Predicting Proximity to Pathology for Single-Cell Data in Alzheimer's Disease

Galya Vergieva

Supervisor(s): Marcel Reinders, Roy Lardenoije, Timo Verlaan, Gerard Bouland

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Galya Vergieva
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Timo Verlaan, Roy Lardenoije, Gerard Bouland, Ricardo Marroquim

Abstract

Alzheimer's Disease is a complex neurodegenerative disorder marked by the abnormal build-up of proteins in the brain. As no cure currently exists, understanding the disease's cellular mechanisms is essential for advancing diagnostics and treatment. To this end, single-cell RNA sequencing (scRNA-seq) is a method that offers detailed information about the gene activity of individual cells but lacks their spatial context. Conversely, spatial transcriptomics technology preserves the localization of the cells but provides more limited transcriptomic information. To resolve this, we provide a model that predicts a cell's distance to pathology from single-cell RNA-sequencing data. Additionally, we identify *APOE*, *LYVE1*, and *SLC17A7* as genes potentially associated with AD-related microglial clustering around plaques.

1 Introduction

As individuals age, the risk of neurodegenerative diseases increases, posing a serious challenge to today's society [1, 2]. Among these conditions, Alzheimer's disease (AD) is emerging as one of the most lethal and impactful disorders, affecting one in ten individuals aged 65 and older [2, 3]. Characterized by the gradual loss of neurons in the brain and the accumulation of *amyloid plaques*[1] and *tau tangles*[2], AD leads to cognitive decline and impairments in daily life and social interactions [1]. Despite recent scientific advances on the inner mechanisms of AD, no *disease-modifying treatments*[3] are currently available [4], thus further research is required on the cellular level.

To provide more insight into biological systems, single-cell technologies offer high-resolution analysis of individual cells [5]. However, performing *single-cell RNA sequencing*[4] (scRNA-seq) requires isolating individual cells through tissue disassociation [6]. This process destroys the information on the cells' location within the tissue and their proximities to each other, which limits our understanding of cell-cell interactions [5, 7]. In contrast, *spatial transcriptomics*[5] technologies, which offer less detailed information about gene activity per individual cell, preserve the localization of cells within the tissue, offering insights into the complex interplay between cell types and their roles in AD pathology [1, 5]. Thus, the integration of single-cell and spatial transcriptomics could provide a more comprehensive view of the inner workings of the disease.

In particular, the distance of a cell to pathology is a biologically meaningful factor, as it may reflect a cell's level of exposure to the disease processes. The distance metric could help the analysis of how molecular changes in cells, e.g. inflammation, vary with proximity to pathology, offering insights into disease progression and potential treatment strategies.

While the integration of spatial transcriptomics and scRNA-seq data has been extensively studied [1, 5, 7–10], research focuses on broad-level cell mapping [1, 7] or *gene imputation*[6] [5, 9, 10], rather than on specific spatial features relevant to disease pathology. A recent study [1], for example, generated a spatial map of the prefrontal cortex by integrating single-cell and spatial data. While this provides a valuable insight into cell-type differences between healthy and diseased individuals, it does not address spatial proximity to pathology. Moreover, the study did not use single-cell resolution spatial transcriptomic data, and thus lacked precise location of individual cells. Another approach to localizing scRNA-seq data is Seurat's spatial reconstruction algorithm [7], which infers spatial positions of cells based on *gene expression*[7] patterns. However, it does not incorporate annotated *histopathological*[8] features which is of interest to our study. On the other hand, a study [11] into the spatial organization within the *amyloid plaque niche*[9] analyzes cells' distance to plaque but does not integrate spatial transcriptomics with scRNA-seq data or attempt to build predictive models. Thus, there is a need to closely examine and predict distance to pathology for single-cell data, particularly for applications in AD research.

Our work aims to help fill this gap by integrating spatial histopathological context into scRNA-seq analysis by developing a model that predicts individual cells' distance to nearest plaque. We chose to focus specifically on *microglia cells*[10], as their spatial behavior is closely tied to the progression of AD [12] and they actively respond to amyloid plaque formations, clustering around these pathological structures [12]. This spatial behavior makes microglia biologically meaningful for studying spatial proximity to AD pathology.

Using spatial transcriptomics as ground truth, we create the distance prediction model by exploring three different approaches for improving its performance. Firstly, we built a model that predicts the distance using genes shared between scRNA-seq and spatial transcriptomics data. Then, we compared the results with those of a second model that uses the full gene set from scRNA-seq to predict the distance. Lastly, to improve the distance predictions, we attempted to classify a cell's overlap with a plaque area.

Finally, we provide the model with the best performance for predicting distance to pathology. Additionally, we identified the genes contributing most strongly to the spatial localization.

2 Materials and Methods

This section describes the methodology employed throughout the study. It covers materials and the key techniques used to prepare and analyze the data, ensuring that each step is clearly documented for reproducibility.

2.1 Datasets

For the experiments conducted in this study, we have utilized the following datasets:

ROSMAP Microglia dataset [13] - a scRNA-seq dataset from ten brain donors, the gene expression of a total of 86,612 cells from the prefrontal cortex is measured, for 19,183 genes. Study was provided by the Religious Orders

Study and Rush Memory and Aging Project (ROSMAP https://adknowledgeportal.org).

Xenium dataset - a single-cell spatial transcriptomics dataset from one brain donor, the gene expressions of 93,258 cells, for 266 genes. This dataset has been provided by Gonçalves lab at TU Delft (https://goncalveslab.tudelft.nl/) and is not publicly available.

2.2 Data Preprocessing

The Xenium dataset was provided as already preprocessed. We further filtered it to include only microglia cells (2,160 cells), as they are of primary interest in this study. Then we observed the distribution of cell distances (Figure 1), which indicates that the vast majority of data points are close to plaques. To focus the analysis on the relevant biological range, we excluded distant outliers. Outlier detection was performed using the interquartile range (IQR) method. We computed the first quartile (Q_1) and the third quartile (Q_3) of the *distance to pathology* values and IQR was then calculated as the difference between these two values. We defined an upper boundary beyond which data points were considered outliers [14] as follows:

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} \tag{1}$$

Cells with a distance to pathology value greater than this upper bound were excluded from the analysis. The resulting matrix has dimensions of 1,965 cells by 266 genes.

We log-normalized the ROSMAP Microglia dataset with ScanPy.

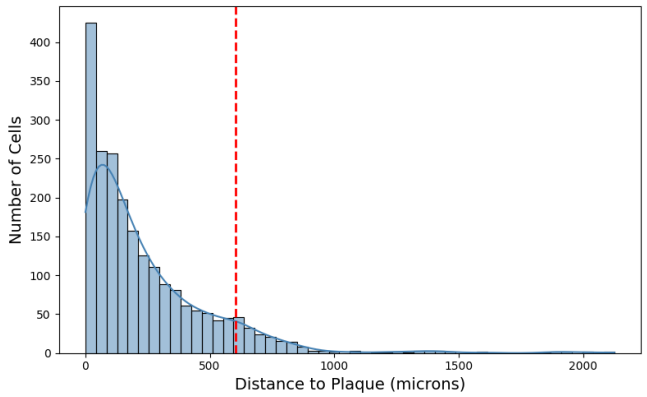


Figure 1: Distribution of distances to plaque of the microglia cells in the Xenium dataset. The red line annotates the upper boundary at which the outliers were removed.

2.3 Distance Prediction using Shared Genes

We trained a weighted k-nearest neighbors (k-NN) model to predict the distance to nearest plaque using the set of shared genes between the scRNA-seq and spatial transcriptomics datasets.

Aligning scRNA-seq and spatial transcriptomics data

Let $\mathbb{R}_{(n\times g)}$ denote the gene expression matrix of the preporcessed scRNA-seq data, where n is the number of cells and g the number of genes. Similarly, let $\mathbb{Q}_{(m\times h)}$ denote the gene expression matrix of the preprocessed spatial transcriptomics data, with m cells and h genes. Using the set of shared genes $p = g \cap h$, we subset the spatial transcriptomics dataset to ensure compatibility between modalities, resulting in a filtered gene expression matrix: $\mathbb{Q}'_{(m\times p)} \in \mathbb{Q}_{(m\times h)}$. Since the number of shared genes is large (265), we performed Principal Component Analysis (PCA) on $\mathbb{Q}'$ to reduce noise and capture the key biological variation. Thus, we reduce the matrix to $\mathbb{Q}''_{(m\times d)}$, where d is the number of principle components.

Weighted k-Nearest Neighbour (k-NN) Regression Model

We implemented a weighted k-nearest neighbour (kNN) regression model that identifies the k-nearest neighbours for each cell x_i where $i \in \{1, \ldots, m\}$. Given a test point x_i, this returns a set of distances $\{d_{i1}, d_{i2}, \ldots, d_{ik}\}$ and corresponding indices $\{j_1, j_2, \ldots, j_k\}$, where each distance is defined as the Euclidean distance between the test point and the given neighbour. If a neighbour is exactly identical to the cell (distance = 0), it is assigned a full weight of 1, and all other weights are set to 0 to avoid division by zero later. Otherwise, for each neighbour j of a test cell i, the weight w_{ij} is computed as the inverse of the distance:

$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} = 0 \\ \frac{1}{d_{ij}} & \text{otherwise} \end{cases} \tag{2}$$

All weights are then normalized to sum to 1:

$$\tilde{w}_{ij} = \frac{w_{ij}}{\sum_{k\in\mathcal{NN}(i)} w_{ik}} \tag{3}$$

where $\mathcal{NN}(i)$ denotes the set of nearest neighbours for cell x_i.

Finally, the predicted distance $\hat{y}_i$ for test cell x_i is computed as the weighted sum of the distances from its k-nearest neighbours:

$$\hat{y}_i = \sum_{j\in\mathcal{N}(i)} \tilde{w}_{ij} \cdot d_i j \tag{4}$$

2.4 Gene Imputation with SpaGE

To impute the expression of genes not in the Xenium panel, **SpaGE** [5] was employed - a machine learning method designed for the integration of scRNA-seq and spatial transcriptomics data. The source code is available at https://github.com/tabdelaal/SpaGE. The version employed in this study was retrieved on 5 May 2025.

We applied SpaGE using the following parameters: the spatial transcriptomics dataset was set to the gene expression matrix of the preporcessed Xenium dataset $\mathbb{Q}_{(m\times h)}$, with m cells and h genes. The scRNA-seq data was the gene expression matrix of the preprocessed ROSMAP Microglia data $\mathbb{R}_{(n\times g)}$, where n is the number of cells and g the number of genes. The number of principal components was set to $\texttt{n\_pv} = 30$, following the original implementation in SpaGE. The set of genes to impute was defined as $imp = h - g$. After constructing the imputed gene expression matrix $\mathbb{I}_{(m\times imp)}$ we filtered out NaN genes (in our case only one) which left 18,917 genes.

2.5 Predicting distance with scRNA-seq Genes

We first performed gene imputation on the spatial transcriptomics data and then trained a model to predict the distance to plaque using the complete set of genes expressed in the scRNA-seq data.

Using the set of shared genes $p = g \cap h$, we subset the spatial transcriptomics dataset, resulting in the gene expression matrix: $\mathbb{Q}'_{(m\times p)} \in \mathbb{Q}_{(m\times h)}$. To construct the input for the actual prediction model, we concatenated $\mathbb{I}_{(m\times imp)}$, obtained from the Gene Imputation with SpaGE, with $\mathbb{Q}'_{(m\times p)}$. This operation is expressed as:

$$\mathbb{C}_{(m\times(imp+p))} = [\mathbb{I}\mathbb{Q}'] \tag{5}$$

where $\mathbb{C}$ denotes the combined matrix formed by concatenating $\mathbb{I}$ and $\mathbb{Q}'$ along the gene (column) axis. The complete number of genes is 19,182.

We perform PCA on $\mathbb{C}_{(m\times(imp+p))}$ to reduce the matrix to $\mathbb{C}'_{(m\times d)}$, where d is the number of principal components. Then we used $\mathbb{C}'_{(m\times d)}$ as input for the model, proceeding with the steps described in Weighted k-Nearest Neighbour (k-NN) Regression Model.

2.6 Gene Identification

We identified the key genes that contributed to the model's predictions to later analyze their potential association with AD. We first extracted the PCA loadings, which represent the weights of each gene in each principal component. Since each principal component explains a different amount of variance in the data, we weighted the absolute loadings by the proportion of variance explained by each component. Next, we summed these weighted loadings across all considered components for each gene to obtain an overall gene importance score. This score reflects how strongly each gene influences the patterns observed in the data. Finally, we ranked the genes based on their scores.

2.7 Plaque Area Overlap Classification

We trained a model to classify whether a cell overlaps with pathology, using only the genes shared between the scRNA-seq and spatial transcriptomics datasets. For this, we employed the procedure described in Aligning scRNA-seq and spatial transcriptomics data. Then, as the cells within the plaque area are still substantially less than those outside of it (310 positive class, 1,655 negative class), we performed oversampling with SMOTE [15]. Additionally, we manually adjusted the classification probability threshold for positive predictions to 0.4. We chose to train a Random Forest Classifier as it accounts for non-linear relations and imbalanced data.

2.8 Cross Validation

We split the data into 80% train and 20% test set, partitioning $\mathbb{Q}''$ from Aligning scRNA-seq and spatial transcriptomics data into $X_{\text{train\_full}}$, $y_{\text{train\_full}}$ and $X_{\text{test\_final}}$, $y_{\text{test\_final}}$, where X is the input gene expression matrix and y is the target feature - distance to pathology. To perform hyperparameter tuning, we employed 10-fold cross-validation only on the training data ($X_{\text{train\_full}}$, $y_{\text{train\_full}}$). In each iteration, one fold was held out as the evaluation set while the model was trained on the remaining of the training set. Formally, for each fold $i \in \{1, \ldots, k\}$, we partitioned $X_{\text{train\_full}}$ into $X_{\text{train(i)}}$ and $X_{\text{val(i)}}$, and $y_{\text{train\_full}}$ into $y_{\text{train(i)}}$ and $y_{\text{val(i)}}$. After each split, we evaluated the model on the evaluation set.

2.9 QuPath

We used QuPath [16], an open-source software for digital pathology image analysis, to process a pathological image from the prefrontal cortex of the Xenium patient and determine amyloid plaque areas relevant for our predictions. The image and the staining were provided by the Gonçalves lab at TU Delft. We set the radius around the plaque as follows: `Objects > Annotations... > Expand annotations`.

2.10 Implementation details

We used the following Python Packages for this study: ScanPy version 1.11.1 [17], AnnData version 0.11.4 [18], Pandas version 2.2.3 [19], Scikit-learn version 1.5.2 [20], SciPy version 1.15.2 [21], NumPy version 2.2.5 [22], Seaborn version 0.13.2 [23], Matplotlib version 3.10.1 [24], Imbalanced-learn version 0.13.0 [15].

3 Results

In this study, we investigated whether a cell's spatial proximity to the nearest plaque could be accurately predicted from scRNA-seq data. To evaluate this, we trained and tested models using Xenium spatial transcriptomics data and ROSMAP scRNA-seq data.

3.1 Predicting Distance with Shared Genes

We began by aligning the two datasets based on shared genes. Dimensionality reduction was then performed using PCA to reduce noise and simplify the input space. Finally, we employed a weighted k-nearest neighbors (k-NN) algorithm to predict distance to pathology. We chose to implement a weighted k-NN model, inspired by the approach used in SpaGE, but with some modifications. This choice was motivated by the algorithm's computational efficiency, which removed the need for high-performance computing resources. Additionally, several state-of-the-art integration methods, such as Seurat [7] and StPlus [9], also utilize k-NN models.

We measure the performance of the model by calculating the mean absolute error (MAE) in microns, reflecting the difference between predicted and true distances. We first considered Spearman correlation, as it is commonly used to assess the performance of integration methods in spatial transcriptomics. However, those methods aim to reconstruct the entire spatial organization of cells, whereas our approach is specifically designed to predict distance to pathology. Therefore, Spearman correlation would not have provided a directly comparable measure of performance. MAE, by contrast, offers a simple and interpretable measure in physical units that allows us to place the results in a biologically relevant context.

For this implementation we had two hyperparameters - the number of neighbours and the number of principal

components. We tuned our model as described in Cross Validation by comparing the average MAE of different values for the hyperparameters. Thus, we selected 10 principal components and 40 neighbours (Figure 2a).

After training the model, we evaluated its performance on the test set by calculating MAE of the predicted distances. As MAE provides a physical unit of the error, to interpret it we need to place this number in its biological context. We established a biologically meaningful threshold to define when the predictions are considered too far from the true values. We set this threshold at 200 µm, as *paracrine signaling*[11] between cells typically occurs within a spatial range of up to 200 µm [25], making this distance biologically meaningful for understanding intercellular communication. Therefore, we considered an average prediction error below 200 µm acceptable for the purposes of this study, as it would imply that the predicted cell location remains within the original signaling environment defined by this radius.

We then defined the following hypothesis to assess our model:

H1: A cell's distance to pathology can be predicted using only genes shared between the single-cell and spatial transcriptomics datasets, with an average absolute error less than 200 µm.

We computed the MAE of the model's predictions, which yielded a value of 118.98 µm. While this error falls far below the threshold of 200 µm, and is thus considered small, we performed a permutation test to determine the statistical significance of this result. By randomly shuffling the predicted values 10,000 times, we created a distribution of MAEs expected by chance. The p-value was then calculated as the proportion of shuffled MAEs that were less than or equal to the observed MAE. The test yielded a p-value of 0.0001. Given that $p < 0.05$, we conclude that the model predicts the distance to pathology with a MAE significantly below 200 µm. This demonstrates that the model can predict a cell's distance to pathology based on shared gene expression patterns between single-cell and spatial transcriptomics data, with sufficient accuracy for the goals of this study.

To further analyze the model's predictions, Figure 3 presents the distributions of the predicted and the true values. The model's predictions exhibit a narrower range, struggling to capture distances that are either very close to or far from pathology. This limitation stems from the nature of the k-NN algorithm, which averages the distances of nearby cells. As a result, predictions are biased towards the center, with fewer values near the extremes.

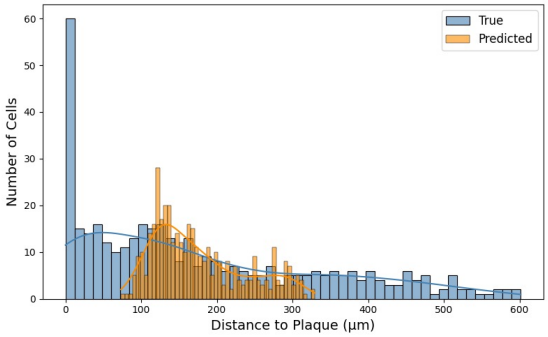
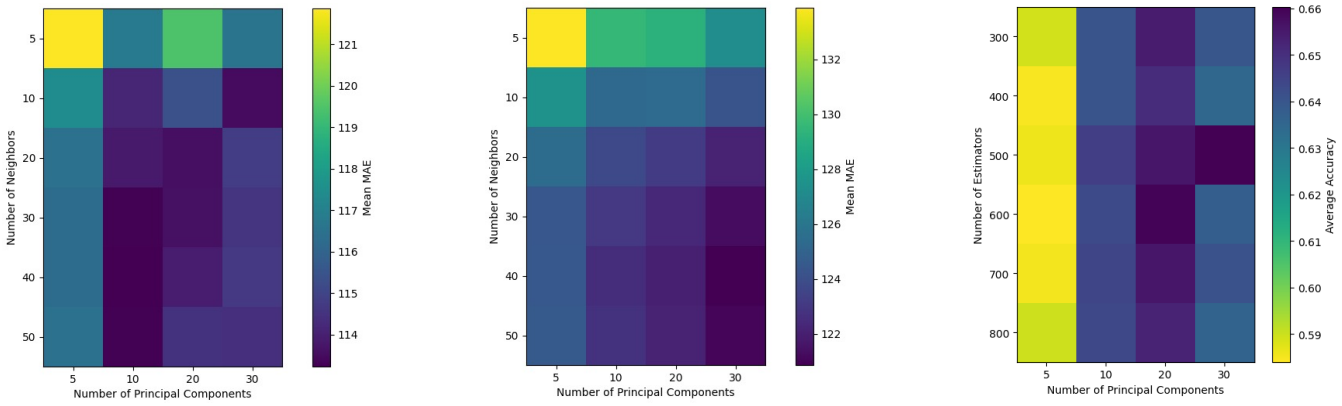


Figure 3: Distribution of actual vs. predicted distances to plaque. The range of predictions is smaller and mainly captures mid-distance values.

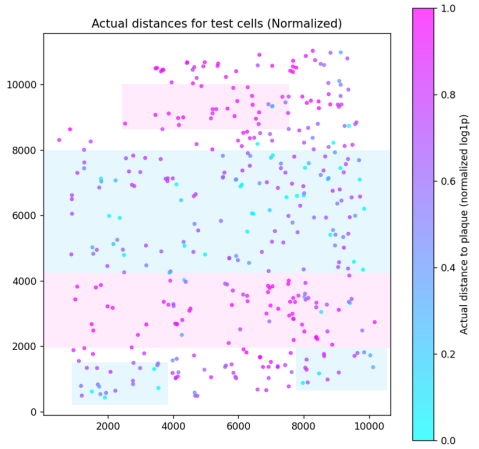
However, to take a closer look, we visualized the test cells in their original spatial coordinates, colored by both the actual (Figure 4a) and predicted (Figure 4b) distances to nearest plaque. To enable a direct comparison of spatial patterns, we applied a log-transformation to the distances and we scaled them independently to a [0, 1] range. This



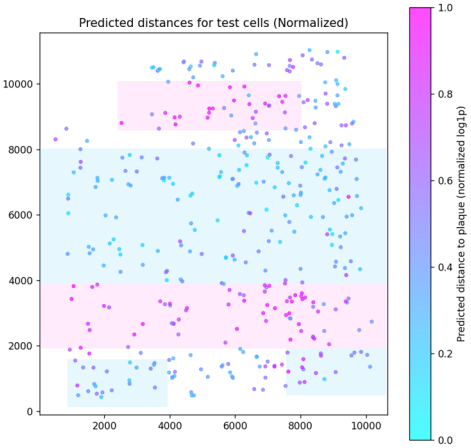
(a) (b) (c)

Figure 2: Hyperparameter tuning of the three models through 10-fold cross-validation. The dark colours present better performance. (a) The average Mean Absolute Error of predicting distance based on shared genes for different numbers of neighbours and principal components. (b) The average Mean Absolute Error of predicting distance based on imputed and shared genes for different number of neighbours and principal components. (c) The average accuracy of predicting overlap with pathological region for different number of estimators and principal components.

normalization ensures the reflection of relative distances within each distribution, to abstract away from the differing value ranges of actual and predicted distances. Although the predicted distances cover a narrower range, we can observe that the model retains the overall spatial pattern - cells predicted to be closer or farther from plaques keep their relative positions.



(a) Normalized actual distances to plaque.



(b) Normalized predicted distances to plaque.

Figure 4: Spatial distribution of test cells shown in their original coordinate space, colored by (a) actual and (b) predicted distances to the nearest plaque. Distances were log-transformed and min-max normalized to enable direct visual comparison. Colored rectangles highlight regions dominated by cells at similar distances - blue indicates closer to plaques, while pink denotes more distant cells. The close correspondence in spatial gradients between the two plots illustrates that the model preserves spatial patterns, despite compressing the range of predicted values.

We further investigated the most contributing genes to the model's predictions to determine what their relevance to AD is. This was done following the procedure described in Gene Identification. In Figure 5, we present the most prominent genes ordered according to their importance score. Notably, the genes *APOE*, *LYVE1*, and *SLC17A7* stand out as the biggest contributors.

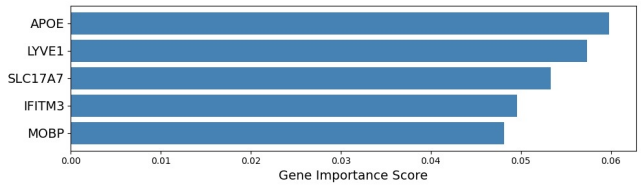


Figure 5: Most contributing genes to shared gene the predictions, ordered according to their importance score.

3.2 Predictions with Shared Genes vs All ScRNA-seq Genes

After implementing the above-mentioned model, we explored alternative ways to improve predictions. Instead of relying solely on shared genes, we proposed imputing the missing genes from the spatial transcriptomics data and combining them with the shared ones to train the k-NN model. We explored the idea that additional gene expression information, particularly from scRNA-seq specific genes, might carry spatial signals relevant to a cell's proximity to pathology.

Therefore, we first performed gene imputation with SpaGE. The choice of this method is based on a recent literature review [8], which offers a comprehensive comparison of state-of-the-art techniques for integrating scRNA-seq and spatial transcriptomics. We selected only methods compatible with our study, resulting in Table 1. From those methods, Seurat [7] was deemed unsuitable as it does not perform gene imputation, while GimVI [10] is too computationally expensive. Ultimately, we chose to utilize SpaGE for its simplicity, accuracy, and greater computational efficiency. Then, we combined the imputed and the shared genes to use as input for the weighted k-NN regressor. After performing Cross Validation, we set the number of neighbors to 40 and the number of principal components to 30 (Figure 2b). To evaluate whether this new model performs better than the previous one, we formulated the following hypothesis:

H2: Distance predictions based on imputed and shared gene expression data have a lower average absolute error than when using only shared genes.

$$H_2 : \mathrm{MAE}_{\mathrm{H2}} < \mathrm{MAE}_{\mathrm{H1}}$$

We calculated a MAE of of 125 µm, which, while still below the 200 µm threshold and thus considered a small error, indicates worse performance compared to the first model. To assess the significance of this observation, we performed a permutation test, which gave a p-value of 0.99. Since $p > 0.05$, we cannot conclude that using the full gene set improves prediction accuracy compared to using only shared genes. The p-value even suggests that the opposite of H2 might be true - including imputed genes could degrade model performance rather than improve it. Analyzing the boxplots of the predictions from the models (Figure 6), we notice that the first model captures more variance, which might

Table 1: Comparison of Integration Methods for scRNA-seq and Spatial Transcriptomics (extracted from [8])

Method	Description	Limitations / Notes
Seurat [7]	Comprehensive data analysis pipeline with integrated algorithms	Only available for certain types of ST platforms
SpaGE [5]	Domain adaptation model aligning ST and scRNA-seq to a common space; efficient for large datasets	Only includes genes shared by both datasets
StPlus [9]	Reference-sequence-based; improved accuracy and reduced resource usage	Applied only to image-based sequencing data
GimVI [10]	Uses variational autoencoders to improve biological interpretation with platform-specific patterns	Slower than benchmarked tools

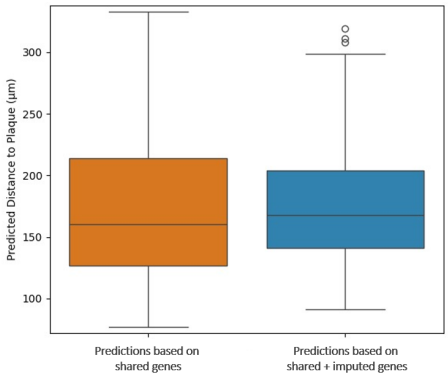


Figure 6: Comparison of the distributions of the predictions from the two models. Observed wider range of variance for the first model which uses only shared genes.

account for the slightly better results. We further derived the primary genes contributing to the predictions of this model (see Gene Identification) and thus identified *FCN1*, *VCAN-AS1*, and *CD300E* (Figure 7). Notably, all of those genes are exclusively found in the scRNA-seq data.

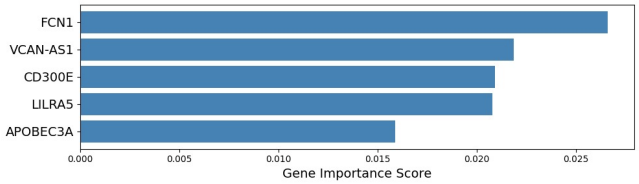


Figure 7: Most contributing genes to the predictions of the second model chosen from the full gene set in the scRNA-seq dataset. The genes are not shared between the scRNA-seq and the spatial transcriptomic data set.

Classification of Plaque Area Overlap

As shown in Figure 6, both distance prediction models struggle to accurately estimate values for cells located close to plaques. To address this limitation, we attempted to build a classifier for a cell's overlap with a pathological region. We reasoned that if overlap can be reliably predicted, this information could be fed back into the original distance prediction model and help refine close-range predictions. Since the first distance prediction model demonstrated overall better performance, we focused on improving that approach and accordingly based the classifier on the set of shared genes.

Given the limited number of cells directly overlapping with pathology (Figure 1), we also included cells closely clustered around plaques in the positive class. Therefore, we defined a radius for the plaque area. Initially, we considered a radius of 10 µm, as microglial cells are highly overrepresented within this distance from plaques [11]. However, to select a radius more appropriate for our specific study, we analyzed an image of the patient's prefrontal cortex using QuPath (Figure 8). Thus, we determined that a radius of 20 µm best captures the clustering of cells around plaques. We

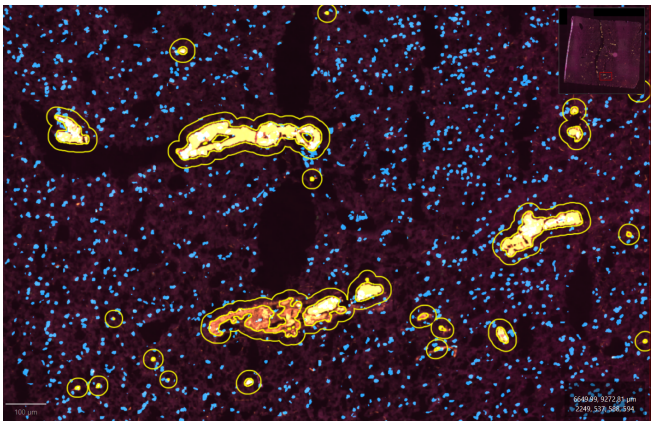


Figure 8: Image of the prefrontal cortex from the patient, displaying individual cells in blue, amyloid plaques in bright yellow, and a 20-micron radius around each plaque indicating the extended pathological region considered in the analysis.

evaluated performance of this model by measuring the prediction accuracy, expressed as the percentage of correctly classified cells. Furthermore, we performed hyperparameter tuning for the number of estimators and number of principal components in the Random Forest Classifier and the PCA respectively. We averaged the accuracies across all folds in the Cross Validation and selected 500 estimators and 30 principal components (Figure 2c).

The model achieved an overall accuracy of 79.13%. As illustrated by the confusion matrix (Figure 9), it performs

well in predicting non-overlapping cells, but struggles to correctly identify overlapping ones. Specifically, the recall is 88% for the negative class and only 37% for the positive class. This imbalance in performance is likely due to the skewed distribution of training labels, with False instances significantly outnumbering True ones. As a result, the predictions are biased and not reliable for further use.

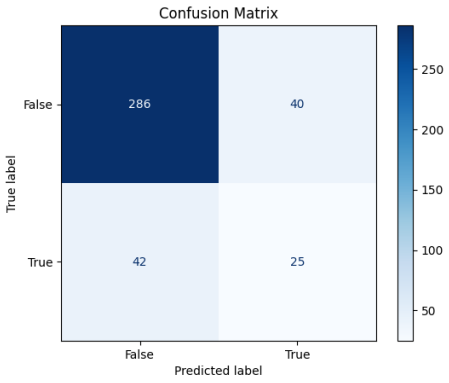


Figure 9: Confusion matrix showing the classification results for predicting cell overlap with a 20-micron dilated area around the plaque. High misclassification for the True class (overlap).

4 Discussion

In this study, we investigated whether the distance to pathology for cells in scRNA-seq data can be predicted. To answer this question, we built and evaluated two distance prediction models, comparing their MAEs to identify the better approach. Furthermore, we attempted to classify cells' overlap with a pathological region to improve the aforementioned distance predictions.

4.1 Predicting Distance to Nearest Plaque

The two distance prediction models differed solely in the gene sets used for training - one was trained using genes shared between the spatial transcriptomics and scRNA-seq datasets, while the other used shared genes and imputed genes in the spatial transcriptomics data. After analyzing the results, we concluded that the model relying solely on shared genes has better performance. Since the key difference between the two models was the gene sets used, we investigated the most influential genes in each model's predictions to determine their relevance to AD. In the first model, we identified the genes *APOE*, *LYVE1*, and *SLC17A7* as the biggest contributors.

APOE is a gene involved in cholesterol transport within the bloodstream and is widely recognized as a major genetic risk factor for AD [26]. Its expression in microglia, along with its direct interaction with amyloid β plaques [27], likely explains its prominent influence on our model's predictions.

LYVE1, a gene responsible for lymphatic drainage, has no direct link to AD. However, dysfunction in the brain's lymphatic drainage has been shown to accelerate amyloid plaque accumulation [28]. Moreover, research suggests that regions enriched with plaques often exhibit reduced *LYVE1* expression [29], hinting that *LYVE1* may be a significant factor in determining a cell's proximity to plaque. The Spearman correlation we observed between *LYVE1* expression and the predicted distances was -0.38, indicating that our model tends to predict shorter distances to plaque for cells with lower *LYVE1* expression, aligning with the above-mentioned research.

Lastly, *SLC17A7* is a gene involved in neurotransmission in the brain. It encodes VGLUT1, a protein whose expression is reduced in the presence of amyloid β plaques [30]. The Spearman correlation between *SLC17A7* expression and the predicted distances is -0.24, confirming this theory. Furthermore, this gene is considered a potential marker for early subtypes of AD [31].

On the other hand, the primary genes in the predictions of the second model are *FCN1*, *VCAN-AS1*, and *CD300E*. These genes are part of the imputed set, originally present only in the scRNA-seq data. Current research does not indicate an association between them and AD. This could explain the decreased performance of the second model, as its most influential features are not directly linked to the spatial context of amyloid plaques or AD pathology. Nevertheless, these genes play a role in inflammatory responses, which may account for the model's prediction error still falling below the 200 µm threshold.

Our study suggests a relationship between the expression of *APOE*, *LYVE1* and *SLC17A7* and the proximity to AD pathology. These findings offer evidence for the genes' potential roles in disease progression and can serve as a foundation for future research focused on these genes.

Additionally, we implemented a classifier to predict whether a cell overlaps with a 20-micron plaque area. However, this model often misclassified overlapping cells, which is due to the imbalance in the training data. Despite efforts to address this during model tuning, the imbalance persisted and negatively impacted the model's performance. This issue is particularly concerning as the goal of the experiment was to enhance the prediction of cells' proximity to plaque, especially at close-range distances. Therefore, the low sensitivity to True labels limits the usefulness of this approach, and currently, it cannot be used to improve distance predictions.

As the distance prediction model based on shared genes demonstrated better performance, we selected it as our final approach. We provide it as a method that predicts a cell's distance to the nearest plaque by taking as input two single-cell datasets in AnnData format - one containing ground truth distances to pathology and another without such annotations, along with parameters for the number of principal components and neighbors. Notably, this method can be used to predict distance to pathology not only for scRNA-seq data, but also for other single-cell datasets. Furthermore, the overall approach for predicting distance to pathology could, in principle, be adapted to estimate other spatial features. However, we cannot guarantee its performance in those contexts without further validation.

Finally, we applied the method to the Xenium and

ROSMAP Microglia datasets using 40 neighbors and 10 principal components. The resulted prediction distribution can be seen in Figure 10.

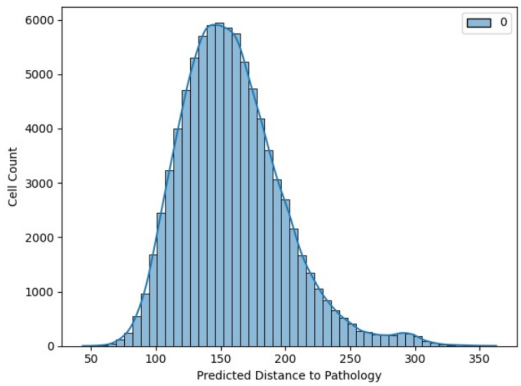


Figure 10: Distribution of predicted distances for the ROSMAP Microglia dataset.

4.2 Limitations

During the course of this research, we have encountered various issues that may limit the validity and accuracy of our results. Those factors are discussed in the following section.

Xenium dataset Limitations

This study relies on the Xenium dataset for both training and evaluation. We acknowledge that the dataset represents tissue from a single patient, which limits the generalization of our findings to the broader population. Additionally, our focus on microglia cells significantly reduces the number of available samples, which may have constrained the model's performance. With access to a larger and more diverse dataset, the predictive accuracy could potentially improve. Finally, the donor, whose data we train our models on, has Cerebral Amyloid Angiopathy, a condition related to amyloid deposits in the blood vessels, commonly found in AD patients. This may have influenced the identified spatial patterns and gene expressions, thereby affecting the results.

K-Nearest-Neighbour Regressor

For the distance prediction models we use a k-Nearest Neighbors regressor. Due to its nature, the predicted values are based on averaging the outcomes of the k nearest neighbors, which limits the range of predicted distances. As a result, the model struggles to capture the extremes. This averaging behavior reduces prediction variance and makes it difficult for the model to accurately identify cells that are either very close to or very far from pathological regions. We had also tried an MLP and a Random Forest Regressor, however, they did not perform better. Thus, we would suggest possibly trying another type of neural network model with more data to train on.

Imbalance in data

We aimed to classify cells based on whether they fall within a 20-micron radius of the nearest plaque or not. However, the number of cells located within this defined region is significantly smaller than the number outside of it, resulting in a tangible imbalance between the positive and negative labels. Despite applying various strategies to mitigate this issue, the skewed distribution continues to heavily impact the model's performance. Consequently, the classifier fails to accurately predict the cells that overlap, thus reducing the overall reliability of the results.

Dependency on SpaGE

We used SpaGE to perform gene imputation to prepare the inputs for our distance prediction model. Since SpaGE does not achieve perfect accuracy in its predictions, we acknowledge that the performance of our model is dependent on the quality of SpaGE's imputations. Consequently, any inaccuracies in SpaGE's predictions may directly impact the results and interpretations of our model.

Stratification

It must be noted that we did not apply stratification during cross-validation. This decision was based on the nature of our models: for the regression models, stratification is generally less applicable, as it is primarily used for maintaining class distributions in classification problems. Moreover, since the k-NN algorithm is highly sensitive to the distribution of training data, using stratification could potentially introduce more bias. For the classification model, we also opted not to use stratification, as we addressed class imbalance through oversampling techniques. While oversampling can help mitigate imbalance during training, the absence of stratification may still influence model evaluation. Due to time constraints, we were unable to explore how stratification might impact model predictions.

5 Conclusions and Future Work

In this study, we investigated whether the distance of a cell to nearby plaque, learned from spatial transcriptomics data, can be predicted in single-cell RNA-seq data. To address this, we developed a predictive model focused on microglia cells, that uses only genes shared between the spatial and single-cell datasets for its predictions. Our model achieves sufficient accuracy to support the hypothesis that spatial proximity to pathology can be inferred from single-cell data. Furthermore, we investigated the genes contributing most significantly to spatial localization and identified *APOE*, *LYVE1*, and *SLC17A7* as potentially associated with Alzheimer's Disease and microglial clustering around plaque.

Building upon the findings of this work, future research could focus on improving the classification of overlap with plaque areas, as the model currently has poor performance. In particular, using a larger and more balanced dataset could help address the bias the model currently exhibits toward the negative class. Additionally, exploring other models - such as a neural network - could possibly give better results. Ideally, the improved classification could help refine the distance prediction model, particularly enhancing accuracy for cells closer to plaque. This feedback loop could help resolve current limitations in close-range predictions and lead to more accurate results.

Lastly, evaluating the model on larger datasets that include more patients could uncover additional important genes and spatial patterns that may not have been detected in the single patient data we used.

6 Responsible Research

Ensuring responsible research practices is essential in studies involving human data, especially when working with sensitive clinical and genomic information. Ethical integrity not only protects the rights and privacy of individuals but also reinforces the credibility, reproducibility, and long-term impact of scientific research. In biomedical fields, mishandling data can have serious ethical, legal, and clinical consequences.

This study was conducted with adherence to ethical research standards and transparency of the methods used.

Data privacy concerns

This study utilized the Xenium spatial transcriptomics dataset to train and evaluate the proposed model. As this dataset contains patient data that is not publicly available and is intended solely for research purposes, it is used exclusively within the scope of this Bachelor's Research Project. To ensure compliance with patient privacy regulations, no data is shared or distributed. All data remnants will be securely deleted from local machines upon completion of the project.

Additionally, the ROSMAP dataset was employed, which is publicly accessible under a Controlled Access agreement. This access model ensures that sensitive human data is used under ethical guidelines to protect the rights of human subjects [13].

Future Intended Use

The results of this research are intended to contribute to the growing understanding of single-cell gene expression in Alzheimer's disease; however, they are not clinically validated. These results must not be adopted in other studies or used for diagnosing patients without thorough validation and confirmation by qualified professionals. Further analysis and replication are necessary before these findings can be generalized or integrated into future research.

Furthermore, the findings of this study are preliminary and may be subject to bias. For example, the training data originates from a specific population and therefore may not be generalizable to broader or more diverse demographic groups.

Reproducibility

All datasets (where permissible), methods, and model parameters used in this study have been carefully documented to ensure transparency and reproducibility of the analysis. However, as stated in Data privacy concerns, access to the Xenium dataset is restricted. Thus, the reproducibility of the study depends on other researchers first obtaining access to the dataset through the appropriate data request procedures.

As this is a Bachelor's Research Project, the source code for this study is available on the GitLab repository https://gitlab.ewi.tudelft.nl/goncalveslab/bachelor-projects/bsc-rp-2425-galya-vergieva and accessible only to users authorized by TU Delft. The ROSMAP dataset can be downloaded from the AD Knowledge Portal.

7 Acknowledgements

We extend our gratitude to the Gonçalves lab (https://goncalveslab.tudelft.nl/) at TU Delft, for providing us with the Xenium dataset and for their valuable support throughout the development of this work. Their work was supported by NIH common fund project number 1U54EY032442-01, NIH T32 DK101003, NIH project number 3OT2OD033759-01S1 subaward number 1090719-473495, and National Institute on Aging (NIA) R01AG078803.

We are also grateful for the data provided by the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) cohort at Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Their work was funded by NIH grants U01AG061356 (De Jager/Bennett), RF1AG057473 (De Jager/Bennett), and U01AG046152 (De Jager/Bennett) as part of the AMP-AD consortium, as well as NIH grants R01AG066831 (Menon) and U01AG072572 (De Jager/St George-Hyslop). The results published here are in whole or in part based on data obtained from the AD Knowledge Portal - https://adknowledgeportal.org.

References

[1] Xing Fan and Huamei Li. Integration of Single-Cell and Spatial Transcriptomic Data Reveals Spatial Architecture and Potential Biomarkers in Alzheimer's Disease. *Molecular Neurobiology*, 62(5):5395–5412, May 2025.

[2] Y. Hou, X. Dan, M. Babbar, et al. Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology*, 15:565–581, 2019.

[3] Philip Scheltens, Bart De Strooper, Miia Kivipelto, Henne Holstege, Gael Chételat, Charlotte E Teunissen, Jeffrey Cummings, and Wiesje M van der Flier. Alzheimer's disease. *Seminars in Neurology*, 397:1577–1590, April 2021. Accessed 2025-04-24.

[4] Qi Wang, Kewei Chen, Yi Su, Eric M. Reiman, Joel T. Dudley, and Benjamin Readhead. Deep learning-based brain transcriptomic signatures associated with the neuropathological and clinical severity of Alzheimer's disease. *Brain Communications*, 4(1), January 2022. Publisher: Oxford Academic.

[5] Tamim Abdelaal, Soufiane Mourragui, Ahmed Mahfouz, and Marcel J T Reinders. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Research*, 48(18):e107, October 2020.

[6] Sarah K. Longo, Michael G. Guo, Andrew L. Ji, and Paul A. Khavari. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, 22(11):627–644, 2021.

[7] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015. Publisher: Nature Publishing Group.

[8] Mingming Ma, Qiao Luo, Liangmei Chen, Fanna Liu, Lianghong Yin, and Baozhang Guan. Novel insights into kidney disease: the scRNA-seq and spatial transcriptomics approaches: a literature review. *BMC Nephrology*, 26(1):181, April 2025.

[9] Shengquan Chen, Boheng Zhang, Xiaoyang Chen, Xuegong Zhang, and Rui Jiang. stplus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics*, 37(Supplement_1):i299–i307, July 2021.

[10] Biyue Li, Wenjun Zhang, Chuang Guo, Jie Li, Xuran Wang, Zhen Liu, Yuhan Lin, Guoqiang Yang, Jun Liu, Qiang Liu, Xingqi Zhang, Jianfang Fan, Ying Wang, and Shihua Zhang. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19:662–670, 2022.

[11] Anna Mallach, Magdalena Zielonka, Veerle van Lieshout, Yanru An, Jia Hui Khoo, Marisa Vanheusden, Wei-Ting Chen, Daan Moechars, I. Lorena Arancibia-Carcamo, Mark Fiers, and Bart De Strooper. Microglia-astrocyte crosstalk in the amyloid plaque niche of an Alzheimer's disease mouse model, as revealed by spatial transcriptomics. *Cell Reports*, 43(6):114216, June 2024.

[12] David V. Hansen, Jesse E. Hanson, and Morgan Sheng. Microglia in Alzheimer's disease. *Journal of Cell Biology*, 217(2):459–472, December 2017.

[13] AD Knowledge Portal. ROSMAP Study - AD Knowledge Portal, 2024. Accessed: 2025-05-12.

[14] H. P. Vinutha, B. Poornima, and B. M. Sagar. Detection of outliers using interquartile range technique from intrusion dataset. In Suresh Chandra Satapathy, Joao Manuel R.S. Tavares, Vikrant Bhateja, and J. R. Mohanty, editors, *Information and Decision Sciences*, pages 511–518, Singapore, 2018. Springer Singapore.

[15] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[16] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Paul D Dunne, Stephen McQuaid, Roland T Gray, Liam J Murray, Hugh G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):16878, 2017.

[17] Fabian A. Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. Accessed using Scanpy version 1.11.1.

[18] Isaac Virshup, Sergey Rybakov, Fabian J. Theis, Philipp Angerer, and Fabian A. Wolf. AnnData: a Python package for handling annotated data matrices. *Bioinformatics*, 37(18):2911–2912, 2021. Accessed using anndata version 0.11.4.

[19] Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Adam Cloud, Simon Hawkins, Gfyoung, Sinhrks, Stephan Klein, et al. pandas-dev/pandas: Pandas. *Zenodo*, 2020. Accessed using pandas version 2.2.3.

[20] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. Accessed using scikit-learn version 1.5.2.

[21] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

[22] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[23] Michael Waskom. Seaborn: statistical data visualization, 2021.

[24] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[25] Sarah K. Longo, Ming G. Guo, Anna L. Ji, and Paul A. Khavari. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, 22(10):627–644, 2021.

[26] Yun Chen, M. Seda Durakoglugil, Xue Xian, and Joachim Herz. Apolipoprotein e: Structural insights and links to alzheimer disease pathogenesis. *Neuron*, 109(2):205–221, 2021.

[27] Alberto Serrano-Pozo, Subhojit Das, and Bradley T. Hyman. Apoe and alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches. *The Lancet Neurology*, 20(1):68–80, 2021.

[28] Dscr1 upregulation enhances dural meningeal lymphatic drainage to attenuate amyloid pathology of alzheimer's disease.

[29] Maya Karam, Hugo Janbon, Guy Malkinson, and Isabelle Brunet. Heterogeneity and developmental dynamics of lyve-1 perivascular macrophages distribution in the mouse brain. *Journal of Cerebral Blood Flow Metabolism*, 42(10):1797–1812, 2022.

[30] Marta Rodriguez-Perdigon, Rosa M. Tordera, Francisco J. Gil-Bea, Gema Gerenu, Maria J. Ramirez, and Maite Solas. Down-regulation of glutamatergic terminals (vglut1) driven by a in alzheimer's disease. *Hippocampus*, 26(10):1303–1312, 2016. Epub 2016 Jun 21.

[31] Wenxu Wang, Jincheng Lu, Ningyun Pan, Huiying Zhang, Jingcen Dai, Jie Li, Cheng Chi, Liumei Zhang, Liang Wang, and Mengying Zhang. Identification of early alzheimer's disease subclass and signature genes based on panoptosis genes. *Frontiers in Immunology*, Volume 15 - 2024, 2024.

Appendix

Glossary of Terms

1. **Amyloid plaques** – harmful proteins that build up around neurons, thus disrupting normal neuron functions.
2. **Tau tangles** – abnormal clumps of proteins that build up inside neurons and interfere with normal neuron functions.
3. **Disease modifying treatment (DMT)** – Therapy that targets the root cause of a disease to slow, stop, or reverse its progression, unlike symptom treatments that only manage effects.
4. **Single-cell RNA sequencing (scRNA-seq)** – a technique that measures gene expression in individual cells. It enables the identification of rare cell types and the detailed study of specific cells.
5. **Spatial transcriptomics** – A method that measures and maps gene expression in its spatial tissue context.
6. **Gene imputation** – a technique used to infer missing genetic information.
7. **Gene expression** – the process by which information from a gene is used to make a product, like a protein, that performs functions in a cell.
8. **Histopathological** – related to the study of diseases of the tissues with a microscope.
9. **Amyloid plaque niche** – the immediate environment around the amyloid plaque.
10. **Microglia cells** – specialized immune cells in the central nervous system.
11. **Paracrine signaling** - a form of cell-to-cell communication where a cell releases signals that affect nearby cells.