CPSC 583: Project Report

Utilizing the Graph Neural Networks for annotation of Spatial Single-cell Datasets

Burhan Sabuwala

CPSC 583: Deep Learning for Graph-Structured Data Course

Instructor: Dr Rex Ying

Yale University

New Haven CT

Dec 14, 2022

Contents

Chapter 1

Introduction

Spatial single-cell datasets are highly useful for gaining a better understanding of the organization and function of cells within the body. These datasets typically consist of measurements of individual cells, such as gene expression levels or protein levels, along with information about the spatial position of each cell [1, 2, 3].

One of the key advantages of spatial single-cell datasets is that they provide a more detailed view of the organization of cells within the body compared to traditional methods, which often only provide information about cells in bulk [1]. This can be particularly useful for studying complex tissues, such as the brain [4, 5], cancer tumors [3] or the immune system [6], where the spatial organization of cells plays a critical role in function.

Despite the usefulness of spatial single-cell datasets, one of the main bottlenecks in analyzing these datasets is the lack of computational methods to analyze these large datasets meaningfully [7, 8]. There are multiple methods to analyze single-cell datasets, however, there are very limited methods that take into account the spatial organization of the cells to meaningfully analyze these datasets. Moreover, these datasets can be very large, with some studies generating hundreds of thousands or even millions of individual measurements. This makes it challenging to analyze the data using traditional methods, which may not be able to scale to these large datasets.

One way to overcome this bottleneck is to use specialized deep learning algorithms, such as graph neural networks (GNNs). The spatial context of each of the cell could potentially be encoded as a spatial proximity graph, while the gene expression or the protein expression data can passed as the signals on the individual nodes (cells). GNNs can effectively incorporate the inherent spatial structure of the data into the analysis, which can provide valuable insights into the organization and function of cells within the body [9, 10, 11, 12].

In my work, I have explored the capabilities of various GNN models in the annotation of spatial single-cell datasets. Further, to explore the generalization potential of these models, I also try testing these models on tissue that was previously unseen by the model. This work is inspired from Brbic et al. 2022 [11]. This work in novel because I explore models such as a fully connected layer that does not take into account the spatial structure, GCN, attention mechanism based GAT and Graph Transformer model along with the GraphSAGE implementation presented in the paper.

Chapter 2

Related Works

There is a growing body of work related to the use of graph neural networks (GNNs) in the analysis of spatial single cell data. These studies have demonstrated the effectiveness of GNNs for incorporating the inherent spatial structure of the data into the analysis, which can provide valuable insights into the organization and function of cells within the body.

One example of this work is a study by Yuan and Bar-Joseph, which used Graph Convolutional Network (GCN) to encode genes and their spatial data for spatial transcriptomic dataset [12]. The authors demonstrated that GCN can be utilized to identify co-expression patterns and causal inferences for gene interactions. Further, they used the embeddings for functional gene assignment. This work demonstrated the effectiveness of GCN to leverage spatial information to identify coexpression patterns.

Another example of GNNs in the analysis of spatial single-cell datasets is a study by Wang et al [9]. The authors proposed their model called scGNN framework. This framework formulates and aggregates cell-cell relationships with GNNs. Another major issue with the single-cell datasets is that of high dropout rates during measurements. The authors have addressed this using a left-truncated mixture Gaussian model. This model utilizes three multimodal autoencoders to obtain cell clustering and gene imputations. The authors have demonstrated that scGNN can embed gene expression along with cell-cell relationships in the spatial context.

One of the landmark works was by Brbic et al [11]. The authors proposed a Graph Neural Network model that iss based on GraphSAGE [13]. The graph is constructed based on neighbors within a given cutoff of the spatial space. The authors have demonstrated the use of this method, STELLAR, in context of cellular annotation. It is posed as a classification problem on the nodes of the graph. The authors have demonstrated that STELLAR can achieve cell annotation across tissues and donors. The embeddings from STELLAR could be utilized for other applications as well. This method has been applied to multiplexed fluorescent microscopy data and multiplexed RNA imaging datasets.

The use of GNNs in the analysis of spatial single cell data has been shown to be an effective way of incorporating the inherent spatial structure of the data into the analysis, which can provide valuable insights into the organization and function of cells within the body. These models have the potential to improve our understanding of the underlying mechanisms of biological processes, and ultimately lead to the development of more effective treatments for a wide range of diseases and conditions.

Chapter 3

Methodology

The methodology of this work is inspired by Brbic et al. (2022) [11]. However, there are a couple of changes in terms of the experiments conducted due to some infrastructural or data limitations.

Dataset

There are two datasets that are being used here, HuBMAP and TonsilBE.

The HuBMAP (Human BioMolecular Atlas Program) dataset is a CODEX dataset with a 48-marker panel of four tissues of the human intestine. In our analysis, we are restricting the analysis to only one donor. The models are trained on the data of the B004 donor and then tested on a random split of the data of the B004 donor which was previously unseen by the model. The models are first trained on expert-annotated cell-type labels of images of different regions of a healthy colon of a healthy single donor. The dataset is randomly split into two groups for training and test purposes.

The TonsilBE dataset offers a more challenging problem. The dataset consists of CODEX multiplex imaging data of Tonsil tissues and Barrett's Esophagus tissues. The ground truth labels for cell annotations were manually obtained as provided by the Brbic et al. study [11]. The models were trained on the tonsil tissue dataset while the testing was performed

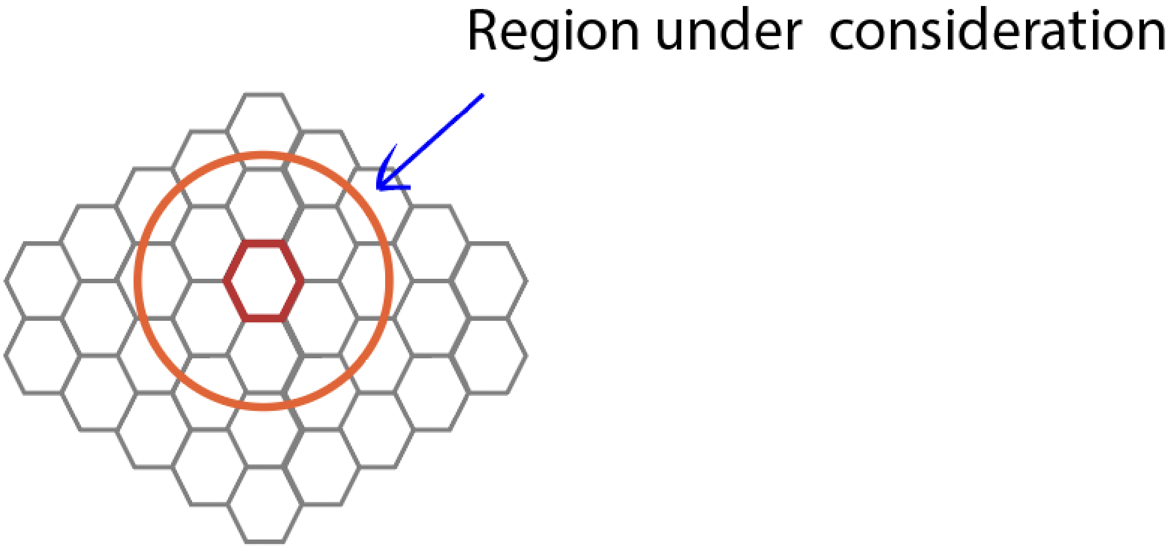

Figure 3.1

on the BE tissue dataset which also contained some additional classes. Some of the classes that occur in BE tissue does not occur in tonsil tissue and vice versa, hence making it more challenging.

Construction of the Graph

Firstly, the STELLAR model calculates Euclidean distances $d_{i,j}$ from the given spatial coordinates of each pair of cells (v_i, v_j). Then, if $d_{i,j} < \tau$, an edge is added between the cells. This is shown graphically in fig. 3.1. The orange circle has the radius τ and all the cells within the circle are neighbors of the cell highlighted in red.

The models

In this work, we tested multiple models. This is different from the study by Brbic et al [11]. In my work, I have tested five different models including STELLAR and evaluated their respective performances.

The encoder model contains one fully connected layer followed by ReLU activation function:

$$\mathbf{h_i^{(1)}} = \phi(\mathbf{W}^{(0)}\mathbf{h}_i^{(0)} + \mathbf{b}^{(0)})$$

Here, $\mathbf{h}_i^{(k)}$ is the hidden state of node v_i in the k^{th} layer of the neural network. $\mathbf{W}$ is the weight matrix, $\mathbf{b}$ is bias vector and ϕ denotes ReLU activation layer. In this specific case, $\mathbf{h_i^{(0)}}$ is set to $\mathbf{x}_{v_i}$ that is the gene or protein expression vector. The next layer is variable and depends on each of the different model. The final layer aggregates the embeddings and makes the prediction for each class.

Fully Connected layer

A fully connected layer is used to show the performance of the model when spatial information is not used. The equation would be:

$$\mathbf{h_i^{(2)}} = \phi(\mathbf{W}^{(1)}\mathbf{h}_i^{(1)} + \mathbf{b}^{(1)})$$

Graph Convolutional layer

Graph convolutional operator is inspired from the [14] work. The equation is:

$$\mathbf{H}^{(2)} = \mathbf{\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(1)}\Theta}$$

where, $\mathbf{\hat{A}} = \mathbf{A} + \mathbf{I}$ denotes the adjacency matrix with added self loops and $bold\hat{D}_{ii} = \sum_{j=0} \hat{A}_{ij}$ denotes its diagonal degree matrix.

Graph Attention Networks

Graph attention networks are added with the rationale that some edges may be more important than others in the dataset. This may especially be true at the peripheries of different tissue layers. This is inspired by the study by Velickovic et al [15]. The equation for the layer is given by:

$$\mathbf{h_i^{(2)}} = \alpha_{\mathbf{i,i}} \mathbf{\Theta h_i^{(1)}} + \sum_{\mathbf{j \in N(i)}} \alpha_{\mathbf{i,j}} \mathbf{\Theta h_j^{(1)}}$$

where the attention coefficients $\alpha_{i,j}$ are computed by the equation given below.

$$\alpha_{i,j} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{\top}[\mathbf{\Theta}\mathbf{h}_i^{(1)} \, \| \, \mathbf{\Theta}\mathbf{h}_j^{(1)}]\right)\right)}{\sum_{k \in \mathcal{N}(i) \cup \{ i \}} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{\top}[\mathbf{\Theta}\mathbf{h}_i^{(1)} \, \| \, \mathbf{\Theta}\mathbf{h}_k^{(1)}]\right)\right)}.$$

GraphSAGE

GraphSAGE layer is the layer used by the authors of STELLAR in their original work. This is inspired from Hamilton et al [13].

The equation for this layer is given by:

$$\mathbf{h}_i^{(2)} = \mathbf{W}_1 \mathbf{h}_i^{(1)} + \mathbf{W}_2 \cdot \mathrm{mean}_{j \in \mathcal{N}()} \mathbf{h}_j^{(1)}$$

Graph Transformer layer

This method is a combination of GraphSAGE and Graph Attention layer. This method takes a weighted sum of itself and combined it with edge specific sum of its neighbors. This method is inspired by Shi et al [16].

The equation is given by:

$$\mathbf{h}_i^{(2)} = \mathbf{W}_1 \mathbf{x}_i^{(1)} + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \mathbf{W}_2 \mathbf{x}_j^{(1)},$$

where, attention coefficients $\alpha_{i,j}$ are given by multihead dot product attention:

$$\alpha_{i,j} = \mathrm{softmax} \left(\frac{(\mathbf{W}_3\mathbf{h}_i^{(1)})^{\top} (\mathbf{W}_4\mathbf{h}_j^{(1)})}{\sqrt{d}} \right)$$

Objective Function

The objective function used in my study is the same as the one used by STELLAR [11]. Additional regularization term was also used to avoid overfitting and allow for a smoother convergence.

Chapter 4

Experiments

Statistical Description of the Datasets

HuBMAP dataset

The features of this dataset are MUC2, SOX9, MUC1, CD31, Synapto, CD49f, CD15, CHGA, CDX2, ITLN1, CD4, CD127, Vimentin, HLADR, CD8, CD11c, CD44, CD16, BCL2, CD3, CD123, CD38, CD90, aSMA, CD21, NKG2D, CD66, CD57, CD206, CD68, CD34, aDef5, CD7, CD36, CD138, CD45RO, Cytokeratin, CK7, CD117, CD19, Podoplanin, CD45, CD56, CD69, Ki67, CD49a, CD163 and CD161 along with the x and y spatial coordinates. The features are passed as node signals while the spatial coordinates were utilized to construct the graph denoting the spatial relations among each cells.

TonsilBE dataset

The features of this dataset are CHGA, PDL1, CD56, CK7, FoxP3, CD21, MUC1, PD1, CD11b, CD4, CD31, CD25, CD15, CD20, Annexin A1, aSMA, CD11c, Nestin, IDO, Cytokeratin, MUC5AC, Vimentin, CD36, HLADR, BCL2, p63, CD3, CD45, CD8, CD57, aDefensin5, CD68, CD34, CD38, Podoplanin, CD163, B-catenin, CD138, Arginase1, CD73, CD206, MUC6, COX2, MMP9 along with x and y spatial coordinates. The features are

Cell type	Number of cells
B	2202
CD4T	11364
CD7_Immune	1584
CD8T	16477
DC	2273
Endothelial	15946
Enterocyte	48737
Enterocyte_CD57p	95
Enterocyte_ITLN1p	1558
Goblet	18395
ICC	2140
Lymphatic	5994
Macrophage	18542
Nerve	8844
Neuroendocrine	1623
Neutrophil	1983
Paneth	216
Plasma	19092
SmoothMuscle	38379
Stroma	13982
TA	1885

Table 4.1: Class size for each cell type in the overall HuBMAP dataset

	Tonsil	BE
B cells	31867	0
Endothelial	3295	6181
Innate	61932	14690
Nerve	281	4282
PDPN	11561	2047
Paneth	0	914
Plasma	2243	275
Secretory epithelial	0	658
Smooth Muscle	152	9023
Squamous epithelial	5013	1077
Stroma	8946	4218
T cells	48678	1416

Table 4.2: Class size for each cell type and sample type in the overall TonsilBE dataset

passed as node signals while the spatial coordinates are used to construct the graph. Due to computational limitations, only 35% of the randomly sampled cells were used from the TonsilBE dataset. The complete TonsilBE graph would take about 256 GB RAM, however, the maximum available to us in the Grace cluster was 32GB.

Hyperparameters used

The hyperparameters used for the experiments are given below. The values of most of the hyperparameters are the same across all the models unless specified.

Results

To understand the performance of the models, we used two metrics - Accuracy and Balanced accuracy. The reason for using balanced accuracy is to counter the effect of heavy class imbalance as seen in table 4.3 and 4.2.

Name of the Hyperparameter	Value
Epoch	30
batch size	1
distance threshold (τ)	50
learning rate	10^{-3}
weight decay	$5 \cdot 10^{-2}$
Number of dimensions in hidden layer (FC)	128
Number of dimensions in hidden layer (GraphSAGE)	128
Number of dimensions in hidden layer (GAT)	32
Number of dimensions in hidden layer (GCN)	128
Number of attention heads in GAT	8
Number of dimensions in hidden layers (Transformer)	128
Number of attention heads in Graph Transformer	1

Table 4.3: List of hyperparameters

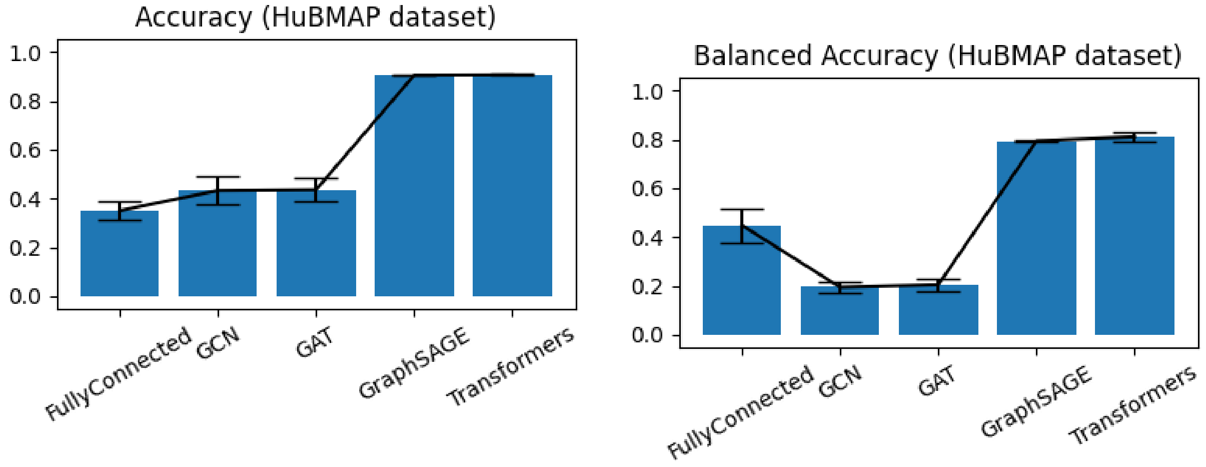

Figure 4.1: Performance of Fully Connected model, GCN model, GAT model, GraphSAGE model, Transformer model on the HuBMAP dataset based on accuracy and balanced accuracy

HuBMAP

The results of the performance of different models on the HuBMAP dataset are shown in fig. 4.1. In terms of accuracy, it is seen that the fully connected layer fares the worst. This highlights the importance of spatial information that rest of the models are able to utilize and gain higher accuracy. However, in terms of balanced accuracy, GCN and GAT models are not able to utilize this information effectively to their advantage. Hence, Fully connected layer does better than GCN and GAT. It is noteworthy that GraphSAGE and Transformer model, both are able to achieve similar performance. Similarly, the performance of GCN and GAT is also similar. This shows that the neighbors are equal for most of the cells in the HuBMAP dataset.

TonsilBE

The larger trend is the same in the TonsilBE dataset as seen in the HuBMAP dataset. The models overall have quite poor performance compared to the HuBMAP perforamnce as seen in the fig. 4.2. This could potentially be attributed to the downsampling done on the training data to the TonsilBE dataset. This downsampling was needed to run the model on the Grace cluster. This could be one of the potential reasons why the performance of STELLAR (GraphSAGE) is not the same as reported in the paper. Additionally, this was a more challenging task as compared to HuBMAP as the test data was a completely different tissue.

It is seen that the Fully Connected layer seems to have better performance than GCN and GAT layers. However, GraphSAGE and Graph Transformer models are able to utilize spatial information effectively. It is also seen that GraphSAGE and Graph Transformers have comparable performance while GCN and GAT also have comparable performance. Thus, raising the question of whether there is a need to add an attention mechanism to these graphs. This also suggests that all neighbors are equal for most of the cells in the spatial graph.

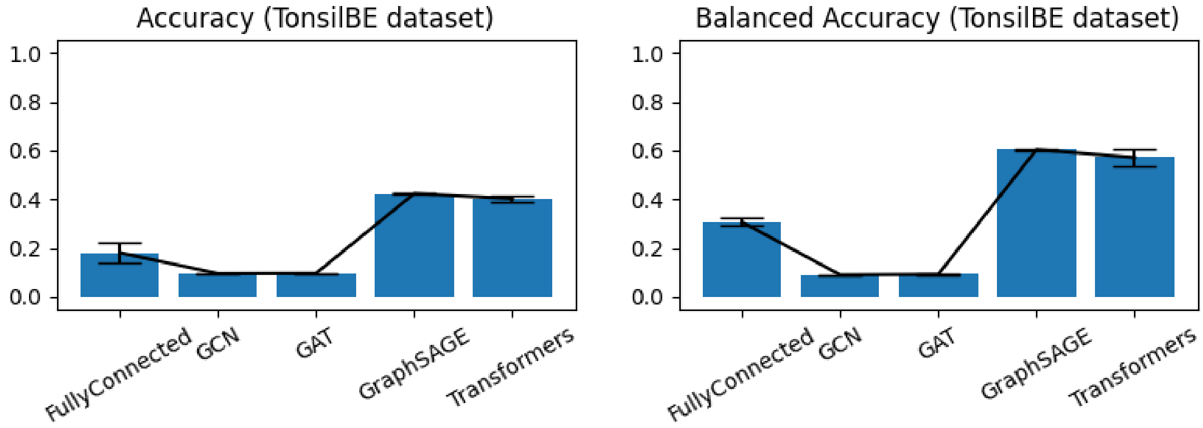

Figure 4.2: Performance of Fully Connected model, GCN model, GAT model, GraphSAGE model, Transformer model on the TonsilBE dataset based on accuracy and balanced accuracy

Chapter 5

Conclusion

This work highlights the potential for using Graph Neural Networks to leverage the spatial information in the spatial single cell datasets analysis. Additionally, based on the results it can also be concluded that for most of the cells, all neighbors are equal. This is demonstrated by similar performances of the models even after the addition of attention mechanisms.

Chapter 6

Code and Data availability

The code with appropriate readme.md, requirements.txt and documentation can be found here: `https://github.com/BurhanSabuwala/CPSC583-Project`.

The dataset used can be found here: `https://datadryad.org/stash/share/1OQtxew0Unh3iAdP-ELew-ctwuPTBz6Oy8uuyxqliZk`. This url is taken from the study conducted by Brbic et al [11].

Bibliography

[1] S. K. Longo, M. G. Guo, A. L. Ji, and P. A. Khavari, “Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics,” vol. 22, no. 10, pp. 627–644. Number: 10 Publisher: Nature Publishing Group.

[2] C. G. Williams, H. J. Lee, T. Asatsuma, R. Vento-Tormo, and A. Haque, “An introduction to spatial transcriptomics for biomedical research,” vol. 14, no. 1, p. 68.

[3] R. Ahmed, T. Zaman, F. Chowdhury, F. Mraiche, M. Tariq, I. S. Ahmad, and A. Hasan, “Single-cell RNA sequencing with spatial transcriptomics of cancer tissues,” vol. 23, no. 6, p. 3042.

[4] J. F. Quintana, P. Chandrasegaran, M. C. Sinton, E. M. Briggs, T. D. Otto, R. Heslop, C. Bentley-Abbot, C. Loney, L. de Lecea, N. A. Mabbott, and A. MacLeod, “Single cell and spatial transcriptomic analyses reveal microglia-plasma cell crosstalk in the brain during trypanosoma brucei infection,” vol. 13, no. 1, p. 5752. Number: 1 Publisher: Nature Publishing Group.

[5] M. Ratz, L. von Berlin, L. Larsson, M. Martin, J. O. Westholm, G. La Manno, J. Lundeberg, and J. Frisén, “Clonal relations in the mouse brain revealed by single-cell and spatial transcriptomics,” vol. 25, no. 3, pp. 285–294. Number: 3 Publisher: Nature Publishing Group.

[6] A. R. Colombo, M. Hav, M. Singh, A. Xu, A. Gamboa, T. Lemos, E. Gerdtsson, D. Chen, J. Houldsworth, R. Shaknovich, T. Aoki, L. Chong, K. Takata, E. A. Chavez,

C. Steidl, J. Hicks, P. Kuhn, I. Siddiqi, and A. Merchant, “Single-cell spatial analysis of tumor immune architecture in diffuse large b-cell lymphoma,” vol. 6, no. 16, pp. 4675–4690.

[7] I. Kleino, P. Frolovaitė, T. Suomi, and L. L. Elo, “Computational solutions for spatial transcriptomics,” vol. 20, pp. 4870–4884.

[8] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S.-O. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. d. Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer, I. Khatri, S. M. Kielbasa, J. O. Korbel, A. M. Kozlov, T.-H. Kuo, B. P. Lelieveldt, I. I. Mandoiu, J. C. Marioni, T. Marschall, F. Mölder, A. Niknejad, L. Raczkowski, M. Reinders, J. d. Ridder, A.-E. Saliba, A. Somarakis, O. Stegle, F. J. Theis, H. Yang, A. Zelikovsky, A. C. McHardy, B. J. Raphael, S. P. Shah, and A. Schönhuth, “Eleven grand challenges in single-cell data science,” vol. 21, no. 1, p. 31.

[9] J. Wang, A. Ma, Y. Chang, J. Gong, Y. Jiang, R. Qi, C. Wang, H. Fu, Q. Ma, and D. Xu, “scGNN is a novel graph neural network framework for single-cell RNA-seq analyses,” vol. 12, no. 1, p. 1882. Number: 1 Publisher: Nature Publishing Group.

[10] Y. Li, S. Stanojevic, and L. X. Garmire, “Emerging artificial intelligence applications in spatial transcriptomics analysis,” vol. 20, pp. 2895–2908.

[11] M. Brbić, K. Cao, J. W. Hickey, Y. Tan, M. P. Snyder, G. P. Nolan, and J. Leskovec, “Annotation of spatially resolved single-cell data with STELLAR,” vol. 19, no. 11, pp. 1411–1418. Number: 11 Publisher: Nature Publishing Group.

[12] Y. Yuan and Z. Bar-Joseph, “GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data,” vol. 21, no. 1, p. 300.

[13] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs.”

[14] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks.”

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks.”

[16] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, “Masked label prediction: Unified message passing model for semi-supervised classification.”