Improving Classification of Cell Types in Acute Myeloid Leukemia with Self-guided Masking Technique

Amirreza Naziri
Department of Computer Science
York University
Toronto, Canada
naziriam@yorku.ca

Arash Asgari
Department of Computer Science
York University
Toronto, Canada
arashasg@yorku.ca

Eleftherios Sachlos
Department of Mechanical Engineering
York University
Toronto, Canada
sachlos@yorku.ca

Aijun An
Department of Computer Science
York University
Toronto, Canada
aan@yorku.ca

Laleh Seyyed-Kalantari
Department of Computer Science
York University
Toronto, Canada
lsk@yorku.ca

Abstract

Acute myeloid leukemia (AML) is a rare but important disease. Because it has many different features and behaviors, classifying its cell types with traditional methods is both difficult and costly. Transformer-based foundation models (FMs) are useful for analyzing biological data, and they usually use random masking during training. But normal uniform random masking selects genes without checking how important they are. To solve this, we propose a self-guided masking method. This method learns which gene positions are most useful to mask at each training step. We show that our approach improves FM training and performs better than uniform masking in cell-type classification for AML.

1 Introduction

Acute Myeloid Leukemia (AML) represents a complex, aggressive, diverse, and extremely rare group of blood cancers. Its clinical relevance stems not only from its severity but also from its remarkable heterogeneity [15]. Conventional methods for identifying and classifying cell types in AML often fall short, as they are time-consuming, costly, and unable to capture the underlying diversity of patterns fully [4, 13, 6]. To overcome these limitations, advances in high-throughput technologies have provided new avenues for dissecting tumor complexity. In particular, Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by measuring transcriptomes at the resolution of individual cells [10, 11].

In parallel, transformer-based foundation models (FMs) are emerging, which are large neural networks pretrained in an unsupervised manner on massive unlabeled data and fine-tuned for diverse

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

downstream tasks. To train FMs, masked language modeling, which masks a small fraction of tokens and reconstructs them from context to capture bidirectional dependencies, has been used to let the model learn the dataset. [9].

Despite the success of masked and autoregressive language modeling in natural language and vision domains, applying these paradigms to scRNA-seq data remains challenging. scRNA-seq data are characterized by extreme sparsity, high dimensionality (tens of thousands of genes), and substantial technical noise [18, 8], which complicate tasks such as cell-type classification [3]. In the case of masked language modeling, uniform random masking often selects non-informative genes that contain limited biological signal. As a result, uniform random masking reduces the learning efficiency of the model in this domain.

In this work, we propose a novel self-guided masking approach, tailored to the unique challenges of scRNA-seq data. Our method can mask genes based on their importance in the dataset, enabling efficient, robust scaling of FMs on scRNA-seq data. Our contributions are as follows:

- We introduce the self-guided masking approach, a dynamic, efficient, attention-guided masking strategy for sparse and noisy data, and compare it to uniform random masking.
- We show our approach outperforms random masking in training and improves cell type classification performance.
- We use biological insights to show the self-guided masking's ability to select important genes.

2 Related works

Foundation models for single-cell RNA data: Recent work has begun to translate FMs paradigm into transcriptomics. For example, scBERT [17], an encoder-only FM, adapts the BERT masked-language framework to gene expression by learning contextualized embeddings of genes and cells through token-level reconstruction on discretized counts. scGPT [1], a decoder-only FM, learns joint embeddings of cells and genes via masked language modeling on multi-omic single-cell data; xTrimoGene [5] exploits sparse cross-attention to process all 20,000 genes without manual discretization, completing that, scFoundation [7] scales to 100 M parameters by incorporating read-depth–aware objectives over 50 M cell profiles. The core idea in both xTrimoGene and scFoundation is skipping zero and masked tokens in encoder layers, which can significantly reduce the computing needs without losing performance. To the best of our knowledge, scFoundation is the only encoder-decoder FM model in this area. All of these models use random masking; thus, it is not possible to investigate masked tokens and interpret how masking information is utilized throughout the layers.

Self-guided masking in self-supervision: The paper by [16] introduces Self-guided Masked Autoencoders (SMA) across image, text, chemical-graph, and particle-physics modalities, using the model's own early attention maps to steer mask selection. However, SMA relies on a complex teacher–student framework and was evaluated on relatively dense, low-dimensional inputs; it does not address the severe zero/non-zero imbalance in scRNA-seq, where most masks still fall on uninformative zeros. The ADIOS approach [12] further couples an adversarially trained masking network with a Siamese encoder to produce maximally challenging masks, yielding strong gains on vision benchmarks. Yet its adversarial objective and multi-mask sampling incur substantial complexity and do not target the sparsity characteristics of single-cell data. To the best of our knowledge, none of these methods is explicitly designed for the high-dimensional, noise-prone sparsity of scRNA-seq. In contrast, our masking approach provides a lightweight, attention-driven mechanism specifically designed to focus on informative genes in single-cell transcriptomes.

3 Method

In this paper, we introduce a novel masking approach that utilizes the attention mechanism to learn which tokens (in our case, genes) are more suitable for masking during the pretraining step. Our method uses learnable parameters for learning and storing the critical features. Then, using attention and multinomial sampling, it selects the desired tokens for masking based on the stored features in learnable parameters. Below are the details of our method:

Latent Cross-Attention Module: We consider an input dataset consisting of gene expression profiles from single cells. Each cell is treated as a separate data instance, represented by a sequence of gene tokens. For a given cell, the input sequence is a vector of gene expression values ordered by genes. We introduce $Z \in \mathbb{R}^{N \times D_z}$, a trainable set of N latent vectors, each of embedding dimension D_z. These latent vectors attend to input token sequences. The ultimate goal is to extract a global summary that guides downstream masking decisions.

Self-guided Masking: To find the critical tokens for masking, we use per-head attention weights. Then, we aggregate attention weights across the H cross-attention heads by summing. This collapses the per-head weights into a single attention score for each latent n and token position l. Attention maps often exhibit heavy overlap. This makes it challenging to construct diverse or informative masking patterns when using all attention heads or latent vectors simultaneously, resulting in overly concentrated, redundant masking. To address this problem, we focus only on a subset of latent vectors. We randomly select $R \subset \{1, \ldots, N\}$. Only the R latents are used to drive masking per cell, reducing noise from less-informative latents. We then collapse these selected latents' scores into token-level importance by summing across selected latents. Then, we use this importance to find which tokens are better to be masked.

Dataset: Hematopoietic Niche scRNA-seq (Acute Myeloid Leukemia) dataset [2] includes scRNA-seq profiles of 350K cells collected from approximately 50 human donors. Each cell is measured across around 16K genes. This dataset captures a diverse range of biological conditions and is particularly well-suited for evaluating models on complex, heterogeneous transcriptomic data spanning multiple individuals and cell types. See appendix Fig. 1 for more details and cell types.

Model: Due to the heavy computation cost of pretraining FMs, without loss of generality, we only used scFoundation [7] as our backbone, which is the most recent and powerful one among other FMs [7]. Additionally, we ensured that the dataset used in our experiments does not overlap with those in the pretraining corpus of the scFoundation model [7].

4 Experiments

To validate the self-guided masking algorithm, we employ a two-stage training protocol: self-supervised pretraining and supervised finetuning.

Self-supervised Pretraining Setup: In this step, the network learns to reconstruct the masked gene expression values using the proposed attention-driven masking. We used pretrained weights of our backbone [7] because training from scratch requires extensive computing resources. We train the parameters using Adam with two learning-rate groups: $\text{LR}_{\text{new}} = 10^{-4}$ for layers introduced by our method and $\text{LR}_{\text{backbone}} = 10^{-5}$ for the rest. For more details, please refer to Appendix Section C.

Cell Type Classification Setup: A lightweight classification head is added on top of the pretrained encoder and can be trained on different downstream tasks. We append a two-layer MLP on top of the pooled encoder output. Also, we freeze all latents, cross-attention weights, and all adapter layers during finetuning, as they are only responsible for self-supervised training. For more details, please refer to Appendix Section C. We evaluate partial fine-tuning and linear probing schemes to assess the usefulness of the pre-trained encoder. In partial finetune, trainable parameters are: token embeddings, positional embeddings, the *last* transformer block of the encoder, and the classification head. On the other hand, for linear probing, the trainable parameters are only the classification head. It is worth mentioning that all masking and cross-attention modules were frozen during finetuning. This is because all those modules are solely added to make the pretraining more robust.

Biological Evaluation: Finally, to show that our approach masks the most important genes, we tried two biological experiments: First, we analyzed the relation between the average of expression values and the number of times that each gene is masked. This can help us understand the pattern of masked genes. Second, we attempted Gene Ontology (GO) enrichment [14], which identifies the biological processes of a set of genes. Using GO enrichment, we can determine if the masked gene is informative or not.

Table 1: Partial-finetune and Linear-probe cell-type classification, comparing self-guided masking (Ours) F1-score vs random masking. (FT: finetune, LP: linear-probe, P: pretraining, Orig. = original backbone weights [7]. **Bolded numbers** are the best. Results are reported as the mean ± confidence interval from 5-fold cross-validation on the test set, with each fold held out once.

Experiment	Masking	Train	F1-score
Partial-finetune	Random	Orig.[7] + FT	80.04 ± 0.001
	Random	P+FT	79.33 ± 0.001
	Self-guided (Ours)	P+FT	**81.33 ± 0.002**
Linear-probe	Random	Orig.[7] + LP	**77.36 ± 0.002**
	Random	P+LP	26.51 ± 0.002
	Self-guided (Ours)	P+LP	71.54 ± 0.001

5 Results

Self-guided masking outperforms random masking: The results in Table 1 demonstrate that self-guided masking consistently outperforms uniform random masking across both partial finetune (P+FT) and (Orig.+FT). In the P+FT setting, random masking (P+FT) reaches only 79.33% (versus 80.04% for Orig.+FT), whereas self-guided masking attains 81.33%, a +2% absolute gain. For linear-probing (LP), as demonstrated in Table 1, self-guided masking outperforms uniform masking across (P+LP). Under P+LP, random masking collapses to 26.51%; However, self-guided masking recovers to about 71.54%, nearly matching the 77.36% original backbone.

Partial-finetuning (P+FT) vs. linear probing (P+LP): Pretraining with uniform random masking shifts the model's weights away from its original optimum, degrading downstream f1-score. For example, in Table 1 (Orig+FT), F1-score drops from 80.04% to (P+FT) 79.33%. Also, in Table 1, (Orig+LP) from 77.36% to (P+LP) 26.51%. In other words, the original FM with FT/LP (Orig+FT/LP) has better performance compared to the pretrained model of the new dataset with FT/LP (P+FT/LP).

In contrast, self-guided masking steers the model toward representations that remain compatible with or even superior to the initial backbone (Orig+FT/LP). As it is shown in Table 1, there is a gap between (Orig+LP) and self-guided masking (P+LP). The potential reason is that self-guided masking (P+LP) deviates the model parameters from the optimal (Orig+LP). However, since all encoder layers are frozen in LP, self-guided masking (P+LP) has limited trainable parameters to adjust to the downstream task, leading to lower performance. In self-guided masking (P+FT), since both MAM learnable parameters and model learnable parameters have the capacity to update, the model does not face this problem, and self-guided masking (P+FT) can consistently outperform all cases, including (Orig+FT).

Analyzing Learned Masked Tokens: In this part, we examined how often each gene was chosen for masking. Specifically, for every gene in the training set, we (i) count the total number of masking events across all iterations and (ii) compute that gene's mean expression level conditional on being masked. Genes with higher expression levels often carry more biological information. If these genes are more frequently masked, it's a sign that the model may be learning to focus on informative features. Plotting these two quantities against one another (see Fig. 2 in Appendix section D) reveals a pronounced positive correlation: genes that register non-zero expression, and especially those with higher expression magnitudes, are selected for masking far more frequently than genes that sit at or near zero. This pattern is exactly what we would expect if self-guided masking is successfully identifying the most information-rich coordinates of the input profile.

Biological meaning of the masked gene: Here, we want to explore why self-guided masking improves model performance. Fig. 3 (Appendix section E) shows the top enriched Gene Ontology (GO) [14] biological processes among the most frequently masked genes by our approach. The enrichment shows strong overrepresentation of transcriptional regulation processes, including regulation of DNA-templated transcription (GO:0006355), regulation of transcription by RNA polymerase II (GO:0006357), and both positive (GO:0045893) and negative (GO:0045892, GO:0000122) regulation of transcription. Additional enriched pathways include regulation of apoptosis (GO:0042981) and cell population proliferation (GO:0042127). These biological processes are highly relevant to leukemogenesis, as they control cell identity, survival, and uncontrolled proliferation in AML. By masking such biologically meaningful genes during training, our approach (self-guided masking) is pushed to reconstruct signals that reflect core aspects of AML.

6 Discussion and Conclusion

We introduced the self-guided masking, an approach that enables effective self-supervised pretraining on sparse scRNA-seq AML data by focusing masking on informative genes. We showed that our approach not only avoids the degradation induced by uniform random masking but also consistently improves AML cell-type classification under both partial fine-tune and linear-probe protocols.

Limitations and future work: Here, we provide limitations and potential future work. Our results rely on an existing, well-trained FM. A poorly pretrained backbone or domain-mismatch between pretraining and the downstream task impacts the total performance. Therefore, pretraining an FM from scratch using our approach can be beneficial. In addition, while self-guided masking adds minimal overhead, the large underlying FMs contain millions of parameters. Finetuning such a large model on full-length gene profiles required multi-day runs, limiting our ability to perform exhaustive hyperparameter tuning.

7 Acknowledgments

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant, the Connected Minds initiative funded by the Canada First Research Excellence Fund (CFREF), the Center for AI & Society at York University (AI-EGS), and the Vector Institute for providing access to high-performance computing resources.

References

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

[2] Sarah Ennis, Alessandra Conforte, Eimear O'Reilly, Javid Sabour Takanlu, Tatiana Cichocka, Sukhraj Pal Dhami, Pamela Nicholson, Philippe Krebs, Pilib Ó Broin, and Eva Szegezdi. Cell-cell interactome of the hematopoietic niche and its changes in acute myeloid leukemia. *Iscience*, 26(6), 2023.

[3] Nafiseh Erfanian, A Ali Heydari, Adib Miraki Feriz, Pablo Iañez, Afshin Derakhshani, Mohammad Ghasemigol, Mohsen Farahpour, Seyyed Mohammad Razavi, Saeed Nasseri, Hossein Safarpour, et al. Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomedicine & Pharmacotherapy*, 165:115077, 2023.

[4] E. E. Genç, İ. S. Saraç, H. Arslan, and A. E. Eşkazan. Diagnostic and treatment obstacles in acute myeloid leukemia: Social, operational, and financial. *Oncology and Therapy*, 11(2):145–152, 2023.

[5] Jing Gong, Minsheng Hao, Xingyi Cheng, Xin Zeng, Chiming Liu, Jianzhu Ma, Xuegong Zhang, Taifeng Wang, and Le Song. xtrimogene: an efficient and scalable representation learner for single-cell rna-seq data. *Advances in Neural Information Processing Systems*, 36:69391–69403, 2023.

[6] F. Guijarro, M. Garrote, N. Villamor, D. Colomer, J. Esteve, and M. López-Guerra. Novel tools for diagnosis and monitoring of aml. *Current Oncology (Toronto, Ont.)*, 30(6):5201–5213, 2023.

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

[8] Yusuke Imoto, Tomonori Nakamura, Emerson G Escolar, Michio Yoshiwaki, Yoji Kojima, Yukihiro Yabuta, Yoshitaka Katou, Takuya Yamamoto, Yasuaki Hiraoka, and Mitinori Saitou.

Resolution of the curse of dimensionality in single-cell rna sequencing data analysis. *Life Science Alliance*, 5(12), 2022.

[9] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[10] 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.

[11] Antoine-Emmanuel Saliba, Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic Acids Research*, 42(14):8845–8860, 07 2014.

[12] Yuge Shi, N. Siddharth, Philip H. S. Torr, and Adam R. Kosiorek. Adversarial masking for self-supervised learning, 2022.

[13] Y. Tazi, J. E. Arango-Ossa, Y. Zhou, et al. Unified classification and risk-stratification in acute myeloid leukemia. *Nature Communications*, 13:4622, 2022.

[14] Paul D. Thomas, Dustin Ebert, Anushya Muruganujan, Tremayne Mushayahama, Laurent-Philippe Albou, and Huaiyu Mi. Panther: Making genome-scale phylogenetics accessible to all. *Protein Science*, 31(1):8–22, 2022.

[15] Anusha Vakiti, Samuel B. Reynolds, and Prerna Mewawalla. *Acute Myeloid Leukemia*. StatPearls Publishing, Treasure Island (FL), 2025. [Updated 2024 Apr 27].

[16] Johnathan Xie, Yoonho Lee, Annie S. Chen, and Chelsea Finn. Self-guided masked autoencoders for domain-agnostic self-supervised learning, 2024.

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

[18] Yi Zhang, Yin Wang, Xinyuan Liu, and Xi Feng. Pbimpute: Precise zero discrimination and balanced imputation in single-cell rna sequencing data. *Journal of Chemical Information and Modeling*, 65(5):2670–2684, 2025. PMID: 39957720.

A Data Preparation

We perform standard quality control on the single-cell dataset to ensure that only high-quality cells and informative genes are retained. Low-complexity cells (those with very few detected genes) and genes expressed in only a handful of cells are filtered out. Next, we partition the cleaned data into training, validation, and test sets (train: 80%, validation: 10%, test: 10%) in a way that avoids leaking information across splits. Rather than splitting at the level of individual cells, we split by donor—so that all cells from a given individual end up in the same subset. To maintain balance, we stratify the donors by their predominant combinations of cell type and timepoint, thereby ensuring that each subset reflects the overall diversity of the study. All details and codes to replicate data loading and processing are available in the provided code.

B UMAP visualization of Hematopoietic Niche (AML) dataset

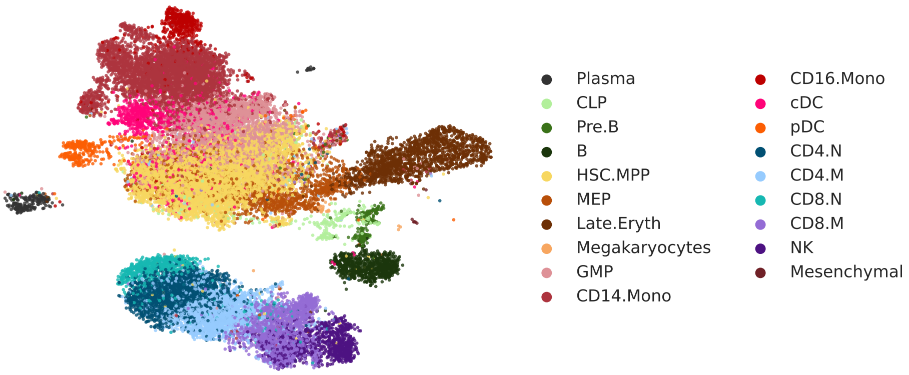

Figure 1: UMAP visualization of Hematopoietic Niche scRNA-seq (Acute Myeloid Leukemia) [2]. Each point represents a single cell colored by (top) cell type.

C Hyper-parameters

For self-supervised pretraining, we used a latent dimension $D_z = 256$ and $N = 128$ latents. The model employs 8 cross-attention heads, and gradient clipping was set to 1.0. The learning rate was configured as 1×10^{-4} for newly added layers and 1×10^{-5} for the backbone. We applied a weight decay of 5×10^{-4} and set $\lambda_{\text{attn}} = 1 \times 10^{-5}$. Training used a physical batch size of 1 with 5 accumulation steps, and early stopping was triggered after 5 epochs without improvement.

For supervised finetuning, we maintained the latent dimension $D_z = 256$ and the number of latents $N = 128$. The learning rate was set to 1×10^{-4}, and gradient clipping remained at 1.0. We used batch sizes of 1 and 8 for different stages (partial and linear), along with 5 accumulation steps. The early-stopping patience was again set to 5 epochs. A weight decay of 5×10^{-4} was applied, and we used 8 cross-attention heads. Additionally, the hidden dimension of the linear head (h_l) was set to 256.

D Gene Masking Frequency versus Average Expression

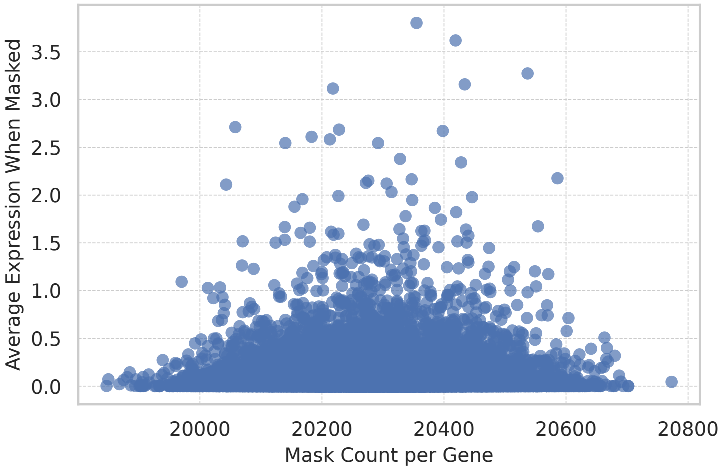

Figure 2: Scatter plot of each gene's masking frequency against its mean expression when masked, illustrating that self-guided masking preferentially targets genes with higher expression levels.

E GO Enrichment of Masked Genes

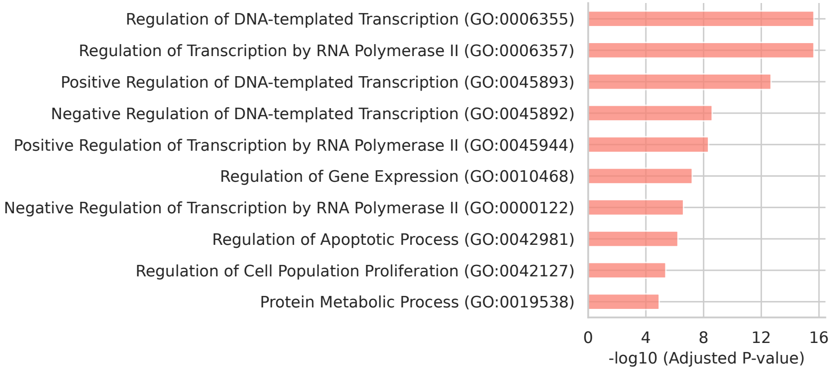

Figure 3: Top enriched GO biological processes among the most frequently masked genes by the self-guided masking approach. The x-axis demonstrates the significance of each biological process (y-axis) derived from masked genes.