CLM-Access: A Specialized Foundation Model for High-Dimensional Single-Cell ATAC-Seq Analysis

Ziqiang Liu[1*], Bowen Li[1*], Zhenyu Xu[12*], Yantao Li[1], Junwei Zhang[1], Chulin Sha[1†], Xiaolin Li[1†]

[1]Hangzhou Institute of Medicine, Chinese Academy of Sciences
[2]Faculty of Health Science, University of Macau
liuzq_dlmu@163.com, libowen20@mails.ucas.ac.cn, xuzy1992.joey@gmail.com, liyantao@him.cas.cn, zhangjunwei@him.cas.cn, shachulin@him.cas.cn, xiaolinli@ieee.org

Abstract

Inspired by the success of large language models (LLMs) in natural language processing, cell language models (CLMs) have emerged as a promising paradigm to learn cell representations from high-dimensional single-cell data—particularly transcriptomic profiles from scRNA-seq. These foundation models have shown remarkable potential across a variety of downstream applications. However, there remains a lack of foundation models for scATAC-seq data, which measures chromatin accessibility at single-cell level and is critical for decoding epigenetic regulation. Developing such model is considerably more challenging due to the unique characteristics of scATAC-seq data, including the vast number of chromatin regions, lack of standardized annotations, extreme sparsity, and near-binary distributions. To address these challenges, we systematically explore various strategies and propose CLM-Access, a specialized foundation model for scATAC-seq data. CLM-Access incorporates three main innovations: (1) an unified data processing pipeline that maps 2.8 million cells onto an unified reference of over 1 million chromatin regions; (2) a specialized patching and embedding strategy to effectively manage high-dimensional inputs; and (3) a tailored masking and loss function design that preserves fine-grained regional information while enhancing training efficiency and representation quality. With comprehensive benchmarks, we show that CLM-Access significantly outperforms existing methods in key downstream tasks, including batch effect correction, cell type annotation, RNA expression prediction, and multi-modal integration. This work establishes a scalable and interpretable foundation model for single-cell epigenomic analysis and expands the application of CLMs in single-cell research.

Code — https://github.com/HIM-AIM/CLM-Access

Introduction

The rapid advancement of single-cell omics technologies has resulted in an explosion of high-dimensional data, posing challenges in effectively delineating meaningful biological insights. The scATAC-seq (single-cell Assay for Transposase-Accessible Chromatin using sequencing) data analysis is a typical example. This technique is used to assess open chromatin regions of a cell(Buenrostro et al. 2015), which enables the localization of active cis-regulatory elements (CREs) such as promoters, enhancers, etc., thereby elucidating the dynamics of gene regulation in various cell types(Zhang et al. 2021; Zu et al. 2023; Cusanovich et al. 2018). In recent years, over a few large-scale scATAC-seq data atlases have been generated(Domcke et al. 2020; Li et al. 2023b), offering resources for investigating CREs and their regulating genes during key biological processes such as embryonic development and brain regional differentiation (Klemm, Shipony, and Greenleaf 2019; Tsompana and Buck 2014; Minnoye et al. 2021). However, scATAC-seq data analysis encounters several primary challenges: First, the scale of candidate CREs (cCREs) reaches 10^5-10^6 orders of magnitude, resulting in a drastic increase in information encoding dimension; Second, the data exhibit extreme sparsity, leading to a discrete distribution of signals across the genome; Third, the binary nature of the signals (accessible/inaccessible) makes it difficult to directly reflect the interaction hierarchies of the CREs(Ji et al. 2020; Pijuan-Sala et al. 2020; Chiou et al. 2021). In addition, unlike scRNA-seq data where standard gene reference is available, scATAC-seq data lacks universally defined genome locations or annotations of cCREs, which makes integrating scATAC-seq data from different experiments particularly challenging(González-Blas et al. 2019; Xiong et al. 2019). These characteristics collectively constrain the scATAC-seq data analysing methods development in revealing gene regulatory networks(Zeng et al. 2024; Li et al. 2021; Tang et al. 2024).

Current scATAC-seq data analytical tools can be categorized into task-specific methods and integrated analytical pipelines(Stuart et al. 2021; Granja et al. 2021). Task-specific methods exhibit limited generalizability, requiring substantial reconfiguration of model architectures to adapt to new scenarios(Danese et al. 2021; Fang et al. 2021). Whereas integrated methods despite incorporating multi-module functionalities, often involve a high proportion (typically exceeding 80%) of pre-filtering of cCREs(Xiong et al. 2022; Yuan and Kelley 2022). This result in dimensionality

[*]These authors contributed equally.
[†]Corresponding author

reduction and information entropy loss in the data, thereby undermining the model's capacity to decipher complex regulatory relationships(Cui et al. 2024b; Chen, Chen et al. 2022).

More recently, several cell language foundation models have been successfully developed for scRNA-seq data, including scBERT(Yang et al. 2022), Geneformer(Theodoris et al. 2023), scGPT(Cui et al. 2024a), and scFoundation(Hao et al. 2024). These approaches treat genes as tokens and transform scRNA-seq data into a language-like format. Leveraging large-scale datasets and self-supervised pretraining strategies, these models can learn generalizable representations of cells that facilitate data integration and biological interpretation. They also exhibit strong potential in downstream tasks such as cell type annotation, biomarker discovery, and in silico gene perturbation analysis, etc. These studies provide encouraging evidence that a foundation model tailored for scATAC-seq data could address key limitations of existing methods and offer a more powerful framework for interpreting epigenetic regulation at single-cell resolution. However, there is still much to explore in terms of how to best represent, train, and scale models that are well-suited for the complexity of scATAC-seq data.

In response to these unmet needs, we present CLM-Access—a Transformer-based cell language foundation model specifically designed for scATAC-seq data. We begin by establishing a standardized data processing pipeline to construct an integrated Human-scATAC-seq dataset encompassing approximately 2.8 million cells for model pretraining. To address the challenges inherent in scATAC-seq data, we systematically explored multiple strategies for tokenization, masking, and loss function design. To handle the high dimensionality, we partitioned accessible chromatin regions into patches, each consisting of a fixed number of peaks (corresponding to cCREs mapped in the genome), and treated each patch as a token. The model inputs combine token embeddings with peak-level representations and are processed through a Transformer architecture to perform masked peak reconstruction, optimized using binary cross-entropy (BCE) loss. This design allows the model to incorporate all cCREs as input, effectively accommodate data sparsity, meanwhile ensure high training efficiency and robust learning performance. Hence, the final CLM-Access architecture incorporates a specialised patching and embedding module, with a customised masking and training design. With fine-tuning, CLM-Access achieves considerably better performance than existing methods on key tasks including batch effect correction, cell type annotation, RNA expression prediction and multi-modal integration, providing a powerful tool for scATAC-seq data research.

Related Work

Deep-learning Methods for Analyzing scATAC-seq Data

The more recent deep learning methods for scATAC-seq data analysis are often based on variational autoencoder (VAE) architecture (Xiong et al. 2019; Ashuach et al. 2022), and some rely on scRNA-seq datasets for cell-type annotation (Lin et al. 2022). Methods such as MultiVI and scButterfly (Ashuach et al. 2023; Cao et al. 2024) also adopt VAE frameworks to perform multi-omics integration, but they pre-filter the input by retaining only highly variable peaks to lower the data dimensionality. More recently, scCLIP has emerged as an innovative Transformer-based model for integrating scRNA-seq and scATAC-seq data, and partitions peaks into patch regions for feature extraction. However, it is trained on a relatively small data(Xiong, Chen, and Kellis 2023).

Foundation Models for scRNA-seq Data

Geneformer, scGPT, and scBERT are foundational models pre-trained on millions of single-cell RNA sequencing (scRNA-seq) profiles, demonstrating exceptional performance in tasks such as cell type annotation and gene network inference (Theodoris et al. 2023; Cui et al. 2024a; Yang et al. 2022). Both Geneformer and scBERT draw inspiration from the BERT model architecture, of which Geneformer innovatively incorporates gene ordering processing. In contrast, models like scGPT, utilize a generative pretraining approach for training. All these models treat cells as "gene sentences" to predict the expression of hidden genes. However, the direct transfer of models trained on scRNA-seq data to single-cell chromatin accessibility sequencing (scATAC-seq) data presents challenges due to the vast number of cCREs, extreme sparsity, and ambiguous regulatory relationships inherent in scATAC-seq data.

Methodology

Overview of CLM-Access

As illustrated in Figure 1, this study adopts a BERT-inspired Transformer architecture for scATAC-seq analysis, leveraging self-supervised pre-training to capture long-range chromatin interactions (Devlin et al. 2019; Vaswani et al. 2017). To address the high dimensionality and extreme sparsity of scATAC-seq data (often exceeding millions of features), we redesign BERT's tokenization strategy by partitioning the data into fixed-size patches, each encoded as a token. This approach compresses high-level peak information into low-dimensional embeddings, enabling efficient processing while preserving full data integrity—a critical advantage over traditional methods like TF-IDF transformation or zero-value pruning, which introduce information loss and scalability limitations (e.g., recalculating TF-IDF weights for new training batches). Our architecture provides a robust framework for scATAC-seq analysis, outperforming conventional approaches in both representational fidelity and computational efficiency.

Data Processing and Tokenization

We constructed the Human-scATAC database using 2.8M scATAC-seq data entries, preprocessed into a sparse cell-by-cCRE count matrix$X \in R^{N \times P}$, where N and P denote the number of cells and the number of cCREs, respectively. More details on how the database is constructed are provided in the supplementary file. After unified peak calling and mapping, we obtained 1.15 million cCREs per cell,

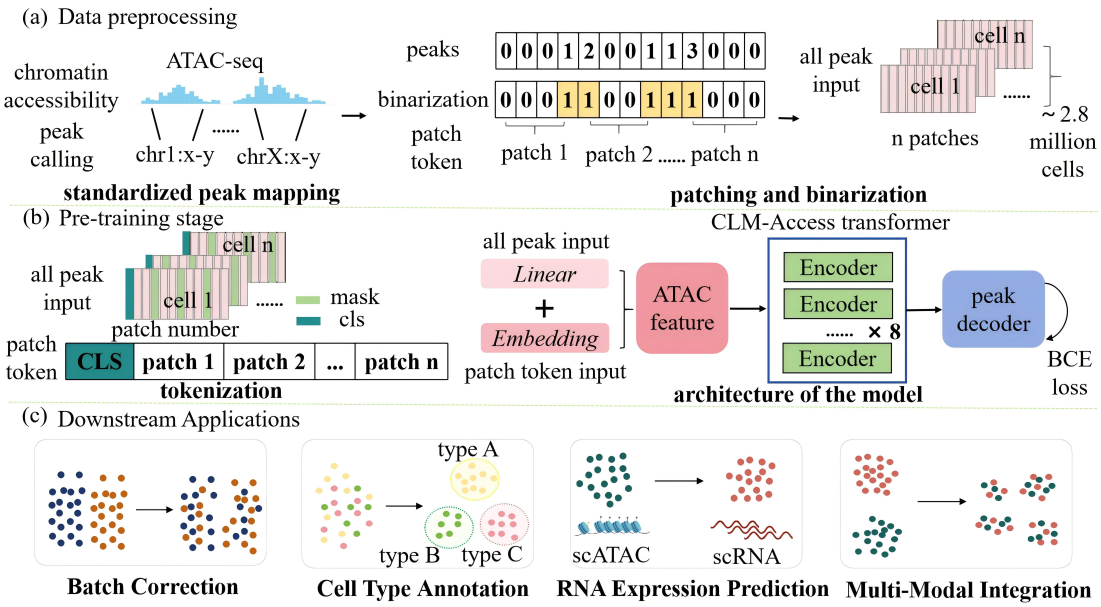

Figure 1: Overview of CLM-Access. **a**, The tokenization process in CLM-Access involves processing the accessible chromatin open regions from scATAC-seq data and dividing them into multiple regions to form cellular sentences. **b**, The model architecture of CLM-Access comprises an embedding module, a CLM-Access transformer, and a peak decoder. **c**,The downstream tasks of CLM-Access include batch effect removal, cell type annotation, RNA expression prediction, and multi-modal integration.

which were divided into 2,000 patches (≈575 cCREs each). When a patch spanned chromosomes, cCREs from the former chromosome were discarded and completed with those from the next, thus yielding the final 1,998 valid patch tokens.

$$[C_i = [CLS, token_1, token_2, ..., token_n] \quad (1)$$

Directly modeling all 1.15M cCREs as tokens would inflate parameters and hinder training. To resolve this, we partitioned each cell into these n (in this case, 1,998) patches, treating each patch as a language model token to aggregate regional cCRE signals. Additionally, four special tokens were added — [CLS], [mask], [pad], and [eoc] — forming the vocabulary for patch tokens. [CLS] represents global features and is introduced at the cell's start to capture its global chromatin accessibility context, [mask] and [pad] are used during pre-training, and [eoc] is reserved. The model thus processes each cell as a fixed-length patch sequence, structured accordingly.

In the preprocessing step prior to patch division, all cCREs are systematically ordered based on their chromosomal positions. This sorting ensures a consistent and biologically meaningful arrangement of the cCREs, facilitating subsequent analysis and modeling steps. After determining the patch regions, we proceed to input all corresponding peak values within each patch into the model. To ensure uniformity across patches, we standardize the length of the peak values within each patch. Specifically, if the number of peak values in a patch is less than the predetermined uniform length, we pad the patch with additional values to meet this length requirement. The structure of each patch, including the handling of peak values and padding, is formally described by the following formula:

$$P_i = \{patch_1, patch_2, ..., patch_n\} \quad (2)$$

$$Patch_i = \begin{cases} [Peak_1, Peak_2, \ldots, Peak_L] \\ \text{if number of peaks} \geq L \\ \\ [Peak_1, Peak_2, \ldots, Peak_k, Pad_{k+1}, \ldots, Pad_L] \\ \text{if number of peaks} < L \end{cases} \quad (3)$$

Here, $patch_i$ represents the i-th patch, L is the predetermined uniform length for all patches, $Peak_j$ denotes the j-th peak value within the patch, and Pad_j represents the padding value used to fill the patch to the desired length when the number of peak values is less than L. This approach ensures that each patch input to the model has a consistent structure, facilitating efficient processing and analysis. Finally, considering that the majority of peak values are either 0 or 1, with rare occurrences of values exceeding 2, a binarization process is applied to all peak values to facilitate model training. This step is taken to simplify the distribution of peak values, making it easier for the model to learn and generalize from the data. By converting the peak values into a binary form, we aim to enhance the training efficiency and effectiveness of the model.

Model Architecture of CLM-Access

This subsection primarily delineates the pre-training architecture of the model, wherein downstream tasks are adapted

through minor modifications and fine-tuning on the pre-trained framework. For detailed information, please refer to the supplementary materials.

Peak Embedding Module The embedding module comprises two primary components: the patch token embedding layer and the peak value embedding layer. The patch token embedding layer takes as input the processed sequence of patch tokens corresponding to each cell. These tokens represent distinct features or segments within the cell data. On the other hand, the peak value embedding layer receives the binarized and padded peak values as input. These peak values are organized into a matrix with a number of columns equal to the length of the cCRE index sequence for each cell, ensuring consistency in dimensionality across different inputs. Subsequently, the embedding module combines the outputs from these two embedding layers through element-wise addition. This integration can be formally represented as follows:

$$E_i = emb\_layer_{patch}(C_i) + emb\_layer_{peak}(P_i) \quad (4)$$

$$E_i = (e_{[cls]}, e_1, e_2, ..., e_n) \quad (5)$$

In the model, $emb\_layer1(\cdot)$ and $emb\_layer2(\cdot)$) are the cCRE and ranking embedding layers, respectively. The patch token embedding layer mirrors NLP's word embedding approach, but introduces a peak value embedding layer to enhance ATAC-seq data utilization. The embedding module uses a 256-dimensional space: the patch token layer encodes cCRE markers and special tokens, while the peak value layer employs a linear transformation to align cellular matrix P_i with the patch token embedding dimension.

Peak Encoder Module The CLM-Access Transformer, the core module of the Transformer-based foundational model, comprises 8 stacked blocks. These blocks employ a bidirectional attention mechanism to model global dependencies in high-dimensional data, capturing epigenetic regulatory networks across cells. Each block has an embedding dimension of 256, uses 8 attention heads, and processes both inputs and outputs, with their mathematical formulation as follows:

$$E_i^{(l)} = block(E_i^{l-1}) = (e_{i,[cls]}^{(l-1)}, e_{i,1}^{(l-1)}, ...e_{i,n}^{(l-1)}) \quad (6)$$

Here,$E_i^{(0)} = E_i$, and $E_i^{(l)}$ is the output of the l-th Transformer block, with $block(\cdot)$ as a single block. The Transformer captures patch interactions via bidirectional attention, described as:

$$Q = E_i^{(l-1)} W_Q^{(l)}, K = E_i^{(l-1)} W_K^{(l)}, V = E_i^{(l-1)} W_V^{(l)} \quad (7)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (8)$$

Wherein, $W_Q^{(l)}$, $W_K^{(l)}$, and $W_V^{(l)}$are trainable weight matrices utilized to compute the Query (Q), Key (K), and Value (V) respectively, with d_k denoting the dimension of the Key vector. To enhance the representational capability of the Transformer, each Transformer block employs a multi-head attention mechanism by computing multiple attention heads in parallel. Each attention head possesses its own set of $W_Q^{(l)}$, $W_K^{(l)}$, and $W_V^{(l)}$, adhering to the same formulation as single-head attention. The outputs from all attention heads are concatenated and subsequently transformed linearly using a learnable matrix $W_O^{(l)}$, yielding the output embeddings. The output of the final Transformer block is denoted as $E_i^{(L)} = (e_{i,[cls]}^{(L)}, e_{i,1}^{(L)}, ..., e_{i,n}^{(L)})$, where $e_{i,[cls]}^{(L)}$represents the cell embedding for cell i. The CLM-Access Transformer comprises approximately 12 million parameters and is implemented using the Flash Attention v2 framework to ensure efficient training and inference.

Peak Decoder Module The peak decoder utilizes a single-layer, fully connected neural network to transform the cell embeddings into the values of all masked peaks for each cell. This process can be formulated as follows:

$$peak_{mask} = MLP(E_i^{(L)}) \quad (9)$$

Finally, the predicted masked peak values and the true values are input into a Binary Cross-Entropy (BCE) loss function to train the model.

Experiment

This section explores the impact of various training strategies on pre-training performance. During the research, it was discovered that directly employing the training approaches of traditional scRNA-seq foundation models and conventional scATAC-seq deep learning models for training scATAC-seq foundation models is infeasible. Neither retaining only accessible peaks nor filtering for highly variable peaks as input can successfully train the model.

To this end, we dedicated substantial efforts to investigating training strategies and data input designs, leading to a series of meaningful insights that could serve as a reference for future development of foundational models of scATAC-seq data. Building upon the successful pre-training of CLM-Access, we evaluated its performance across a range of downstream tasks, including batch effect mitigation, cell type annotation, RNA expression prediction, and cross-modality integration. It is noteworthy that although some foundation models utilize scATAC data, they incorporate information from other sources, and thus are not included in this comparison (Fu et al. 2025).

Experiment Settings

In our experimental design, we employed models with varying parameter scales and utilized datasets of distinct sizes. For the pre-training phase, all models were uniformly trained over 15 epochs with a batch size of 8. In contrast, the fine-tuning configurations for downstream tasks were tailored to each task's specific requirements, leading to task-dependent adjustments. A comprehensive and detailed account of these experimental settings and their rationales can be found in the supplementary materials.

metrics	processing	patch 1k	patch 2k	patch 5k	TAD(14k)
ARI	sum	0.0007	0.0033	0.0001	0.0030
	binarization	0.0113	0.0459	0.1688	**0.2397**
NMI	sum	0.0139	0.0212	0.0087	0.0206
	binarization	0.0513	0.1617	0.4781	**0.5752**

Table 1: The influence of data processing and patch size choice on the pre-trained model

metrics	loss function	patch 1k	patch 2k	patch 5k
ARI	patch-level MSE	0.0113	0.0459	0.1688
	patch-level BCE	0.0122	0.0394	0.0854
	peak-level BCE	0.2120	**0.2511**	0.1304
NMI	patch-level MSE	0.0513	0.1617	0.4781
	patch-level BCE	0.0537	0.1479	0.2560
	peak-level BCE	0.5078	**0.5813**	0.3648

Table 2: The influence of loss function choice on the pre-trained model

metrics	mask forms	patch 1k	patch 2k	patch 5k
ARI	mask token and value	0.0019	0.0016	0.0048
	only mask value	0.2120	**0.2511**	0.1304
NMI	mask token and value	0.0181	0.0181	0.0294
	only mask value	0.5078	**0.5813**	0.3648

Table 3: The influence of mask forms on the pre-trained model

Influence		NMI	ARI
different data	scCLIP(0.33M)	0.5813	0.2510
	CATlas(1.3M)	0.5881	0.2628
	all data(2.8M)	**0.6236**	**0.2785**
different model size	small model	0.6236	0.2785
	medium model	**0.6618**	**0.3024**
	large model	0.6444	0.2958

Table 4: Model performances under different sizes of datasets and model parameters

Systematically Analysis of CLM-Access Pre-training Model Design

This section primarily focuses on the critical pre-training factors of the CLM-Access model, including whether to apply binarization, the selection of loss functions, and the forms of masking. We systematically investigated their impacts on the pre-training performance. In the experiment, we uniformly utilized a dataset of 370,000 single cells processed through the scCLIP pipeline (Xiong, Chen, and Kellis 2023). This dataset was divided into a training set and a test set. We then calculated the Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) using the cell types in the test set for evaluation purposes.

Binarization Processing and Patch Number Effects We evaluated the performance of the data under different patch sizes and with the presence or absence of Topologically Associating Domains (TADs), comparing cases where the data was binarized versus when it was not. The experimental results are presented in Table 1. When only summing the intensity values of all peak signals without applying binarization processing, the model exhibited significant convergence difficulties during training and struggled to effectively capture the internal feature information of each patch region. In contrast, the introduction of binarization processing markedly reduced the model's learning complexity. Further analysis revealed that as the granularity of patch regions decreased, the model retained more complete original information during the feature extraction process and demonstrated superior performance.

Loss Function Comparisons In this study, we designed and compared two data processing strategies for plaque segmentation. In Strategy 1, peak signal intensities within each patch are summed then binarized, and the model is optimized using MSE and BCE loss functions, respectively, at the patch level. In Strategy 2, individual peak signals are binarized directly and input into the model on a patch-by-patch basis, with BCE loss applied at the peak level. We conducted a systematic comparison of the performance of these two approaches.

Initial results showed that reducing patch area improved model performance, but summing peaks within patches led to significant information loss. Increasing patch numbers mitigated this but raised computational costs due to longer input sequences. Hence, we adopted Strategy 2, binarizing all peak signals and processing them patch-by-patch.As Table 2 indicates, Strategy 2 achieved optimal performance with an input sequence length of only 2000. This approach not only enhanced model performance but also reduced training and inference time, effectively balancing computational efficiency and feature preservation.

Impact of Forms of Masking Through systematic experimental design, this study explored the impact of different masking strategies on model performance, proposing two approaches: jointly masking patch tokens and peak signals, or masking only peak signals. We conducted comparative experiments to assess their differential effects on the model's feature learning capabilities.

As shown in Table 3, the joint masking strategy substantially increased model complexity, hindering effective extraction of key biological features from highly masked input sequences. Consequently, subsequent experiments adopted the peak signal-only masking approach. This method maintained model convergence stability, reduced information loss, and enabled the model to focus on learning local features and global distribution patterns of peak signals. This optimization strategy provided crucial support for enhancing model performance in subsequent experiments.

During pre-training, we evaluated the model's zero-shot learning capability by varying parameter and dataset sizes. As Table 4 illustrates, models with moderate parameter counts exhibited optimal performance. Increasing parameters further yielded diminishing returns, suggesting a potential link to the pre-training dataset's size and distribu-

tion. With limited data, excessively large models risk overfitting, thereby limiting generalization. Hence, we selected a medium parameter count for the Transformer component, balancing computational efficiency and model capability. We also conducted pre-training experiments with datasets of varying sizes. Table 4 shows that, with a consistent model architecture, increasing pre-training data volume led to a clear improvement in downstream task performance, approximately linearly correlated with data size. This finding underscores the enhancing effect of large-scale data on deep learning models' feature representation capability, providing a theoretical foundation for dataset construction strategies in subsequent experiments.

CLM-Access Excels in Key Downstream Tasks of scATAC-seq Analysis

CLM-Access is a foundational model specifically tailored for scATAC-seq data, boasting high adaptability. We next fine tuned the model for a few typical downstream tasks in scATAC-seq analysis, including batch effect correction, cell type annotation, prediction of scRNA expression, and multimodal integration with scRNA-seq data.

Batch Effect Correction Batch effects stem from systematic technical biases introduced during experimental processing, which may obscure biological differences among single cells and compromise the accuracy of downstream analyses. The pre-training data encompass samples from multiple batches, allowing the model to implicitly integrate batch relationships among cells. In this study, we take 4 PBMC scATAC-seq datasets (Hu et al. 2024) to validate the performance of batch effect removal, and compare CLM-Access against two conventional methods PCA and harmony. As illustrated in Figure 2, in the zero-shot scenario, the CLM-Access method outperforms traditional approaches such as PCA and Harmony (Korsunsky et al. 2019). After fine-tuning with batch information incorporated, the model's performance is further significantly enhanced, surpassing its zero-shot performance. CLM-Access is capable of constructing cell representations that are biologically meaningful and free from batch-related biases, all without relying on prior knowledge of cell types. This demonstrates its robust independence and broad applicability across various scenarios.

Cell Type Annotation Cell type annotation is a fundamental task in single-cell analysis, playing a crucial role in accurately characterizing the cellular composition and heterogeneity of biological samples. To rigorously assess the performance of CLM-Access on cell type annotation, we selected two large-scale datasets—GSE219281 and GSE181346—each comprising approximately 70,000 cells (Li et al. 2023a; Ameen et al. 2022). Traditional tools often struggle with cell type annotation at this scale, as many lack the capacity to efficiently process such large and complex datasets. These two datasets were therefore chosen to provide a challenging and representative benchmark for evaluating CLM-Access. We compared CLM-Access with a conventional annotation tool scATAnno(Jiang et al. 2024) as well as the state-of-the-art deep learning-based method Cellcano (Ma, Lu, and Wu 2023). As illustrated in Table 5, the fine-tuned CLM-Access model demonstrates outstanding performance across multiple key evaluation metrics, including accuracy, precision, recall, and micro-averaged F1 score, underscoring its remarkable efficacy and effectiveness in cell type annotation tasks.

dataset	model	accuracy	precision	recall	macro_f1
GSE219281	CLM-Access	**0.7635**	**0.6687**	**0.4795**	**0.5437**
	scATAnno	0.7283	0.5366	0.3338	0.3610
	Cellcano	0.7064	0.4685	0.3309	0.3434
GSE181346	CLM-Access	**0.7470**	**0.8335**	**0.6186**	**0.6742**
	scATAnno	0.6834	0.7129	0.5221	0.5473
	Cellcano	0.6286	0.5069	0.3871	0.3955

Table 5: The performance of the CLM-Access model in cell type annotation

Model	Pearson	Rmse
CLM-Access	**0.9175**	**1.4185**
BABEL	0.9101	1.5316
MultiVI	0.8845	3.5755

Table 6: The performance of the CLM-Access model in gene expression prediction

RNA Expression Prediction The prediction of RNA expression from scATAC-seq not only broadens the utility of chromatin accessibility data in transcriptome-level analyses, but also serves as a direct reflection of the model's capacity to extract and interpret meaningful regulatory signals from peak-level information. In this study, we use a paired scRNA-seq and scATAC-seq data (Hu et al. 2024) to examine the performance of CLM-Access on RNA expression task, and compare it with two well-known methods Babel (Wu et al. 2021) and MultiVI(Ashuach et al. 2023).

As illustrated in Table 6, Our fine-tuned CLM-Access model exhibits great performance in gene expression prediction tasks, with the predicted transcriptomic profiles showing strong Pearson correlation of 0.9175 with the corresponding scRNA-seq data. Furthermore, CLM-Access consistently outperforms Babel and MultiVI, demonstrating superior performance over current state-of-the-art methods.

Multi-modal Integration The integration of RNA-seq and ATAC-seq represents a cornerstone in single-cell multi-omics analysis, enabling a more comprehensive dissection of cellular characteristics and regulatory mechanisms compared to single-modality approaches. In this study, we utilized scATAC-seq and scRNA-seq data (Hu et al. 2024) as inputs and conducted fine-tuning training to fuse modality-specific features. For scRNA-seq data, after normalization and log-transformation, we selected 2,000 highly variable genes to focus on key expression features. During the feature fusion phase, we concatenated the linearly mapped scATAC-seq feature vectors with the processed vectors of highly variable genes to construct a joint feature representation.

To enhance the model's capacity for modality discrimina-

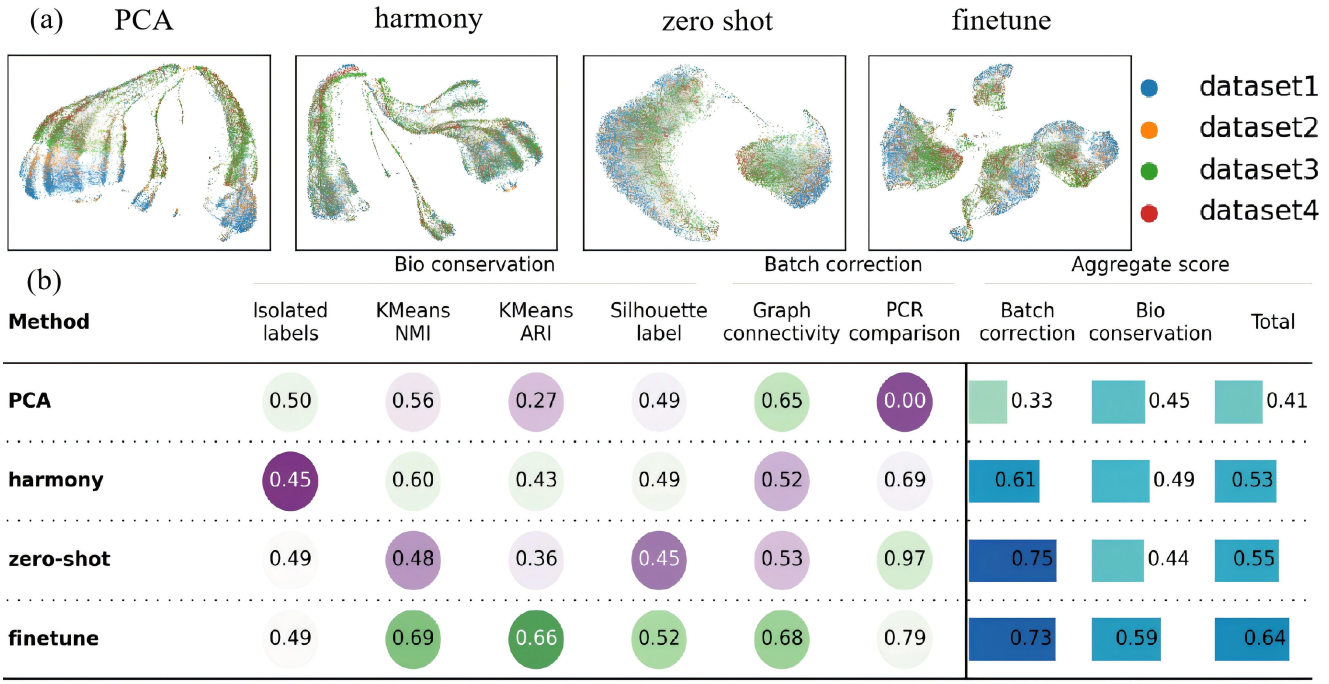

Figure 2: Performance comparison of CLM-Access with other models in the task of removing batch correction

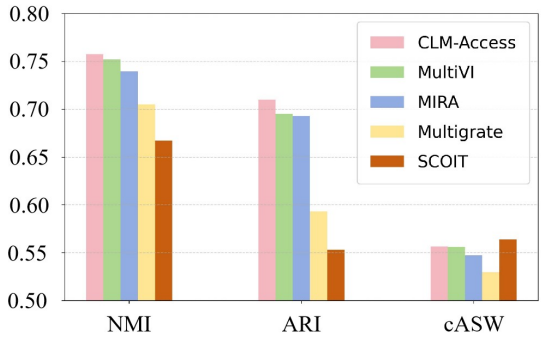

Figure 3: Performance comparison of CLM-Access with other models in the task of multi-modal integration

tion, we introduced a conditional marking mechanism incorporating a special marker token (CLS), along with scATAC-seq and scRNA-seq region markers, facilitating clear modality differentiation in the feature space. In training task design, while maintaining consistency with the ATAC pre-training task, we innovatively introduced a task based on CLS token features to recover the expression of 2,000 highly variable genes from scRNA-seq. By weighted summation of the loss functions from this task and the original task, we constructed a multi-task joint optimization objective to promote the model's collaborative learning of multi-modal data.

As shown in Figure 3, comparative results with four state-of-the-art benchmark models on four multi-modal datasets demonstrated that our method achieved superior overall performance (Ashuach et al. 2023; Lynch et al. 2022; Wang, Wang, and Li 2023; Lotfollahi, Litinetskaya, and Theis 2022).

Discussion

We designed CLM-Access, a foundation model specifically for scATAC-seq data. Unlike previous foundation models for transcriptomes, CLM-Access introduced a novel integration strategy embedded within the Transformer framework: the genomic peaks are partitioned into 2,000 fixed patches, with each patch treated as an independent token. The encoder comprehensively encodes the peak information within each patch into low-dimensional vectors. After fusing the embeddings of each token with the low-dimensional representations of the peaks as model input, this input is fed into the Transformer for deep learning to reconstruct masked peaks. This approach provides a new perspective for integrating scATAC-seq data into the Transformer architecture. It not only preserves all peak information without loss but also significantly enhances model performance and training/inference efficiency, overcoming the limitations of existing scATAC-seq-based methods, such as weak framework generalizability, inadequate utilization of large-scale datasets, and the inability to capture all accessible cis-regulatory elements.

Limitations and Future Work: Our current work studied only data of human cells. In our future work, we will further integrate multi-source information to unify cCREs across multiple species within a shared embedding space. To explore the significance and potential of multi-omics data, we will incorporate more omics modalities, such as single-cell proteomics, in our future research.

Acknowledgments

This work is supported by 100 Talents Programme of The Chinese Academy of Sciences, the National Natural Science Foundation of China (Grant No. 32200524), the National Key Research and Development Program of China (2022YFC3600902), and the Key Research and Development Program of Zhejiang (2025C01129).

References

Ameen, M.; Sundaram, L.; Shen, M.; et al. 2022. Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease. *Cell*, 185(26): 4937–4953.

Ashuach, T.; Gabitto, M. I.; Koodli, R. V.; et al. 2023. MultiVI: deep generative model for the integration of multimodal data. *Nature methods*, 20(8): 1222–1231.

Ashuach, T.; Reidenbach, D. A.; Gayoso, A.; et al. 2022. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell reports methods*, 2(3): 100182.

Buenrostro, J. D.; Wu, B.; Litzenburger, U. M.; et al. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561): 486–490.

Cao, Y.; Zhao, X.; Tang, S.; et al. 2024. scButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nature communications*, 15(1): 2973.

Chen, X.; Chen, S.; et al. 2022. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. *Nature Machine Intelligence*, 4: 116–126.

Chiou, J.; Zeng, C.; Cheng, Z.; et al. 2021. Single-cell chromatin accessibility identifies pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. *Nature genetics*, 53(4): 455–466.

Cui, H.; Wang, C.; Maan, H.; et al. 2024a. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature methods*, 21(8): 1470–1480.

Cui, X.; Chen, X.; Li, Z.; et al. 2024b. Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity. *Nature computational science*, 4(5): 346–359.

Cusanovich, D. A.; Hill, A. J.; Aghamirzaie, D.; et al. 2018. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, 174(5): 1309–1324.

Danese, A.; Richter, M. L.; Chaichoompu, K.; et al. 2021. EpiScanpy: integrated single-cell epigenomic analysis. *Nature communications*, 12(1): 5228.

Devlin, J.; Chang, M.-W.; Lee, K.; et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1: 4171—-4186.

Domcke, S.; Hill, A. J.; Daza, R. M.; et al. 2020. A human cell atlas of fetal chromatin accessibility. *Science*, 370(6518): eaba7612.

Fang, R.; Preissl, S.; Li, Y.; et al. 2021. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature communications*, 12(1): 1337.

Fu, X.; Mo, S.; Buendia, A.; et al. 2025. A foundation model of transcription across human cell types. *Nature*, 637(8047): 965–973.

González-Blas, C. B.; Minnoye, L.; Papasokrati, D.; et al. 2019. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature methods*, 16(5): 397–400.

Granja, J. M.; Corces, M. R.; Pierce, S. E.; et al. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics*, 53(3): 403–411.

Hao, M.; Gong, J.; Zeng, X.; et al. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8): 1481–1491.

Hu, Y.; Wan, S.; Luo, Y.; et al. 2024. Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nature Methods*, 21(11): 2182–2194.

Ji, Z.; Zhou, W.; Hou, W.; and Ji, H. 2020. Single-cell ATAC-seq signal extraction and enhancement with SCATE. *Genome biology*, 21(1): 161.

Jiang, Y.; Hu, Z.; Jiang, J.; et al. 2024. scATAnno: Automated Cell Type Annotation for single-cell ATAC Sequencing Data. *bioRxiv : the preprint server for biology*.

Klemm, S. L.; Shipony, Z.; and Greenleaf, W. J. 2019. Chromatin accessibility and the regulatory epigenome. *Nature reviews. Genetics*, 20(4): 207–220.

Korsunsky, I.; Millard, N.; Fan, J.; et al. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, 16(12): 1289–1296.

Li, J.; Jaiswal, M. K.; Chien, J.-F.; et al. 2023a. Divergent single cell transcriptome and epigenome alterations in ALS and FTD patients with C9orf72 mutation. *Nature communications*, 14(1): 5714.

Li, Y. E.; Preissl, S.; Miller, M.; et al. 2023b. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science*, 382(6667): eadf7044.

Li, Z.; Kuppe, C.; Ziegler, S.; et al. 2021. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nature communications*, 12(1): 6386.

Lin, Y.; Wu, T.-Y.; Wan, S.; et al. 2022. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nature biotechnology*, 40(5): 703–710.

Lotfollahi, M.; Litinetskaya, A.; and Theis, F. 2022. Multigrate: single-cell multi-omic data integration. *BioRxiv*, 484643.

Lynch, A. W.; Theodoris, C. V.; Long, H. W.; et al. 2022. MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nature methods*, 19(9): 1097–1108.

Ma, W.; Lu, J.; and Wu, H. 2023. Cellcano: supervised cell type identification for single cell ATAC-seq data. *Nature communications*, 14(1): 1864.

Minnoye, L.; Marinov, G. K.; Krausgrube, T.; et al. 2021. Chromatin accessibility profiling methods. *Nature reviews. Methods primers*, 1: 10.

Pijuan-Sala, B.; Wilson, N. K.; Xia, J.; et al. 2020. Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nature cell biology*, 22(4): 487–497.

Stuart, T.; Srivastava, A.; Madad, S.; et al. 2021. Single-cell chromatin state analysis with Signac. *Nature methods*, 18(11): 1333–1341.

Tang, S.; Cui, X.; Wang, R.; et al. 2024. scCASE: accurate and interpretable enhancement for single-cell chromatin accessibility sequencing data. *Nature communications*, 15(1): 1629.

Theodoris, C. V.; Xiao, L.; Chopra, A.; et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965): 616–624.

Tsompana, M.; and Buck, M. J. 2014. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*, 7(7): 33.

Vaswani, A.; Shazeer, N.; Parmarand, N.; et al. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, R. H.; Wang, J.; and Li, S. C. 2023. Probabilistic tensor decomposition extracts better latent embeddings from single-cell multiomic data. *Nucleic Acids Research*, 51(15): e81.

Wu, K. E.; Yost, K. E.; Chang, H. Y.; et al. 2021. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15): e2023070118.

Xiong, L.; Chen, T.; and Kellis, M. 2023. scclip: Multi-modal single-cell contrastive learning integration pretraining. In *NeurIPS 2023 AI for Science Workshop*.

Xiong, L.; Tian, K.; Li, Y.; et al. 2022. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nature communications*, 13(1): 6118.

Xiong, L.; Xu, K.; Tian, K.; et al. 2019. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nature communications*, 10(1): 4576.

Yang, F.; Wang, W.; Wang, F.; et al. 2022. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4: 852–866.

Yuan, H.; and Kelley, D. R. 2022. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nature methods*, 19(9): 1088–1096.

Zeng, Y.; Luo, M.; Shangguan, N.; et al. 2024. Deciphering cell types by integrating scATAC-seq data with genome sequences. *Nature computational science*, 4(4): 285–298.

Zhang, K.; Hocker, J. D.; Miller, M.; et al. 2021. A single-cell atlas of chromatin accessibility in the human genome. *Cell*, 184(24): 5985–6001.

Zu, S.; Li, Y. E.; Wang, K.; et al. 2023. Single-cell analysis of chromatin accessibility in the adult mouse brain. *Nature*, 624(7991): 378–389.