The Application of Graph Neural Networks to Drive Spatially Aware Molecular Discoveries in Cancer Transcriptomics Research

香 港 大 學
THE UNIVERSITY OF HONG KONG

LAIDLAW FOUNDATION

Yu Wing Tung, BSc (Bioinformatics), Year 3

Supervised by Professor Wong Wing Hon Jason

School of Biomedical Sciences, Faculty of Medicine, The University of Hong Kong

HKU Med LKS Faculty of Medicine The University of Hong Kong 香港大學李嘉誠醫學院

Introduction

- Spatial transcriptomics (ST) is a recent advancement that enables quantification of mRNA in pathology imaging slides while retaining spatial information. Latest breakthroughs in technologies that increased resolution of molecular profiling to single cell level prompted development of computational methods for contextualized data analysis.

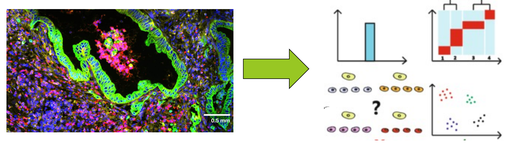

Fig 1. Single cell molecular imaging of formalin-fixed paraffin embedded (FFPE) or fresh tissue [1] and data analysis.

- Many deep learning architectures, particularly Graph Neural Networks (GNN) were developed to extract features with respect to the topology of cells, but their usage in clinical and translational research was minimal due to the current absence of appropriate downstream analytical frameworks.

- This project aims to build a computational pipeline to bridge the gap between methods and biological interpretation with three main objectives. First, to explore metrics to validate GNN discoveries; second, to link results to functional molecular characterization in the context of cancer; third, to streamline the data analysis workflow.

Materials and Methods

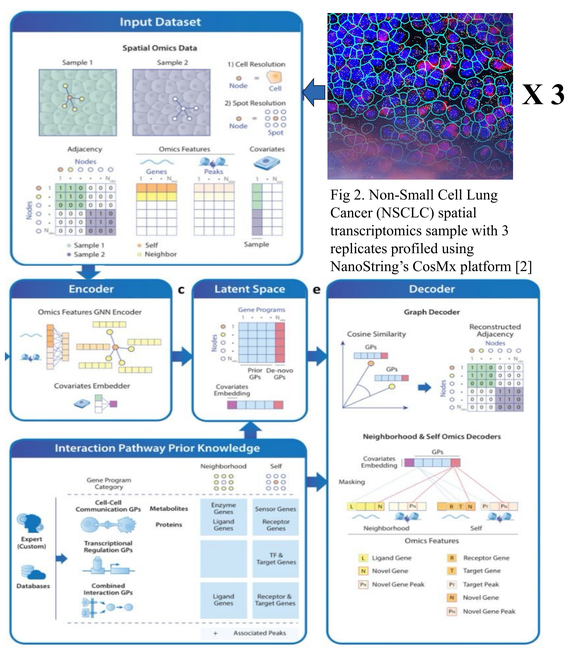

Fig 2. Non-Small Cell Lung Cancer (NSCLC) spatial transcriptomics sample with 3 replicates profiled using NanoString's CosMx platform [2]

Fig 3. Schematic overview of the architecture of NicheCompass [3], the GNN model applied in this study.

Results

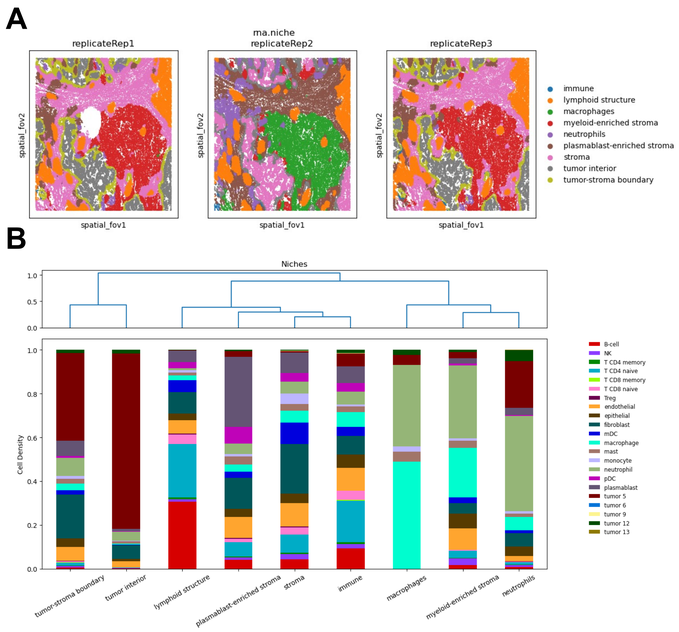

Fig 4. A) Official spatial domain annotation of CosMx NSCLC dataset consisted of immune-enriched, neutrophils-enriched lymphoid structure, macrophage-enriched, myeloid-enriched stroma, plasmablast-enriched stroma, stoma, tumour interior and tumour-stroma boundary spatial domains. B) Hierarchical clustering of spatial domains by cell type composition revealed ambiguity spatial domain labelling.

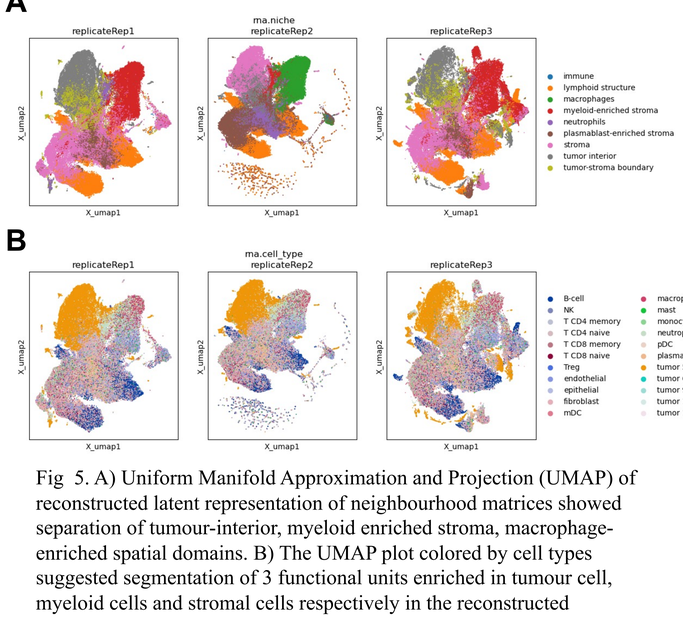

Fig 5. A) Uniform Manifold Approximation and Projection (UMAP) of reconstructed latent representation of neighbourhood matrices showed separation of tumour-interior, myeloid enriched stroma, macrophage-enriched spatial domains. B) The UMAP plot colored by cell types suggested segmentation of 3 functional units enriched in tumour cell, myeloid cells and stromal cells respectively in the reconstructed representation.

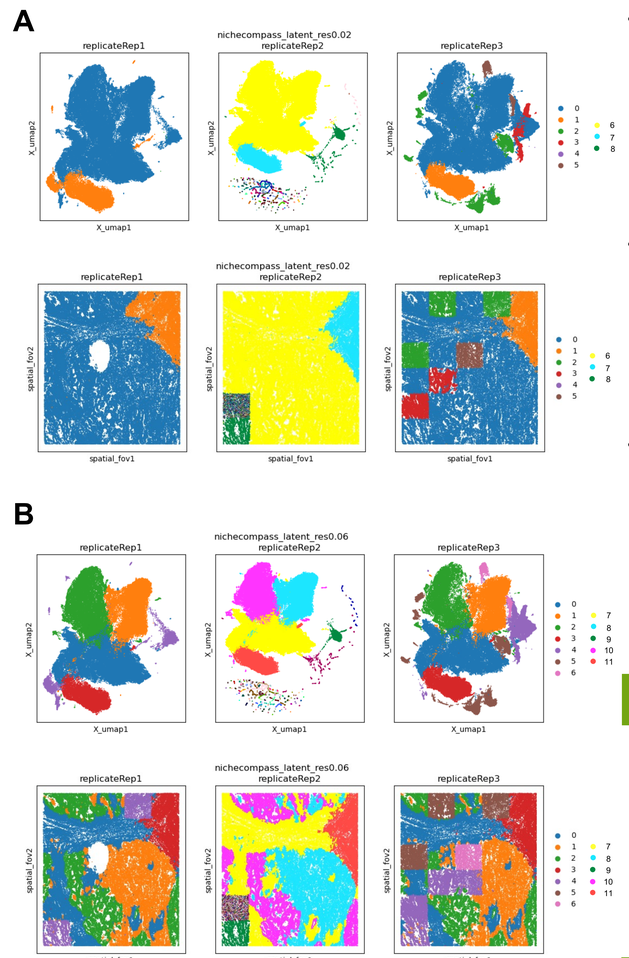

Fig 6. A) Top: UMAP of latent representation colored by Leiden clustering of reconstructed neighbourhood matrices. 8 major clusters were identified at Leiden clustering resolution = 0.02 where cluster 6 and 7 were replicate 2 specific. Bottom: Leiden clusters in physical coordinates revealed cluster 2, 5 and 8 were specific to field of view (FOV). B) Top: UMAP of latent representation colored by Leiden clustering of reconstructed neighbourhood matrices. 11 major clusters were identified at Leiden clustering resolution = 0.06 where cluster 7 to 11 were replicate 2 specific. Bottom: Leiden clusters in physical coordinates revealed cluster 4, 5, 6 and 9 were specific to field of view (FOV). Results indicated the model's inability to correct for batch effects across field of views despite incorporating FOV in the co-embedded space.

Discussions

- The analysis performed in this study illustrated the analytical challenge of applying deep learning methods in characterizing spatially dependent gene expression patterns at single cell level. In absence of anatomically defined regions for validation, the spatial domains identified by GNN reconstructed embedding followed by unsupervised clustering could only be arbitrarily labelled depending on the context of research.

- The latent feature space combining gene expression, covariates and spatial distribution also lacked direct interpretability due to the implicit feature extraction methods in deep learning models. Without canonical features to explain both gene expression and topology of the cells, results vary significantly based on hyperparameter tuning of the GNN model and the unsupervised clustering algorithm

- Due to the timeframe of the study and the lack of appropriate validation metrics, the data characteristics specific to the CosMx platform and how it interplays with the model performance remained largely unexamined. While the research is still on-going, the next step is to explore how factors such as data-preprocessing, spatial graph construction and hyperparameter-tuning would affect performance of the model and define suitable evaluation metrics for systematic benchmarking of the available GNN models.

Conclusion

The application of GNN models in application to spatial transcriptomics analysis is limited by the availability of validation datasets that effectively defines domains with respect to both their gene expression patterns and spatial distribution. Before establishing linkage to functional characterization of domains in diseases context, an analytical framework to benchmark and validate the results from GNN modeling should be established.

Acknowledgements

Throughout the summer research period, Professor Wong has provided immense support in guiding project development and providing valuable feedback for improvements. I am deeply grateful for his guidance and support through my research journey.

References

1. NanoString. What is single-cell imaging? [Internet]. NanoString; 2023 [cited 2024 Sept 28]. Available from: https://nanostring.com/blog/what-is-single-cell-imaging/
2. COSMX SMI NSCLC FFPE Dataset [Internet]. 2024 [cited 2024 Sept 16]. Available from: https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/nsclc-ffpe-dataset/
3. Birk S, Bonafonte-Pardàs I, Feriz AM, Boxall A, Agirre E, Memi F, et al. Quantitative characterization of cell niches in spatial atlases. BioRxiv 581428 [Preprint]. 2024 [cited Sept 16]. Available from https://www.biorxiv.org/content/10.1101/2024.02.21.581428v1