Leveraging Single-Cell ATAC-Seq for Genomic Language Models and Multimodal Foundation Models

by

Dong Young Kim

S.B. Computer Science and Engineering, Massachusetts Institute of Technology, 2024

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored by:	Dong Young Kim Department of Electrical Engineering and Computer Science January 8, 2025
Certified by:	Lee Zamparo Research Engineer at InstaDeep, Thesis Supervisor
Certified by:	Siniša Hrvatin Assistant Professor of Biology, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair Master of Engineering Thesis Committee

Leveraging Single-Cell ATAC-Seq for Genomic Language Models and Multimodal Foundation Models

by

Dong Young Kim

Submitted to the Department of Electrical Engineering and Computer Science
on January 8, 2025 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) has emerged as a powerful tool for profiling chromatin accessibility at single-cell resolution. By capturing epigenomic landscapes, scATAC-seq provides critical insights into the regulatory elements that govern gene expression. However, the sparsity of scATAC-seq data, resulting from its low sequencing depth relative to the genome's potential complexity, poses significant challenges for effective and accurate modeling. To advance the utility of scATAC-seq in modern biology, we explore its integration into deep learning frameworks through two innovative applications. First, we demonstrate how incorporating scATAC data enhances the performance of existing genomic language models by providing complementary context about chromatin accessibility. Specifically, we introduce scATAC to improve SegmentNT, a DNA segmentation model that leverages the Nucleotide Transformer (NT) to predict 14 types of genomic and regulatory elements from DNA sequences up to 30kb at single-nucleotide resolution. Second, we introduce a novel multimodal foundation model that extends existing scRNA-seq foundation models by integrating scATAC-seq data. This model captures cross-modal relationships between gene expression and chromatin accessibility, establishing a unified framework that can be fine-tuned for diverse downstream tasks, including cell type classification and cross-modal imputation. Our work highlights the potential of incorporating scATAC-seq data into existing genomics deep learning strategies, providing a framework for integrating regulatory DNA analysis more seamlessly into genomic modeling.

Thesis supervisor: Lee Zamparo
Title: Research Engineer at InstaDeep

Thesis supervisor: Siniša Hrvatin
Title: Assistant Professor of Biology

Acknowledgments

First and foremost, I give all glory and thanks to God. From a middle school boy whose only dream was to play soccer to the writer of this thesis today, I owe everything to God, who has guided me through every step of my journey.

Furthermore, I want to express my sincere gratitude for my co-workers at InstaDeep. Specifically, I want to thank my mentor, Lee Zamparo, for his overwhelming support and guidance throughout my entire internship. I also want to thank Ameya Joshi and Juanjo Garau for always being open to provide feedback and help me work through numerous technical obstacles. I'd also like to thank Thomas Pierrot for his management of the amazing team.

I would also like to thank my advisor Siniša Hrvatin, who was also my first and last Biology professor during my undergrad years at MIT. His biological insights and ample experience in biology research helped me shape my research into a more meaningful and impactful work. I also want recognize the MIT Department of EECS and 6A Program for providing me this amazing opportunity to complete my fifth year Masters program. I have grown tremendously over my years at MIT, both academically and socially, and this place will hold a special place in my heart.

Last but not least, I want to thank family, especially my mother, and friends I have met at MIT. My life has been and will continue to be so special because of each one of them. I cannot envision where I would have been today if it wasn't for their love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The regulation of gene expression is fundamental to cellular function, development, and disease. At its core, this regulation is largely determined by chromatin accessibility—the degree to which specific regions of DNA are available for interaction with transcription factors and other regulatory proteins. While our understanding of these regulatory mechanisms has grown substantially in recent years, the ability to precisely map and analyze chromatin accessibility at the single-cell level has remained a significant challenge in molecular biology.

The emergence of single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) has revolutionized our ability to study chromatin accessibility, providing granular insights into cellular heterogeneity and regulatory dynamics. This technology enables researchers to profile the accessibility landscape of individual cells, revealing cell-specific patterns that were previously masked in bulk sequencing approaches. However, the computational integration of scATAC-seq data presents unique challenges, including significant data sparsity and the lack of standardized formats across datasets.

This thesis addresses these challenges by exploring novel approaches to integrate scATAC-seq data into deep learning frameworks for genomic analysis. We investigate two complementary strategies: enhancing existing genomic language models with chromatin accessibility information, and developing new multi-modal foundation models that can leverage both transcriptomic and chromatin accessibility data. Through these approaches, we aim to advance the field's ability to extract meaningful biological insights from single-cell chromatin accessibility data, ultimately contributing to a more comprehensive understanding of gene regulation at the cellular level.

1.1 Motivation

Despite its promise, integrating scATAC-seq data with existing models remains challenging. Large language models (LLMs), which have demonstrated remarkable success in natural language processing and biological sequence modeling, offer a potential avenue for incorporating scATAC-seq data. However, this approach faces a few technical hurdles. Primary among them is data sparsity—caused by limited sequencing depth relative to the genome's complexity—which complicates the identification of cell-specific regulatory elements and functional predictions. Additionally, scATAC-seq data lacks standardized formats across datasets; scATAC-seq features (peaks) are unique to each dataset with no universal reference framework for peak identification.

These challenges underscore the need to develop methods for effectively integrating scATAC-seq data into existing models. In this work, we explore strategies to leverage scATAC-seq's complementary information to enhance model performance on biological tasks. Specifically, we seek to incorporate scATAC-seq data into models trained on other modalities and evaluate the impact of this integration on downstream performance. This study aims to lay the groundwork for incorporating chromatin accessibility data into foundational genomic models, advancing our understanding of cellular regulation and enabling more comprehensive analyses.

1.2 Problem Statement and Objectives

Although scATAC-seq provides valuable insights into chromatin accessibility, its integration into computational models for biological tasks remains largely unexplored. Challenges such as data sparsity and the lack of standardized formats across datasets make it difficult to directly model scATAC-seq data without preprocessing and alignment strategies. This thesis aims to address these challenges by investigating effective methods to leverage the complementary information provided by scATAC-seq alongside other modalities. We outline three primary objectives for this exploration.

First, we aim to incorporate scATAC-seq data into existing language models trained on complementary modalities. For instance, genomic language models trained on DNA sequences can benefit from scATAC-seq, as chromatin accessibility regions are directly tied to genomic loci. Similarly, transcriptomic language models are promising candidates, given recent advancements in multimodal sequencing technologies that simultaneously profile scRNA-seq and scATAC-seq data at the single-cell level.

Second, we seek to experiment with diverse model architectures and training strategies. While our primary focus is on transformer-based language models, we will explore variations in model scale, architecture, and head design to optimize their integration with scATAC-seq.

Additionally, we plan to investigate different training paradigms, including transfer learning, multi-modal training, and pre-training foundation models. Given the nascent stage of this field, a broad and systematic approach is essential to identify effective methods for utilizing scATAC-seq data.

Third, we want to rigorously quantify the impact of scATAC-seq on model performance. Beyond exploring integration strategies, our ultimate goal is to demonstrate tangible improvements in existing models through the inclusion of scATAC-seq data. This involves isolating and measuring the specific contributions of scATAC-seq to model behavior and performance metrics. By doing so, we hope this study serves as a foundational effort that inspires future research into the optimal use of scATAC-seq in computational biology.

1.3 Contributions

Many existing models incorporate chromatin accessibility data either in isolation or as a secondary consideration, limiting the full potential of scATAC-seq data in genomic modeling. Those strategies have prevented the seamless integration of chromatin accessibility into foundational genomic applications.

To address these challenges and meet our objectives, we adopt two complementary approaches:

1. **Enhancement of Existing Genomic Language Models:** We explore how integrating scATAC-seq data can augment the performance of established models. Using SegmentNT as a case study, we demonstrate that incorporating chromatin accessibility data enhances DNA segmentation tasks, particularly in predicting regulatory elements.

2. **Development of a Novel Multi-modal Foundation Model:** We extend current scRNA-seq foundation models by incorporating scATAC-seq data during pretraining, creating a unified framework capable of addressing diverse downstream applications, such as cell type classification.

In both applications, we introduce innovative methods for integrating scATAC-seq data in a multi-modal context. To assess the effectiveness of these methods, we conduct rigorous evaluations, comparing our approaches against multiple baseline models. Alongside quantitative results, we provide detailed performance analyses to offer insights that can guide future research into leveraging scATAC-seq data for various applications.

1.4 Thesis Structure Overview and Organizing Principles

This thesis is organized into several chapters:

- **Chapter 1: Introduction -** This chapter provides an overview of the thesis, including the background, motivation, and contributions of the work.

- **Chapter 2: Background and Related Work -** This chapter discusses the relevant technical background in both biology and machine learning. It also reviews related works in the field, highlighting previous approaches and their limitations.

- **Chapter 3: Augmenting Existing Genomic Language Models with scATAC -** As the first part of our project, this chapter discusses our effort in leveraging scATAC-seq to improve an existing genomic language model on DNA segmentation task. This chapter contains its own introduction, methods, results, and discussion sections.

- **Chapter 4: Developing a Multimodal Single-Cell Foundation Model with scATAC -** As the second part of our project, this chapter shows how we have developed a multi-modal foundation model that can flexibly deal with scATAC and/or scRNA data. This chapter also contains its own introduction, methods, results, and discussion sections.

- **Chapter 5: Discussion and Conclusion -** We end our thesis by summarizing the research and its key findings, as well as its broader implications for future work in scATAC-seq modeling. This chapter also suggests open questions and directions for future research.

Chapter 2

Background and Related Work

In this section, we provide background on the biology and deep learning techniques that builds the foundation of our work. This background is followed by two sections, each of which examines earlier approaches that informed the two studies conducted in our research, respectively. We outline previous works' limitations and how our work seeks to extend and enhance them.

2.1 Backgrounds on Biology and Deep Learning

The foundations of this work span both experimental biology and computational methods. We begin by examining key sequencing technologies in biology, with particular emphasis on the evolution of chromatin accessibility profiling from traditional ATAC-seq to single-cell resolution (scATAC-seq). Following this context, we introduce fundamental deep learning concepts essential to our research, focusing on foundation models and their extension to multi-modal training paradigms.

2.1.1 Sequencing in Biology

Advances in sequencing technologies have revolutionized our understanding of biology, providing insights into the molecular mechanisms underlying cellular and organismal processes. Genome sequencing, one of the earliest and most foundational sequencing methods, involves determining the complete DNA sequence of an organism's genome. This approach has enabled the identification of genetic variants, the annotation of functional genomic elements, and the study of evolutionary relationships across species. High-throughput methods such as next-generation sequencing (NGS) have further accelerated genome sequencing, reducing costs and increasing accessibility for a broad range of applications [1].

In the realm of cellular biology, single-cell RNA sequencing (scRNA-seq) has emerged as a pivotal tool for studying gene expression at the single-cell level. Unlike bulk RNA sequencing, which provides an averaged transcriptomic profile of a cell population, scRNA-seq captures the heterogeneity of individual cells, revealing distinct cell types, states, and developmental trajectories. This capability has proven invaluable in fields such as immunology, neurobiology, and cancer research, where cellular diversity plays a critical role. By profiling the transcriptome of individual cells, scRNA-seq enables researchers to map cell-specific functions and interactions with high resolution.

Single-cell ATAC sequencing (scATAC-seq) complements scRNA-seq by profiling chromatin accessibility, which is a key determinant of gene regulation. By identifying open chromatin regions where transcription factors can bind, scATAC-seq provides insights into the regulatory landscape of cells. This method is particularly useful for studying cell-specific regulatory elements and understanding the epigenetic mechanisms that underlie gene expression. Combined with scRNA-seq, scATAC-seq facilitates multi-modal analyses that link gene expression to chromatin accessibility, offering a more comprehensive view of cellular function.

Together, these sequencing technologies represent significant milestones in molecular and cellular biology. While genome sequencing provides the foundational blueprint of an organism, single-cell sequencing methods uncover the dynamic regulatory and functional characteristics of individual cells. However, challenges remain in integrating these diverse data types, particularly in the development of models that leverage their complementary strengths for downstream biological applications. These challenges underscore the need for innovative approaches to model and analyze sequencing data, which is a central focus of our work.

2.1.2 History of ATAC and scATAC

Chromatin accessibility profiling has played a pivotal role in understanding gene regulation and epigenetic landscapes. The advent of ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) in 2013 by Buenrostro et al. revolutionized this field by enabling genome-wide identification of accessible chromatin regions with minimal input material [2]. This technique leveraged the Tn5 transposase to preferentially insert sequencing adapters into open chromatin, providing a more straightforward and efficient alternative to earlier methods such as DNase-seq and FAIRE-seq.

While bulk ATAC-seq offered a high-resolution view of chromatin accessibility across populations of cells, it suffered from a key limitation: it provided averaged measurements, potentially obscuring cell-specific regulatory dynamics in heterogeneous samples. Recognizing this limitation, scientists sought to adapt ATAC-seq for single-cell resolution. In 2015, Buenrostro and colleagues introduced scATAC-seq, combining ATAC-seq with microfluidics to profile chromatin accessibility at the level of individual cells [3]. This advancement enabled researchers to dissect cellular heterogeneity and study regulatory mechanisms in unprecedented detail.

Since its inception, scATAC-seq has seen significant improvements in throughput, sensitivity, and data quality. Innovations like combinatorial indexing [4], droplet-based methods [5], and multi-modal sequencing technologies such as multiome sequencing have expanded the scope of scATAC-seq. These methods not only increased the number of cells profiled in a single experiment but also allowed for simultaneous measurement of transcriptomics and chromatin accessibility in individual cells.

Despite its success, analyzing scATAC-seq carries challenges stemming from the data's sparsity and the lack of standardized annotations. These challenges yet also the opportunities form the foundation for exploring how to integrate scATAC-seq into advanced computational models and with other modalities, a central focus of this thesis.

2.1.3 Foundation Models and Its Presence in Biology

Language foundation models, a stastical method for modeling language, have revolutionized natural language processing (NLP). Train from broad, unlabeled datasets and adaptable for a wide range of predictive tasks, language foundation models offer novel capabilities for tasks like translation, summarization, and question-answering. These models, such as BERT [6] and GPT [7], rely on architectures like transformers [8], which excel at processing sequential data and capturing contextual relationships. Their success stems from pre-training on vast corpora using self-supervised objectives, enabling the extraction of meaningful representations that generalize across diverse downstream tasks.

The adaptability of language foundation models has inspired their application in fields beyond traditional NLP, including biology. Genomic and transcriptomic data, like language, consist of sequential information encoded in DNA, RNA, or protein sequences. This similarity has motivated researchers to repurpose language models to interpret biological information. For example, foundational models like DNABERT [9] and ESM [10] have been tailored to genomic and proteomic sequences, respectively, learning representations that can predict regulatory elements, identify protein structures, or annotate functional domains.

In single-cell biology, the concept of foundation models has gained traction with tools like scBERT [11] and scGPT [12], which leverage the sequential structure of transcriptomic and multi-modal data. These models pre-train on gene expression profiles or chromatin accessibility data, offering powerful representations that can be fine-tuned for tasks such as cell type classification, regulatory element prediction, and multi-modal integration.

Despite these advancements, foundational models in biology are still in their infancy compared to their NLP counterparts. Challenges like data sparsity, noise, and the lack of standardized benchmarks in biological datasets persist. Additionally, the integration of multi-modal data remains an open area of exploration. The rapid evolution of language models, however, continues to push the boundaries of what is possible in computational biology, heralding a new era of discovery driven by AI.

2.1.4 Multi-modal Training and Fine-tuning

Multi-modal learning refers to the training of machine learning models using multiple types of data or "modalities," each providing distinct but complementary information about the underlying phenomena. In genomics, these modalities might include gene expression data, chromatin accessibility profiles, or DNA sequence information. The fundamental challenge in multi-modal learning lies in effectively integrating these diverse data types while accounting for their different statistical properties, scales, and sparsity patterns.

The integration of multiple modalities can occur at two distinct phases of foundation model development: pre-training and fine-tuning. Pre-training with multiple modalities involves exposing the model to different data types during its initial learning phase, allowing it to develop fundamental representations that capture cross-modal relationships. This approach typically leads to more robust and generalizable features, as the model learns to understand the intrinsic connections between modalities from the ground up. In contrast, fine-tuning involves taking a pre-trained model (often trained on a single modality) and adapting it to incorporate additional modalities for specific tasks. While this approach can be more computationally efficient and easier to implement, it may limit the model's ability to fully capture deep relationships between different data types, as the fundamental representations are already established during the single-modality pre-training phase.

2.2 Related Work on Genomic Language Models and Modeling scATAC

This section reviews the literature relevant to our first project: leveraging scATAC-seq to enhance existing genomic language models. We begin by examining foundational works that inform the methods and approach of our study. Additionally, we discuss previous efforts addressing similar challenges, highlighting their strengths and identifying their limitations.

2.2.1 Modeling scATAC-seq Data with Fragment Counts

scATAC-seq data can be accessed in either read counts, which measure the total number of individual sequencing reads mapped to a region, or fragment counts, which measure the number of unique DNA fragments (typically pairs of reads) that originated from a single transposition event. Previous research [13] has demonstrated that scATAC-seq data is best modeled using fragment counts rather than read counts or binary accessibility indicators. This is because scATAC-seq data provides nuanced information beyond simple binary accessibility, and fragment counts, which follow a Poisson distribution, can be modeled more properly than read counts. Following this insight, we fine-tune our model to predict fragment counts from scATAC-seq data, building on the datasets used in earlier studies that overlap with those in our own work.

2.2.2 Genomic Language Models and the Nucleotide Transformer (NT)

Foundation models are large-scale machine learning models that are trained on diverse, large datasets in an unsupervised or self-supervised manner. Once trained, foundation models can be fine-tuned or adapted to perform specific tasks with task-specific data. Foundation models have revolutionized fields such as natural language processing (NLP) [6] [14] [7] [15], and more recently, these models have also shown promise in genomics applications [9] [16]. Our work builds upon the Nucleotide Transformer (NT) [17], a foundation model ranging from 50M to 2.5B parameters, trained on diverse genome datasets. Notably, the NT model's transformer [8] architecture has been shown to identify key genomic elements without supervision. We fine-tune pre-trained SegmentNT models that use NT encoders of various sizes for scATAC-seq predictive tasks before further refining them for DNA segmentation tasks.

2.2.3 DNA Segmentation Models and the SegmentNT

DNA segmentation involves partitioning the genome into distinct segments and assigning each segment one or more labels from a finite set. Earlier approaches employed non-machine learning techniques for this task, such as Segway [18] that utilizes a dynamic Bayesian network, and ChromHMM [19] that applies a multivariate hidden Markov model (HMM). Recent advancements in deep learning and foundation models have led to significant progress in DNA segmentation. One notable model, SegmentNT [20], is a DNA segmentation model that integrates a pre-trained NT model with a 1D U-Net architecture [21]. SegmentNT processes DNA sequences up to 30kb in length and predicts 14 distinct genomic elements at single-nucleotide resolution. These 14 types of genomic elements include gene elements (e.g., protein-coding genes, lncRNAs, UTRs, exons, introns, splice sites) and regulatory elements (e.g., polyA signals, promoters, enhancers, CTCF-bound sites). Our work builds upon SegmentNT and aims to improve its performance on the DNA segmentation task. Specifically, we enhance the NT component by fine-tuning it with a scATAC-seq predictive task to incorporate chromatin accessibility information, which provides important contexts that the model can learn to improve its performance on DNA segmentation tasks.

2.2.4 Using ATAC Data for Genomic Elements Classification

Previous studies have explored the integration of genomic and chromatin data, such as BPNet [22], which utilizes DNA sequences to predict chromatin immunoprecipitation-nexus binding profiles of pluripotency transcription factors. A key initiative in leveraging ATAC-seq data for genomic element classification is CoRE-ATAC [23]. CoRE-ATAC uses DNA sequences along with ATAC-seq cut sites and read pileups to predict *Cis*-Regulatory elements (*cis*-REs), such as promoters, enhancers, and insulators, at base-pair resolution in open chromatin regions. CoRE-ATAC demonstrated the potential of integrating genome and chromatin accessibility information to predict *cis*-RE functions across diverse cell types. While CoRE-ATAC focuses on predicting *cis*-REs specifically within open chromatin regions, our work extends the scope by classifying entire genome sequences into 14 genomic element types, incorporating chromatin accessibility data for broader predictive capabilities.

2.3 Related Work on Transcriptomics and Multi-omics Models

Since the introduction of scATAC-seq, numerous efforts have focused on modeling chromatin accessibility data. With the advent of multiome sequencing, which enables simultaneous profiling of scRNA and scATAC data, many studies have worked toward leveraging the multi-modal features of cell samples for various downstream applications. This section reviews works that have attempted to integrate scRNA and scATAC data, emphasizing transcriptomics and multi-omics models that have utilized multiome data. Additionally, we highlight how our approach distinguishes itself from prior efforts in this field.

2.3.1 Transcriptomics Foundation Model

Transformer-based foundation models have found notable applications in single-cell transcriptomics [24], where they are used to extract biologically meaningful representations of cells based on their gene expression profiles. These representations can then be fine-tuned for various downstream tasks. A key example is scBERT [11], which leverages BERT's masked language modeling approach [6] to pre-train on scRNA-seq data. In scBERT, each gene is represented as the sum of two embeddings: one that represents log-normalized, binned log-scale gene expression levels and the other with gene identity through gene2vec [25] embedding. Much of our model's architectural design is inspired by scBERT, which also serves as a baseline for performance comparisons in our study.

2.3.2 Integrating scRNA and scATAC

With the advent of multiome sequencing, numerous efforts have been made to integrate multiome data effectively. One common approach involves aligning the embedding spaces of scRNA and scATAC modalities. For instance, scCLIP [26] employs a CLIP-like method, utilizing contrastive learning with cosine similarity to align multi-modal embeddings. Similarly, Multivi [27], a variational autoencoder (VAE), encodes each modality, combines them into a shared representation, and decodes each modality from the combined embedding; its variant, Poisson-Multivi, further models ATAC fragment counts using a Poisson distribution, following the key principles discovered in previous studies [13]. While these methods excel at aligning modalities, they do not necessarily enhance performance on downstream tasks. Indeed, neither scCLIP nor Multivi functions as a foundation model, as their primary objective is alignment rather than task improvement. In contrast, our approach prioritizes improving task performance, even if it comes at the cost of imperfect alignment between modalities.

2.3.3 Multi-omics Foundation Models

In addition to aligning the representations of scRNA and scATAC modalities, other studies have approached multiome data by developing multi-omics foundation models. Notable among these are scGPT [12], scMoformer [28], and scooby [29]. Each of these works provides unique methodologies for modeling multiome data, but they also present limitations that our research seeks to address. In this section, we briefly overview these models, highlighting their strengths and identifying the gaps our work aims to improve upon.

scGPT is primarily a single-cell transcriptomics foundation model with the capability to incorporate scATAC data. It adopts an architecture and pre-training process similar to scBERT, implementing generative masked pre-training with a causal masking strategy inspired by OpenAI's GPT models [7]. scGPT introduces trainable gene identity embeddings and a "conditional embedding" to capture meta-information specific to each gene. Although pre-training is conducted exclusively with scRNA data, scGPT accommodates scATAC data during fine-tuning by integrating modality and peak embeddings. As a result, while scGPT supports multi-modal data, it is not a true multi-modal foundation model because it incorporates multiple modalities only in the fine-tuning stage, not during pre-training.

scMoformer addresses multi-modal training by integrating transformer and graph neural network (GNN) architectures. It incorporates three primary modalities: DNA, RNA, and protein measurements. The model first employs modality-specific transformer encoders to embed data points unique to each modality. Next, it constructs a GNN to connect these embeddings using edges informed by prior biological knowledge, such as the STRING database [30]. The training objective focuses on cross-modality prediction tasks, such as using gene expression to predict protein levels. While scMoformer's innovative combination of transformers and GNNs enables seamless integration of scRNA and scATAC data, its application is limited to supervised cross-modality prediction rather than broader downstream tasks.

scooby is a recent model designed to predict scRNA-seq coverage and scATAC-seq insertion profiles along the genome at single-cell resolution. Its architecture builds on the Borzoi [31] foundation model, incorporating data-specific parameters through LoRA [32]. The adapted sequence embeddings are combined with a single-cell decoder, which integrates scRNA and scATAC profiles using Poisson-Multivi [27] to produce a final sequence representation. This representation is then used to predict RNA coverage and ATAC insertion profiles. While scooby effectively integrates scRNA and scATAC data to enhance performance on downstream tasks, its reliance on dataset-specific adjustments and the additional effort needed to adapt RNA coverage predictions into gene expression predictions limit its flexibility.

Chapter 3

Augmenting Existing Genomic Language Models with scATAC

3.1 Introduction

Foundation models, which are large models trained on broad datasets and adaptable for a wide range of predictive tasks, have demonstrated success in artificial intelligence (AI). One such field that has seen recent successes of foundation models is genomics. For instance, SegmentNT [20] is a DNA segmentation model that processes DNA sequences (up to 30kb length) and predicts 14 distinct classes of genomic elements at a single nucleotide level; among these 14 are 8 gene elements (protein-coding genes, lncRNAs, 5'UTR, 3'UTR, exon, intron, splice acceptor and donor sites) and 6 regulatory elements (polyA signal, tissue-invariant and tissue-specific promoters and enhancers, and CTCF-bound sites). The architecture of SegmentNT combines the pre-trained DNA foundation model, Nucleotide Transformer (NT) [17], as an embedding layer with a 1D U-Net [21] architecture for classification.

Despite its competitive nucleotide accuracy, SegmentNT faces a critical limitation—it lacks direct exposure to additional biological information from the sequenced genomic DNA. Its backbone foundation model, NT, is trained following the BERT methodology [6], predicting masked nucleotides in unlabeled DNA data. Furthermore, SegmentNT itself is trained to segment 14 genomic elements in input DNA sequences of 3kb, where the label annotations were derived from GENCODE [33] and ENCODE [34]. At no point during these training processes is SegmentNT exposed to information beyond the DNA sequences and their corresponding labels. This omission could exclude crucial biological insights necessary for more accurate DNA segmentation.

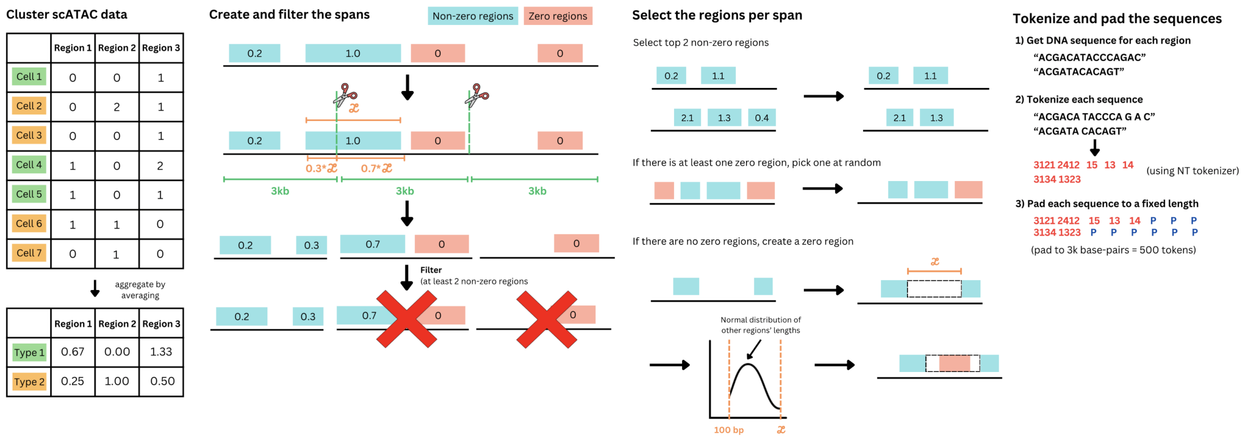
Chromatin accessibility, which reflects the degree to which nuclear macromolecules can physically interact with chromatinized DNA, represents one such missing piece. Integrating chromatin accessibility information as a form of scATAC into SegmentNT's training may enhance the model's performance in two main ways. First, chromatin accessibility data provides more intrinsic, direct boundaries for regulatory elements such as promoters, enhancers, and insulators from the DNA sequences themselves. It could reduce SegmentNT's dependence on genomic element boundary labels that are from prior genomic segmentation models or hand-annotated and contain unmodeled noise. Second, introducing this new form of biological data could unlock further potential in SegmentNT by moving beyond the limitations of genome annotation data alone. Training the model with additional reference genomes may result in diminishing returns due to the high similarity between genomes already in use. By incorporating multi-modal fine-tuning with chromatin accessibility data, we can better inform SegmentNT to identify the boundaries of regulatory elements more accurately. These hypotheses guide the objectives of this study as follows:

- **Develop a model that directly links genomic sequences with their chromatin accessibility profiles.**
- **Transfer this model's knowledge to SegmentNT, enabling it to perform DNA segmentation with additional context.**
- **Validate the benefits of chromatin accessibility fine-tuning in SegmentNT with its original DNA segmentation task.**

In this work, we show that incorporating chromatin accessibility data can improve SegmentNT's performance on DNA segmentation tasks, specifically for its classifications on regulatory elements. We first train the model's embedding layer using a scATAC predictive task. Then, we evaluate the performance of SegmentNT with this enhanced embedding layer against the original model. Overall, our study demonstrates that exposing genomics models to additional biological context can improve their performance on downstream predictive tasks.

3.2 Datasets

Raw ATAC datasets, while rich in information, are often unstructured and require extensive preprocessing to extract meaningful features and prepare them for downstream analyses. In this section, we introduce the datasets we used in our study as well as their sources. Then we cover the 4-step procedure we followed to process the raw data into a format that can be used to train our model.


Cluster scATAC data
aggregate by averaging
Create and filter the spans
Non-zero regions
Zero regions
0.2
1.0
0
0
$\mathcal{X}$
0.3*$\mathcal{X}$
0.7*$\mathcal{X}$
3kb
3kb
3kb
0.2
0.3
0.7
0
0
Filter
(at least 2 non-zero regions
Select the regions per span
Select top 2 non-zero regions
0.2
1.1
2.1
1.3
0.4
If there is at least one zero region, pick one at random
If there are no zero regions, create a zero region
Normal distribution of other regions' lengths
100 bp
$\mathcal{X}$
Tokenize and pad the sequences
1) Get DNA sequence for each region
"ACGACATACCCAGAC"
"ACGATACACAGT"
2) Tokenize each sequence
"ACGACA TACCCA G A C"
"ACGATA CACAGT"
3121 2412 15 13 14
3134 1323
(using NT tokenizer)
3) Pad each sequence to a fixed length
3121 2412 15 13 14 P P P
3134 1323 P P P P P P
(pad to 3k base-pairs = 500 tokens)

	Region 1	Region 2	Region 3
Cell 1	0	0	1
Cell 2	0	2	1
Cell 3	0	0	1
Cell 4	1	0	2
Cell 5	1	0	1
Cell 6	1	1	0
Cell 7	0	1	0

	Region 1	Region 2	Region 3
Type 1	0.67	0.00	1.33
Type 2	0.25	1.00	0.50

Figure 3.1: **4-step procedure for data processing.** Our data processing consists of four steps: **1)** Cluster scATAC data by cell type or cluster annotations. **2)** Filter for fixed-length DNA spans containing at least two regions with non-zero ATAC fragment counts. **3)** Select two non-zero regions and one zero region per span. **4)** Tokenize and pad the DNA sequences.

3.2.1 scATAC Dataset Processing and Summary

We used human scATAC-seq datasets with the reference genome of hg38 and hg19, which was aligned to hg38 with UCSC lifeover [35]. Raw published data for these datasets are available from 10XGenomics and the GEO under accession codes GSE194122, GSE129785, and GSE149683. We then processed these raw data to filter the peaks to only that were detected in at least 1% of the cells in their respective sample. Furthermore, based on the previous literature that led to improved performance [13], we converted scATAC-seq data representation from raw read counts to estimated fragment counts before presenting them to our model. To estimate the fragment counts from the read counts, we rounded up all the read counts to their next highest even number and halved the resulting read counts. After these filtering steps, we prepare scATAC-seq datasets in a four-step processing procedure. This four-step procedure is visually illustrated in Figure 3.1.

3.2.2 Step 1: Cluster scATAC data by cell type or cluster.

Unlike bulk ATAC-seq, which averages chromatin signals across all cells in a sample, scATAC-seq offers single-cell resolution, allowing for more detailed insights into chromatin accessibility. However, the data's sparsity and susceptibility to noise complicate analysis and interpretation. To mitigate these challenges while retaining the resolution benefits of scATAC-seq, we aggregated the data by cell type or cluster annotations provided by each dataset and created a new, 'sudo-bulk ATAC-seq data. Specifically, we averaged the fragment counts among all the cells that belong to the same cell type (or cluster). Then, we convert all the fragment counts of less than 0.1 into 0 to minimize the impact of noises produced in the single-cell sequencing procedure.

3.2.3 Step 2: Filter for fixed-length DNA spans containing at least two regions with non-zero ATAC fragment counts.

We divide the reference genome into fixed length of 3k base-pairs, matching the context length used in the SegmentNT model; we call each of this fixed length data points as “spans." When a cutting boundary falls within a region, we split the region and allocate fragment counts proportionally to the new lengths. Spans with fewer than two regions with non-zero ATAC fragment counts are filtered out to ensure clear ranking of regions. This process allows us to focus only on spans that have sufficient non-zero data points to work with. To enhance data generation, we apply a referential shift, shifting cutting sites across the genome. Using a shift factor of 10, we create 10 times the number of spans we would have created without a reference shift, significantly increasing the number of usable spans.

3.2.4 Step 3: Select two non-zero regions and one zero region per span.

From each filtered span, we select two regions with non-zero fragment counts and one region with a zero count to ensure that the regions being ranked share the similar genomic context, which allows the model to better isolate the effects of genomic location on its performance. If a span has exactly two non-zero regions, they are selected by default. For spans with more than two non-zero regions, two are chosen at random. For the zero region, if one or more exist, we select one at random. If not, we look at the span's longest empty range that doesn't overlap with any regions from the ATAC data. From that empty range, we randomly pick a region with a length that is comparable to those from other regions in the ATAC dataset. The length of this region is modeled by the modified truncated normal distribution [36] of the formula:

$$\psi(\mu, \sigma, a, b; x) = \begin{cases} a & \text{if } x \leq a \\ \mathcal{N}(\mu, \sigma; x) & \text{if } a < x < b \\ b & \text{if } x \geq b \end{cases}$$

Here, $\mathcal{N}$ is the normal distribution. μ and σ is the average and the standard deviation of all the regions in the dataset. We set the minimum length, a, to be 100 base-pairs and the maximum length, b, to be the length of the empty region identified.

3.2.5 Step 4: Tokenize and pad the DNA sequences.

Finally, we tokenize the corresponding genomic sequences of the regions using the same 6-mer tokenizer employed by SegmentNT. To standardize input lengths, we pad each sequence to a fixed length. In the first iteration of our study, padding tokens are added to the end of each tokenized sequence. In the second iteration, padding is distributed randomly between the front and back, ensuring the total token length matches the desired fixed length.

After applying these four steps, we split the datasets into train, validation, and test sets based on the chromosomes. Specifically, we use the same split as in the Nucleotide Transformer paper [17]: chromosomes 1 through 19, X, and Y for the train set, chromosome 22 for the validation set, and chromosomes 20 and 21 for the test set.

3.3 Methods

In this section, we describe the methodology that aims to achieve the specific objectives outlined in Section 3.1. In particular, we cover the framework that is used to multi-modal fine-tune SegmentNT with scATAC-seq data, transfer of knowledge from the chromatin accessibility fine-tuning to the primary DNA segmentation task, and different design choices and baseline models tested.

3.3.1 scATAC Prediction Task Formulation

We fine-tune the original SegmentNT model with scATAC-seq data by extracting out the NT encoder from its architecture and training this encoder on the ATAC prediction task. We frame this auxiliary task as a list-wise ranking problem [37], the Region Ranking task. In this setup, the model receives three DNA regions as input. These regions are subsets of the fixed-length “spans," which are continuous segments of DNA whose lengths are determined by the base-pair length required for the SegmentNT model in use. Two of the regions are labeled with non-zero ATAC fragment counts, while one is labeled with a zero count. The model assigns a score representing the chromatin accessibility of each region, which is evaluated based on the relative ranking of their actual fragment counts. Importantly, the model never directly observes the fragment counts nor predicts them outright. Instead, it learns to rank the regions concordant to their relative chromatin accessibility. For this ATAC regions ranking task, we designed a new architecture that resembles that of SegmentNT but with the DNA segmentation head (1-D U-Net for SegmentNT) replaced with chromatin accessibility scorer (CAS) head (a of Figure 3.2).

This ranking-based approach simplifies the task compared to directly predicting fragment counts with methods like Poisson regression. While precise count predictions are ideal, they are impractical due to missing biological factors influencing chromatin accessibility, such as nucleosome positioning, histone modifications, and DNA-binding proteins. Additionally, the sparsity and variability of regions across datasets limit the model's ability to generalize fragment count predictions across the genome. Although sparsity could be addressed by focusing only on non-zero regions, as done in scBERT [11], this approach sacrifices valuable information provided by zero regions. The ranking-based approach, on the other hand, provides a framework that is generalizable to different datasets while also balancing out the model's learning space with both zero and non-zero regions.

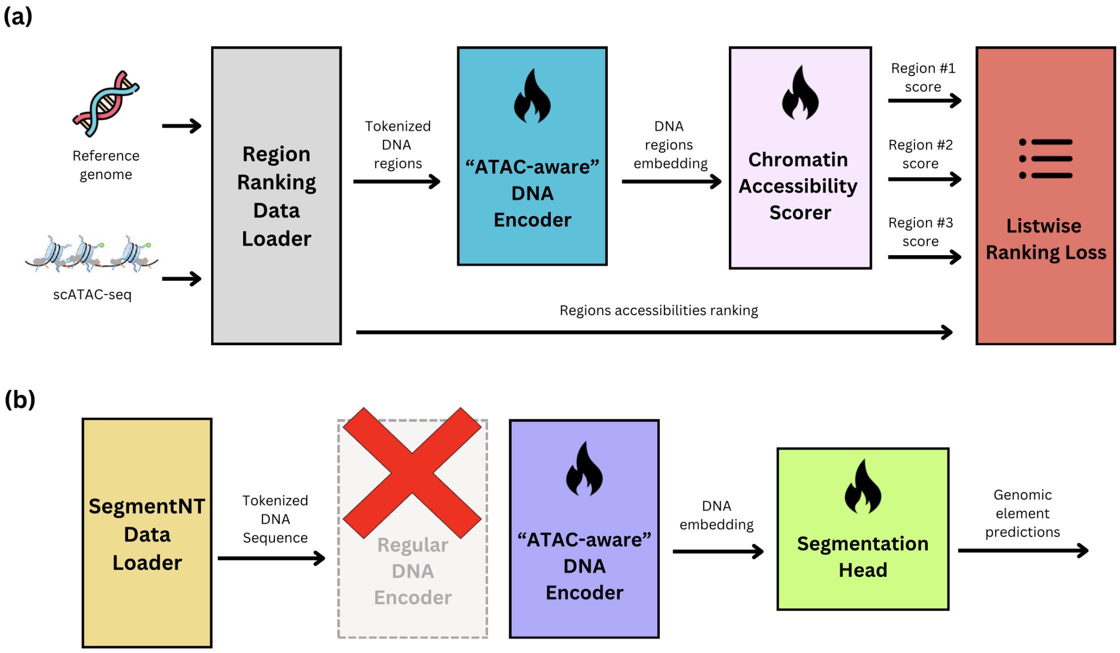


Figure 3.2: **Process of multi-modal fine-tuning of SegmentNT with scATAC-seq data. a)** Our proposed new architecture takes the DNA encoder of SegmentNT to train it with the Region Ranking task. Both the encoder and CAS head are unfrozen and updated through the listwise ranking loss optimization process. **b)** "ATAC-aware" encoder that was used for the Region Ranking task is brought back to the SegmentNT model to replace the original encoder. SegmentNT is further fine-tuned to leverage the "ATAC-aware" encoder for its DNA segmentation task.

3.3.2 Chromatin Accessibility Scorer (CAS) Heads

To train a DNA encoder capable of extracting chromatin accessibility information from genomic sequences, we attach a task-specific head, referred to as the "chromatin accessibility scorer (CAS)," to the DNA encoder. This component takes a fixed-length DNA segment as input and outputs a score representing the segment's chromatin accessibility level. As outlined in 3.3.1, we adopt a ranking-based approach to train both the encoder and the CAS. To do so, we utilize the Rax [38] library, which implements Learning-to-Rank (LTR) in Jax. Specifically, for the three (two non-zero and one zero) regions in a given span, CAS scores each region. These scores are then used to compute the softmax loss relative to the ground-truth ranks of the chromatin accessibility levels for the three regions, which the model uses to train the DNA encoder and the CAS.

To ensure that most of the knowledge is retained in the DNA encoder, it is essential that we keep CAS small and simple. Hence, we test out four versions of CAS—linear, transformer, U-net, and convolutional. Each of these scorers resembles a simple architecture of its respective key design point. Linear CAS head converts the DNA embedding to the chromatin accessibility scores through a fully-connected layer with a single outpue activation. Transformer CAS does so through two simple self-attention-based layers with a single head and positional embedding, followed by a linear layer at the end that outputs a single score for each region. U-net scorer employs the same 1D U-Net [21] architecture as in SegmentNT but with a flattening and linear layer at the end. Convolutional scorer is structured with two 2D convolutional layers, each followed by 2D average pooling. The output is then passed through a flattening layer and subsequently through two fully connected layers.

3.3.3 "ATAC-aware" Encoder Transferred to DNA Segmentation

After training the DNA encoder with the CAS head on the Region Ranking task, the encoder's parameters should now reflect its knowledge about how to infer chromatin accessibility levels from DNA sequences. The NT paper [17] demonstrated that the attention layers in the foundation model could detect key genomic elements and learn nuanced biological features—such as splice junctions and transcription factor motifs—in an unsupervised manner. Similarly, we deduce that our encoder has become "ATAC-aware," meaning it has learned to identify genomic features indicative of ATAC fragment count levels across different sequences.

To take advantage of this capability, we replace the original encoder in SegmentNT with our “ATAC-aware" encoder (see panel b of Figure 3.2). Once replaced, the modified SegmentNT is retrained on its original DNA segmentation task. We use the same formulation for the DNA segmentation task as used in [20], in which we predict 14 types of genomics elements at single nucleotide resolution. This transfer learning process allows the model to utilize its enhanced understanding of chromatin accessibility to improve segmentation accuracy. We call this modified SegmentNT model, “ATAC-aware" SegmentNT.

3.3.4 “ATAC-aware” v.2 with Random Padding and No Duplicating Sequences

We identified two modifications to the data preprocessing pipeline described in Section 3.2 that could enhance the model’s performance.

First, we introduced randomized padding for region sequences. In the original approach, sequences were padded to a fixed length by appending padding tokens at the end. This fixed padding creates an arbitrary constraint, which limits chromatin-accessible regions to the beginning of each sequence. To address this, we randomized padding placement by adding a variable number of padding tokens at the start and end of each sequence, ensuring the total token length remains consistent. This approach allows the “ATAC-aware” encoder to learn patterns independent of sequence location. Second, we removed spans containing regions that overlap with others in the dataset. Due to the clustering by cell types and the application of referential shifts during dataset creation, duplicate DNA sequences were prevalent. To mitigate this redundancy, we filtered out spans that included already-seen regions, despite a significant reduction in dataset size.

The original “ATAC-aware” SegmentNT was trained using data processed as described in Section 3.2, while the “ATAC-aware” SegmentNT with random pad and no duplicates utilized data preprocessed with these two additional steps.

3.4 Results and Analysis

In this section, we present and analyze the experimental results from our study. We evaluate the model’s performance on the Region Ranking task, demonstrating that the encoder effectively learns the relationship between DNA sequences and their chromatin accessibility. Furthermore, we compared the performance of our “ATAC-aware” SegmentNT models in the DNA segmentation task with baseline models.

3.4.1 CAS Heads Comparison

We train an encoder to extract chromatin information from DNA sequences by attaching CAS heads and optimizing its parameters through the Region Ranking task. There are two key objectives during this phase of the study. First, we seek a CAS architecture that excels in the Region Ranking task, using its performance as a proxy for the model’s ability to learn chromatin-accessibility features from DNA sequences. Second, we ensure the encoder independently captures information linking DNA sequences to chromatin accessibility, avoiding over-reliance on the CAS heads. Since knowledge transfer to the DNA segmentation task depends solely on the encoder, it is crucial that the encoder independently generates biologically enriched representations of DNA sequences.

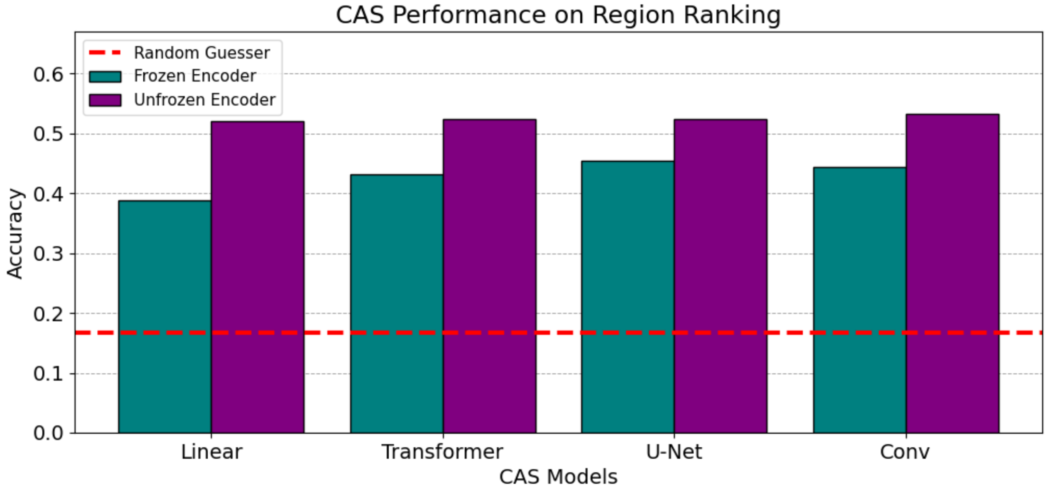


Figure 3.3: **CAS heads comparison on the Region Ranking task.** We compare the performance of four types of CAS on the Region Ranking task. Purple bars display the accuracy of the model with an unfrozen encoder. The Teal bars display the accuracy of the model with a frozen encoder. Red dotted line shows the accuracy a random guesser would achieve.

To validate these objectives, we experiment with four CAS types (detailed in Section 3.3.2) combined with two DNA encoder configurations: one with frozen weights and another with trainable weights. Both encoders are initialized with pre-trained NT-v2 50M model weights [17]. Model performance is evaluated based on accuracy in the Region Ranking task, defined as correctly ordering three regions in a span by their ground-truth chromatin accessibility levels. With six possible ranking permutations, a random model achieves an accuracy of $\frac{1}{6}$. We compare our models against this baseline and each other to assess performance improvements.

From the result of this experiment, the result of which is visualized in Figure 3.3, we can draw a few conclusions.

1. **Every CAS head significantly outperforms the random guesser.** For all types of CAS heads, both with frozen and unfrozen encoders, they achieve an accuracy significantly higher than that of the random guesser. This shows that our model

learns to identify biologically intrinsic features in DNA sequences that are related to chromatin-accessible regions.

2. **The model performs better when the encoder is trained together.** In all CAS head types, the model performs significantly better when trained with an unfrozen encoder compared to a frozen one. This suggests that the encoder actively learns to construct DNA sequence representations that encapsulate the chromatin accessibility information.

3. **The encoder retains most of the necessary information.** When the encoder is frozen, more complex CAS heads, such as Transformer, U-Net, and Conv, outperform the simpler Linear CAS head. However, this performance gap vanishes when the encoder is unfrozen. This indicates that, with an unfrozen encoder, the crucial chromatin information is retained by the encoder itself rather than relying on the CAS heads.

3.4.2 Performance Comparison in DNA Segmentation

Following training on the Region Ranking task, we transfer the DNA encoder to the DNA segmentation task by attaching it to a 1D U-Net segmentation head. The segmentation head mirrors the architecture used in the original SegmentNT paper [20], consisting of two down-sampling and two up-sampling convolutional blocks. Each block comprises two convolutional layers with 2,048 and 4,096 kernels, respectively. The model is trained to predict the probabilities of 14 genomic element classes at single-nucleotide resolution, using focal loss with $\gamma = 2$.

For the encoder, we use the model that performed best in the Region Ranking task—

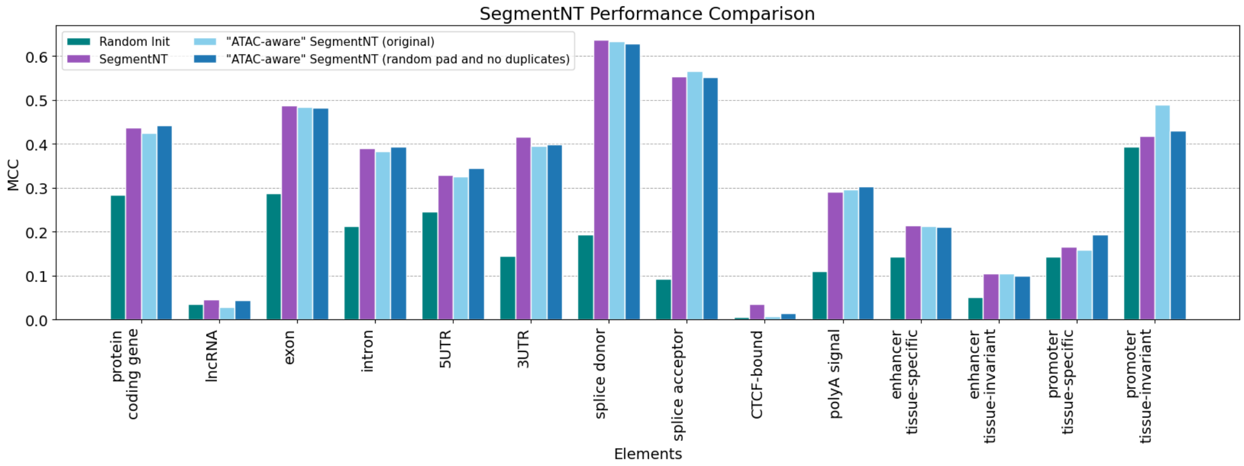


Figure 3.4: **SegmentNT Comparison on the DNA Segmentation Task.** We used the MCC metric to compare the performance of different SegmentNT models on the DNA segmentation task. MCC metric was calculated for each of the 14 types of genomic and regulatory elements.

specifically, the DNA encoder trained with the convolutional CAS head. This encoder, paired with a randomly initialized 1D U-Net head, is used to train two versions of our “ATAC-aware" SegmentNT models: “original", which follows the original data preprocessing pipeline described in Section 3.2, and “random pad and no duplicates", incorporating modifications detailed in Section 3.3.4.

We then compare the performance of these models against two baseline SegmentNT models using Matthews correlation coefficient (MCC). The first baseline uses a randomly initialized DNA encoder, highlighting the benefits of pre-training the encoder. The second baseline employs an encoder initialized with pre-trained NT-v2 50M weights [17], illustrating the impact of transferring knowledge from the Region Ranking task on the model’s performance on the DNA segmentation task. Results from this comparative study are presented in Figure 3.4.

By comparing the performance of these four different types of SegmentNT models, we can draw a few key conclusions.

1. **Using pre-trained NT encoder is essential for effective fine-tuning.** The randomly initialized model not only took about 4 times the training iterations to reach convergence when compared with the normal SegmentNT, it also significantly underperformed any other SegmentNT models. Specifically, the randomly initialized model achieved average MCC of 0.17 across genomics elements, compared with the normal SegmentNT that achieved average MCC 0.32. This discrepancy in performance affirms and underscores the importance of DNA foundation models for solving challenging tasks in genomics.

2. **“ATAC-aware" SegmentNT achieves similar mean MCC across elements as normal SegmentNT.** The overall performance between SegmentNT and “ATAC-aware" SegmentNT models were in-line. Specifically, all three models achieved average MCC of 0.32 across genomics elements, although their performance on each element varied.

3. **“ATAC-aware" SegmentNT performs significantly better in some regulatory elements.** Although “ATAC-aware" SegmentNT models displayed in-line overall performance with the normal SegmentNT, they outperformed in a few elements. Notably, “ATAC-aware" SegmentNT (original) and (random pad and no duplicates) significantly outperformed on tissue-invariant and tissue-variant promoters, respectively, when compared with the normal SegmentNT. This result confirms our initial hypothesis that exposing our model with scATAC data will primarily boost the model's performance for identifying regulatory elements, whose genomic boundaries are intrinsically related

with the chromatin accessibility information.

4. **Both versions of "ATAC-aware" SegmentNT outperform for the regulatory elements but in different ways.** "ATAC-aware" SegmentNT (original) primarily excelled at tissue-invariant promoters, whereas "ATAC-aware" SegmentNT (random pad and no duplicates) demonstrated superior performance with tissue-specific promoters. Despite both targeting regulatory elements, the two model versions achieved their respective strengths in different types of regulatory elements.

3.5 Discussion and Key Findings

Our study introduced a new approach to augmenting genomic language models by integrating single-cell ATAC-seq (scATAC) data into the SegmentNT model. Specifically, we used transfer learning for the SegmentNT genomics encoder. We developed a unique list-wise ranking approach called Region Ranking Task for learning chromatin accessibility, which overcomes limitations in direct fragment count prediction. By presenting the model with two non-zero and one zero regions, we created a robust framework for learning intrinsic genomic features. We then systematically explored four different architectures (linear, transformer, U-Net, and convolutional) on top of the DNA encoder to score chromatin accessibility. We then transferred the knowledge learned from the scATAC prediction task to the original DNA segmentation task by replacing SegmentNT's encoder with an "ATAC-aware" encoder.

From the results we observed on the Region Ranking Task, we examined that the encoder significantly outperformed random guessing in the Region Ranking task, indicating its ability to capture intrinsic chromatin accessibility features from the DNA sequences. Furthermore, training with an unfrozen encoder consistently improved performance, suggesting the encoder learned productive representations from the chromatin data. Finally, complex CAS heads did not provide substantial advantages when the encoder was trainable, implying that the encoder itself can effectively learn and encode chromatin accessibility information.

The performance of the "ATAC-aware" SegmentNT models on the DNA segmentation task highlighted several important observations as well. While the models maintained a comparable overall performance to the original SegmentNT across genomic elements, they significantly outperformed the normal SegmentNT models in regulatory elements, particularly promoters (both tissue-invariant and tissue-variant). Different data pre-processing strategies showed nuanced performance gains in specific types of promoters.

3.6 Limitations and Future Work

While our study demonstrates that exposing genomic models to additional biological context can influence their predictions in biologically meaningful ways, several limitations warrant consideration. This section outlines these limitations and suggests potential directions for future research to address them.

1. **Multispecies extension:** This study focuses exclusively on the human genome, using a single reference genome (hg38/hg19). Future work could evaluate the approach's applicability across species. Given the scarcity of scATAC-seq data for non-human organisms, advanced data augmentation and sampling strategies may be necessary to simulate comparable datasets.

2. **Scaling genomic context and model size:** Due to computational constraints,

our analysis was limited to a 3kb genomic context using the smallest NT encoder, NT-Multispecies-v2 (50M). Exploring longer genomic ranges (e.g., 10 kb, 30 kb) and larger encoders with more parameters could better capture the relationship between DNA and chromatin accessibility.

3. **Model interpretability:** While "ATAC-aware" SegmentNT models showed improved performance on specific regulatory elements, the source of these improvements remains unclear. Future studies could employ interpretability techniques, such as analyzing attention maps, to identify regions of importance and compare them with known transcription factor binding sites linked to regulatory elements.

4. **Cell-type specific scATAC analysis:** The current preprocessing pipeline pseudo-bulked scATAC-seq data based on cell types, but this information was not explicitly incorporated into the model. Developing cell-type-specific models could uncover nuanced variations in chromatin accessibility between cell types and enhance predictive accuracy.

Chapter 4

Developing a Multimodal Single-Cell Foundation Model with scATAC

4.1 Introduction

Single-cell ATAC-seq (scATAC-seq) has emerged as a transformative technology for profiling chromatin accessibility, providing detailed insights into gene regulation, cellular differentiation, and the dynamics of the genomic landscape. Advances in sequencing techniques now enable the simultaneous profiling of chromatin accessibility and gene expression at single-cell resolution, producing multiome data that integrates scATAC and scRNA reads. These multi-modal profiles offer a promising avenue for enhancing existing models in genomics and transcriptomics by capturing complementary layers of biological information. For instance, in our initial study (Section 3), we demonstrated that incorporating scATAC predictive tasks into the genomics language model SegmentNT improved its performance in segmenting regulatory elements from the genome.

Beyond our work, other notable studies have explored leveraging multiome data for multi-modal modeling. scGPT [12], for example, integrates scATAC-seq data into a transcriptomics foundation model by introducing modality- and peak-specific tokens. scooby [29] extends Borzoi [31], a sequence to functional assay model trained on bulk data, with a cell-specific decoder to achieve state-of-the-art performance in predicting scRNA-seq coverage and scATAC-seq insertion profiles along the genome.

While these approaches successfully incorporate multiome data, they share a key limitation: scATAC-seq data is only introduced during the fine-tuning phase of an already pre-trained model. This task-specific integration could restrict the generalizability of the model's newly acquired capacities to other tasks. Introducing scATAC-seq data during the pre-training phase could enable the model to learn epigenetic features in a more universally applicable manner. Furthermore, early exposure to scATAC-seq data eliminates task- and model-specific biases introduced during fine-tuning, offering a clearer understanding of the benefits of multi-modal training and optimizing its implementation. This study is driven by the following key objectives:

- **Develop a multi-modal foundation model exposed to both scRNA and scATAC data during the pre-training phase.**
- **Isolate the impact of scATAC-seq data by maintaining other factors consistent with a baseline transcriptomics foundation model.**
- **Evaluate the performance of the multi-modal model against existing baseline models on downstream tasks.**

To this end, we propose a multimodal single-cell foundation model that integrates scRNA and scATAC data during pre-training. Our model architecture builds on scBERT [11], incorporating additional components for multi-modal pre-training. We systematically compare models trained on different modality combinations and benchmark them against baseline transcriptomics foundation models and simpler predictive baselines. While our approach did not yield significant improvements in cell-type classification tasks, it introduces a novel framework for seamlessly incorporating scRNA and scATAC data that is generalizable across datasets and downstream applications.

4.2 Datasets

In our study, we utilize three main categories of datasets: scRNA-only, scATAC-only, and multiome. Multiome datasets, which combine both scRNA and scATAC profiles, are processed by applying pre-processing steps specific to each modality. This section outlines the modality-specific pre-processing strategies employed to align the datasets, followed by the filtering and binning methods applied uniformly across all datasets, regardless of their modality.

4.2.1 Preprocessing and Aligning Datasets

Our study included 16 datasets containing gene expression profiles across scRNA and multiome data. To ensure compatibility with our model, which requires a consistent set of features, we standardized the datasets by aligning their features to a shared set of genes. Specifically, we extracted Ensembl gene IDs [39] from each dataset and selected only those genes present in all datasets, yielding 10,784 unique gene IDs. While this step excluded some gene expression data from the raw sources, it minimized the impact of inherent cross-dataset differences. Further refinement involved removing gene IDs associated with multiple gene symbols, reducing the total to 10,733 unique genes

Our preliminary results (discussed in Section 4.4.1) indicate that the model performs optimally when the number and order of features (genes or peaks) are consistent. To achieve this alignment for scATAC data, we needed to match the number of peaks to the genes in the scRNA datasets following preprocessing. However, this alignment poses two main challenges. First, scATAC datasets contain significantly more peaks than genes, making simple 1-1 filtering impractical due to substantial data loss. Second, while gene annotations are standardized across scRNA datasets, Tn5 transposase insertion sites vary across scATAC datasets, even for the same regulatory element, complicating peak alignment.

To address these challenges, we align the scATAC datasets based on genetic loci of the genes used in our study. Specifically, for each gene, we define three regions relative to its position: 0-10kb, 10kb-20kb, and 20kb-50kb upstream and downstream from its starting genetic locus, including the gene body. Peaks within these regions were aggregated based on fragment counts, with overlapping peaks assigned to the region closest to the gene. To create gene-specific scATAC profiles, we computed a weighted sum of aggregated fragment counts for the regions, assigning weights of $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{6}$ to the 0–10kb, 10–20kb, and 20–50kb regions, respectively. This approach not only equalized the number of genes and peaks but also integrated peak data with genes in a biologically meaningful manner. The peak alignment process is illustrated in Figure 4.1.

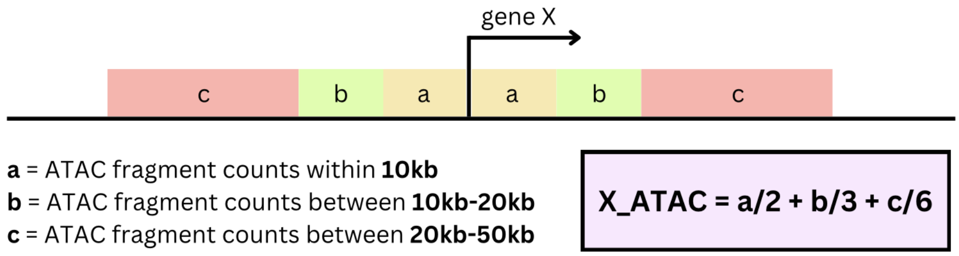

gene X
c
b
a
a
b
c
a = ATAC fragment counts within 10kb
b = ATAC fragment counts between 10kb-20kb
c = ATAC fragment counts between 20kb-50kb
X_ATAC = a/2 + b/3 + c/6

Figure 4.1: **Peak-alignment process relative to a gene.** For a given gene X, we align the scATAC data by aggregating fragment counts from nearby peaks and calculating a weighted sum based on their distances from the gene.

4.2.2 Dataset Filtering and Binning

To enable cross-modality alignment based on genetic loci, we used datasets referenced to hg38 and hg19, aligning hg19 data to hg38 using UCSC LiftOver [35]. For scATAC data, we converted raw read counts to estimated fragment counts, following established methods shown to improve performance [13]. After aligning the reference genomes and preprocessing scRNA and scATAC data as described in 4.2.1, we filtered cells, retaining only those with at least 1% of genes expressed and 1% of non-zero peak counts.

Gene expression and ATAC count values were normalized across genes or peaks and log-transformed to reduce skewness and facilitate comparisons. Subsequently, we applied an additional normalization step across cells, similar to that employed in [40], to downweight genes or peaks that are ubiquitously expressed or accessible. This emphasized features expressed or accessible at lower levels but more informative for distinguishing cell states. Finally, non-zero entries were percentile-binned into five buckets before being passed into the model.

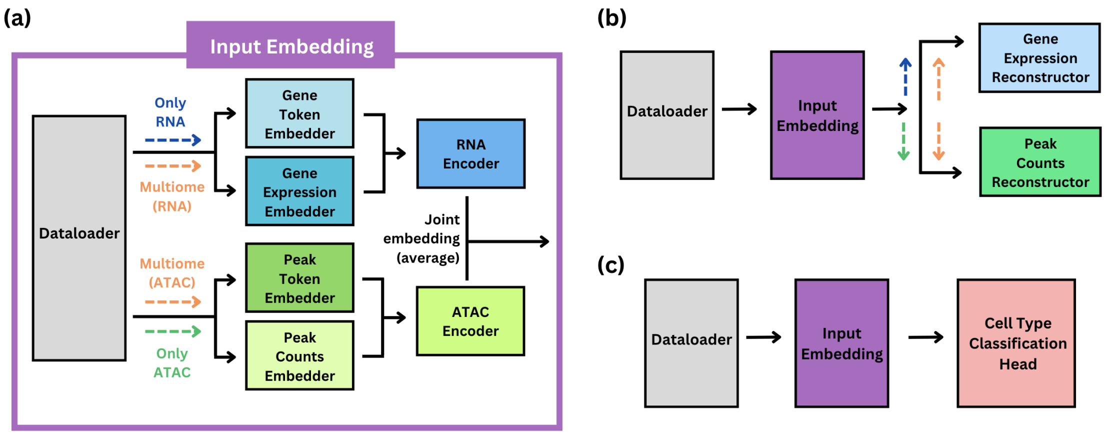


Figure 4.2: **Model architecture overview and the general workflow. a)** The embedding layer is designed to accept scRNA and/or scATAC input, comprising gene/peak tokens and their associated expression/count values. For multi-modal input, the model generates a joint embedding by averaging the embeddings from both modalities. **b)** Generative pre-training is conducted by attaching reconstructor heads to the embedding layer and training it on masked cell atlas data. **c)** For downstream applications, we take the pre-trained embedding layer and evaluate its performance on a cell type classification task.

4.3 Methods

This section details the architecture of our model, designed to flexibly handle both single- and multi-modal datasets. Following the architectural overview, we describe the pre-training task used to train the foundation model and the fine-tuning task employed to evaluate its performance. We also cover how each stage of training involves different training techniques designed to take advantage of different modalities of the data given.

4.3.1 Architecture Outline and Description

The primary objective of our foundation model is to learn an embedding mechanism capable of transforming input data into biologically meaningful representations applicable across various downstream tasks. To achieve this, the model architecture is divided into two encoding paths, each tailored to a specific modality: one for scRNA data and another for scATAC data.

- **Single-modality input:** When provided with only scRNA data, the input is processed exclusively through the scRNA encoding path. Similarly, scATAC-only input follows the scATAC encoding path.

- **Multi-modal input:** For multiome input, the scRNA portion is processed via the scRNA encoding path, and the scATAC portion through the scATAC encoding path. The final embedding is derived by averaging the representations from both modalities.

Each encoding path accepts modality-specific feature tokens and binned values: the scRNA encoding path uses gene tokens and expression values, while the scATAC path uses peak tokens and count values. Tokens are uniquely assigned to individual genes or peaks, ensuring no overlap. The feature tokens and binned values are processed through separate embedding layers, each mapping them to fixed-length embedding vectors of dimension $D = 128$, implemented using the Haiku embedding module (https://dm-haiku.readthedocs.io/en/latest/api.html). These token and value embeddings are summed to create a unified representation, which is then passed into an encoder.

Given the significant memory demands of Transformer models, especially with input feature sizes exceeding 10,000, we employed Performer—a matrix decomposition variant of Transformer—as the encoder. The model uses a Performer with six attention layers and ten attention heads, mirroring the structure of scBERT [11]. This design enables efficient processing while maintaining high representational capacity.

4.3.2 Pre-training Task: Generative Imputation

We pre-trained our foundation model using a masked reconstruction strategy, inspired by the masked language modeling approach from BERT [6]. In our adaptation, we randomly masked non-zero gene expression or peak count values and tasked the model with reconstructing these masked values based on the unmasked genes or peaks. The reconstruction loss was computed using cross-entropy loss:

$$L = -\frac{1}{M}\sum_{i=1}^{M}(\frac{1}{N}\sum_{j=1}^{N} y_{i,j}\log(p_{i,j}))$$

Here, M represents the number of cells and N represents the number of masked values; $y_{i,j}$ and $p_{i,j}$ denote the true and predicted values, respectively, for gene or peak j in cell i. This self-supervised learning strategy allowed the model to learn rich representations of gene expression and peak count patterns from large volumes of unlabeled data.

For single-modality inputs, the pre-training loss was calculated using the corresponding modality's cross-entropy loss. For multi-modal inputs, the model used the same joint embedding for both the RNA and ATAC encoders to reconstruct gene expression and peak counts. The final pre-training loss was computed as the average of the reconstruction losses for the two modalities. By sharing the same embedding for both reconstructions, the model was encouraged to learn representations that captured meaningful biological patterns relevant to both scRNA and scATAC data.

To enhance flexibility and robustness, we also implemented a mixed training regime, exposing the model to single-modality and multi-modal inputs during the same training run. Specifically, we alternated between three input types (scRNA only, scATAC only, and multiome) at each training step, with each batch containing 256 samples of the respective modality. This mixed training approach allowed the model to adapt to different modalities during fine-tuning and addressed the challenge of limited multiome data by leveraging the more abundant single-modality datasets.

4.3.3 Fine-tuning Task: Cell Type Classification

To evaluate the effectiveness of the foundation model, we trained the model on supervised learning task of cell type classification. To do so, we put a cell type classification head on a 128-dimensional embedding produced by the embedding layer. This classification head is composed of a one-dimensional convolution layer followed by a three-layer linear neural network that computed the probability for each cell type. For the training objective, we used the cross-entropy loss:

$$L = \frac{1}{M} \sum_{i=1}^{M} z_i \log(q_i)$$

where M represents the number of cells, and z_i and q_i are the ground-truth cell type label and predicted label for cell i, respectively.

4.4 Results and Analysis

In this section, we first cover the result from the preliminary studies we conducted that were informative for making critical design decisions of our model. Following these preliminary results are how our the transcriptomics vs. multiome versions of the model perform on both pre-training and fine-tuning tasks. Within the analysis, we also compare how our model performs relative to other baseline models, including scBERT and linear regression.

4.4.1 Feature Ordering Study

The model's input consists of two primary data types: unique tokens representing features (genes or peaks) and their corresponding expression or count values. The manner in which these features are ordered before being presented to the model plays a critical role in learning. We explored three distinct feature ordering strategies:

1. **Fixed order across cells:** In this approach, all cells share a consistent feature order. While this eliminates per-cell unique ordering and reduces the risk of overfitting, it closely resembles traditional positional encoding, which may limit its capacity to capture novel information.

2. **Cell-specific fixed order:** Here, feature order is randomized for each cell at the start of training and remains fixed throughout. This creates unique feature orderings for individual cells, which could lead to overfitting as the model may memorize the ordering instead of learning biologically relevant patterns.

3. **Dynamic randomization:** Feature order is randomized at every training step for each cell. This method prevents cell memorization entirely and pushes the model to learn robust, feature-specific embeddings independent of positional encoding. However, the added complexity might hinder the model's ability to learn effectively.

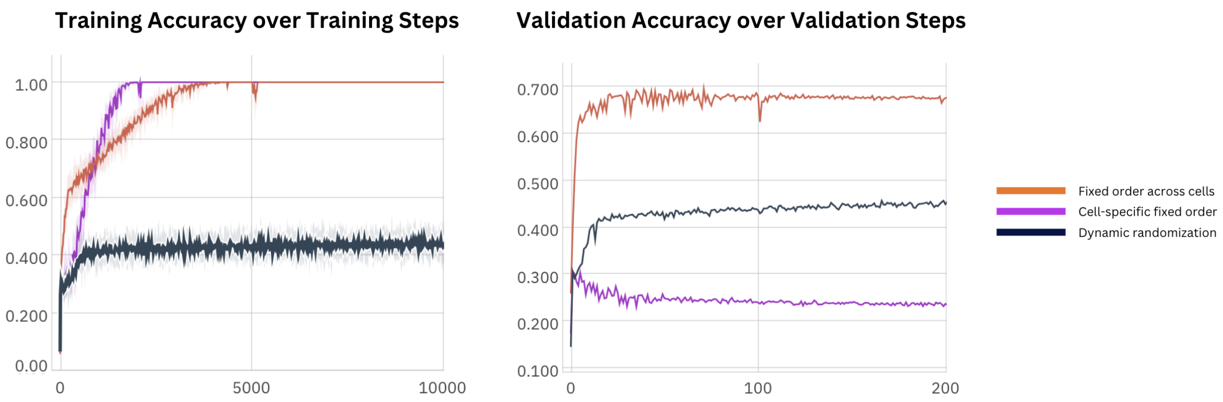


Figure 4.3: **Training and validation performance on cell type classification for different feature ordering mechanisms.** The two graphs show how each feature ordering mechanism impacted the model's performance on the cell type classification task. The orange line uses "fixed for all cells" ordering, the purple line uses "randomized initially" ordering, and the black line uses "randomized every time" ordering.

To evaluate the impact of feature ordering on model performance, we pre-trained three scRNA-only models, each employing one of the previously described ordering strategies. The models were then fine-tuned on cell type classification task using the Zheng68k (only scRNA) dataset.

As shown in the training and validation curves in Figure 4.3, feature ordering significantly affects performance. The "fixed order across cells" strategy achieved the best validation accuracy. Despite its similarity to traditional positional encoding, this approach provides a consistent feature representation, preventing overfitting while retaining sufficient information for effective learning. In contrast, when each cell has a unique, stationary feature ordering, the model severely overfits. Training accuracy (purple line) rapidly converges to perfection, but validation performance deteriorates over time, reflecting the model's inability to generalize. Lastly, introducing new feature orderings for every training batch presents excessive complexity. The model struggles to learn meaningful patterns, as evidenced by its poor performance on both training and validation sets (black lines). While the results may vary with changes in data size or model scale, our findings indicate that maintaining a consistent feature order across all cells is optimal for balancing generalization and learning efficiency for our model.

4.4.2 Results with Only scRNA

Before exploring multiome datasets and mixed training, we first evaluated our model's performance using only scRNA data for both pre-training and fine-tuning. To assess this, we benchmarked our model on a cell type classification task against two baseline models: logistic regression and scBERT. These baselines were selected to provide a comprehensive comparison. Logistic regression represents a simple, interpretable approach that does not rely on the scale or complexity of deep learning. In contrast, scBERT serves as a comparable foundation model with a similar architecture, allowing us to fairly evaluate the relative performance of our model. Both our model and scBERT were pre-trained on the complete set of scRNA-only training datasets available. Subsequently, all three models—our model, scBERT, and logistic regression—were fine-tuned and evaluated on the Zheng68k dataset.

The key performance metrics and confusion metrices of test results of our and baseline models are shown in Table 4.1 and Figure 4.4. From these results, we can draw some conclusive insights.

1. **Models' performance varies significantly across cell types.** Each model demonstrates a stark disparity between the cell type it classifies most accurately (near 90% accuracy) and the one it performs worst on (near 0% accuracy). This discrepancy can be attributed to the highly imbalanced nature of the Zheng68k dataset, where models

	Accuracy	MCC	F1-score
Logistic Regression	**0.725**	**0.656**	**0.715**
scBERT	0.718	0.649	0.708
Our model	0.706	0.635	0.705

Table 4.1: **Performance summary of our and baseline models on Zheng68k dataset.** This table summarizes the key metrics achieved by our model and the baseline models on the cell type classification task using only scRNA data. Notably, the logistic regression model outperformed the others, achieving the best results despite its simplicity and lower computational demands

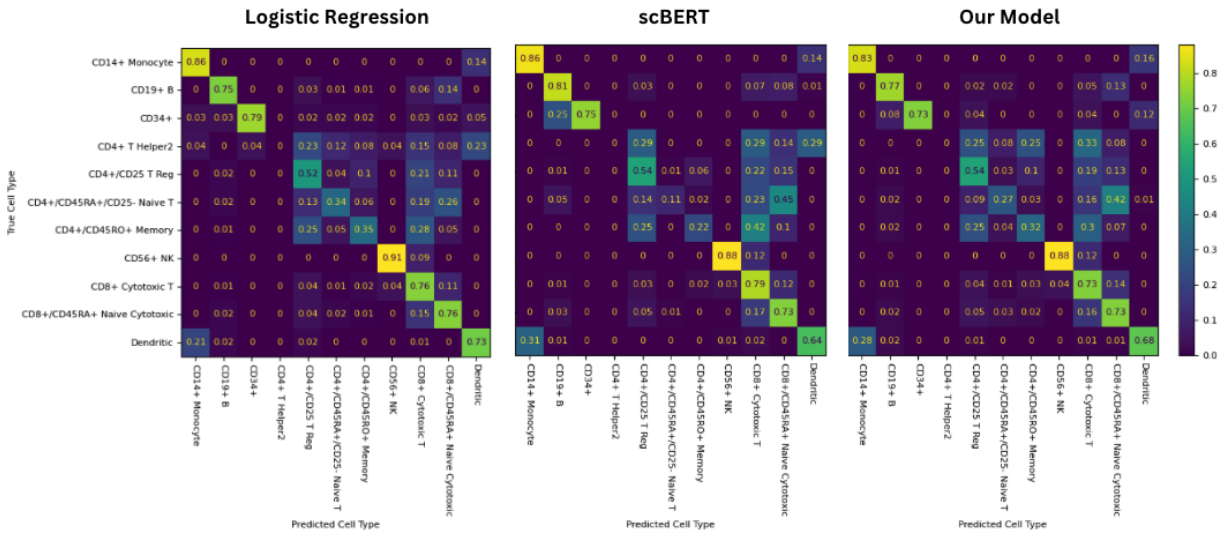


Figure 4.4: **Confusion matrices of the test results on Zheng68k dataset.** The confusion matrices in this figure illustrate the cell type predictions made by the tested models compared to the true labels. True cell type labels are represented by rows, while predicted cell types are represented by columns.

tend to focus on classifying the most abundant cell types. For instance, CD8+ cytotoxic T cells, which all models classify with the highest accuracy, constitute 30.3% of the dataset, whereas CD4+ T Helper2 cells, classified with near-zero accuracy, represent only 0.1%.

2. **Our model does not outperform baseline models with scRNA data alone.**

When trained solely on scRNA data without multiome inputs or mixed training, our model does not demonstrate a performance advantage over baseline models like scBERT or logistic regression. Nevertheless, it achieves comparable results across key metrics and does not exhibit significant underperformance.

3. **Logistic regression achieves the best performance despite its simplicity.** With no pre-training requirements and a straightforward parameter update mechanism, the logistic regression model outperforms both our model and scBERT. This finding raises critical questions about the advantages of foundation models for cell type classification tasks using gene expression data.

4.4.3 Results with Multiome and Mixed Training

In Section 4.4.2, we observed that our model, when trained exclusively on scRNA data, did not outperform the baseline models. However, the true advantage of our model lies in its ability to process multiome datasets and flexibly handle both single- and multi-modal inputs during training. In this section, we delve into the impact of mixed training, a capability unique to our model, and compare its performance on fine-tuning datasets of various modalities with that of baseline models.

During pre-training, the models were trained using either only scRNA datasets or a mixed training regime. Since scBERT is designed solely for gene expression data, it was pre-trained exclusively on scRNA datasets. To evaluate the effects of incorporating multi-modal datasets, we trained multiple versions of our model using both pre-training strategies.

For fine-tuning, we utilized the NeurIPS dataset, a single-cell multiome dataset with annotated cell types. First, all three baseline models were fine-tuned and evaluated using only

Model Name	**Training Mechanism**	
	Pre-training Strategy	Fine-tuning Modality
Logistic Regression (scRNA)	N/A	scRNA
Logistic Regression (multiome)	N/A	multiome
scBERT	only scRNA	scRNA
Our Model (scRNA)	only scRNA	scRNA
Our Model (mixed & scRNA)	mixed training	scRNA
Our Model (mixed & multiome)	mixed training	multiome

Table 4.2: **Summary of the models for NeurIPS dataset cell type classification study.** This table outlines the six models employed in the study. For pre-training, models were either not pre-trained ("N/A"), trained solely on scRNA data ("only scRNA"), or trained using a mixed regime ("mixed training"). During fine-tuning, the models utilized either gene expression data only ("scRNA") or both gene expression and chromatin accessibility features ("multiome") for the cell type classification task.

the scATAC features from this dataset. Then, while both our model and a logistic regression baseline could also be trained with multi-modal data, scBERT could not, as it is restricted to gene expression inputs. In total, we trained and compared six versions of our model and the baselines, enabling a comprehensive analysis of performance across different modalities and training strategies. The summary of these six models are shown in Table 4.2.

The key performance metrics and confusion metrices for NeurIPS cell type classification study are illustrated in Table 4.1 and Figure 4.4. These results lead to some key conclusions.

1. **Models perform better on NeurIPS dataset than Zheng68k dataset despite there being more labels.** Models demonstrated higher accuracy on the NeurIPS dataset, even though it includes more cell type labels. This improved performance likely stems from the NeurIPS dataset being less imbalanced, highlighting the impact of dataset balance on classification challenges.

	Accuracy	MCC	F1-score
Logistic Regression (scRNA)	0.837	0.819	0.828
Logistic Regression (multiome)	**0.862**	**0.846**	**0.855**
scBERT	0.840	0.824	0.836
Our Model (scRNA)	0.820	0.802	0.815
Our Model (mixed & scRNA)	0.840	0.825	0.838
Our Model (mixed & multiome)	**0.862**	0.844	0.852

Table 4.3: **Performance summary of our and baseline models on NeurIPS dataset.** This table summarizes the key metrics achieved by our model and the baseline models on the cell type classification task on NeurIPS dataset. Across different versions of the models tested, we see the general trend where mixed pre-training and multi-model fine-tuning helps the overall performance.

2. **Generally, more modalities seen in pre-training and fine-tuning helps the model.** Both pre-training and fine-tuning with multiple modalities improved model performance. Logistic regression benefited from using both scRNA and scATAC features, and our model achieved better results with mixed pre-training and multiome fine-tuning compared to single-modal approaches. These results emphasize the value of multi-modal datasets in enriching feature representations and enhancing downstream task performance.

3. **Among the deep learning foundation models, our model with mixed pre-training and multiome fine-tuning performed the best.** When trained and fine-tuned solely with scRNA data, scBERT outperformed our model. However, with the inclusion of mixed pre-training and multiome fine-tuning, our model surpassed scBERT. This underscores the flexibility of our model in leveraging multiome data, offering advantages over other transcriptomics foundation models in multi-modal contexts.

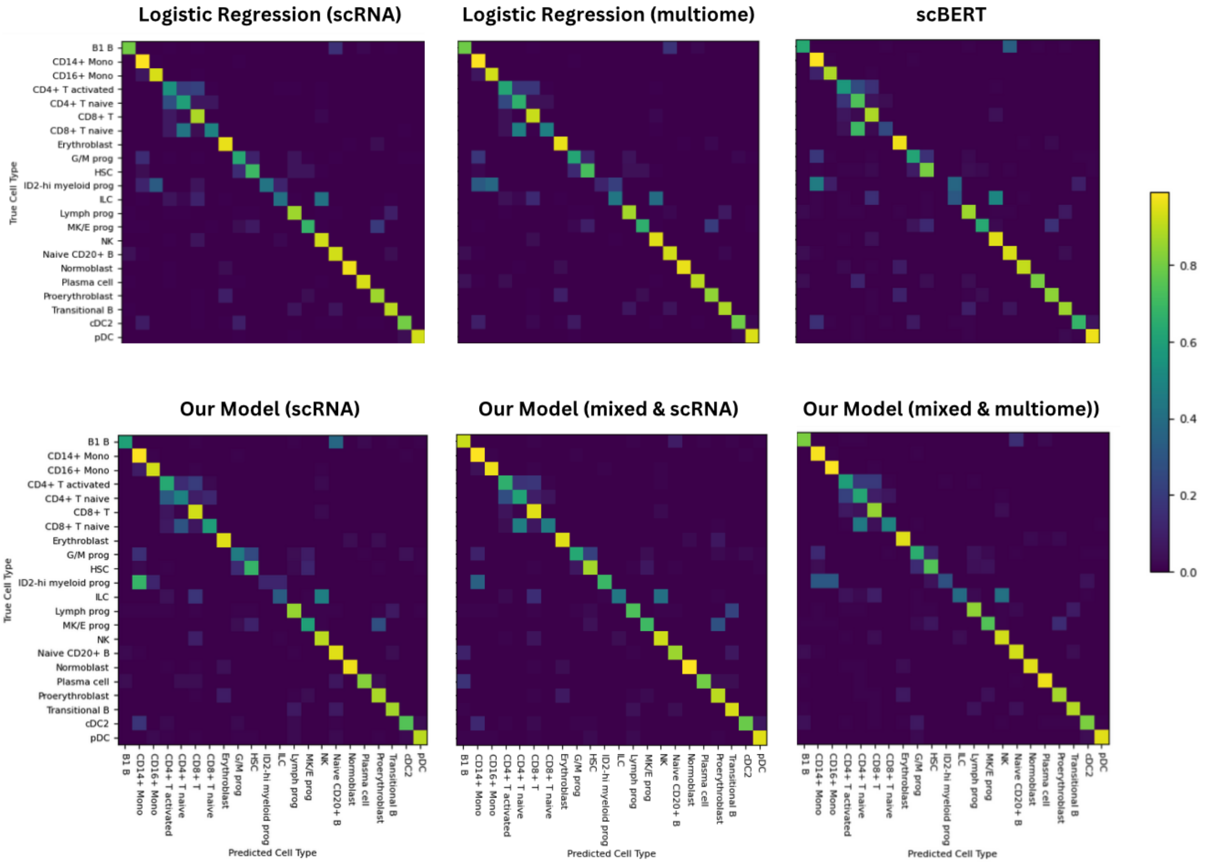


Figure 4.5: **Confusion metrices of the test results on Zheng68k dataset.** On the top row are the results from logistic regression with scRNA only, logistic regression with multiome, and scBERT, from left to right. On the bottom row are the results from different versions of our mode: scRNA pre-train and fine-tune, mixed pre-train and scRNA fine-tune, and mixed pre-train and multiome fine-tune, from left to right.

4. **Logistic regression still achieves the best performance despite its simplicity.**

   Consistent with findings from the Zheng68k dataset study, logistic regression achieved the best performance, outperforming both foundation models. This result challenges the necessity of deep learning-based foundation models for cell type classification tasks; it also challenges us to explore what types of tasks could benefit from foundation models. We dive deeper into these questions in subsequent sections.

4.5 Discussion and Key Findings

In this study, we developed a foundation model designed to flexibly handle both single- and multi-modal data during pre-training and fine-tuning phases. Our model is capable of training on scATAC, scRNA, or combined scATAC and scRNA data within the same pre-training phase, which has not been achieved before to the best of the author's knowledge. To enable this capability, we processed scRNA datasets by standardizing gene features across datasets and aligned scATAC peaks based on their genetic proximity to corresponding genes. These features, along with their binned values, were fed into the model, which employed masked language modeling to learn biologically meaningful cell representations derived from gene expression and chromatin accessibility data.

When evaluated on gene expression alone, our model did not outperform—and in some cases underperformed—existing foundation models like scBERT. However, when leveraging multi-modal data for both pre-training and fine-tuning, our model demonstrated superior performance compared to scBERT. While this improvement cannot be universally generalized, the results indicate the potential of multi-modal data in enhancing foundation model capabilities over traditional single-modal approaches.

Surprisingly, despite the sophistication of deep-learning-based foundation models, simple logistic regression consistently outperformed these models in cell type classification tasks across datasets. This aligns with prior findings that transformer-based models, while successful in NLP and computer vision, have yet to consistently yield superior performance in biological applications. As noted by prior work [41], the benefits of these models in biology remain under debate.

Our findings prompt critical reflection on the current trajectory of foundation models in biology. Future research should explore tasks where these models could excel, such as few-shot and zero-shot learning or scenarios with sparse labels. Additionally, more robust evaluation frameworks are needed to accurately quantify the utility of foundation models in this field.

4.6 Limitations and Future Work

Our study carries certain limitations that warrant discussion. For each limitation, we suggest potential avenues for future research to mitigate its effects.

1. **Model and dataset scale:** Computational and memory constraints restricted the experiments we could conduct. For instance, we were unable to test larger encoder architectures, higher-dimensional representations, or alternative mechanisms for joint embedding computation, such as cross-attention. Additionally, the limited availability of multiome datasets at the time constrained the breadth of our analyses. Future research could address these issues by leveraging increasing computational resources and the growing scale of multiome datasets, potentially uncovering performance improvements with expanded model and dataset scales.

2. **Different dataset pre-processing techniques:** In our study, we grouped scATAC peaks based on genetic distances relative to their corresponding genes. While effective, this method led to some loss of features. Exploring alternative pre-processing techniques

that retain more information or apply different grouping criteria could enhance model performance.

3. **Fine-tuning task diversity:** We evaluated our models primarily using two cell type classification tasks. Expanding the scope of downstream tasks, including more biologically diverse applications, could yield deeper insights into the models' strengths and weaknesses. Additionally, incorporating zero-shot or few-shot tasks—more challenging scenarios by nature—could help better examine the advantages of foundation models over simpler alternatives.

4. **More diverse baseline models:** Our comparisons were limited to scBERT as the primary deep-learning foundation model baseline. Including a wider variety of baseline models that employ different architectures and techniques would provide a more comprehensive assessment of our model's relative performance within the state of the art.

5. **Less dependency on the classification head:** Reducing dependency on the classification head represents a crucial direction for future research. Our experiments revealed that simple logistic regression can achieve competitive performance in cell type classification, suggesting that the model may overly rely on the classification layer rather than learning rich representations during pre-training. While evaluating performance on more complex biological tasks offers one path forward, an alternative approach involves architectural modifications to encourage deeper feature learning during the pre-training phase. Several potential strategies include implementing low-rank classification layers, introducing sparsity constraints, or employing differential learning

rates to limit classification head optimization relative to the core model parameters.

Addressing these limitations in future research will provide a clearer understanding of the potential and limits of foundation models in the biological domain, paving the way for their more effective application in real-world scenarios.

Chapter 5

Discussion and Conclusion

5.1 Summary of Research and Key Findings

This thesis investigated the integration of single-cell ATAC sequencing (scATAC-seq) data into deep learning frameworks for genomic analysis. By providing chromatin accessibility profiles at single-cell resolution, scATAC-seq offers valuable complementary information to existing genomic and transcriptomic data modalities. We explored this integration through two distinct approaches: enhancing genomic language models and developing a novel multi-omic foundation model.

In the first phase, we augmented SegmentNT's DNA segmentation capabilities by incorporating scATAC-seq data into a pre-trained DNA encoder. Our results demonstrated two key findings: first, the pre-trained DNA encoder significantly outperformed conventional genome representation methods, validating the foundation model's capacity to capture intrinsic biological features. Second, the integration of chromatin accessibility data improved the model's ability to identify regulatory elements, consistent with the known biological relationship between chromatin accessibility and regulatory function.

The second phase focused on developing a flexible foundation model that extends beyond traditional transcriptomics by incorporating multiome data during pre-training. Our architecture supports three distinct modality combinations—scRNA only, scATAC only, or both—enabling comprehensive evaluation of multi-modal integration benefits. While the model showed comparable performance to existing foundation models when using gene expression data alone, it demonstrated superior capabilities when leveraging multi-modal data. Notably, however, a simple logistic regression model outperformed all tested language models in cell type classification, suggesting that the advantages of foundation models may become more apparent in complex tasks requiring deeper biological understanding.

5.2 Final Remarks

The findings of this study do not necessarily provide a definitive answer the question: *"What is the best way to leverage scATAC-seq data to improve existing models?"* The vast search space of potential architectural designs and data processing methods leaves this question largely unexplored. However, our work serves as an essential foundation, offering insights that could guide future research toward addressing this challenge. We observed that integrating scATAC-seq data enhanced existing models in some aspects while presenting limitations in others. The key contribution of this study lies in our novel strategies to utilize scATAC-seq effectively by capitalizing on its biological strengths while mitigating its inherent drawbacks.

In summary, this thesis underscores the transformative potential of scATAC-seq data, particularly when combined with complementary modalities like scRNA-seq, to enhance genomic and transcriptomic models. By extending these models' capabilities, we move closer to developing a more quantitative understanding of gene expression and cell function. Moreover, accurate models informed by scATAC-seq could illuminate how DNA mutations across evolutionary timescales influence gene expression and regulatory mechanisms. This understanding is pivotal not only for decoding natural gene regulation but also for designing synthetic regulatory elements for therapeutic purposes, such as directing payload expression in gene therapies. As the field progresses, the insights and methodologies presented in this work provide a robust framework for optimizing the integration of scATAC-seq data, paving the way for impactful advances in biological research and biomedical applications.

Appendix A

Supplementary Materials for Augmenting Existing Genomic Language Models with scATAC

	Source
Neurips	GSE194122
Satpathy	GSE129785
Trapnell	GSE149683
10X	10X Genomics

Table A.1: **scATAC-seq datasets and sources.** In total, we used 4 different datasets that contained scATAC-seq data. Their names and sources are described in the table. The sources include both multiome and scRNA-only datasets.

	Chromosomes	Number of Files	Total Number of Spans
Train	chr1-19, X, Y	397	182,794
Validation	chr22	19	4,648
Test	chr20-21	38	7,377

Table A.2: **Pre-processed dataset summary.** This table shows a summary of the pre-processed datasets for the train, validation, and test splits. Note that the numbers in this table are before any referential shifts get applied.

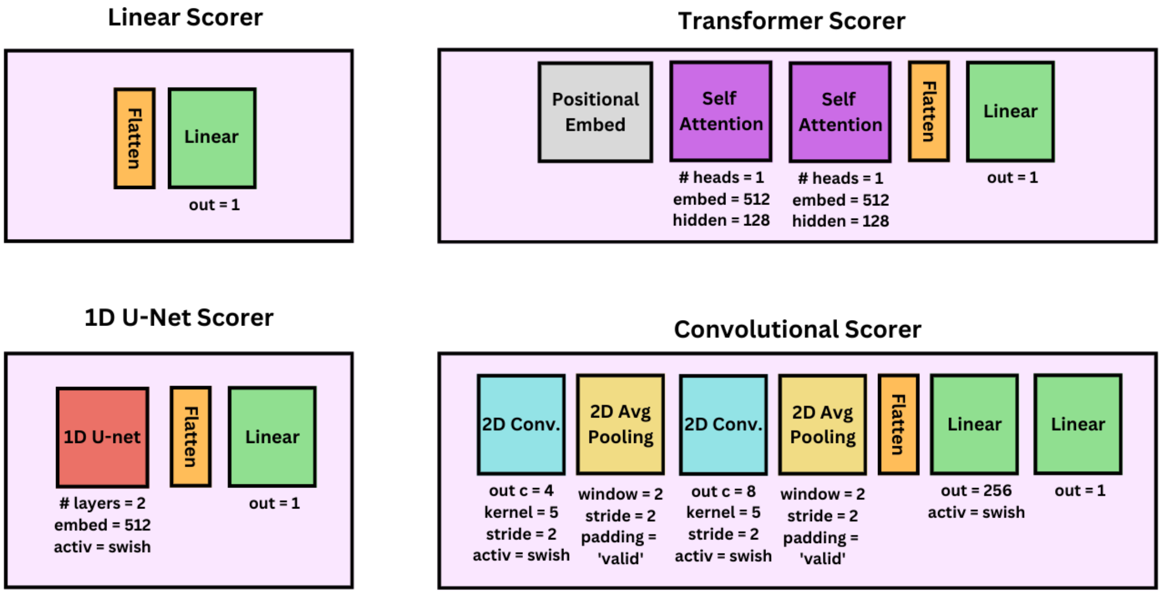


Figure A.1: **Chromatin accessibility scorer diagram.** The diagram visualizes the four types of chromatin accessibility scorers used in our study: linear, transformer, 1D U-Net, and convolutional.

Head Type	**Chromatin Accessibility Head Features**	
	Number of Layers	Number of Parameters
Linear	1	256,513
Transformer	18	2,881,793
1D U-Net	11	16,778,241
Convolutional	4	2,033,041

Table A.3: **Chromatin accessibility scorer heads summary.** Here, we summarize the sizes of each chromatin accessibility head by looking at its number of layers and parameters.

	Region Ranking Task	DNA Segmentation
learning rate	0.00005	0.00005
optimizer	adam	adam
batch size	64	256
regularization lambda	0.001	N/A
gamma (for focal loss)	N/A	2

Table A.4: **Hyperparameters summary.** This table shows the key sets of hyperparameters used in our study. The names of the columns indicate the training phase at which the recorded hyperparameters were used.

Appendix B

Developing a Multimodal Single-Cell Foundation Model with scATAC

	Source	**# Cells**	**scRNA /scATAC /Multiome**	**Cell Type Labels**	**Training Phase**
Neurips	GSE194122	69,249	Multiome	Yes	Fine-tune
PBMC 10K	10X Genomics	10,970	Multiome	No	Pre-train
Jejunum 10K	10X Genomics	10,640	Multiome	No	Pre-train
Lymph Node 10K	10X Genomics	14,566	Multiome	No	Pre-train
Kidney 10K	10X Genomics	22,772	Multiome	No	Pre-train
Brain 10K	10X Genomics	3,233	Multiome	No	Pre-train
Swanson	GSE158013	16,724	Multiome	No	Pre-train
Dogma	GSE156478	39,374	Multiome	No	Pre-train
Pituiary	GSE178454	15,024	Multiome	No	Pre-train
Cortex	GSE162170	8,944	Multiome	No	Pre-train
Panglao	PanglaoDB	1.356M	scRNA	No	Pre-train
Zheng68k	10X Genomics	68,579	scRNA	Yes	Fine-tune
PBMC 5K	10X Genomics	4,585	scATAC	No	Pre-train
Satpathy	GSE129785	63,882	scATAC	Yes	Fine-tune
Trapnell	GSE149683	719,109	scATAC	Yes	Pre-train

Table B.1: **Pre-processed dataset summary.** This table summarizes the datasets used in our study after pre-processing. This table shows each dataset's source, number of samples (cells), modality, if it contains cell type labels, and training phase at which it was used.

	Pre-train	Fine-tune
learning rate	0.00005	0.0005
optimizer	adamw	adamw
weight decay lambda	0.0001	0.0001
batch size	256	256
embedding size	256	N/A

Table B.2: **Hyperparameters summary.** This table shows the key sets of hyperparameters used in our study. The names of the columns indicate the training phase at which the recorded hyperparameters were used.

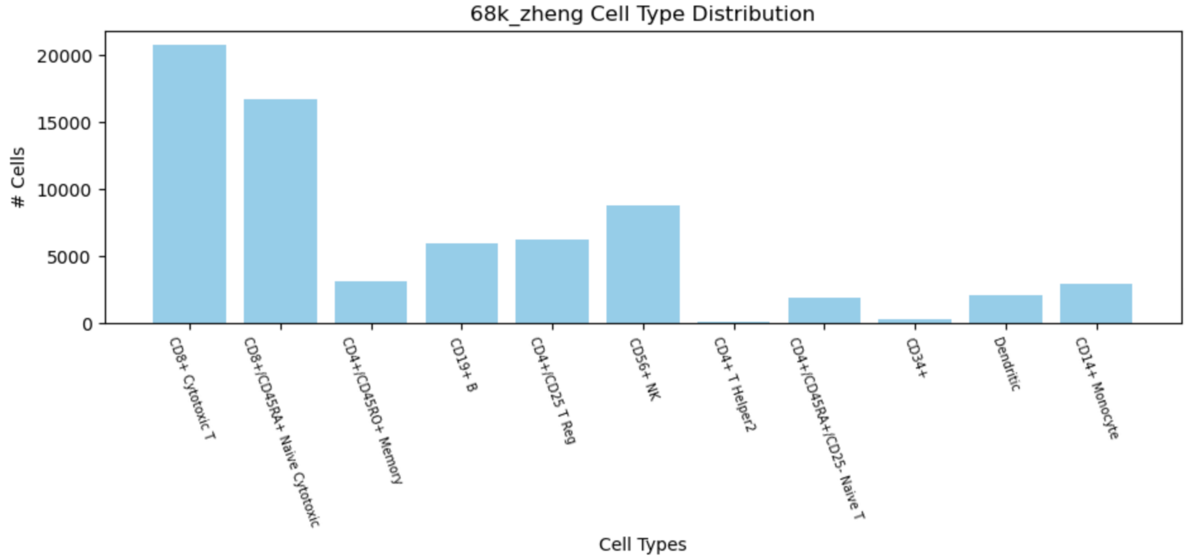


Figure B.1: **Zheng68k dataset cell type distribution.** This graph shows the cell type distribution of the Zheng68k dataset used in the fine-tuning phase.

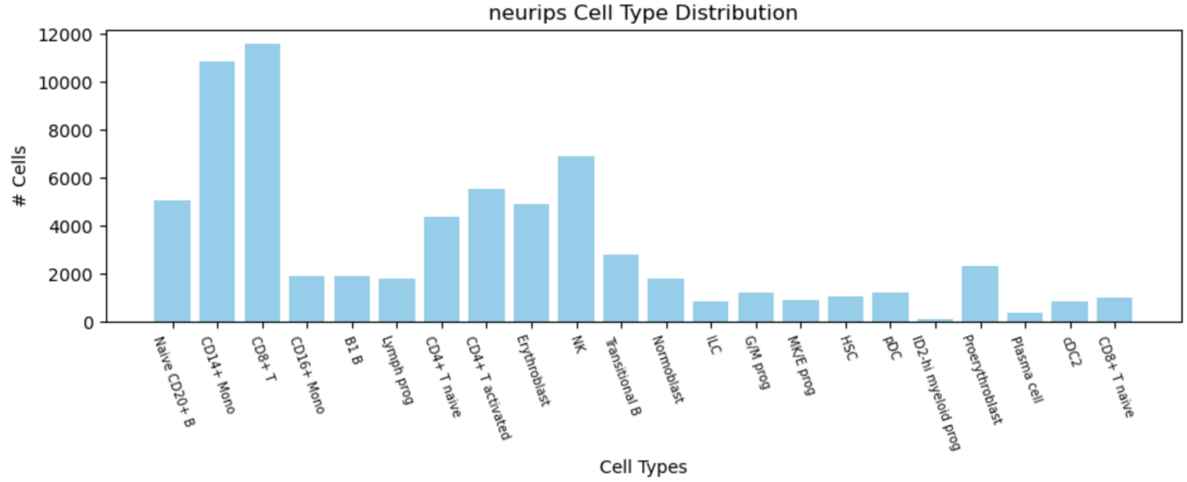


Figure B.2: **Neurips dataset cell type distribution.** This graph shows the cell type distribution of the Neurips dataset used in the fine-tuning phase.

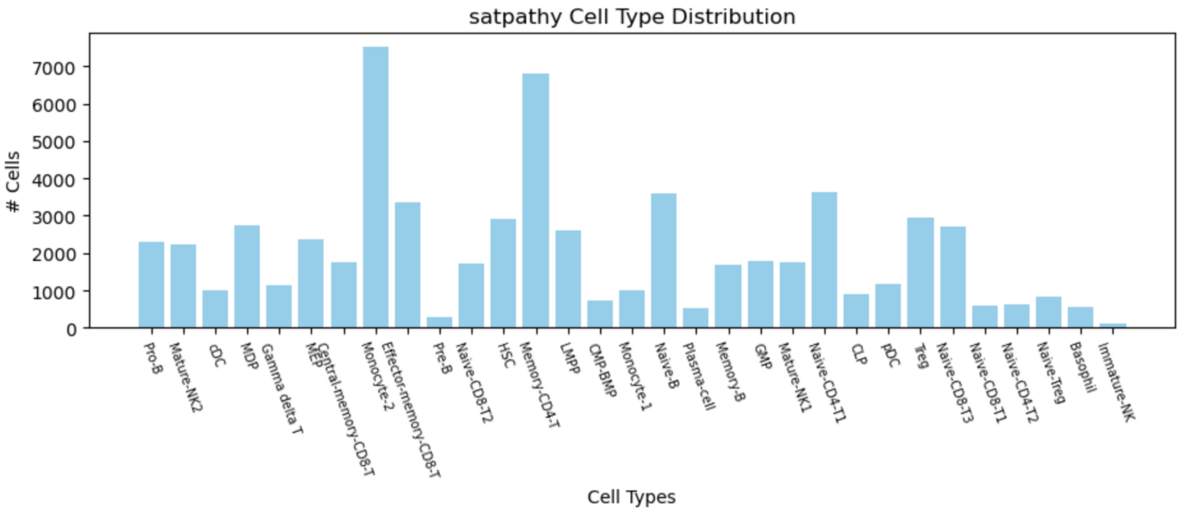


Figure B.3: **Satpathy dataset cell type distribution.** This graph shows the cell type distribution of the Satpathy dataset used in the fine-tuning phase.

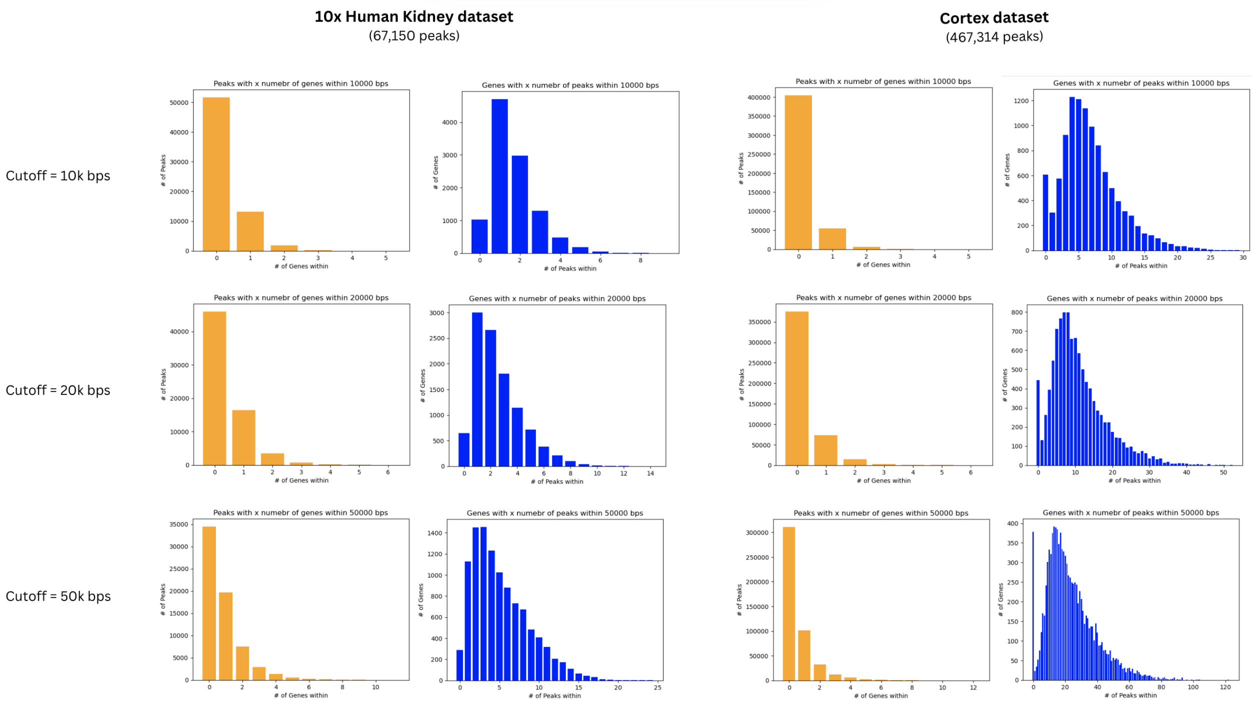


Figure B.4: **Genes and peaks distance study.** Our alignment process associates peaks with genetic loci, meaning peaks not closely located to any genes are excluded from our analysis. To assess the extent of this exclusion, we visualize the distribution of peaks and genes based on their relative distances. Specifically, we examine two metrics: the distribution of peaks per gene and the distribution of genes per peak. This study focuses on two datasets—the human kidney and cortex—which represent opposite extremes in terms of the number of available peaks. We evaluate three distance thresholds for peak-gene association: 10 kb, 20 kb, and 50 kb.

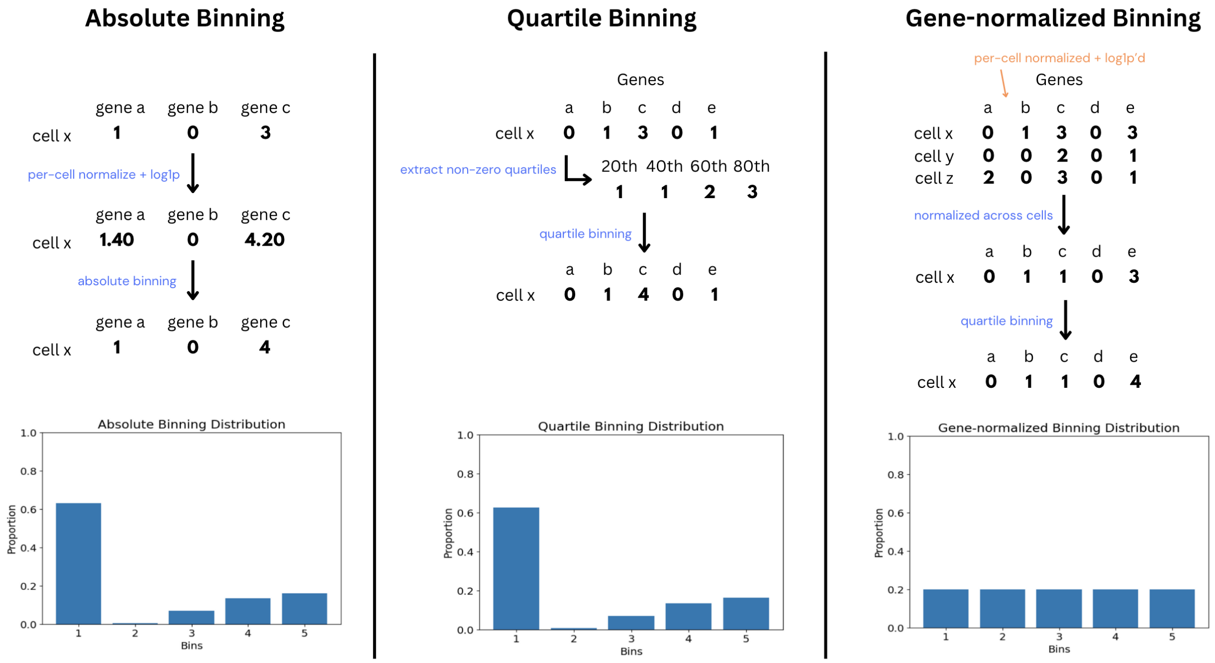


Figure B.5: **Binning strategies comparison.** We compare three binning strategies: absolute, quartile, and gene-normalized. To illustrate their differences, we visualize both the binning process and the distribution of the resulting binned values. The gene-normalized binning strategy, which we adopt in our study, produces binned values that are nearly evenly distributed.

References

[1] E. R. Mardis. “The impact of next-generation sequencing technology on genetics”. In: *Trends in genetics* 24.3 (2008), pp. 133–141.

[2] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature methods* 10.12 (2013), pp. 1213–1218.

[3] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf. “Single-cell chromatin accessibility reveals principles of regulatory variation”. In: *Nature* 523.7561 (2015), pp. 486–490.

[4] D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure. “Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing”. In: *Science* 348.6237 (2015), pp. 910–914.

[5] A. T. Satpathy, J. M. Granja, K. E. Yost, Y. Qi, F. Meschi, G. P. McDermott, B. N. Olsen, M. R. Mumbach, S. E. Pierce, M. R. Corces, et al. “Massively parallel single-cell

chromatin landscapes of human immune cell development and intratumoral T cell exhaustion". In: *Nature biotechnology* 37.8 (2019), pp. 925–936.

[6] J. Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[7] T. B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[9] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu. "Dnabert-2: Efficient foundation model and benchmark for multi-species genome". In: *arXiv preprint arXiv:2306.15006* (2023).

[10] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, et al. "Language models of protein sequences at the scale of evolution enable accurate structure prediction". In: *BioRxiv* 2022 (2022), p. 500902.

[11] F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao. "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data". In: *Nature Machine Intelligence* 4.10 (2022), pp. 852–866.

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

[13] L. D. Martens, D. S. Fischer, V. A. Yépez, F. J. Theis, and J. Gagneur. "Modeling fragment counts improves single-cell ATAC-seq analysis". In: *Nature Methods* 21.1 (2024), pp. 28–31.

[14] H. Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

[15] A. Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*. 2022. arXiv: 2204.02311 [cs.CL]. URL: https://arxiv.org/abs/2204.02311.

[16] E. Nguyen, M. Poli, M. Faizi, A. Thomas, M. Wornow, C. Birch-Sykes, S. Massaroli, A. Patel, C. Rabideau, Y. Bengio, et al. "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution". In: *Advances in neural information processing systems* 36 (2024).

[17] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. L. Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, et al. "The nucleotide transformer: Building and evaluating robust foundation models for human genomics". In: *BioRxiv* (2023), pp. 2023–01.

[18] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. "Unsupervised pattern discovery in human chromatin structure through genomic segmentation". In: *Nature methods* 9.5 (2012), pp. 473–476.

[19] J. Ernst and M. Kellis. "Chromatin-state discovery and genome annotation with ChromHMM". In: *Nature protocols* 12.12 (2017), pp. 2478–2492.

[20] B. P. de Almeida, H. Dalla-Torre, G. Richard, C. Blum, L. Hexemer, M. Gélard, J. Mendoza-Revilla, P. Pandey, S. Laurent, M. Lopez, et al. "SegmentNT: annotating

the genome at single-nucleotide resolution with DNA foundation models". In: *bioRxiv* (2024), pp. 2024–03.

[21] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[22] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, et al. "Base-resolution models of transcription-factor binding reveal soft motif syntax". In: *Nature genetics* 53.3 (2021), pp. 354–366.

[23] A. Thibodeau, S. Khetan, A. Eroglu, R. Tewhey, M. L. Stitzel, and D. Ucar. "CoRE-ATAC: A deep learning model for the functional classification of regulatory elements from single cell and bulk ATAC-seq data". In: *PLoS Computational Biology* 17.12 (2021), e1009670.

[24] A. Szałata, K. Hrovatin, S. Becker, A. Tejada-Lapuerta, H. Cui, B. Wang, and F. J. Theis. "Transformers in single-cell omics: a review and new perspectives". In: *Nature methods* 21.8 (2024), pp. 1430–1443.

[25] Q. Zou, P. Xing, L. Wei, and B. Liu. "Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA". In: *Rna* 25.2 (2019), pp. 205–218.

[26] L. Xiong, T. Chen, and M. Kellis. "scCLIP: Multi-modal Single-cell Contrastive Learning Integration Pre-training". In: *NeurIPS 2023 AI for Science Workshop*. 2023.

[27] T. Ashuach, M. I. Gabitto, R. V. Koodli, G.-A. Saldi, M. I. Jordan, and N. Yosef. "MultiVI: deep generative model for the integration of multimodal data". In: *Nature Methods* 20.8 (2023), pp. 1222–1231.

[28] W. Tang, H. Wen, R. Liu, J. Ding, W. Jin, Y. Xie, H. Liu, and J. Tang. "Single-cell multimodal prediction via transformers". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 2422–2431.

[29] J. C. Hingerl, L. D. Martens, A. Karollus, T. Manz, J. D. Buenrostro, F. J. Theis, and J. Gagneur. "scooby: Modeling multi-modal genomic profiles from DNA sequence at single-cell resolution". In: *bioRxiv* (2024).

[30] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, et al. "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest". In: *Nucleic acids research* 51.D1 (2023), pp. D638–D646.

[31] J. Linder, D. Srivastava, H. Yuan, V. Agarwal, and D. R. Kelley. "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation". In: *Biorxiv* (2023), pp. 2023–08.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

[33] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, et al. "GENCODE: the reference human genome annotation for The ENCODE Project". In: *Genome research* 22.9 (2012), pp. 1760–1774.

[34] E. P. Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), p. 57.

[35] G. Perez, G. P. Barber, A. Benet-Pages, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, C. M. Lee, et al. “The UCSC Genome Browser database: 2025 update”. In: *Nucleic Acids Research* (2024), gkae974.

[36] J. Burkardt. “The truncated normal distribution”. In: *Department of Scientific Computing Website, Florida State University* 1.35 (2014), p. 58.

[37] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. “Learning to rank: from pairwise approach to listwise approach”. In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 129–136.

[38] R. Jagerman, X. Wang, H. Zhuang, Z. Qin, M. Bendersky, and M. Najork. “Rax: Composable Learning-to-Rank using JAX”. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, pp. 3051–3060.

[39] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. “The Ensembl genome database project”. In: *Nucleic acids research* 30.1 (2002), pp. 38–41.

[40] C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, et al. “Transfer learning enables predictions in network biology”. In: *Nature* 618.7965 (2023), pp. 616–624.

[41] A. DenAdel, M. Hughes, A. Thoutam, A. Gupta, A. W. Navia, N. Fusi, S. Raghavan, P. S. Winter, A. P. Amini, and L. Crawford. “Evaluating the role of pre-training dataset

size and diversity on single-cell foundation model performance". In: *bioRxiv* (2024), pp. 2024–12.