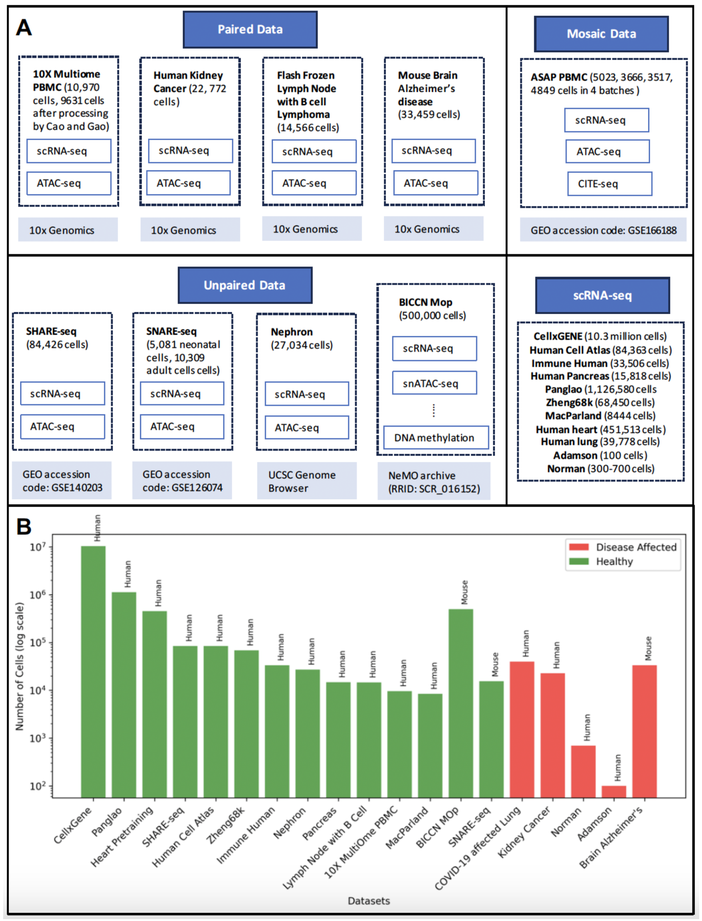
# Unlocking the Potential of Single-Cell Multi-Omics Data with Large-Scale Pre-Trained Transformer Models

Sajib Acharjee Dip, Sindhura Kommu, Xuan Wang

Department of Computer Science, Virginia Tech

## Task 1: Training a Shared Encoder

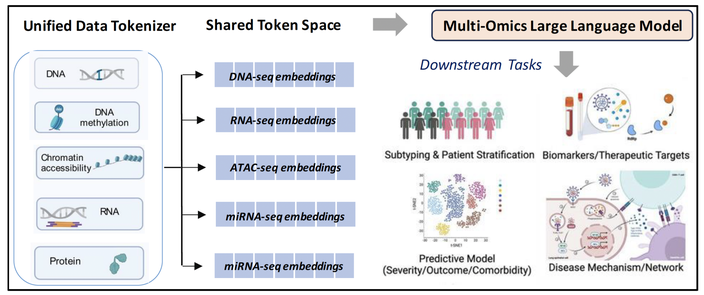
- We've amassed a large collection of single-cell multi-omics datasets with unique characteristics, including paired datasets from 10x Genomics and unpaired datasets such as SHARE-seq, SNARE-seq, and Nephron. These include massive single-omic and multi-omic resources, like the 10.3 million sample CELLxGENE 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.

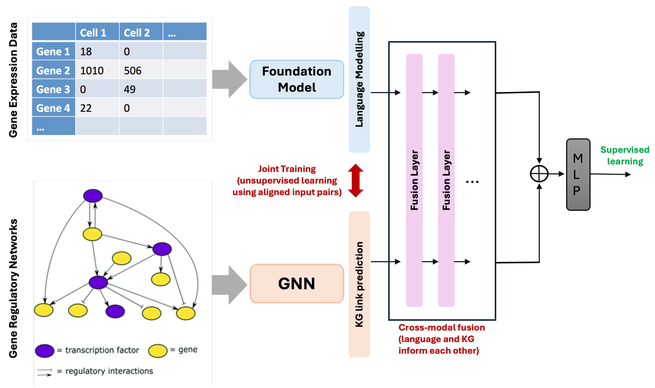
## Task 1: Training a Shared Encoder (continued)

- 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.



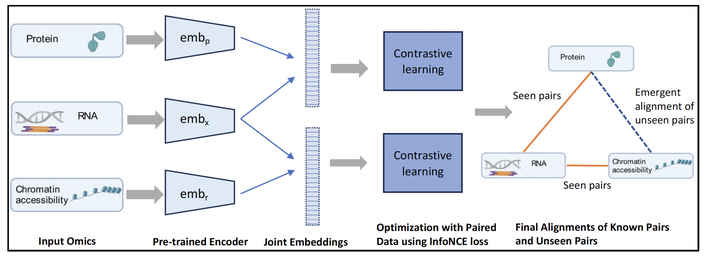

## Task 2: Integrate Gene Regulatory Networks

- 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 enhance 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.




## Task 3: Enhance Single-cell Multi-omics Data Integration

- 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.




- 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.

## Acknowledgement

Our 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.