scMBERT: A Pre-Trained Deep Learning Model for Single-Cell Multiomic Data Representation and Prediction (Student Abstract)

Xiaojian Chen[1, 2], Kuai Yu[1, 2], Min-Zhi Jiang[1], Cihan Xiao[3], Ziqi Fu[4], Weiqiang Zhou*[1]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[2]Department of Biomedical Engineering, Johns Hopkins University
[3]Center for Language and Speech Processing, Johns Hopkins University
[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health
{xchen279, kyu25, mjiang26, cxiao7, wzhou14}@jhu.edu, ziqi_fu@g.harvard.edu

Abstract

Recent advancements in single-cell sequencing technologies enable the measurement of multiple modalities in individual cells, offering insights into the transcriptome and regulome in various biological systems and human diseases in an unprecedented resolution. However, effectively using these ultra-high-dimensional and large-scale multiomic data to understand gene regulation remains challenging. Inspired by the success of adapting large language models into the genomics field, we develop scMBERT, a BERT framework-based pre-trained deep learning model using single-cell multiomic data. We showed that scMBERT increases model flexibility and performance in downstream tasks like cell type annotation and batch-effect correction, demonstrating the potential of leveraging multiomic data to improve single-cell genomic data analyses.

Introduction

Single-cell sequencing technologies such as single-cell RNA-seq have become the leading method for exploring gene expression heterogeneity in various biological systems and human diseases at the individual cell level. Recently, advancements in single-cell multiomic technologies allow for the measurement of a broader array of modalities, enabling deeper insights into the gene regulatory landscape. These advancements extend our understanding beyond the transcriptome which only tells what genes are expressed, to the regulome which shows how these genes are regulated. Hence, offering a multimodal representation of the same cell that provides insights into how different functional genomic layers influence each other. However, the high sparsity, noise level, and dimensionality of these single-cell multiomic data pose many computational and analytical challenges.

Recent developments in large language models (LLMs) have shown promising applications in various fields. Several studies (Cui et al. 2024; Chen and Zou 2024; Hao et al. 2024; Theodoris et al. 2023; Yang et al. 2022) have adapted deep language models to the single-cell genomic field. By mapping high-dimensional scRNA-seq data into low-dimensional embeddings, these pre-trained models can capture hidden relationships among the genes and perform well when fine-tuned for downstream tasks such as cell type annotation. For instance, scBERT (Yang et al. 2022) adapted the Bidirectional Encoder Representations from Transformers (BERT) framework to the scRNA-seq data using fixed input. After pre-training with a large amount of scRNA-seq data, scBERT was fine-tuned for cell type annotation and showed state-of-the-art (SOTA) performance. Cui et al. (Cui et al. 2024) recently introduced scGPT, which adapted the Generative Pre-trained Transformer (GPT) framework to the scRNA-seq data using variable input. However, these meth-

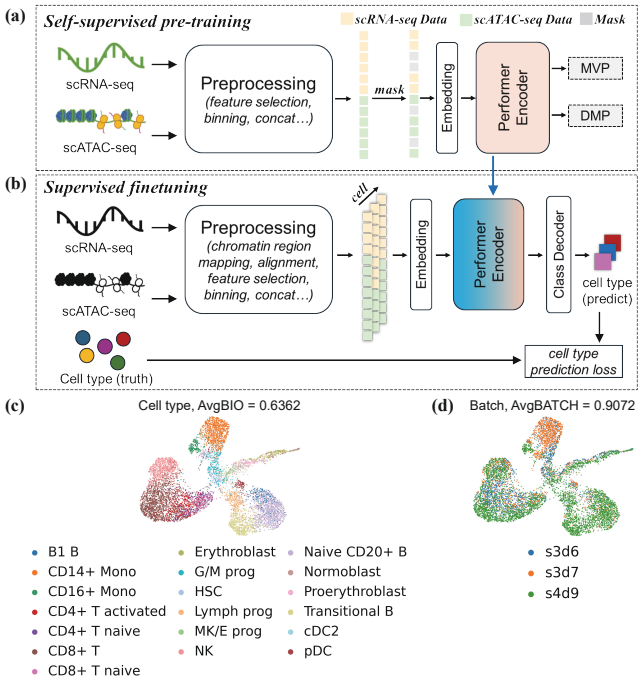

Figure 1: An overview of scMBERT. (a) We pret-rained the model with two tasks: masked value prediction (MVP) and data modality prediction (DMP). (b) The pre-trained model was then fine-tuned for downstream tasks such as cell type annotation. The batch-effect corrected BMMC data annotated by (c) cell types and (d) batch labels.

*Corresponding Author

ods have several limitations. Both models are pre-trained using only scRNA-seq data which lacks the information from the regulome. Although the BERT framework is suitable for unordered input, scBERT is limited to a fixed set of genes as input which can't handle new genes or modalities. While scGPT can handle variable input, the GPT framework is designed for ordered input and requires extra effort to be applied to unordered genomic data.

To address these limitations, we propose a novel framework, scMBERT, which adapts the BERT architecture to single-cell multiomic data using variable input. scMBERT consists of two parts: self-supervised pre-training and supervised fine-tuning (**Fig. 1 (a)-(b)**). We evaluated scMBERT in downstream tasks including cell type annotation and batch-effect correction. We showed that by leveraging multiomic data, scMBERT showed improved performance over using single-omic data alone. scMBERT achieved comparable performance with the SOTA models in cell type prediction and outperformed the SOTA models in batch-effect correction.

Methods

Let $X \in \mathbb{R}^{N \times F \times 3}$ be the raw data matrix, where N denotes the number of cells and F denotes the number of features (i.e., genes or open chromatin regions), with three views of the cell, i.e., gene expression (or peak accessibility) v, gene name (or peak location) g, and data modality m. Given an input sequence $x \in \mathbb{R}^{F \times 3}$ (i.e., a cell), we first preprocess the input by retaining only the non-zero features value, resulting in $x' \in \mathbb{R}^{T \times 3}$, where $T \leq F$ is a variable sequence length. We apply a random downsampling strategy for sequences longer than 20k, and binned the data followed the way of scGPT. Our goal is to model x' using a low-dimensional hidden variable $z \in \mathbb{R}^{T \times H}$, where H denotes the hidden dimension, according to the following two training stages: (1) general-purpose pre-training on large-scale unlabeled data; (2) fine-tuning on smaller datasets for specific applications.

Self-supervised Pre-training on Unlabeled Data. Fig. 1 (a) depicts pre-training using each cell for the following two tasks: mask value prediction and data modality prediction.

#Task 1: We randomly mask the input data value and predict it based on the remaining inputs. Then we utilized cross-entropy loss as the reconstruction loss $\mathcal{L}_{mvp}$, between the true and predicted values.

#Task 2: To regularize the model to prevent overfitting to any single modality, we introduce the task to predict the modality of the reconstructed sequence and compare it to the original modality input with cross-entropy loss $\mathcal{L}_{dmp}$, which encourages the model to preserve critical modality-specific information and learn cross-modal consistency.

Finally, the overall pre-training loss function $\mathcal{L}_p$ can be expressed as $\mathcal{L}_p = \mathcal{L}_{mvp} + \lambda \mathcal{L}_{dmp}$ where λ is a tuning parameter to adjust the relative weight between the loss terms. The model is pre-trained on the ENCODE 4 portal single-cell multiomic datasets (Kagda et al. 2023), which contains two data modalities including scRNA-seq and scATAC-seq. We retrieved the 47 ENCODE datasets covering 13 general tissue types, and merge them together under the unified open chromatin regions of interest (Sheffield et al. 2013).

Method	Accuracy	AvgBIO	AvgBATCH
scBERT (RNA)	75.32%	0.5687	0.8696
scGPT (RNA)	**76.68%**	0.5272	0.8039
scMBERT (RNA)	74.26%	0.6298	0.9013
scMBERT (Multiome)	75.56%	**0.6362**	**0.9072**

Table 1: Evaluation results on the cell type annotation task.

Supervised Fine-tuning for Downstream Tasks. Fine-tuning is designed for different downstream tasks, such as cell type annotation shown in **Fig 1 (b)**. Since the sequence length is variable, we can also add new features or modalities to the vocabulary as alignment. The scATAC-seq data should first be mapped to the open chromatin regions as pretrained. The processed data was embedded using the pre-trained encoder where the initial layer is frozen, and then predict the cell type labels through a classification decoder. We define the cell-type prediction loss $\mathcal{L}_f$ using cross-entropy as $\mathcal{L}_f = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} y_i^{(n)} \log(\hat{y}_i^{(n)})$, where across N cells, $\hat{y}_i$ is the predicted probability for class i, y_i is the true cell type, and C is the total number of classes.

Case Study: Cell Type Annotation

Experimental Setup. To evaluate scMBERT's ability to represent multimodal data, the single-cell multiome dataset from the NeurIPS public challenge (Luecken et al. 2021) is used to fine-tune scMBERT for cell type annotation. The dataset comprises 69,249 bone marrow cells (BMMCs) from 12 healthy human donors. To demonstrate the model's generalizability to new data, we retain 3 donors as an independent testing set. Consequently, the training, validation, and testing sets contain 49,179, 12,295, and 7,775 cells with both scRNA-seq and scATAC-seq data, respectively. An early stopping strategy is applied, where training stops if the validation loss does not improve for 10 consecutive epochs, and the model from the best-performing epoch is used.

Results. We compare scMBERT with scBERT and scGPT. All models are fine-tuned and tested using the same datasets. We adopt the metrics from scGPT to evaluate the models' performance in terms of biological representation (AvgBIO) and batch-effect correction (AvgBATCH). As shown in **Table 1**, scMBERT shows comparable performance with the SOTA models in terms of cell type prediction accuracy and outperforms the SOTA models in batch-effect correction. We also find that scMBERT fine-tuned with both scRNA-seq and scATAC-seq data outperforms scMBERT fine-tuned with only scRNA-seq. **Fig 1 (c)** and **(d)** show the UMAP of cell embedding from scMBERT colored by cell types and batch labels, respectively. The cell embedding shows that our model can eliminate the effect of batch.

Future direction. The current scMBERT model is pre-trained on a relatively small amount of single-cell multiomic datasets from ENCODE. We believe that using more multiomic data or incorporating single-omic data into the pre-training process will further improve scMBERT's performance. Furthermore, testing on other downstream tasks such as missing data prediction and signal enhancement should be conducted to explore scMBERT's full potential.

Acknowledgments

This work is supported by the Johns Hopkins Bloomberg School of Public Health Faculty Innovation Fund. All computations were performed on the Joint High Performance Computing Exchange (JHPCE), organized by the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health, and the Maryland Advanced Research Computing Center (MARCC) Rockfish cluster. We would like to thank all members of the Hongkai Ji group for valuables discussions and advice.

References

Chen, Y.; and Zou, J. 2024. GenePT: a simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv*, 2023–10.

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.

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.

Kagda, M. S.; Lam, B.; Litton, C.; Small, C.; Sloan, C. A.; Spragins, E.; et al. 2023. Data navigation on the ENCODE portal. *arXiv preprint arXiv:2305.00006*.

Luecken, M. D.; Burkhardt, D. B.; Cannoodt, R.; Lance, C.; Agrawal, A.; et al. 2021. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track*.

Sheffield, N. C.; Thurman, R. E.; Song, L.; Safi, A.; Stamatoyannopoulos, J. A.; Lenhard, B.; Crawford, G. E.; and Furey, T. S. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research*, 23(5): 777–788.

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.

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.