A Simple and Comprehensive Benchmark for Single-Cell Transcriptomics

Jiaxin Qi[1*], Yan Cui[2*], Kailei Guo[4], Xiaomin Zhang[4], Jianqiang Huang[1,2,3†], Gaogang Xie[1,3†]

[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Tianjin Medical University Eye Hospital, Tianjin, China
jxqi@cnic.cn, cuiyan.ch@gmail.com, guokailei2019@tmu.edu.cn, xzhang08@tmu.edu.cn, jqhuang@cnic.cn, xie@cnic.cn

Abstract

Single-cell transcriptomics describes complex molecular features at the individual cell level, serving various roles in biological research, such as enhancing gene expression and predicting drug responses. Due to transcriptomic data structurally resembling sequential data, many researchers have trained numerous transformers on extensive transcriptomic datasets. However, they have consistently neglected to explore the intrinsic properties of the data and the appropriateness of their chosen model architecture. In this paper, we carefully investigate the nature of transcriptomics, identifying three overlooked problems: 1) long-tailed data problem, 2) model selection problem, and 3) evaluation problem. Consequently, by applying the weighted sampling strategy, we address the long-tailed data problem and achieve consistent improvement across all settings. By adapting different model structures to transcriptomic data, we discover that transformers are not the only option. By developing three downstream tasks and fair evaluation metrics, we establish a simple and comprehensive benchmark to validate the effectiveness of models for transcriptomics. Through extensive experiments, we clarify the misunderstandings in the traditional methods and provide competitive baselines, thereby paving the way for future research in this field.

Introduction

Single-cell transcriptomics, also known as single-cell RNA sequencing, is a high-throughput technology that sequences and analyzes RNA within individual cells (Kolodziejczyk et al. 2015), which aids many biological tasks, such as cell annotation and gene regulatory discovery. Since the samples of transcriptomics resemble sequential data in form, typically presented as a series of genes and their corresponding expression values: $(\text{gene}_1, \text{gene}_2, \ldots, \text{gene}_n)$ $(\text{value}_1, \text{value}_2, \ldots, \text{value}_n)$, Transformers are naturally introduced into this field. With the development of Transformers, researchers have conducted extensive pre-training on the transcriptomic dataset, which includes about 50 million samples (Cui et al. 2024). However, a critical misunderstanding persists: transcriptomics is inherently not sequential data. The indiscriminate use of Transformers for large-scale pre-training has resulted in substantial consumption of computational resources. Therefore, there is a pressing need for in-depth discussions on data and models within tran-

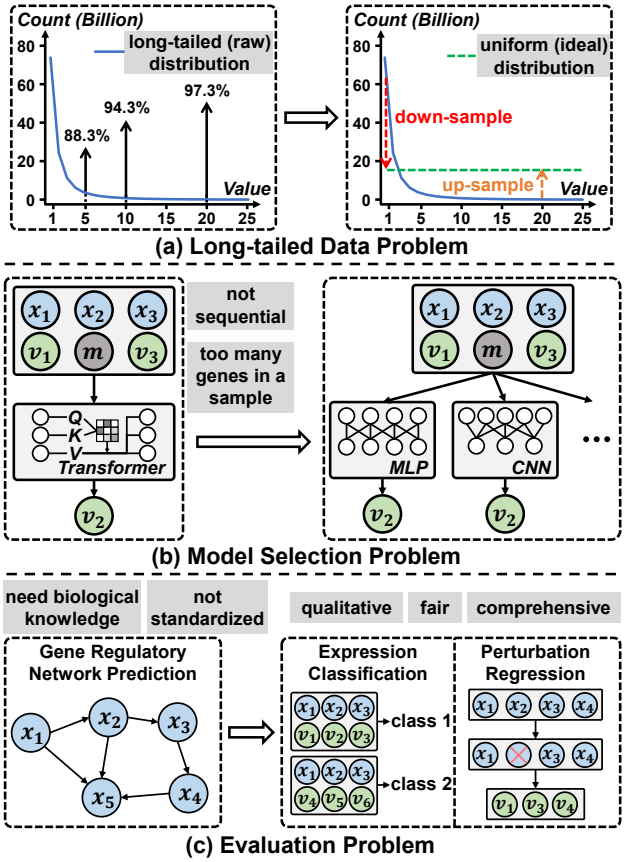

Figure 1: Illustrations of three overlooked problems in transcriptomics. (a) Long-tailed Data Problem. Left: original long-tailed value distribution (blue solid line), where percentages denote proportions of values less than or equal to the corresponding values. Right: ideal uniform distribution (green dashed line) after weighted sampling. (b) Model Selection Problem. Left: the commonly used Transformer. Right: exploring other architectures for transcriptomics. (c) Evaluation Problem. Left: traditional evaluation involving biological knowledge. Right: our proposed simple and quantifiable evaluations with classifications and regressions.

*These authors contributed equally.
†Corresponding authors.

scriptomics. After carefully analyzing the datasets and existing methods, as shown in Figure 1, we identified three consistently overlooked problems in this field: long-tailed data problem, model selection problem, and evaluation problem.

First, as illustrated in Figure 1(a), we performed statistical analyses for each expression value across all samples, deriving comprehensive statistical results that clearly exhibit an extreme long-tailed distribution for these values. For example, the value '1' accounts for over 54% of the data, and values less than '10' constitute over 94%. These findings indicate that traditional training for transcriptomic data will result in models biased toward high-frequency values and neglect the rare ones, i.e., the model is encouraged to learn more '1's to minimize the training loss and overlook less frequent but larger values. However, many biological studies in transcriptomics suggest that larger expression values may represent more significant meanings and should be emphasized (Jessop et al. 2020; Danopoulos et al. 2020). Moreover, the measurements in transcriptomics render smaller values less robust against errors (Karaayvaz et al. 2018; AlJanahi, Danielsen, and Dunbar 2018), which amplifies the long-tailed problem in traditional methods. To address this problem, we implemented a commonly used strategy in long-tail tasks, weighted sampling, which enhances the sampling probability for the rarer samples, i.e., the large values. As illustrated in Table 1 and Table 2, this straightforward method consistently improves performance across all settings.

Second, regarding model selections, because transcriptomic data structurally resemble sequential data, and its training loss is similar to masked language modeling, Transformers are intuitively utilized. This instinct has led researchers to focus solely on Transformers or their variants (Hao et al. 2024), thus neglecting the potential of other architectures. However, it is critical to recognize that transcriptomics fundamentally differs from sequential data, particularly because each gene appears only once per sample and there is no relative order between genes. Furthermore, in transcriptomics, where sequence lengths can reach up to 60,000, Transformers require substantial memory and computational resources because their attention mechanisms scale quadratically with sequence length (Keles, Wijewardena, and Hegde 2023), making them inefficient. Therefore, we explored other foundational architectures, such as Multilayer Perceptrons (MLP) (Haykin 1998), Convolutional Neural Networks (CNN) (Krizhevsky, Sutskever, and Hinton 2012), and Mamba (Gu and Dao 2023), for transcriptomics. With simple adaptation, we found these models could achieve competitive performance with less computational overhead. Additionally, we conducted detailed analyses of various Transformer settings to address some misunderstandings from traditional methods.

Finally, the evaluation problem appears in two aspects. One is that the downstream tasks used to validate pre-trained models are overly specialized, requiring specific biological knowledge and lacking quantitative results, such as the assessments of gene regulatory networks prediction. This hinders the involvement of researchers from other fields, e.g., AI researchers. The other one is that the evaluations of traditional methods are neither fair nor standardized. For example, these methods often fine-tune the pre-trained models in downstream tasks, which impedes an accurate assessment of pre-training performance. Additionally, some experiments are performed only once, leading to non-reproducible results and unreliable conclusions. To address these concerns, in Figure 1(c), we introduced a simple and comprehensive benchmark for large-scale and standardized evaluations, including expression classification across 10 datasets, and perturbation classification and prediction across 3 datasets. We also standardized downstream evaluations by freezing the model backbone, running multiple times to report the average, and utilizing extracted features for k-nearest neighbors classification or training only one linear layer to ensure the rigorous and quantitative evaluation of the pre-trained models. This benchmark provides a fair and comprehensive testbed, enabling more researchers, without specialized biological knowledge, to participate in this field.

Our contributions can be summarized in three aspects:

1. We thoroughly analyzed single-cell transcriptomics and existing methods, revealing three overlooked problems: the long-tailed data problem, the model selection problem, and the evaluation problem, which significantly impair the large-scale pre-training for transcriptomic data.
2. We proposed straightforward yet effective solutions to address these problems: implementing weighted sampling, adapting different model structures, and designing rigorous downstream evaluations with fair comparison criteria.
3. We introduced *a simple and comprehensive benchmark for single-cell transcriptomics*. Through extensive experiments, we validated the effectiveness of our proposed methods. The benchmark also serves as a convenient testbed for other researchers, preventing the blind waste of computational resources on large-scale transformer pre-training, delineating the future direction in this field.

Code — https://github.com/simpleshinobu/scbenchmark

Related Works

Traditional Single-Cell Transcriptomics. Single-cell transcriptomics, also known as single-cell RNA sequencing, i.e., scRNA-seq, was first introduced by Surani Lab (Tang et al. 2009). Since then, significant improvements in sensitivity, speed, and affordability have been achieved (Sasagawa et al. 2013; Macosko et al. 2015), and computational tools and public data resources for scRNA-seq are rapidly expanding (Voigt et al. 2021; Kharchenko 2021). Today, scRNA-seq is extensively used in the field of human health, primarily to characterize cell types in various organs (Voigt et al. 2021; Ramachandran et al. 2020), such as exploring transcriptome heterogeneity across similarly classified cells in different states (Kravets and Benninger 2020; Wheeler et al. 2020) and clarifying temporal processes like human tissue development (Olaniru et al. 2023; Collin et al. 2021). Despite rapid advancements in transcriptomics, there remains a critical need for standardized benchmarks to evaluate the performance of computational tools or models. Efforts to establish benchmarks for transcriptomics analysis have been

made (Tian et al. 2019; Li et al. 2022). However, prior works have primarily focused on traditional biological tools, while in this work, we provide fair and comprehensive benchmarks for large-scale pre-trained deep models.

Large-Scale Pre-training for Transcriptomics. With the development of Transformers (Vaswani et al. 2017), transcriptomics has demonstrated significant interest in large-scale pre-training. scBERT (Yang et al. 2022) was the pioneer to propose the single-cell pre-training framework based on Transformers. Subsequent advancements can be categorized into two camps: expanding training scales and refining algorithms. In the first camp, scGPT (Cui et al. 2024) utilized an increased number of parameters to pre-train on a more extensive dataset, which comprises approximately 33 million cells. Furthermore, GeneCompass (Yang et al. 2023) leveraged a cross-species transcriptomic dataset to access over 130 million samples. In the second camp, scFoundation (Hao et al. 2024) introduced modifications to the Transformer framework for effectively handling long gene sequences. GeneFormer (Theodoris et al. 2023) enhanced efficiency by adopting a gene sequencing approach to eliminate the need for gene embeddings. CellPLM (Wen et al. 2023) and tGPT (Shen et al. 2023) introduced novel structures to analyze intercellular relationships and create an autoregressive gene prediction pipeline, respectively. However, these methods have consistently overlooked the essential problems we identified in transcriptomics, resulting in the blind adoption of Transformers for large-scale pre-training that consumes extensive resources. In this work, we address these problems with a simple and comprehensive benchmark, providing a testbed that supports the development of future pre-training frameworks and innovative algorithms.

Method

Preliminaries. Considering the common training framework in transcriptomics (Yang et al. 2023; Hao et al. 2024), we start with the training set $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{v}_i)\}_{i=1}^{N}$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_l)$ denotes a set of genes, $\boldsymbol{v} = (v_1, v_2, \ldots, v_l)$ denotes the expression values corresponding to each gene, l is the number of genes in this sample (cell). Similar to the masked language modeling (Devlin et al. 2018) in sequential training, transcriptomics adopts masked value prediction, as shown in Figure 2(a), where $\boldsymbol{v}$ is masked as $\tilde{\boldsymbol{v}} = (v_1, [\text{mask}], \ldots, v_l)$, [mask] denotes the masked value token, and the self-supervised loss can be written as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k} \left\| \boldsymbol{v}_{i,k} - (\boldsymbol{e}_i)_k \boldsymbol{w} \right\|^2, \tag{1}$$

$$\boldsymbol{e}_i = \phi(\boldsymbol{E}_x(\boldsymbol{x}_i), \boldsymbol{E}_v(\tilde{\boldsymbol{v}}_i)), \tag{2}$$

where $\boldsymbol{v}_{i,k}$ represents the k-th masked value of sample i and $(\boldsymbol{e}_i)_k \in \mathbb{R}^{1\times d}$ is the gene feature corresponding to $\boldsymbol{v}_{i,k}$, d is the hidden dimension, $\boldsymbol{e}_i \in \mathbb{R}^{l\times d}$ denotes the extracted gene features, $\boldsymbol{w} \in \mathbb{R}^{d\times 1}$ is a linear layer to project $\boldsymbol{e}_i$ into scalars, $\boldsymbol{E}_x$ and $\boldsymbol{E}_v$ are embedding layers for genes and values respectively, and ϕ is the feature extractor.

Since each gene appears only once in a cell, enabling the predictions derived from the product of gene embedding and cell features (Cui et al. 2024), which is the Masked Values prediction loss with Cell features modified from Eq. (1):

$$\mathcal{L}_{\text{MVC}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k} \left\| \boldsymbol{v}_{i,k} - (\boldsymbol{E}_x(x_{i,k}) \cdot \bar{\boldsymbol{e}_i}) \right\|^2, \tag{3}$$

where $x_{i,k}$ is the gene for k-th masked value of cell i, $\bar{\boldsymbol{e}_i} \in \mathbb{R}^{1\times d}$ is the cell feature, which can be implemented as the average of all gene features or a specific token feature (Devlin et al. 2018), and $\cdot$ denote the dot product.

Additionally, due to the broad range of expression values, from zero to potentially millions, large losses will occur that impair the model optimization. Therefore, two common preprocessing methods are adopted. One is *binning*, the default preprocessing (Cui et al. 2024), which divides all values in a sample into several bins, thus producing a sequence of bin indices as the substitute value input. The other one is *logarithmic transformation*, formulated as $v = \log(1+v)$, which effectively constrains the scale of the values.

Long-Tailed Problem. This issue has been extensively explored in long-tailed classification tasks (Tang, Huang, and Zhang 2020; Zhong et al. 2021; Zhang et al. 2023), where the most effective methods involve sampling more low-frequent samples to increase their occurrence. By converting the long-tailed data distribution to the uniform distribution, these methods effectively remove the model bias introduced by the frequent classes, i.e., frequent values in our scenarios. According to Eq. (1), our weighted sampling strategy can be written as:

$$\mathcal{L}_{\text{LT}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k} \alpha_{i,k} \left\| \boldsymbol{v}_{i,k} - (\boldsymbol{e}_i)_k \boldsymbol{w} \right\|^2, \tag{4}$$

where $\alpha_{i,k}$ denotes the resampling weight for value $\boldsymbol{v}_{i,k}$, and $\alpha_{i,k} \propto 1/p(\boldsymbol{v}_{i,k})$ denotes that the greater the probability $p(\boldsymbol{v}_{i,k})$ of a value's occurrence, the smaller the resampling weight for this value. Since $p(\boldsymbol{v}_{i,k}) \propto 1/\boldsymbol{v}_{i,k}$ in transcriptomics, L_{LT} is more focused on optimizing for larger values, which conforms to our motivation.

Model Selection Problem. The prediction of gene expression values relies on correlations between genes, e.g., some genes activate the expressions of others while some suppress them (Cordell 2009). Thus, the model needs to perform interactions among gene features as shown in Figure 2(a), and that is why traditional methods apply Transformers. Due to the computational burden discussed earlier, it is meaningful to adapt other fundamental architectures for transcriptomics. Note that, since Mamba is similar to Transformer, which can directly establish relationships between genes, we will focus on the adaptations for MLP and CNN.

For MLP, the direct implementation only transforms features at the last dimension, which fails to realize interactions among genes, resulting in poor performance as shown in Table 3. Our adaptation, as illustrated in Figure 2(b), uses an additional linear layer to transform the gene dimension, thereby enabling feature interactions between genes, which can be written as:

$$\boldsymbol{e}^{j+1} = \sigma(\boldsymbol{e}^{jT} \boldsymbol{w}_1)^T \boldsymbol{w}_2, \tag{5}$$

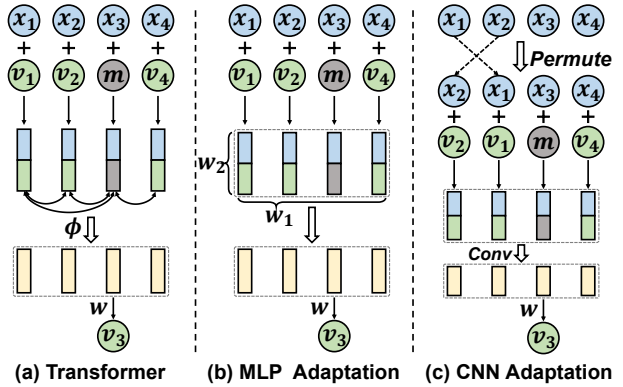

Figure 2: Illustrations of frameworks for Transformer and adapted models. (a) Transformer extracts feature through interactions among genes to predict masked values, i.e., m. (b) The adapted MLP, formulated in Eq. (5), employs $\boldsymbol{w}_1$ to capture gene interactions. (c) The adapted CNN, formulated in Eq. (6), uses gene permutation to overcome the limitations of local receptive fields, i.e., by positioning closely related genes such as x_1 and x_3 to enhance their interactions.

where $\boldsymbol{e}^j \in \mathbb{R}^{l\times d}$ denotes gene features of the j-th layer, $\boldsymbol{w}_1 \in \mathbb{R}^{l\times l}$ and $\boldsymbol{w}_2 \in \mathbb{R}^{d\times d}$ are linear transformations for the last two dimensions, respectively, σ denotes the activation, and ^{T} is the transposition of feature dimensions. Although the input gene length is constrained, it is not a significant issue in transcriptomics, as each gene appears only once per sample and the total number of genes is fixed.

For CNN, the local receptive field hinders its ability to capture the relationships among all genes. Considering that the interactions often occur within specific groups of genes (Funk et al. 2022), re-arranging the input genes, by using prior knowledge or other gene interaction metrics, is a valid adaptation for CNN to transcriptomics. As illustrated in Figure 2(c), our adaptation can be formulated as:

$$\boldsymbol{e}^{j+1} = \sigma(\text{Permute}(\boldsymbol{e}^j) * \boldsymbol{w}), \tag{6}$$

where Permute denotes permuting genes into a pre-defined order, which only happens in the first layer, $\boldsymbol{w}$ is the convolution kernel, and $*$ is the convolution operation. In experiments, we find that the adaptation could improve the direct implementation, making its performance more competitive.

Evaluation Problem. As we have discussed, traditional methods overemphasize specialized biological knowledge and lack standardized evaluations for pre-trained models. Therefore, we designed three large-scale and quantifiable downstream tasks:

1) Expression Classification, where samples are labeled with classes, such as cell types, and the objective is to predict the class label based on gene expressions. The training loss can be written as:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(\frac{\exp(\bar{\boldsymbol{e}}_i \boldsymbol{w})}{\sum_j \exp(\bar{\boldsymbol{e}}_i \boldsymbol{w})_j}\right), \tag{7}$$

where y_i is the one hot class label, $\bar{\boldsymbol{e}}_i$ is the extracted cell features by a frozen pre-trained model, $\boldsymbol{w} \in \mathbb{R}^{d\times c}$ is a learnable linear layer, c is the total number of classes, j indexes all classes. This task can also be realized by the k-nearest neighbors (KNN) strategy based on the extracted cell features $\bar{\boldsymbol{e}}_i$.

2) Perturbation Classification, where a certain gene is perturbed (e.g., knockout (Egorov et al. 2021)) and the resulting expressions are recorded. The objective is to identify which gene was perturbed based on the post-perturbation expressions. Since this is also a classification task, the loss described in Eq. (7) and KNN strategy are applicable. Note that, this task presents greater challenges than the previous one, due to much more class types.

3) Perturbation Regression, where the same perturbation datasets are used and the objective changes to predict all expressions based on the perturbed gene, and the loss function can be written as a regression version:

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N} \left\|\boldsymbol{v}_i - \phi(\boldsymbol{E}_x(\boldsymbol{x}_i), \boldsymbol{E}_x(x_{i,p})\boldsymbol{w})\right\|^2, \tag{8}$$

where $x_{i,p}$ is the perturbed gene in sample i, the pre-trained parameters $\boldsymbol{E}_x$ and ϕ are frozen, $\boldsymbol{w}$ is a trainable parameter to learn the influence of perturbation for $x_{i,p}$.

Experiments

Datasets

Pre-training Dataset. We follow the approach proposed by scGPT (Cui et al. 2024) to assemble a pre-training transcriptomic dataset, containing 54.6 million human cells from the CELLxGENE collection (Biology et al. 2023). This dataset encompasses more than 50 organs (e.g., blood and heart) and tissues across over 400 studies, offering a broad representation of cellular heterogeneity throughout the human body.

Expression Classification Datasets. We collect 10 expression classification datasets following the strategies in scGPT.

- Myeloid (**Myel**) (Cheng et al. 2021) performs a comprehensive pan-cancer analysis of myeloid cells, consisting of 13,178 samples and 21 sub-cancer classes.
- Multiple Sclerosis (**MS**) (Schirmer et al. 2019) reveals specific cellular changes in multiple sclerosis lesions, which consists of 21,312 samples and 18 cell classes.
- Pancreas (**Panc**) (Chen et al. 2023) contains data from five human pancreas studies for cell type annotation tasks, consisting of 14,818 samples and 14 cell classes.
- Checkpoint Inhibitor Colitis (**CIC**) (Thomas et al. 2024) reveals that the crosstalk between circulating T cells and epithelial cells is critical to PD-1/CTLA-4-dependent tolerance and barrier function in this disease, which consists of 118,818 samples and 8 classes.
- Myasthenia Gravis (**MG**) (Zhong et al. 2023) identifies a unique subset of monocytes, displaying significant pro-inflammatory pathways during and after the crisis, which consists of 53,748 samples and 25 classes.
- Systemic Lupus Erythematosus (**SLE**) (Perez et al. 2022) shows increased type 1 interferon-stimulated genes, with fewer naive CD4 T cells and more cytotoxic GZMH CD8 T cells, consisting of 79,322 samples and 14 cell classes.

Config	Sampling	Myel	MS	Panc	CIC	MG	SLE	scF	LIC	SD	LM	Average
Raw Data		14.514	16.519	38.792	33.752	32.913	40.157	21.380	24.386	39.060	19.897	28.137
Baseline	Random	68.462	83.131	91.084	51.821	81.028	74.140	67.943	92.349	82.601	59.650	75.221
	LT_{test}	69.459	83.328	91.453	55.509	84.095	74.436	67.367	94.577	85.116	90.476	79.582
$Baseline_{log}$	Random	70.395	89.080	97.602	53.152	81.595	77.488	68.338	92.994	82.988	66.646	78.028
	LT_{test}	71.684	89.209	97.962	57.212	84.593	77.564	67.832	94.806	85.412	91.596	81.787
MVC_{avg}	Random	70.607	88.930	97.760	55.557	82.508	77.624	67.135	93.558	84.315	70.851	78.885
	LT_{test}	71.351	89.202	98.232	61.971	85.419	77.613	67.347	95.215	86.271	92.635	82.526
MVC_{cls}	Random	71.259	88.789	97.737	55.815	83.131	77.951	68.388	93.797	84.440	70.307	79.161
	LT_{test}	71.755	88.915	98.174	61.072	85.711	77.789	68.327	95.389	86.237	92.134	82.550
LT_{train}	Random	72.858	89.005	97.949	56.295	82.929	78.523	68.388	93.709	83.944	68.986	79.259
	LT_{test}	**73.647**	**89.403**	**98.277**	**63.103**	**86.459**	**78.725**	**68.631**	**95.621**	**86.676**	**92.685**	**83.323**

Table 1: Test Accuracy (%) of the linear classifier on ten expression classification datasets. Raw Data denotes directly using expressions as input. Baseline denotes the reproduced default scGPT (Cui et al. 2024), with *binning* preprocessing, compared to $Baseline_{log}$ using log preprocessing. Sampling denotes the sampling strategy in the testing, including traditional random sampling and weighted sampling LT_{test}. Results are the mean of five independent trials.

Config	Sampling	Myel	MS	Panc	CIC	MG	SLE	scF	LIC	SD	LM	Average
Raw Data		9.408	11.175	12.584	35.834	49.723	16.550	24.989	10.305	47.164	16.488	23.422
Baseline	Random	62.615	75.912	83.761	48.916	78.085	**64.844**	62.345	88.528	80.293	48.592	69.389
	LT_{test}	64.775	75.674	84.314	52.461	81.517	64.743	61.071	91.633	83.340	86.064	74.559
$Baseline_{log}$	Random	67.972	82.265	94.525	50.901	78.912	63.735	63.608	90.143	81.755	58.811	73.263
	LT_{test}	69.332	82.521	94.593	54.223	82.151	63.317	62.911	93.084	83.524	87.753	77.341
MVC_{avg}	Random	68.093	83.897	96.536	53.703	79.831	61.780	61.647	91.317	82.249	60.770	73.982
	LT_{test}	69.029	84.632	96.968	58.452	82.593	61.879	61.991	93.897	83.920	89.850	78.321
MVC_{cls}	Random	69.707	84.589	96.361	54.167	81.025	62.712	63.729	92.936	82.787	62.015	75.003
	LT_{test}	70.020	84.833	97.179	58.280	83.580	62.780	64.709	94.922	84.428	**91.446**	79.218
LT_{train}	Random	71.133	84.992	97.580	54.421	80.924	62.147	64.730	92.126	82.976	60.056	75.108
	LT_{test}	**72.059**	**85.549**	**97.926**	**60.035**	**84.447**	62.404	**64.770**	**94.593**	**85.201**	90.788	**79.777**

Table 2: Test Accuracy (%) of the k-nearest neighbors on ten expression classification datasets. Other settings are the same as those in Table 1. Results are the mean of five independent trials.

- scFoundation (**scF**) (Hao et al. 2024) employed Zheng68K dataset (Zheng et al. 2017), a study for human peripheral blood mononuclear cell, for the expression classification task. We adopt the dataset, which consists of 6,595 samples and 11 classes.
- Leptomeningeal Immune Cells (**LIC**) (Remsik et al. 2023) finds that leptomeningeal-specific IFN-γ signaling is critical for cancer cell growth independent of adaptive immunity, consisting of 20,676 samples and 10 classes.
- Severe Dengue (**SD**) (Ghita et al. 2023) defines the target cells of the dengue virus and the immunological hallmarks of severe dengue (SD) progression in children's blood, consisting of 193,727 samples and 23 classes.
- Leptomeningeal Metastasis (**LM**) (Chi et al. 2020) finds that macrophages produce inflammatory cytokines in the cerebrospinal fluid, supporting cancer cell growth, consisting of 10,650 samples and 13 classes.

Perturbation Datasets. For conducting perturbation classification and regression, we select 3 perturbation datasets.

- **Adamson** (Adamson et al. 2016) applies Perturb-seq to dissect the mammalian unfolded protein response, which consists of 63,585 and 76 classes of perturbations.
- **Dixit** (Dixit et al. 2016) develops Perturb-seq, combining transcriptomic and CRISPR-based perturbations to enable the large-scale analysis of complex phenotypes, which consist of 44,735 samples and 20 classes.
- **Norman** (Norman et al. 2019) reveals the relationship between the set of genes a cell expresses and its phenotype, which consists of 53,508 samples and 101 classes.

Implementation Details

To ensure consistent and fair comparisons, we standardized the experimental setup for all architectures during pre-training and downstream tasks. For pre-training, without specific note, a subset of data was used, with a 6-layer network, hidden dimensions of 256, a batch size of 128, and a gene length of 512. The Adam (Kingma 2014) optimizer with a learning rate of 0.0002 was employed over 10 epochs. For downstream tasks, we split the datasets into 70% training and 30% testing, standardizing the settings to 50 epochs and batch size of 64, with an Adam optimizer and 0.005 learning rate. For classifications, we performed the K-Nearest Neighbors algorithm with 10 neighbors or a single linear layer classifier. For regressions, we used a trainable embedding to learn perturbed genes. Our downstream experiments froze the pre-trained model and were independently repeated to ensure fairness and reproducibility, and thus our baseline is reproduced scGPT (Cui et al. 2024) with their de-

Model	Params	FLOPs	Myel	MS	Panc	CIC	MG	SLE	scF	LIC	SD	LM	Average
Baseline_2	2.44M	2.01G	71.755	88.915	98.174	61.072	85.711	77.789	68.327	95.389	86.237	92.134	82.550
MLP	0.86M	0.40G	39.636	43.165	75.043	46.788	65.889	53.820	45.255	68.373	69.608	54.931	56.251
MLP-Ours	0.64M	0.29G	72.605	89.324	96.446	56.780	84.225	76.209	67.954	95.428	86.088	89.093	81.415
CNN	3.61M	0.60G	72.018	88.086	97.445	55.153	84.080	72.916	66.630	94.977	85.296	90.532	80.713
CNN-Ours	3.61M	0.60G	71.619	88.104	97.436	57.460	84.604	73.636	67.246	95.048	85.952	91.227	81.233
Mamba	2.58M	0.43G	71.765	89.124	97.958	59.171	84.723	73.153	67.984	94.857	85.999	92.034	81.677
Trm^*	21.87M	21.74G	**74.269**	**90.206**	**98.475**	**70.068**	**87.084**	**81.165**	**70.025**	**96.395**	**88.572**	**93.298**	**84.956**

Table 3: Test Accuracy (%) of the linear classification for different architectures on ten expression classification datasets. Baseline_2 denotes MVC_{cls} with LT_{test} in Table 1, offering a competitive alternative to the original one. MLP and CNN denote the direct implementation, and Ours denote the adaptations formulated in Eq. (5) and Eq. (6), respectively. Trm^* denotes the optimal combination of settings selected in Table 4 and Table 5.

Structure	Params	FLOPs	KNN	Linear
Baseline_2	2.44M	2.01G	**79.218**	82.550
(3, 256)	1.25M	1.01G	79.003	82.460
(9, 256)	3.63M	3.02G	78.738	82.360
(12, 256)	4.81M	4.03G	78.856	82.541
(6, 128)	0.61M	0.70G	77.559	80.328
(6, 512)	9.73M	6.44G	77.817	83.239
(6, 768)	21.87M	13.29G	77.605	**83.376**
(12, 512)	19.20M	12.88G	77.736	83.003

Table 4: Test Accuracy (%) of different Transformer structures. Numbers in the Structure column denote the layers and hidden dimensions of the models, respectively. Params and FLOPs are calculated for the encoder, excluding embedding and output layers for clarity. Results are averaged over ten expression classification datasets.

Mask	Seq	KNN	Linear
Baseline_2		79.218	82.550
30%	512	78.704	82.290
50%		**78.887**	**82.589**
60%		78.585	82.534
40%	256	77.748	81.634
	768	**79.713**	**83.073**
	1024	79.395	82.972

Table 5: Test Accuracy (%) of different Transformer settings during pre-training. Seq denotes the training gene length. Other settings are the same as those in Table 4.

fault pre-processing and loss under our rigorous settings for fair comparison.

Results and Analysis

Through the following Q&A, we provide in-depth discussions of the experimental results, mainly reported on large-scale expression classification, pertaining to the three proposed problems and corresponding solutions, offering a comprehensive analysis of our benchmark.

Q1. *How does weighted sampling solve the long-tailed distribution problem?*

A1. As shown in Table 1 and Table 2, the weighted sampling strategy (abbreviated as LT) almost improves performance across all settings. Specifically, during testing, the improvements of LT (highlighted by shading) are method-agnostic. By integrating LT in the training, the best performance is achieved, surpassing the best settings without considering the long-tailed problem, i.e., MVC_{cls} with averaged improvements of 4.2% and 4.8% for Linear and KNN, respectively, and by combining other tricks, surpassing Baseline with significant averaged improvements of 8.1% and 10.4% for Linear and KNN, respectively, thereby establishing a new strong baseline.

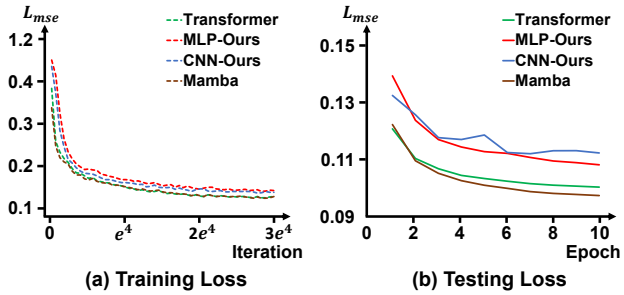

Figure 3: Visualization of training and testing losses in transcriptomics pre-training across four architectures: Transformer, adapted MLP, adapted CNN, and Mamba.

Q2. *Which preprocessing and cell embeddings are better?*

A2. Through Table 1 and Table 2, we find that some default settings in Baseline (Cui et al. 2024) are suboptimal. Improvements can be achieved through simple modifications, such as replacing the *binning* preprocessing with *log* preprocessing, incorporating MVC loss during training, and using *cls* features for cell embeddings instead of averaged token features. By implementing these adjustments, performance is improved by 3.9% and 5.6% for Linear and KNN, respectively, offering valuable insights for future pre-training.

Q3. *Which Transformer setting performs better?*

A3. Traditional methods commonly assume that increasing parameters and input length will always lead to better pre-training results. However, as shown in Table 4 and Table 5, this hypothesis is not supported. Optimal performance may be achieved with settings of 6 layers, 768 hidden dimensions, 50% masking ratio, and 768 input length. These ex-

Type	Size	Myel	MS	Panc	CIC	MG	SLE	scF	LIC	SD	LM	Average
Baseline_2	1.8M	71.755	**88.915**	**98.174**	61.072	85.711	77.789	68.327	95.389	86.237	92.134	**82.550**
Blood	1.8M	70.263	68.355	93.846	**65.169**	**86.413**	**79.521**	**70.409**	95.141	**87.385**	92.272	80.877
Brain	1.8M	63.576	88.108	95.088	52.383	80.372	66.541	62.183	90.073	79.471	83.279	76.107
Heart	1.8M	70.329	71.013	95.700	58.916	82.968	71.307	63.295	93.735	85.355	89.787	78.240
Kidney	0.8M	69.049	70.838	96.176	60.680	83.272	74.069	66.609	94.403	85.523	90.982	79.160
Lung	1.8M	72.054	71.611	97.436	64.040	85.385	77.544	67.984	95.180	86.221	92.278	80.973
PCancer	1.8M	**73.546**	73.162	97.251	60.575	85.763	77.967	68.398	**95.767**	86.039	**93.110**	81.158
Full (20%)	10.8M	72.114	87.529	97.931	59.241	85.159	78.128	68.378	95.505	86.057	92.578	82.262
Full	54.6M	70.830	84.342	97.598	58.563	84.975	74.117	68.055	94.935	85.643	92.059	81.112

Table 6: Test Accuracy (%) of the linear classification for pre-training on different organ datasets. Results are reported on ten downstream expression classification datasets. Size denotes the number of training samples.

Models	Adamson	Dixit	Norman
Baseline	41.586	26.801	22.248
Baseline_2	43.552	**27.265**	37.336
Mamba	42.271	27.190	35.362
MLP-Ours	40.215	27.159	24.799
CNN-Ours	41.081	26.943	31.433
Trm*	**45.841**	27.201	**41.959**

Table 7: Test Accuracy (%) of the linear classification for different pre-training settings and architectures on three perturbation classification tasks.

Models	Adamson	Dixit	Norman
Baseline_2	**0.200**	0.046	0.163
Mamba	0.320	0.130	0.343
MLP-Ours	0.221	0.046	0.166
CNN-Ours	0.284	0.103	0.223
Trm*	0.202	**0.042**	**0.142**

Table 8: Test loss of different pre-training architectures on three perturbation regression tasks, lower is better. The baseline method is not applicable.

periments address the misconceptions in traditional methods and provide suggestions for future research.

Q4. *What is the performance of other architectures?*

A4. As we have noted that transcriptomic data is inherently not sequential, we explore other architectures in Figure 3 and find other architectures that could perform similar training and testing processes. Table 3 shows that our straightforward adaptations for MLP and CNN outperform the direct implementations. These alternatives achieve comparable performance compared to Transformers, with the largest average difference being 1.3%. Although there remains a gap compared to Transformers with optimal settings, the computational efficiency of these architectures suggests they have significant potential for this task.

Q5. *What is the performance on other downstream tasks?*

A5. As shown in Table 7 and Table 8, Transformers achieve the best performance in both perturbation classification and regression tasks. As for classification, the increased number of classes introduces significant challenges, resulting in poor performance across all models. Additionally, the performance gap between other models and Transformers, such as MLP in the Norman classification, and Mamba in regressions, has widened. This indicates the need for further evaluations when designing new architectures.

Q6. *Does more training data lead to better performance?*

A6. A prevailing trend in transcriptomic pre-training is to collect more data, such as GeneCompass (Yang et al. 2023). As shown in Table 6, by comparing different sizes and types of data for pre-training, we find that more data does not necessarily equate to better performance, thereby challenging traditional assumptions. Additionally, we discover that pre-training on specific organ datasets achieves state-of-the-art (SOTA) performance on certain downstream classifications, such as Blood pre-training on SLE classification, and the possible reason may be the classes in SLE related to blood. These findings suggest that for specific downstream tasks, further pre-training on relevant datasets might be beneficial.

Conclusion

Through careful analysis of transcriptomic data and existing methods, we identified three overlooked problems: the long-tailed data problem, the model selection problem, and the evaluation problem. To address them, we introduced weighted sampling, specific architecture adaptations, and fair and quantifiable downstream evaluations, respectively. These strategies enable us to investigate the optimal settings and explore the potential of alternative architectures. Additionally, we resolved several misconceptions common in traditional methods, thus providing better insights for large-scale pre-training. By developing a simple and comprehensive benchmark, we provided researchers with a testbed for analyzing the pre-trained models, with no need for biological knowledge. Besides the above contributions, we suggest two further directions: 1) Implementing more advanced methods to address the long-tail problem, moving beyond the naive weighted sampling; 2) Integrating other architectures with attention mechanisms to establish a novel framework for transcriptomic pre-training.

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA0460205.

References

Adamson, B.; Norman, T. M.; Jost, M.; Cho, M. Y.; Nuñez, J. K.; Chen, Y.; Villalta, J. E.; Gilbert, L. A.; Horlbeck, M. A.; Hein, M. Y.; et al. 2016. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7): 1867–1882.

AlJanahi, A. A.; Danielsen, M.; and Dunbar, C. E. 2018. An introduction to the analysis of single-cell RNA-sequencing data. *Molecular Therapy Methods & Clinical Development*, 10: 189–196.

Biology, C. S.-C.; Abdulla, S.; Aevermann, B.; Assis, P.; Badajoz, S.; Bell, S. M.; Bezzi, E.; Cakir, B.; Chaffer, J.; Chambers, S.; et al. 2023. CZ CELLxGENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *BioRxiv*, 2023–10.

Chen, J.; Xu, H.; Tao, W.; Chen, Z.; Zhao, Y.; and Han, J.-D. J. 2023. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1): 223.

Cheng, S.; Li, Z.; Gao, R.; Xing, B.; Gao, Y.; Yang, Y.; Qin, S.; Zhang, L.; Ouyang, H.; Du, P.; et al. 2021. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell*, 184(3): 792–809.

Chi, Y.; Remsik, J.; Kiseliovas, V.; Derderian, C.; Sener, U.; Alghader, M.; Saadeh, F.; Nikishina, K.; Bale, T.; Iacobuzio-Donahue, C.; et al. 2020. Cancer cells deploy lipocalin-2 to collect limiting iron in leptomeningeal metastasis. *Science*, 369(6501): 276–282.

Collin, J.; Queen, R.; Zerti, D.; Bojic, S.; Dorgau, B.; Moyse, N.; Molina, M. M.; Yang, C.; Dey, S.; Reynolds, G.; et al. 2021. A single cell atlas of human cornea that defines its development, limbal progenitor cells and their interactions with the immune cells. *The ocular surface*, 21: 279–298.

Cordell, H. J. 2009. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6): 392–404.

Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; and Wang, B. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 1–11.

Danopoulos, S.; Bhattacharya, S.; Mariani, T. J.; and Al Alam, D. 2020. Transcriptional characterisation of human lung cells identifies novel mesenchymal lineage markers. *European Respiratory Journal*, 55(1).

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dixit, A.; Parnas, O.; Li, B.; Chen, J.; Fulco, C. P.; Jerby-Arnon, L.; Marjanovic, N. D.; Dionne, D.; Burks, T.; Raychowdhury, R.; et al. 2016. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *cell*, 167(7): 1853–1866.

Egorov, A. A.; Alexandrov, A. I.; Urakov, V. N.; Makeeva, D. S.; Edakin, R. O.; Kushchenko, A. S.; Gladyshev, V. N.; Kulakovskiy, I. V.; and Dmitriev, S. E. 2021. A standard knockout procedure alters expression of adjacent loci at the translational level. *Nucleic Acids Research*, 49(19): 11134–11144.

Funk, L.; Su, K.-C.; Ly, J.; Feldman, D.; Singh, A.; Moodie, B.; Blainey, P. C.; and Cheeseman, I. M. 2022. The phenotypic landscape of essential human genes. *Cell*, 185(24): 4634–4653.

Ghita, L.; Yao, Z.; Xie, Y.; Duran, V.; Cagirici, H. B.; Samir, J.; Osman, I.; Rebellón-Sánchez, D. E.; Agudelo-Rojas, O. L.; Sanz, A. M.; et al. 2023. Global and cell type-specific immunological hallmarks of severe dengue progression identified via a systems immunology approach. *Nature immunology*, 24(12): 2150–2163.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hao, M.; Gong, J.; Zeng, X.; Liu, C.; Guo, Y.; Cheng, X.; Wang, T.; Ma, J.; Zhang, X.; and Song, L. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 1–11.

Haykin, S. 1998. *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Jessop, Z. M.; Al-Sabah, A.; Simoes, I. N.; Burnell, S. E.; Pieper, I. L.; Thornton, C. A.; and Whitaker, I. S. 2020. Isolation and characterisation of nasoseptal cartilage stem/progenitor cells and their role in the chondrogenic niche. *Stem Cell Research & Therapy*, 11: 1–13.

Karaayvaz, M.; Cristea, S.; Gillespie, S. M.; Patel, A. P.; Mylvaganam, R.; Luo, C. C.; Specht, M. C.; Bernstein, B. E.; Michor, F.; and Ellisen, L. W. 2018. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature communications*, 9(1): 3588.

Keles, F. D.; Wijewardena, P. M.; and Hegde, C. 2023. On the computational complexity of self-attention. In *International Conference on Algorithmic Learning Theory*, 597–619. PMLR.

Kharchenko, P. V. 2021. The triumphs and limitations of computational methods for scRNA-seq. *Nature methods*, 18(7): 723–732.

Kingma, D. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolodziejczyk, A. A.; Kim, J. K.; Svensson, V.; Marioni, J. C.; and Teichmann, S. A. 2015. The technology and biology of single-cell RNA sequencing. *Molecular cell*, 58(4): 610–620.

Kravets, V.; and Benninger, R. K. 2020. From the transcriptome to electrophysiology: searching for the underlying cause of diabetes. *Cell Metabolism*, 31(5): 888–889.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, B.; Zhang, W.; Guo, C.; Xu, H.; Li, L.; Fang, M.; Hu, Y.; Zhang, X.; Yao, X.; Tang, M.; et al. 2022. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature methods*, 19(6): 662–670.

Macosko, E. Z.; Basu, A.; Satija, R.; Nemesh, J.; Shekhar, K.; Goldman, M.; Tirosh, I.; Bialas, A. R.; Kamitaki, N.; Martersteck, E. M.; et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5): 1202–1214.

Norman, T. M.; Horlbeck, M. A.; Replogle, J. M.; Ge, A. Y.; Xu, A.; Jost, M.; Gilbert, L. A.; and Weissman, J. S. 2019. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455): 786–793.

Olaniru, O. E.; Kadolsky, U.; Kannambath, S.; Vaikkinen, H.; Fung, K.; Dhami, P.; and Persaud, S. J. 2023. Single-cell transcriptomic and spatial landscapes of the developing human pancreas. *Cell Metabolism*, 35(1): 184–199.

Perez, R. K.; Gordon, M. G.; Subramaniam, M.; Kim, M. C.; Hartoularos, G. C.; Targ, S.; Sun, Y.; Ogorodnikov, A.; Bueno, R.; Lu, A.; et al. 2022. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589): eabf1970.

Ramachandran, P.; Matchett, K. P.; Dobie, R.; Wilson-Kanamori, J. R.; and Henderson, N. C. 2020. Single-cell technologies in hepatology: new insights into liver biology and disease pathogenesis. *Nature reviews Gastroenterology & hepatology*, 17(8): 457–472.

Remsik, J.; Tong, X.; Kunes, R. Z.; Li, M. J.; Osman, A.; Chabot, K.; Sener, U. T.; Wilcox, J. A.; Isakov, D.; Snyder, J.; et al. 2023. Leptomeningeal anti-tumor immunity follows unique signaling principles. *bioRxiv*, 2023–03.

Sasagawa, Y.; Nikaido, I.; Hayashi, T.; Danno, H.; Uno, K. D.; Imai, T.; and Ueda, H. R. 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology*, 14: 1–17.

Schirmer, L.; Velmeshev, D.; Holmqvist, S.; Kaufmann, M.; Werneburg, S.; Jung, D.; Vistnes, S.; Stockley, J. H.; Young, A.; Steindel, M.; et al. 2019. Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature*, 573(7772): 75–82.

Shen, H.; Liu, J.; Hu, J.; Shen, X.; Zhang, C.; Wu, D.; Feng, M.; Yang, M.; Li, Y.; Yang, Y.; et al. 2023. Generative pretraining from large-scale transcriptomes for single-cell deciphering. *Iscience*, 26(5).

Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B. B.; Siddiqui, A.; et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5): 377–382.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in neural information processing systems*, 33: 1513–1524.

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

Thomas, M. F.; Slowikowski, K.; Manakongtreecheep, K.; Sen, P.; Samanta, N.; Tantivit, J.; Nasrallah, M.; Zubiri, L.; Smith, N. P.; Tirard, A.; et al. 2024. Single-cell transcriptomic analyses reveal distinct immune cell contributions to epithelial barrier dysfunction in checkpoint inhibitor colitis. *Nature Medicine*, 1–14.

Tian, L.; Dong, X.; Freytag, S.; Lê Cao, K.-A.; Su, S.; JalalAbadi, A.; Amann-Zalcenstein, D.; Weber, T. S.; Seidi, A.; Jabbari, J. S.; et al. 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature methods*, 16(6): 479–487.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Voigt, A. P.; Mullin, N. K.; Stone, E. M.; Tucker, B. A.; Scheetz, T. E.; and Mullins, R. F. 2021. Single-cell RNA sequencing in vision research: Insights into human retinal health and disease. *Progress in retinal and eye research*, 83: 100934.

Wen, H.; Tang, W.; Dai, X.; Ding, J.; Jin, W.; Xie, Y.; and Tang, J. 2023. CellPLM: pre-training of cell language model beyond single cells. *bioRxiv*, 2023–10.

Wheeler, M. A.; Clark, I. C.; Tjon, E. C.; Li, Z.; Zandee, S. E.; Couturier, C. P.; Watson, B. R.; Scalisi, G.; Alkwai, S.; Rothhammer, V.; et al. 2020. MAFG-driven astrocytes promote CNS inflammation. *Nature*, 578(7796): 593–599.

Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; and Yao, J. 2022. 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.

Yang, X.; Liu, G.; Feng, G.; Bu, D.; Wang, P.; Jiang, J.; Chen, S.; Yang, Q.; Zhang, Y.; Man, Z.; et al. 2023. Genecompass: Deciphering universal gene regulatory mechanisms with knowledge-informed cross-species foundation model. *bioRxiv*, 2023–09.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10795–10816.

Zheng, G. X.; Terry, J. M.; Belgrader, P.; Ryvkin, P.; Bent, Z. W.; Wilson, R.; Ziraldo, S. B.; Wheeler, T. D.; McDermott, G. P.; Zhu, J.; et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1): 14049.

Zhong, H.; Huan, X.; Zhao, R.; Su, M.; Yan, C.; Song, J.; Xi, J.; Zhao, C.; Luo, F.; and Luo, S. 2023. Peripheral immune landscape for hypercytokinemia in myasthenic crisis utilizing single-cell transcriptomics. *Journal of Translational Medicine*, 21(1): 564.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16489–16498.