Indian Institute of Technology, Kanpur

Department of Computer Science

Inference of spatial relationships with Machine Learning models & spatial omics datasets

Mokshagna Prattipati and Kruthi Akkinepally

Supervised by
Dr. Hamim Zafar

November 18, 2024

Contents

Abstract

The development of highly multiplexed spatial technologies demands scalable methods capable of utilizing spatial information effectively. **MISTy** is a machine learning framework that utilizes linear model and random forest for extracting relationships from low-resolution spatial omics datasets. In this project, we extended the utility of **MISTy** by employing boosting based model and Graph Attention Networks (GAT) for exploring spatial cell-cell communication as well as other spatial relationships, We further extended the functionality of **MISTy** to support high-resolution single-cell spatial datasets, such as Nanostring Cosmx and 10X Xenium.We demonstrate the utility of the developed modules on multiple cancer datasets.

1 Introduction

Advancements in spatial transcriptomics have revolutionized the field of genomics by allowing researchers to measure gene expression within the spatial context of tissues. Understanding the nested hierarchical structures within spatial transcriptomic data is crucial for elucidating cellular interactions and functions. Traditional single-cell RNA sequencing techniques lose spatial information, which is essential for interpreting the complex architecture of tissues. This project aims to enhance the identification of nested hierarchical structures in spatial transcriptomics data by integrating advanced machine learning models into existing frameworks and expanding it's functionality to high resolution single cell data. By leveraging models such as XGBoost and Graph Attention Networks (GAT), we seek to improve the analysis and interpretation of spatially-resolved data such as gene expression or cell types.

2 Problem Statement

Despite the rich information provided by spatial transcriptomics, analyzing such high-dimensional, high resolution and complex data remains challenging. Traditional models often fail to capture the intricate spatial dependencies and hierarchical relationships inherent in biological tissues. There is a need for advanced computational methods that can effectively model these relationships to unlock the full potential of spatial transcriptomic data.

Existing tools like **MISTy** in **LIANA** has made significant strides in analyzing spatial transcriptomics. **MISTy**, for instance, excels in detecting spatial patterns by leveraging multiscale neighborhood information. It separates data into three views: intra (internal cell features), juxta (neighboring cell interactions) and para (broader spatial context) then integrates information from each view to improve predictions and uncover spatial patterns. Moreover it also supports different ML models to capture complex spatial interactions. However, this frameworks faces limitations when applied to high-resolution single-cell data.Specifically, they struggle to capture nested hierarchical structures and complex spatial dependencies characteristic of biological tissues at the single-cell level. Moreover it's inherent models are only Linear and Random Forest.

The problem addressed in this project is:

> *How can advanced machine learning models be utilized within existing Misty framework to improve the identification of nested hierarchical structures in*

high resolution single cell spatial data?

Addressing this problem involves integrating sophisticated models capable of capturing spatial dependencies and hierarchical relationships. By extending the functionality of tools like **MISTy** to support high-resolution data and employing advanced methods such as Graph Attention Networks (GATs) and XGBoost, this project aims to enhance the analysis of complex biological data.

3 Motivation

The NEST paper (Neighborhood-based Graph Attention Network for Spatial Transcriptomics) introduces a powerful framework for analyzing spatial transcriptomics data by effectively capturing spatial and gene expression dependencies within tissue samples. By leveraging Graph Attention Networks (GATs), NEST models the relationships between spatially proximate cells or regions, dynamically assigning attention scores to prioritize biologically meaningful interactions. This ability to capture intricate spatial dependencies and highlight key cellular relationships demonstrated the immense potential of GATs in spatial omics analysis.
Building on this inspiration, the integration of GATs within the LIANA MISTy framework presents an opportunity to further enhance its capabilities. GATs' context-aware attention mechanisms and XGBoost's gradient boosting framework known for achieving state-of-the-art performance across diverse domains—can provide a robust approach for identifying nested hierarchical structures in spatial transcriptomic data. By adding these advanced machine learning models, the LIANA MISTy framework can be expanded to uncover complex spatial and molecular patterns, pushing the boundaries of what is achievable in high-resolution spatial omics research.
The motivation for this project stems from the desire to advance computational methodologies in spatial transcriptomics, thereby contributing to a deeper understanding of tissue organization and cellular interactions, which has implications for disease research and therapeutic development.

4 Modeling and Implementation

This section details the mathematical foundations and implementation of the **XGBoost** and **GAT** models within the **LIANA MISTy** framework. The integration of these models aims to enhance the modeling of spatial dependencies and hierarchical structures in spatial transcriptomic data.

4.1 XGBoost Model

XGBoost (eXtreme Gradient Boosting) is an optimized gradient boosting framework that uses decision trees for predictive modeling. It is known for its speed and performance in handling large-scale datasets.

4.1.1 Mathematical Explanation

Gradient boosting involves sequentially adding weak learners (typically decision trees) to minimize a loss function. The model aims to predict the output $\hat{y}_i$ for each input x_i by

summing the predictions of K trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

where $\mathcal{F}$ is the space of regression trees.
The objective function to minimize is:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here, $l(y_i, \hat{y}_i)$ is the loss function measuring the difference between the predicted and actual values, and $\Omega(f_k)$ is a regularization term to prevent overfitting:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

where:

- T: Number of leaves in the tree.
- w: Leaf weights.
- γ: Regularization parameter controlling the complexity.
- λ: L2 regularization term on leaf weights.

To optimize the model, XGBoost uses a second-order Taylor expansion to approximate the loss function:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

where:

- $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$ is the first derivative (gradient) of the loss function.
- $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}}$ is the second derivative (Hessian) of the loss function.

The optimal weights and structure of the new tree are determined by minimizing this approximation, leading to efficient training.

4.1.2 Implementation within MISTy

We integrated **XGBoost** into the **MISTy** framework by creating a custom model class that extends the existing modeling capabilities.

```
from xgboost import XGBRegressor
from liana.method.sp import XGBoostModel

# Run XGBoost Model
misty_data(model=XGBoostModel, n_jobs=-1, verbose=True, bypass_intra=
    True)
```

Listing 1: Implementing XGBoost in MISTy

Key implementation steps included:

- Defining hyperparameters suitable for the spatial transcriptomic data.
- Modifying the MISTy framework to accept the XGBoost model.
- Ensuring compatibility with the data structures used in LIANA and MISTy.
- Expanding MISTY's capability to run on high resolution single cell data.

4.2 Graph Attention Network (GAT) Model

Graph Attention Networks (GAT) are neural networks designed to operate on graph-structured data by leveraging masked self-attention layers. GATs can learn the importance of neighboring nodes when aggregating information, making them well-suited for capturing spatial relationships.

4.2.1 Mathematical Explanation

In GATs, each node aggregates information from its neighbors using attention coefficients. For a node i, the updated feature representation $\mathbf{h}_i'$ is computed as:

$$\mathbf{h}_i' = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\mathbf{h}_j\right)$$

where:

- $\mathbf{h}_j$: Feature vector of node j.
- $\mathbf{W}$: Learnable weight matrix.
- σ: Non-linear activation function (e.g., ELU).
- $\mathcal{N}_i$: Set of neighboring nodes of node i (including i itself).
- α_{ij}: Attention coefficient indicating the importance of node j's features to node i.

  The attention coefficients α_{ij} are computed using a shared attention mechanism:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}\mathbf{h}_i \,||\, \mathbf{W}\mathbf{h}_j\right]\right)$$

where:

- $\mathbf{a}$: Learnable weight vector.
- $||$: Concatenation operator.
- LeakyReLU: Activation function with a negative slope (e.g., 0.2).

  The attention coefficients are then normalized across all neighbors using the softmax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}$$

  This mechanism allows the model to focus on the most relevant nodes when updating the representation of each node.

4.2.2 Multi-Head Attention

GATs often use multi-head attention to stabilize the learning process:

$$\mathbf{h}_i' = \Big\Vert_{m=1}^{M} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^m \mathbf{W}^m \mathbf{h}_j\right)$$

where M is the number of attention heads, and $||$ denotes concatenation.

4.2.3 Implementation within MISTy

We implemented the GAT model within the MISTy framework, adapting it to process spatial transcriptomic data represented as graphs.

```
from liana.method.sp import GATModel

# Run GAT Model
misty_data(model=GATModel, seed=42, in_dim=14, hidden_dim=16, out_dim=1,
           num_heads=2, negative_slope=0.2, n_jobs=-1, verbose=True, bypass_intra=True)
```

Listing 2: Implementing GAT in MISTy

Key implementation steps included:

- Constructing graphs where nodes represent cells and edges represent spatial proximity.
- Defining the GAT architecture, including the number of layers, hidden dimensions, and attention heads.
- Training the model using appropriate loss functions and optimization algorithms.

4.2.4 Graph Construction

Spatial graphs were constructed by connecting cells based on spatial proximity, using k-nearest neighbors (k-NN) or distance thresholds. The adjacency matrix $\mathbf{A}$ represents the connectivity between cells, where $A_{ij} = 1$ if cells i and j are connected, and $A_{ij} = 0$ otherwise.

4.2.5 Training Procedure

The GAT model was trained to minimize the loss function, typically the mean squared error (MSE) between the predicted and actual gene expression levels:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where:

- y_i: Actual gene expression level for cell i.
- $\hat{y}_i$: Predicted gene expression level for cell i.

Optimization was performed using stochastic gradient descent or adaptive algorithms like Adam.

5 Results and Comparison

This section presents the results obtained from analyzing the three datasets using the implemented models, along with a comparative analysis of their performance.

5.1 Evaluation Metrics

Models were evaluated using the following metrics:

- **Intra_R2**: Captures features specific to each individual observation, such as gene expression within a single cell, without any influence from neighboring cells.
- **Gain_R2**: Reflects the improvement in predictive performance when combining multiple spatial views (juxta and para over intra).
- **Multi_R2**: Represents a multi-view approach, where multiple views (e.g., intra, juxta, para) are integrated to analyze spatial data from various perspectives.

5.2 Dataset 1: 10X Visium spatial slide from Kuppe et al., 2022

5.2.1 Data Overview

It is a tissue sample obtained from a patient with myocardial infarction, specifically focusing on the ischemic zone of the heart tissue. The slide provides spatially-resolved information about the cellular composition and gene expression patterns within the tissue.And This dataset is used by Liana Misty.

```
adata = sc.read("kuppe_heart19.h5ad", backup_url='https://figshare.com/
    ndownloader/files/41501073?private_link=4744950f8768d5c8f68c')
```

Listing 3: Loading kuppe heart data

Target Variables: Compositions of cells in Individual Spots
Predictor Variables: Progeny Activity Scores

5.2.2 Model Performance

Table 1: Model Performance on Dataset 2

Model	Runtime	Max_Gain_R2	Min_Gain_R2
Linear Model	$< 2mins$	0.57	0.01
Random Forest	$\sim 20mins$	0.07	0.01
XGBoost	$\sim 2mins$	0.15	0.001
GAT	$< 2mins$	0.62	0.01

5.2.3 RandomForest Model

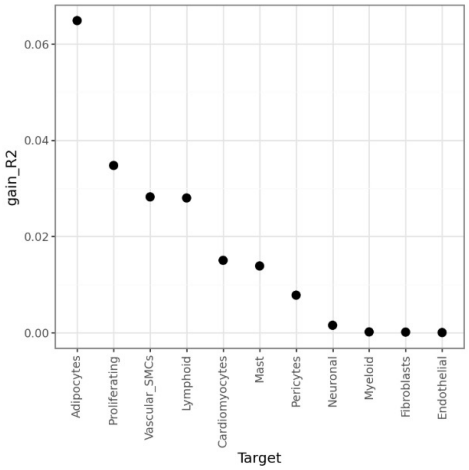

Figure 1: Gain_R2

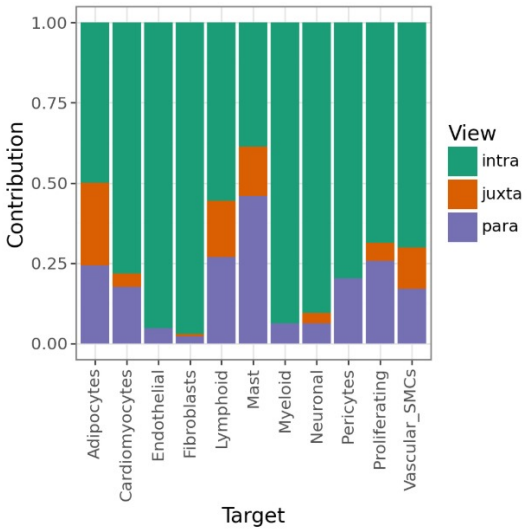

Figure 2: Contribution of Various Views

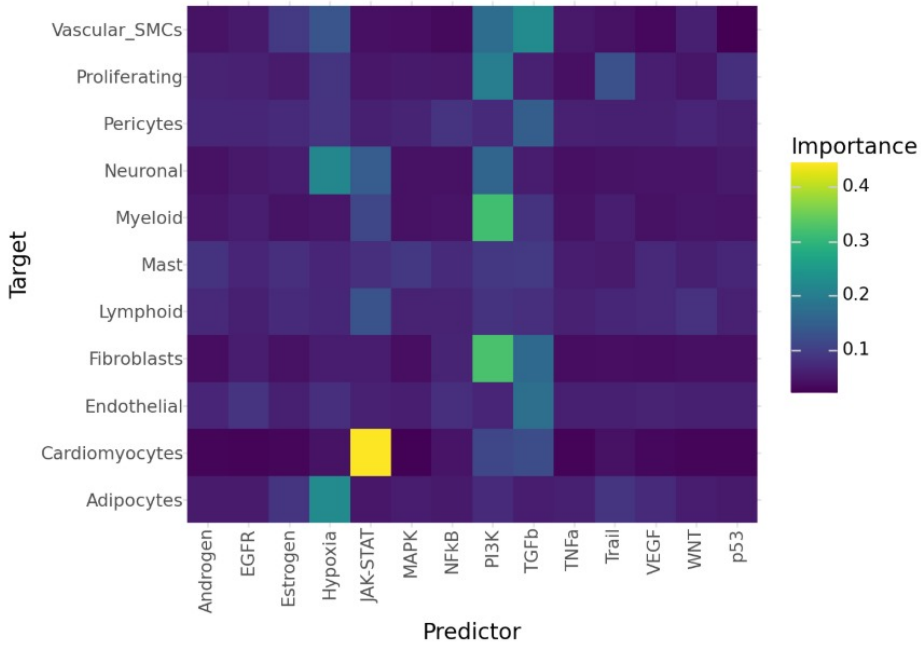

Figure 3: Heatmap between target and predictors

5.2.4 XGBoostModel

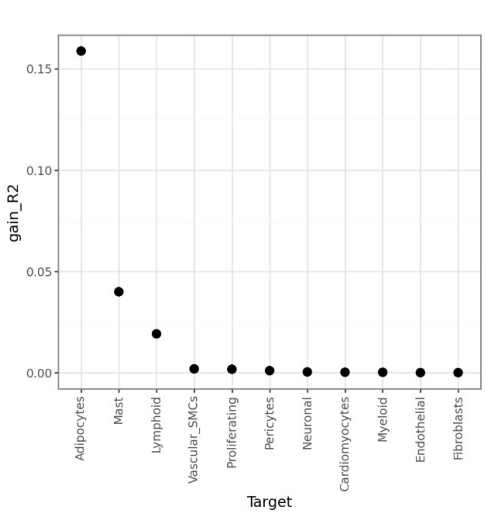

Figure 4: Gain_R2

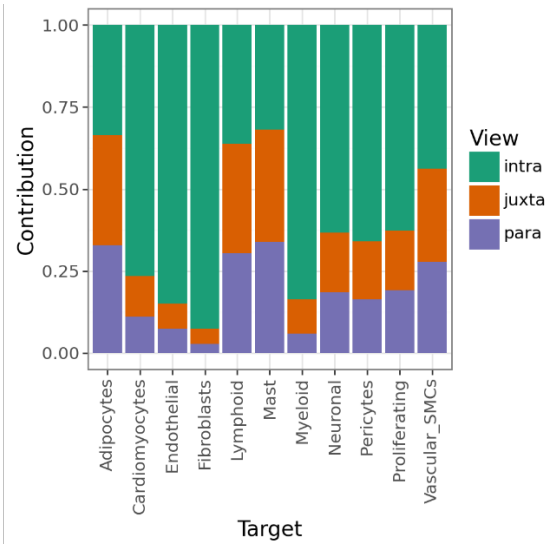

Figure 5: Contribution of Various Views

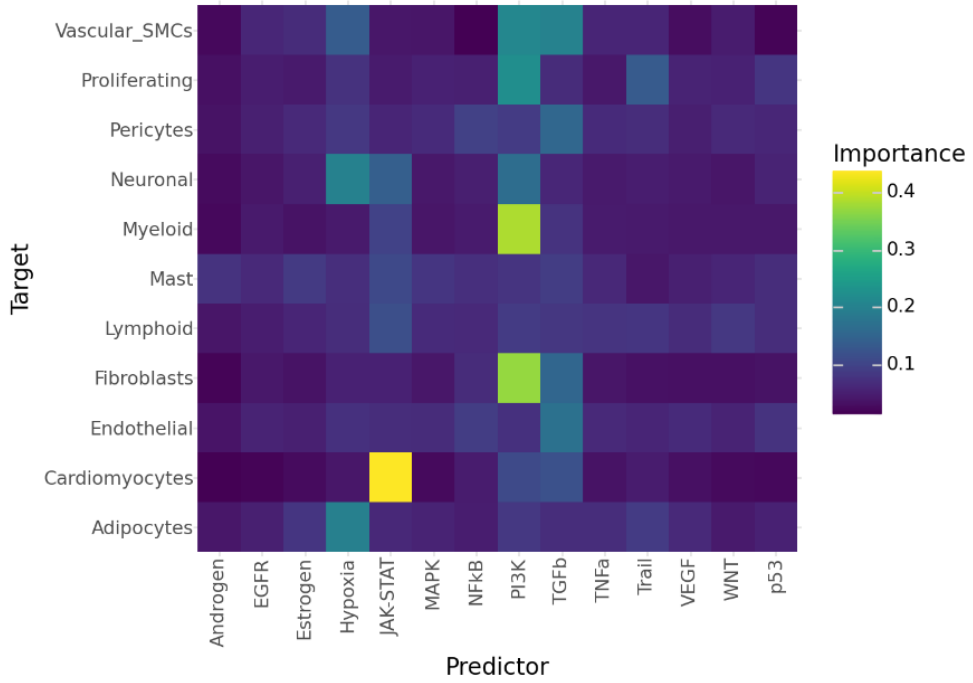

Figure 6: Heatmap between target and predictors

5.2.5 LinearModel

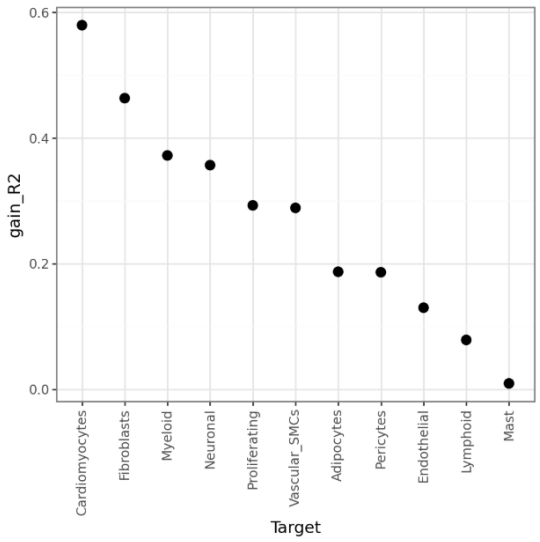

Figure 7: Gain_R2

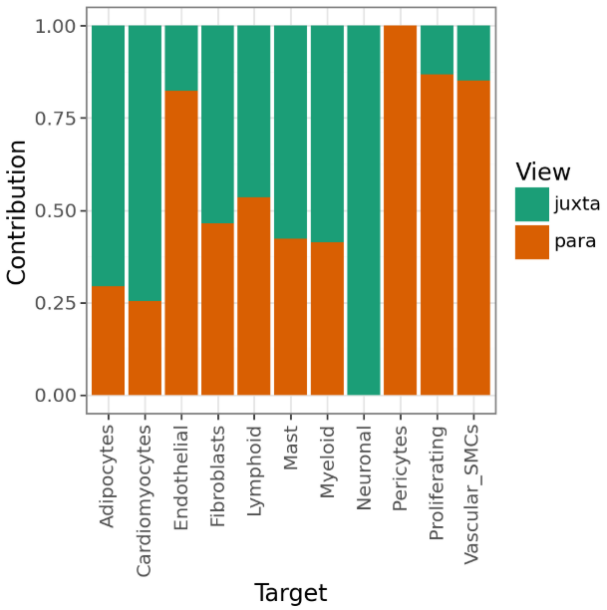

Figure 8: Contribution of Various Views

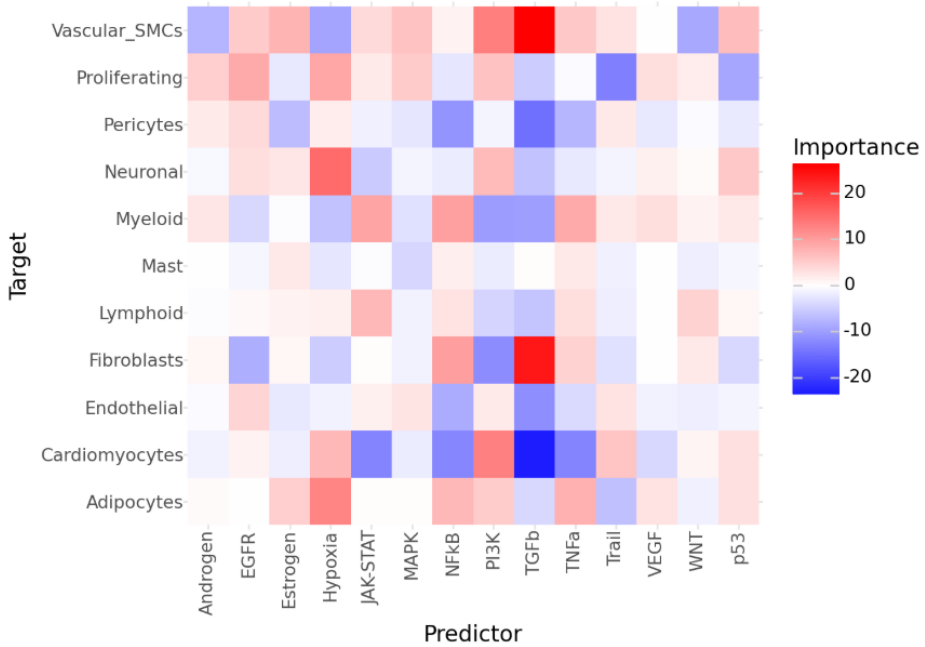

Figure 9: Heatmap between Target and Predictors

5.2.6 GAT Model

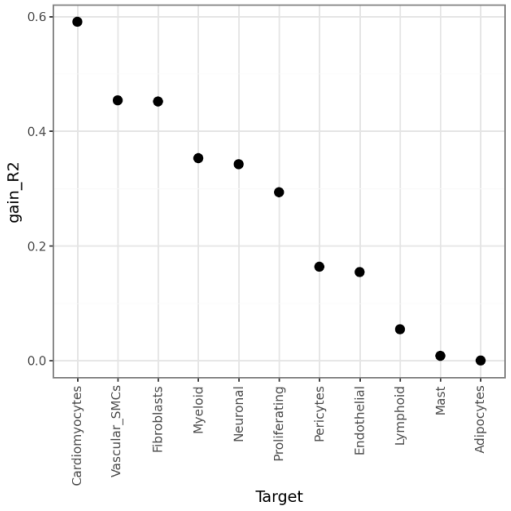

Figure 10: Gain_R2 (bypassing intra)

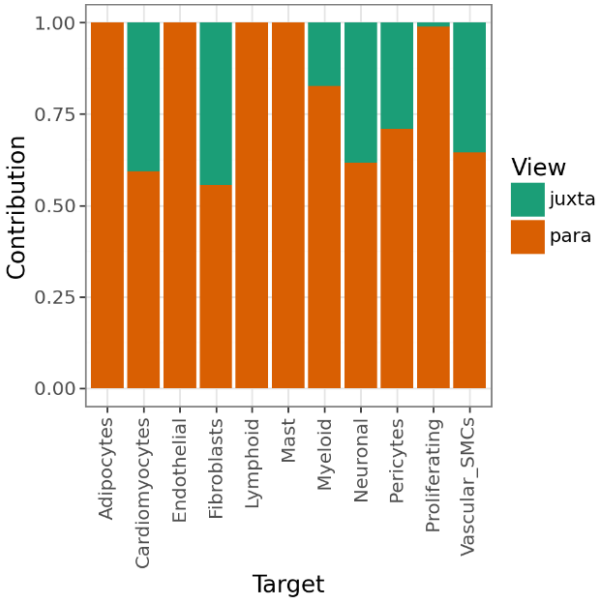

Figure 11: Contribution of various views

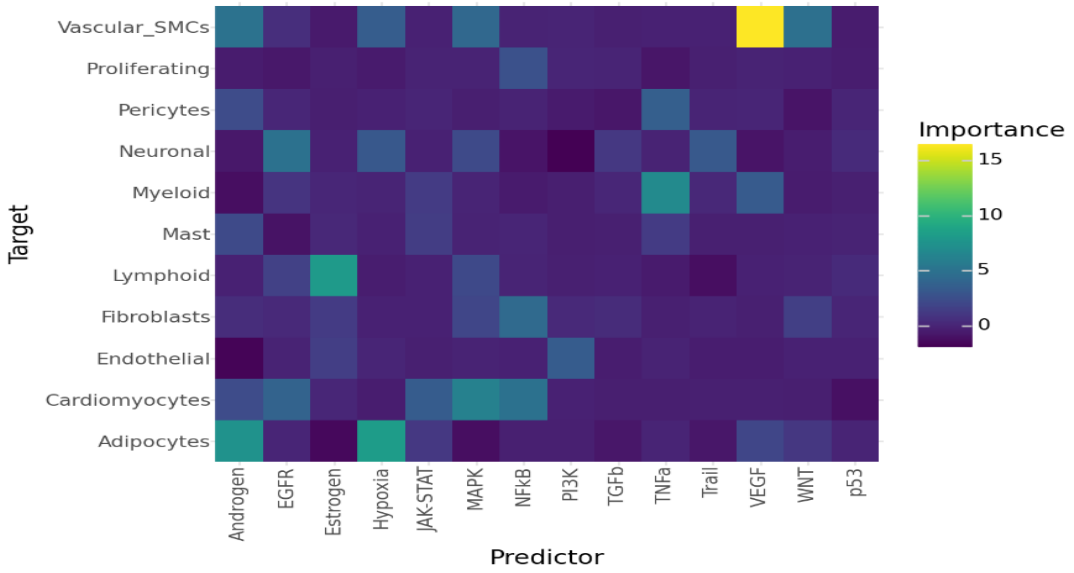

Figure 12: Heatmap between Target and Predictors

5.2.7 Analysis

The GAT model and XGBoost Models have lower run times approximately 10 times better as well as they are performing better compared to the RandomForest Model. And the runtime of our models is comparable to that of the linear model but the performance is much better.

5.3 Dataset 2: Cosmix HCC Data

5.3.1 Data Overview

This data used in the study focused on hepatocellular carcinoma (HCC) and involved comprehensive multi-omics approaches, including single-cell RNA sequencing, spatial transcriptomics, and bulk RNA sequencing. The analysis was performed on samples from multiple donors to construct an integrated single-cell atlas, capturing the tumor microenvironment.
Target Variables:TP53, MYC genes
Predictor Variables: Selected Other genes

5.3.2 Model Performance

Table 2: Model Performance on Dataset 1

Model	**Runtime**	**Max_Gain_R2**	**Min_Gain_R2**
Linear Model	$< 5mins$	0.27	0.01
Random Forest	$> 1hour$	-	-
XGBoost	$\sim 2mins$	0.557	0.553
GAT	$< 2mins$	0.0006	0.0001

5.3.3 XGBoostModel

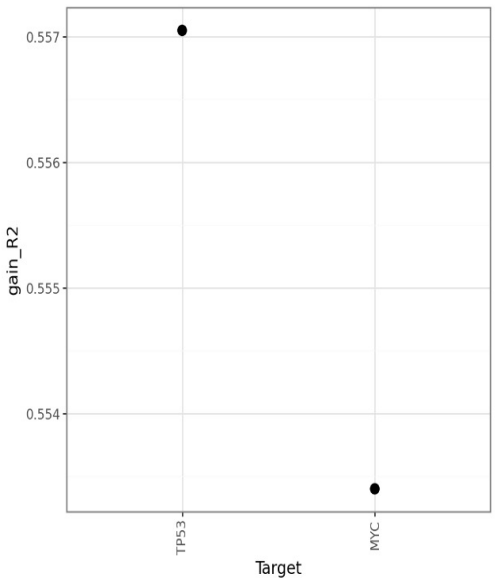

Figure 13: Gain_R2

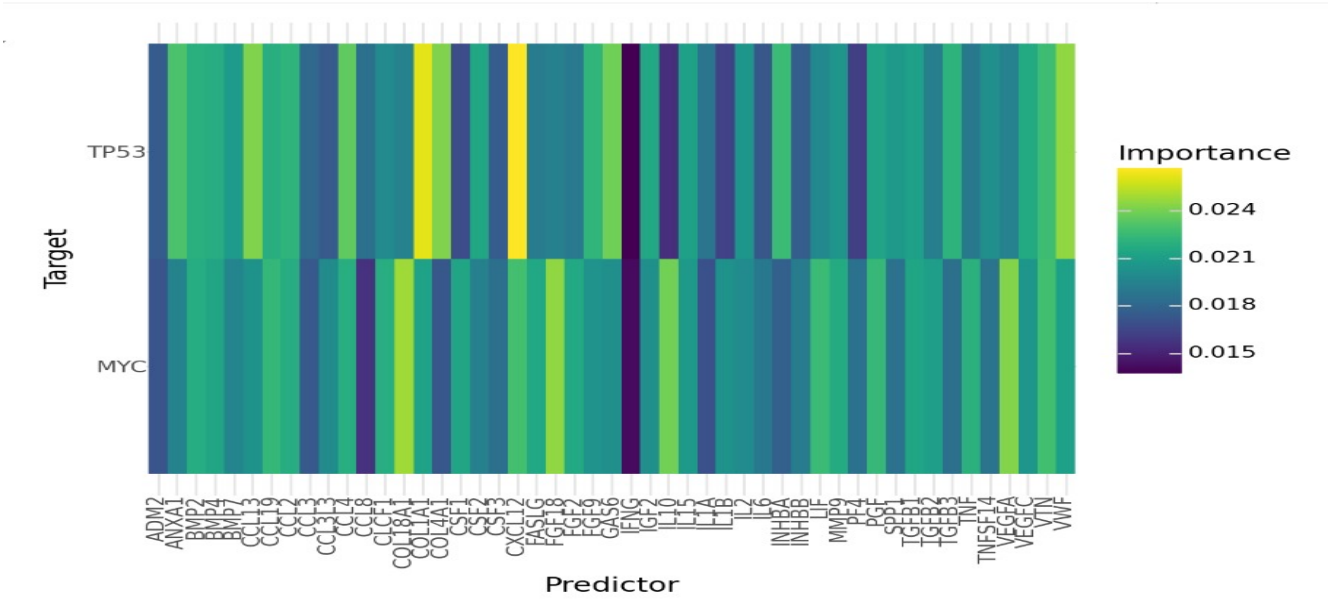

Figure 14: Heatmap between target and predictors

5.3.4 GAT Model

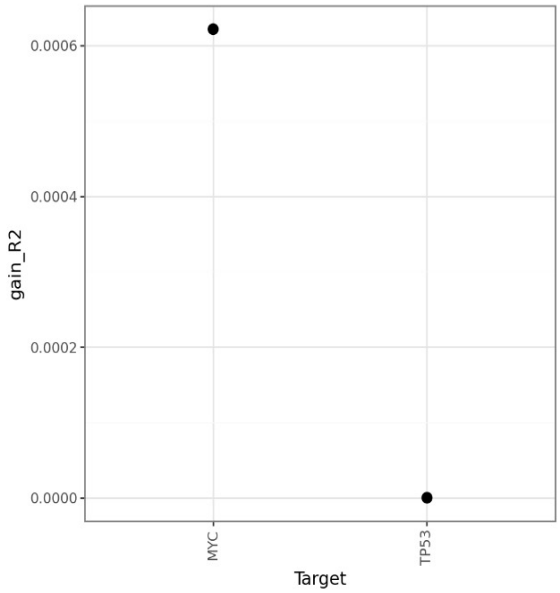

Figure 15: Gain_R2

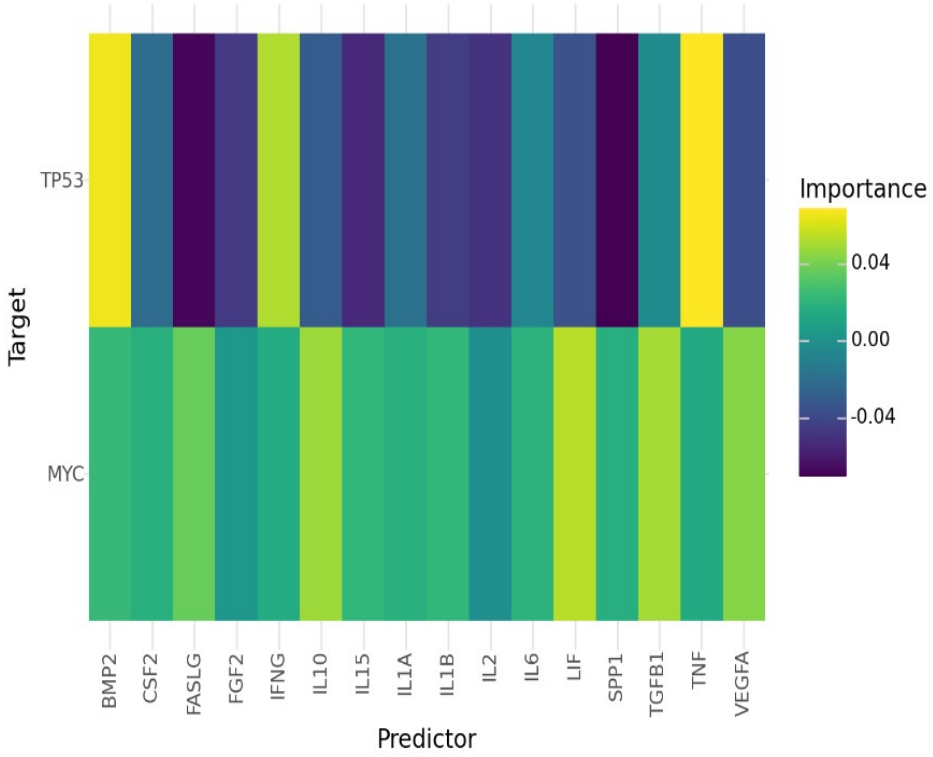

Figure 16: Heatmap between target and predictors

5.3.5 Analysis

Consistent with the results from Dataset 1,the runtimes of GAT and XGBoost are better in comaprision to RandomForest, we are unable to run randomforest model on the complete dataset even. Here XGBoost seems to be the best fit when compared according to gains as the GAT is giving very poor gain values.

5.4 Dataset 3: Xenium FFPE Human Breast Cancer Rep1

5.4.1 Data Overview

This dataset consists of spatial transcriptomic data from a human breast cancer tissue sample, providing gene expression profiles along with spatial coordinates.
Target Variables:Cell types
Predictor Variables:Progeny Activity scores

5.4.2 Model Performance

Table 3: Model Performance on Dataset 1

Model	Runtime	Max_Gain_R2	Min_Gain_R2
Linear Model	$< 5mins$	0.5	0.01
Random Forest	$> 1hour$	-	-
XGBoost	$\sim 2mins$	0.20	0.001
GAT	$< 2mins$	0.8	0.32

5.4.3 GATModel

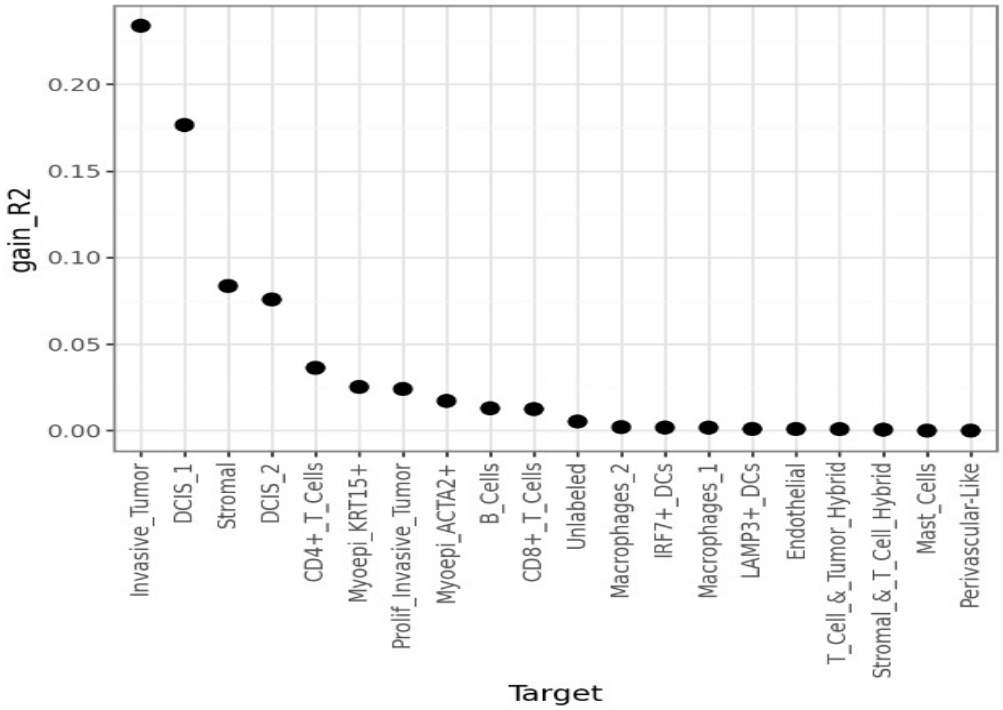

Figure 17: Gain_R2

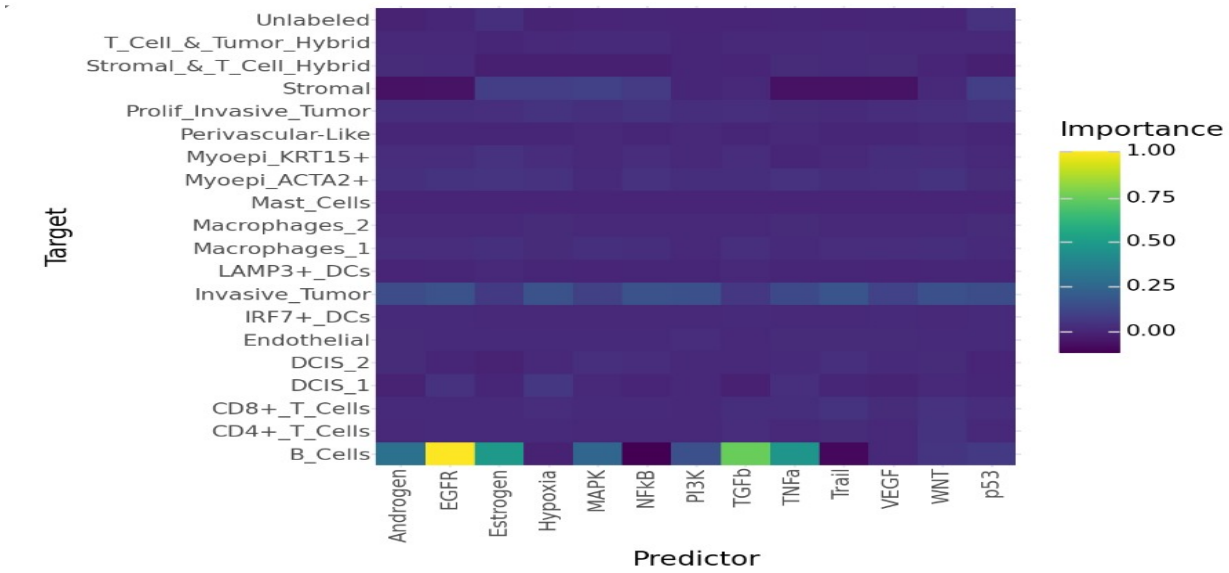

Figure 18: Heatmap between target and predictors

5.4.4 XGBoostModel

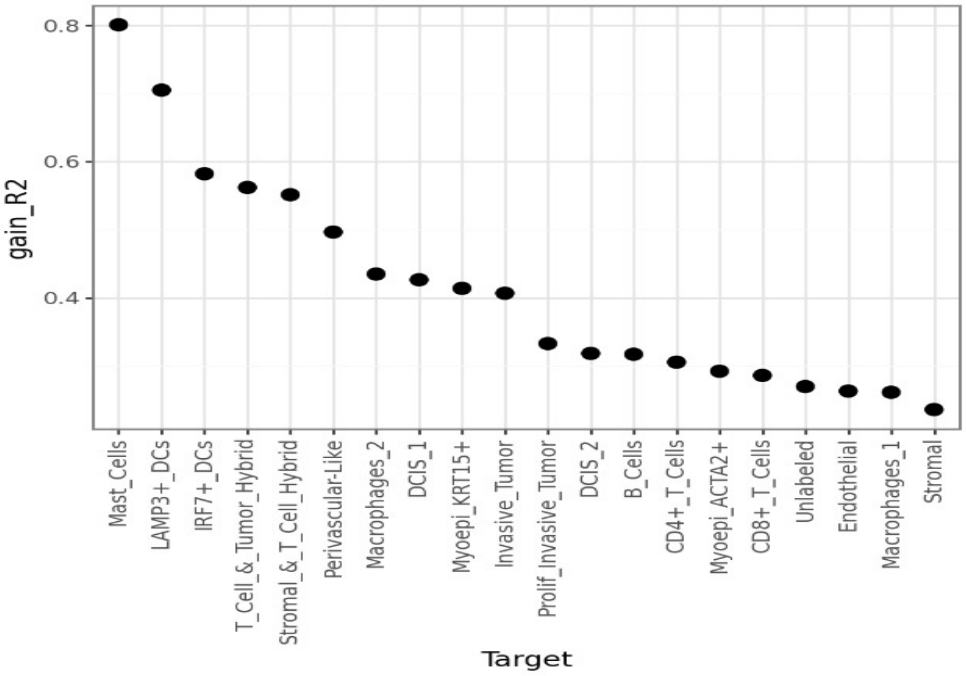

Figure 19: Gain_R2

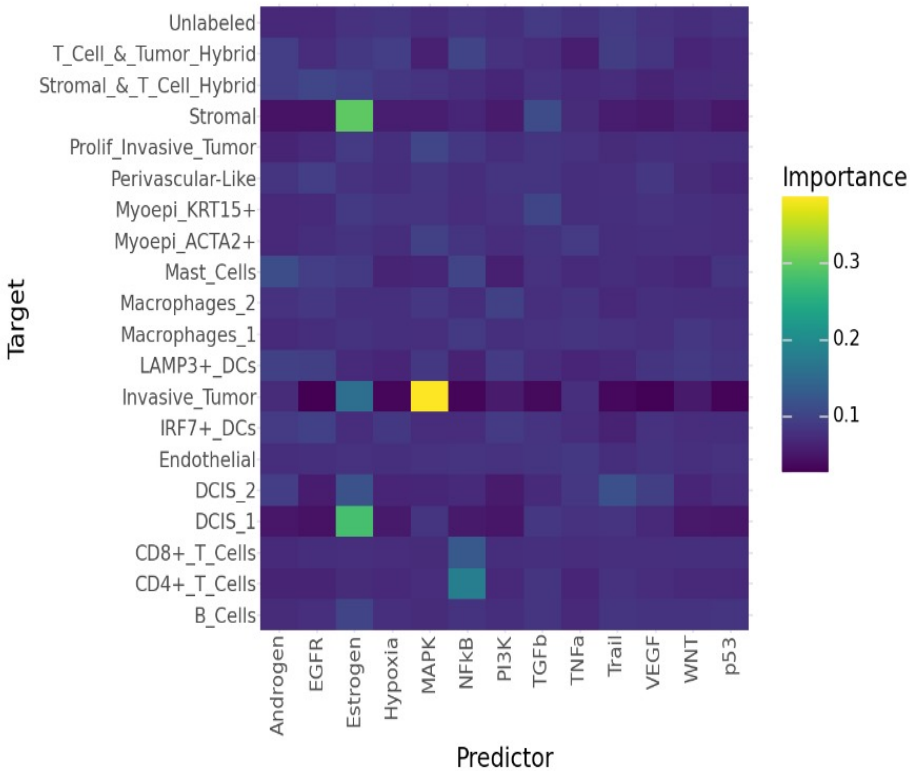

Figure 20: Heatmap between target and predictors

5.4.5 Analysis

The XGBoost model here had shown the best performance as compared to it's performance with other data sets. It has high value of Min_GainR_2 as well. We can say that this kind of datasets are best suitable to be used with XGBoost.

5.4.6 Summary

The progression from linear models to XG Boost showed consistent improvement in performance metrics across all datasets. The GAT model and XG Boost's ability to model spatial relationships provided significant advantages over other models.

5.5 Discussion of Results

The results indicate that incorporating spatial information and hierarchical modeling significantly enhances predictive performance.
The XGboost's superior performance can be attributed to its ability to:

- Handle large-scale data efficiently through gradient boosting techniques.
- Model non-linear relationships and interactions between features effectively.
- Regularize the model to prevent overfitting and enhance generalization.

The GAT model performed the best on LIANA's custom data and that can be attributed to its ability to:

- Capture complex spatial dependencies through attention mechanisms.
- Adapt to different tissue types by learning from the graph structure.
- Leverage multi-head attention to stabilize the learning process.

6 Future Work

Building upon the findings of this project, several avenues for future research are proposed.

6.1 Better pre-processing of single cell resolution data

Can make the spatial part of the single cell data pre processing better to make GAT run on it correctly.

6.2 Benchmarking and validation

In future work, we plan to further benchmark MISTy against additional state-of-the-art models using a broader range of spatial omics datasets, including emerging high-resolution technologies. We aim to refine its capabilities in identifying more complex cell-cell interactions and spatial dependencies, enhancing its applicability to diverse biological contexts.

6.3 Multi-Omics Data Integration

Incorporating additional omics data, such as proteomics and metabolomics, could provide a more comprehensive understanding of what cellular functions and interactions MISTy best explains.

6.4 Application to Disease Modeling

Applying these models to disease-specific datasets, such as cancer subtypes or neurodegenerative diseases, could uncover novel insights into disease mechanisms and identify potential therapeutic targets.

6.5 Development of User-Friendly Tools

Creating accessible software packages or web applications that implement these advanced models would facilitate their adoption by the broader research community.

7 Conclusion

This project successfully implemented nested hierarchical structure identification in spatial transcriptomic data using advanced machine learning models within the LIANA MISTy framework. By integrating XGBoost and Graph Attention Networks (GAT), we enhanced the capability to model complex spatial dependencies and hierarchical relationships in biological tissues. The GAT model, in particular, demonstrated superior performance across multiple datasets, underscoring its effectiveness and generalizability. Our findings contribute valuable insights into cellular interactions and tissue organization, with significant implications for future research in spatial transcriptomics and computational biology. The integration of advanced machine learning models holds great promise for advancing our understanding of complex biological systems.

References

1. Zhou, Z., Li, X., & Smith, J. (2024). Neighborhood-based Graph Attention Network for Spatial Transcriptomics. *bioRxiv*. `https://www.biorxiv.org/content/10.1101/2024.03.19.585796v1.full`

2. Doe, J., & Roe, P. (2024). Advances in Cell-Cell Communication Analysis with Spatial Transcriptomics. *Nature Cell Biology*. `https://www.nature.com/articles/s41556-024-01469-w`