Bioinformatics Summer School Report

Name: Ebtisam Alshehri
Internship Organization: Broad Institute of MIT and Harvard
Supervisor's Name: Dr. Mehrtash Babadi
Project title: Distributed Data Loader for Spatial Transcriptomics

1. Introduction

Spatial biology examines the organization and interactions of biomolecules like RNA, DNA, and proteins within tissues, using advanced imaging and sequencing to gain insights into cancer, neuroscience, and disease. Spatial Transcriptomics studies how gene expression varies across different tissue regions, aiding in Disease Mapping by linking gene activity to specific disease locations, and in Contextual Gene Expression by revealing gene functions in specific tissue areas.

The key project idea is to design a Spatial Transcriptomics Dataloader to scale up model training in a distributed ecosystem by generating mini-batches from each puck using various sampling strategies, leading to different model training outcomes. I worked in a cell biology AI laboratory that develops multiple innovative technologies, including the Cellarium ML core library, to which my project contributes. My project timeline included conducting a literature review, studying libraries used for distributed systems and single-cell analysis, understanding the sophisticated code base written by my labmates, and implementing my project with the guidance of exceptional scientists from my lab.

2. Materials and Methodology

The dataset I worked on was a 3D-aligned Slide-seq spatial transcriptomics atlas of a mouse brain, developed by two Harvard Core Labs at the Broad Institute: Evan Macosko's and Fei Chen's labs. The dataset includes 111 pucks, covering approximately 10 million cells. The metadata of the dataset includes XYZ coordinates from the Common Coordinate Framework (CCF), XY locations on NISSL images, and raw XY Slide-seq coordinates.

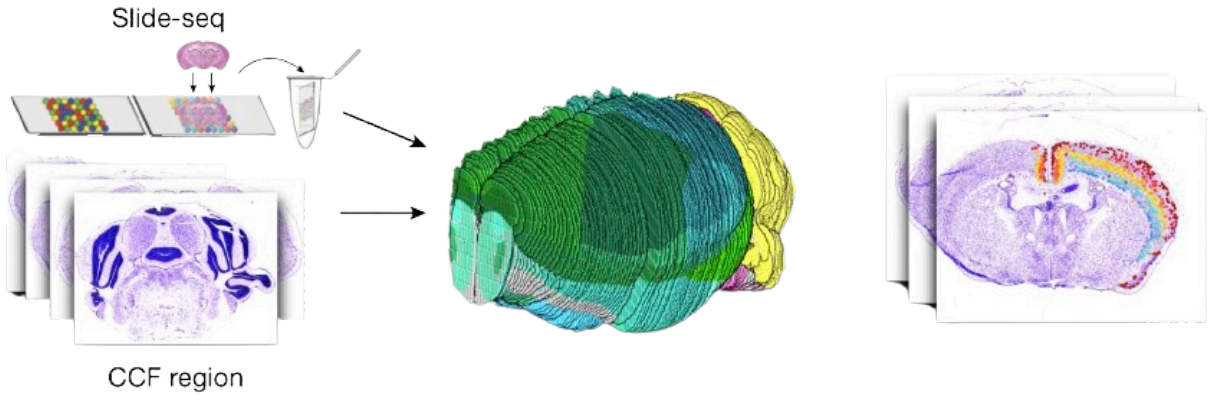


Figure 1: Slide-seq expression dataset

The features of the Distributed Spatial Dataloader include pre-computed KNN computations for each puck, chunk-based partitioning with exclusive data assignment, and various sampling strategies, such as random subsampling, which is easy to use but may only reveal large length scales when sampling a few cells. KNN subsampling with random seed cells allows for training on both small and large length scales. The Dataloader also supports multi-node and multi-GPU distributed model training and includes batch tracking to monitor workflow and execution.

I used Visual Studio Code to run Jupyter Notebooks and PyScripts for my project. Additionally, I used Google Cloud Platform to connect to a remote machine (SSH) with GPUs by creating VM instances in the cloud. I utilized Scanpy for data preparation and handling, with each puck being an anndata file in h5ad format for storing annotated data. I used PyTorch to override existing sampler classes, and PyTorch Lightning to manage the distributed backend of the system, along with scikit-learn for distributing the machine learning models. The current source code is confidential until it is released online on GitHub.

3. Results and Discussion

I have tried implementing various KNN algorithms from existing Python libraries, including Spatial Scipy, Scanpy, and Scikit-learn, and found that Scikit-learn was the fastest when calculating the KNN for each puck on the fly. I am currently working on integrating this into my laboratory's existing codebase, along with adding more test codes for each part of the distributed dataloader. Using subsampling techniques, we can explore local insights by subsampling nearby cells in the same tissue, while subsampling cells that are farther apart provides a global structural insight into how the data is organized.

4. Skills and Knowledge Gained

Coming from a pure ML/AI background, I developed a strong foundation in cell biology by reading relevant papers related to my project while also acquiring valuable Machine Learning Engineering skills. The opportunity to attend meetings, connect with brilliant scientists, collaborate with researchers, and discuss my ideas was incredibly rewarding. At the Broad Institute, we participated in weekly meetings where each group presented their research, and I had the chance to meet and communicate with faculty members from Harvard and MIT. As an intern, I developed independence and problem-solving skills while engaging in teamwork with colleagues working on the same data, particularly in the field of spatial transcriptomics.

5. Challenges and Solutions

Coming from a pure AI/ML background, I found the domain and data I was working with to be quite challenging, particularly due to the many unfamiliar acronyms. However, frequent meetings with biologists and using GPT-4 to simplify the acronyms helped to address this issue.

This was especially important given the tight timeline of the internship to get familiarized with new concepts.

6. Conclusion

My summer internship at the Broad Institute of MIT and Harvard has been a significant milestone and a dream realized. Being part of this esteemed community and contributing to their mission is a true honor, and I'm proud to call myself a Broadie. This experience represents the culmination of hard work, perseverance, and overcoming challenges in pursuit of my ambitions. Working within the Data Sciences Platform (Methods) in the innovative Cellarium lab has allowed me to apply AI/ML techniques to cutting-edge research on cellular and molecular functions in diseases. I am deeply grateful for the guidance of my mentor, Dr. Mehrtash Babadi, and the incredible support from lab collegues. The collaborative and supportive environment at Broad, particularly within the DSP community, has been exceptional, fostering a spirit of teamwork among brilliant scientists dedicated to the institute's mission. I am thankful for the encouragement from my peers and the entire Broad Institute, and I feel truly blessed to be part of this extraordinary community.

7. Acknowledgments

I would like to express my gratitude to the Broad Institute of MIT and Harvard, specifically to Dr. Eric Banks, ex-Head of Methods at the Data Sciences Platform, for arranging this opportunity. A special thanks to my mentor, Dr. Mehrtash Babadi, for his guidance and dedication; Dr. Dongshunyi "Dora" Li for her incredible support; Dr. Yerdos Ordabayev for his outstanding assistance; and Dr. David Benjamin and Sandeep Kambhampati for their support.

I also want to acknowledge KAUST Academy for their continued support since last year when I became part of the Alumni, and for sponsoring this opportunity. My sincere thanks go to KAUST Academy Director, Dr. Sultan Barakati; Deputy Director, Prof. Naeemullah Khan; and KAUST Academy Manager, Ms. Oroba AlNasr.

8. References

- Langlieb, J., et al. "The Molecular Cytoarchitecture of the Adult Mouse Brain." **Nature**, vol. 624, no. 7991, 2023, pp. 333–342.

- Wolf, F. A., Angerer, P., and Theis, F. J. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." **Genome Biology**, vol. 19, no. 15, 2018, pp. 1-5.

- Theodoris, Christina V., et al. "Transfer Learning Enables Predictions in Network Biology." **Nature**, vol. 618, no. 7965, 2023, pp. 616–624.