Unlocking the Potential of Single-Cell Multi-Omics Data with Large-Scale Pre-Trained Transformer Models

***Sindura Kommu**
Virginia Tech
sindhura@vt.edu

***Sajib Acharjee Dip**
Virginia Tech
sajibacharjeedip@vt.edu

Xuan Wang
Virginia Tech
xuanw@vt.edu

Abstract

Integrating single-cell multi-omics data at a large scale is crucial for understanding cellular features and diagnosing complex medical conditions. This research aims to develop large-scale pre-trained transformer models that can effectively utilize single-cell multi-omics data. This work is expected to provide foundational models that can benefit various applications, including biomarker discovery and disease progression prediction.

1 Introduction

Integrating single-cell multi-omics data at a large scale is crucial for understanding cellular features and diagnosing complex medical conditions, like cancers, which exhibit tissue-level or cell-specific heterogeneity. Despite recent advancements in single-cell sequencing technologies, operationalizing large-scale multi-omics databases to reveal cellular features remains challenging. AI, especially large language models like ChatGPT, has shown promise in modeling various data types at scale (Theodoris et al., 2023; Yang et al., 2022). There's a growing interest in leveraging transformer architectures for understanding biological sequencing data from publicly available single-cell sequencing databases. However, fully realizing this potential faces challenges such as effective integration methods, lack of paired data for certain omics, and insufficient consideration for additional knowledge like gene regulatory networks. This research aims to develop large-scale pre-trained transformer models that can effectively utilize single-cell multi-omics data. The focus will be on incorporating gene regulatory networks into the pre-training process to enhance the model's understanding of multi-omics landscapes. This work is expected to provide foundational models that can benefit various applications, including biomarker discovery and disease progression prediction.

2 Method

2.1 Task 1

The objective is to develop a system that learns to integrate various omics data types using a shared encoder, enabling unified analysis without needing all data types to be present or paired. We've amassed a large collection of single-cell multi-omics datasets with unique characteristics, including paired datasets from 10x Genomics and unpaired datasets such as SHARE-seq, SNARE-seq, and Nephron. These include massive single-omic and multi-omic resources, like the 10.3 million sample CELLxGENE and extensive ATAC-seq and gene expression datasets. Additionally, we have disease-specific datasets like COVID-19-affected human lung samples and Human Kidney Cancer, which are vital for understanding cellular differences in health and disease, and for developing multi-omics analysis models. Our multi-omics data processing approach utilizes a shared backbone inspired by multimodal learning, which doesn't rely on paired data. The cell-gene matrix for different omics types is processed using a unified tokenizer that maps data to a common token space, facilitating analysis with a shared token encoder. This shared encoder, part of a large language model, is trained across modalities to extract semantic features for each cell, enabling multi-omic understanding without paired training data. During pre-training, self-training objectives generate labels, and task-specific heads are applied post-training for various biomedical applications.

2.2 Task 2

Gene regulatory networks offer valuable insights into context-specific gene regulation. We propose a novel approach that integrates these networks into the pre-training of transformer models for single-cell multi-omics analysis. Our method involves a joint self-supervised training strategy, aiming to en-

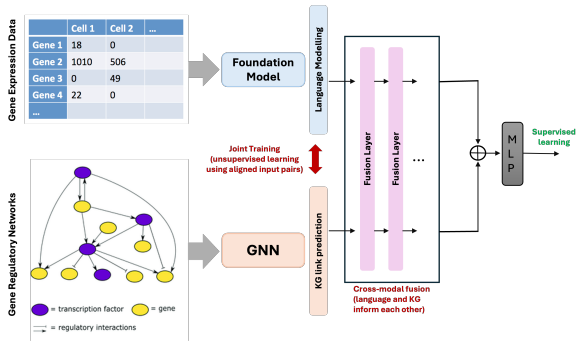

Figure 1: Overview of Our Co-Training Network.

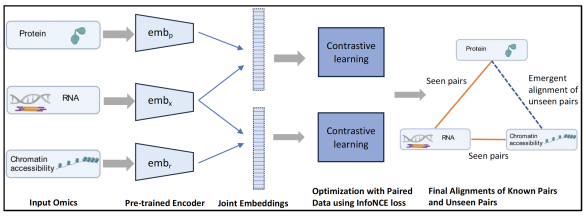

Figure 2: Overview of Our Full Model.

hance the model's understanding of gene characteristics and interactions. Our co-training framework combines single-cell multi-omics data with gene regulatory networks using a multi-omics large language model and Graph Neural Networks (GNNs). This approach unifies masked language modeling and link prediction in the networks, providing a comprehensive understanding of gene interactions beyond simple co-expression relationships. By integrating biological network knowledge into the pre-training process, our approach enhances the identification of functionally related genes based on shared regulatory mechanisms. This work has the potential to significantly improve our understanding of cellular processes and disease mechanisms.

2.3 Task 3

This task enhances single-cell multi-omics data integration by using cross-modal translation and contrastive learning, techniques designed to utilize unpaired data. The goal is to boost the model's capacity to interpret and relate information across various omics modalities for a deeper biological insight. Our approach is focused on aligning unpaired multi-omics data using cross-modal translation, building on previous successful methods like BABEL and scCross that translate between RNA-seq and ATAC-seq data. Our approach expands on this by including a wider variety of omics data and employing a bidirectional transformer for pairwise alignment, aiming to minimize the distance between original and translated data across different omics types, thus enhancing alignment accuracy. The other method includes enhancing the representation learning of a multi-omics large language model using contrastive learning and aligned multi-omics data. By encoding each omics modality with a specific self-attention transformer encoder and optimizing embeddings through contrastive learning with the InfoNCE loss, the goal is to minimize the distance between paired omics, while maximizing it for unpaired ones. This approach aims to improve the model's accuracy in classifying and aligning multi-omics datasets, leveraging both the diversity and the specificities of the data.

3 Conclusion

In this paper, we described the development of large-scale pre-trained transformer models that can effectively utilize single-cell multi-omics data. This work is expected to provide foundational models that can benefit various applications, including biomarker discovery and disease progression prediction.

Acknowledgements

This work is sponsored by the Commonwealth Cyber Initiative, Children's National Hospital, Fralin Biomedical Research Institute (Virginia Tech), Sanghani Center for AI and Data Analytics (Virginia Tech), Virginia Tech Innovation Campus, and a generous gift from the Amazon + Virginia Tech Center for Efficient and Robust Machine Learning. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Amazon + VT Center for Efficient and Robust ML.

References

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

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