" series datasets. Furthermore, those human organ datasets have more ground-truth clusters for testing, which makes it hard to predict the same clustering result as the annotated clusters. In terms of scDeepCluster datasets from 4 different sequencing platforms, contrastive-sc achieves decent results on both 10 PBMC cells and Mouse ES cells. However, the performance of Mouse bladder cells and Worm neuron cells is worse, especially on Worm neuron cells (0.16 ARI, 0.30 NMI, and 0.29 Silhouette scores). Moreover, the baseline achieve a very high Calinski-Harabasz score on 10 PBMC cells, which could be due to the significantly large inter-cluster dispersion distance for all predicted clusters. More importantly, they only took a relatively short gene expression sequence as one anchor sample in contrastive loss for each cell, which leads to bad capacity of modeling interconnections across expressions of long sequences for clustering.

C Teammates & Work Division

Shentong Mo: edit reports, review literature, and implement the proposed mask-sc.

Yixuan Chen: edit reports, review literature, and preprocess 15 scRNA-seq datasets.

Ding Bai: edit reports, review literature, and run baseline model.

D Project Milestone

Week 4 • Submit the project proposal. ✓
Week 11 • Reproduce the baseline and finish the midway report. ✓
Week 13 • Implement our proposed mask-sc and run experiments. ✓
Week 16 • Wrap up experimental results and finish the final report. ✓