Development of a Deep Learning methodology for synthetic Spatial Metabolomics data generation

Background

Spatial metabolomics, based on mass spectrometry imaging (MSI), offers unique insights into tissue metabolite distributions. However, analyzing MSI data is challenging due to its high dimensionality, sparsity, and complex spatial dependencies. While spatial transcriptomics (ST) has benefited from deep learning to address similar challenges, MSI lacks comparable advances due to a scarcity of annotated datasets.

Project

This project is based on the hypothesis that distinct gene expression profiles at the spot level reflect distinct metabolic states, allowing us to predict corresponding metabolite quantifications. To do this we will use paired ST and MSI data from brain cancer samples (Heiland et al., 2022) to **develop a deep learning model predicting metabolite quantification values from spot-level gene expression**. In particular, we will use pre-processed ST data already partitioned into spatial clusters, each characterized by a unique gene expression signature (a defined set of genes).

We propose to employ a two-stage approach.

1. First, a model to predict the *presence/absence* of each metabolite value for each spot, generating a binary vector of length n, where n is the total number of metabolites $(m_1,..., m_n)$. We will explore suitable multi-label classification strategies for this m_i selection step. This can be achieved by training individual binary classifiers (e.g., logistic regression, SVM with one-vs-rest, or single-output neural networks with sigmoid activation) for each m_i value. We will also consider using a neural network with multiple sigmoid outputs, one for each m_i, allowing for simultaneous prediction of all metabolite presence/absence labels. The input for these models will be gene expression vector $(g_1,..., g_k)$ and a signature mask (1 for the signature genes and 0 for the rest) for each spot.
2. Second, a model to predict the *intensities of all metabolite values* (continuous outputs) for each spot based on the gene expression vector $(g_1,..., g_k)$. To address this regression problem, we propose to begin with baseline models (e.g. Linear and/or Ridge Regression, Multilayer Perceptron) to establish a performance benchmark before exploring more complex architectures such as Transformers. To handle the varying sets of selected m_i values across spots, the output (a vector of length n) is element-wise multiplied by the binary presence/absence vector from the first stage, masking the intensities of non-selected m_i values to zero. The input is the gene expression vector but can potentially be combined with positional encodings for m_i.

Model performance will be evaluated using appropriate metrics. For the metabolite selection task, we will use precision, recall, F1-score, and AUC. For the intensity prediction task, we will use mean squared error, R-squared, and Pearson correlation. Cross-validation will be used for model selection and hyperparameter tuning. Regularization techniques will be employed for deep learning models to prevent overfitting. Attention visualization will be explored to provide biological insights into the model's predictions.

Contacts:

Johanna Galvis deisy-johanna.galvis-rodriguez@u-bordeaux.fr

Christer Lohk christer.lohk@u-bordeaux.fr

Macha Nikolski macha.nikolski@u-bordeaux.fr

References:

Yilmaz, M., Fondrie, W.E., Bittremieux, W. *et al.* Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nat Commun* 15, 6427 (2024). https://doi.org/10.1038/s41467-024-49731-x

Pizurica, M., Zheng, Y., Carrillo-Perez, F. *et al.* Digital profiling of gene expression from histology images with linearized attention. *Nat Commun* 15, 9886 (2024). https://doi.org/10.1038/s41467-024-54182-5