Directed Studies Proposal

Title: *Predicting Spatial Gene Expression from Immunofluorescence Imaging using Deep Learning*
Term: Fall/Winter 2025–2026
Supervisor: Amrit Singh
Co-Supervisor: Jiarui Ding
Lab: Computational Biology Lab, Faculty of Medicine

1. Background

Recent advances in deep learning (DL) have demonstrated remarkable capabilities in the field of medical imaging, achieving expert-level performance in tasks such as skin cancer classification and digital pathology. Specifically, deep learning models trained on histopathology images (e.g., H&E-stained slides) have been shown to predict tumor genotype, gene expression, and clinical phenotypes.

The **GeoMx Digital Spatial Profiler (DSP)** enables high-throughput spatial transcriptomic profiling using immunofluorescence-stained tissue sections. However, it currently requires **manual selection of regions of interest (ROIs)**, which presents two major limitations:

1. Selected ROIs may miss regions with highly discriminative gene-expression signatures.
2. Full-slide spatial gene-expression data is not available, limiting the interpretability of disease-specific biological signals.

This project addresses these limitations by leveraging DL to **predict spatial gene expression across the entire tissue section** using paired immunofluorescence (IF) and gene expression data.

2. Objectives

1. **Data Curation:** Collect and curate a large corpus of publicly available datasets from published studies using GeoMx DSP across multiple organs and disease contexts.
2. **Model Development:** Train an autoencoder-based convolutional neural network (CNN) to predict whole-slide gene expression from IF images.
3. **Software Tooling:** Deliver a reproducible Python package for spatial gene-expression imputation from image data.

4. **Visualization & Evaluation:** Identify the most predictable genes and assess model performance across organ types, sex, and cell markers.

3. Student Role

Developer – Bishneet

- Replicate analyses from prior publications (e.g., [HE2RNA](HE2RNA))
- Adapt models for compatibility with GeoMx file formats
- Train an autoencoder to predict ~18,000 protein-coding genes across 5 key organ types
- Build a modular Python package (IF2RNA) with preprocessing, training, evaluation, and visualization functions
- Generate figures and results for publication and presentation

4. Project Timeline

Month	Timeline	Goals
Month 1	Sept 1 – Sept 30	Set up development environment, reproduce baseline (HE2RNA) code, and test on sample datasets
Month 2	Oct 1 – Oct 31	Adapt input pipelines to support GeoMx datasets and ensure compatibility with model
Month 3	Nov 1 – Nov 30	Train base models on 5-organ dataset; begin evaluation of predictive gene performance
Month 4	Dec 1 – Dec 31	Analyze gene prediction performance (by sex, cell type); meet to review initial results
Month 5	Jan 1 – Jan 31	Package model code into a reusable Python module (IF2RNA); refactor and document
Month 6	Feb 1 – Feb 29	Iterate on feature set and evaluation metrics; explore visualization options
Month 7	Mar 1 – Mar 31	Integrate feedback, finalize analysis outputs, and prepare materials for Research Day
Month 8	Apr 1 – Apr 30	Final review, cleanup, and submission of poster and package; wrap-up project deliverables

Note: Lab meetings will occur approximately every 3 weeks. Additional progress will be tracked via GitHub commits and weekly asynchronous check-ins.

5. Deliverables

- **Curated Dataset Table** (Tabular summary of 50+ datasets with metadata)
- **Python Library:** IF2RNA with core functions for image → gene expression prediction
- **Figures for Manuscript:**
  - Table 1: Curated datasets
  - Figure 1: Method overview (IF2RNA pipeline)
  - Figure 2: Gene prediction performance by organ and condition
  - Figure 3: Cross-condition predictions using IF-only data
- **Poster Presentation** at BC Children's Hospital Research Day (April/May 2025)
- **Final Report or GitHub Repository** documenting methodology, results, and reproducibility guidelines

6. Learning Outcomes

- Gain experience in **deep learning, autoencoders, and spatial omics**
- Practice **reproducible research** using Git, GitHub, and Jupyter
- Exposure to **dataset curation** and ethical reuse of public research data
- Develop skills in **model evaluation, error analysis**, and **data visualization**
- Improve **science communication** via poster presentation and lab updates

7. Resources

- HE2RNA GitHub: https://github.com/owkin/HE2RNA_code
- Spatial Atlas: https://nanostring.com/products/spatial-organ-atlas
- Reference paper: https://www.nature.com/articles/s41467-020-17678-4